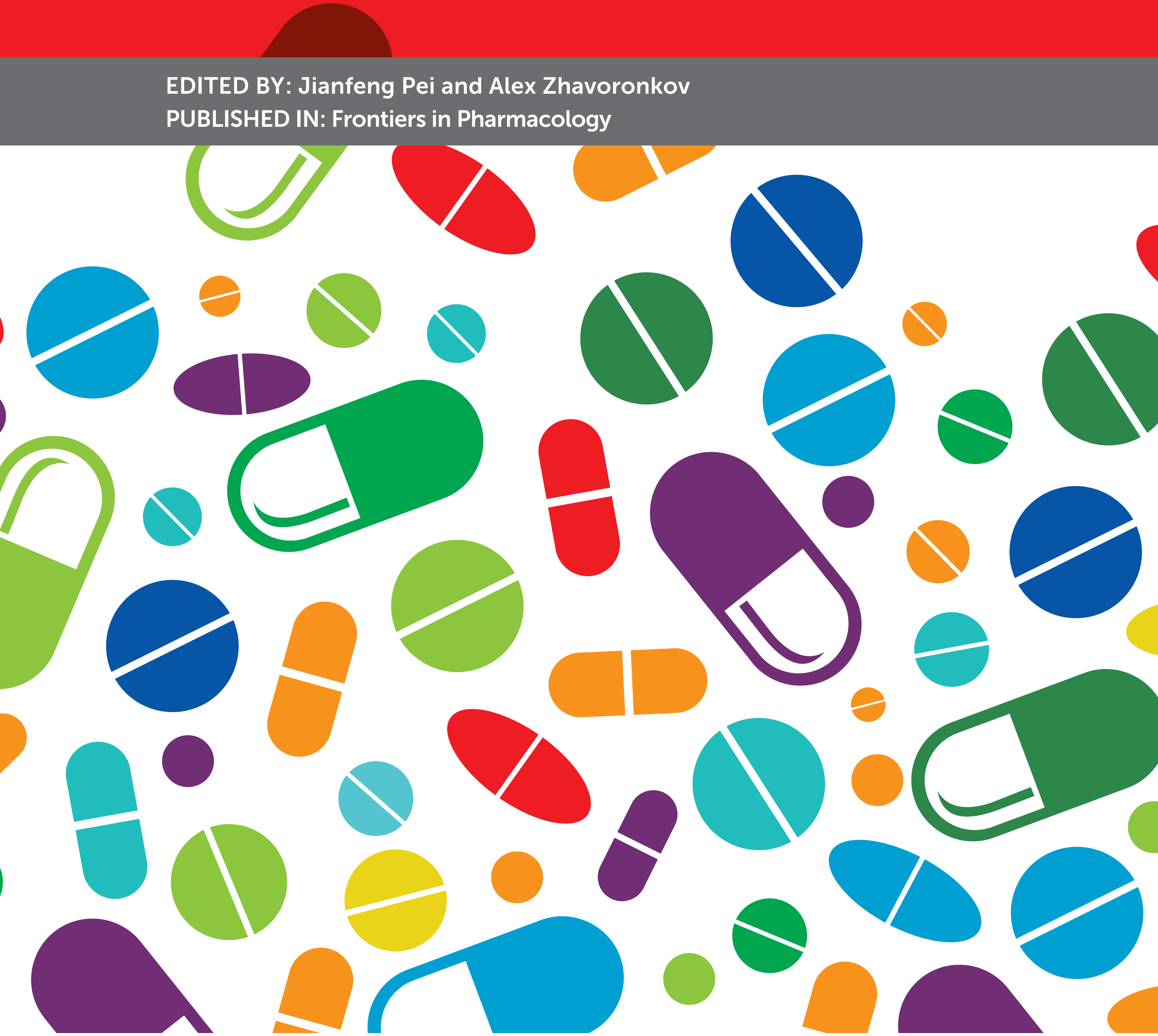
# ARTIFICIAL INTELLIGENCE FOR DRUG DISCOVERY AND DEVELOPMENT

EDITED BY: Jianfeng Pei and Alex Zhavoronkov

PUBLISHED IN: Frontiers in Pharmacology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-649-4
DOI 10.3389/978-2-88971-649-4

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# ARTIFICIAL INTELLIGENCE FOR DRUG DISCOVERY AND DEVELOPMENT

Topic Editors:
**Jianfeng Pei,** Peking University, China
**Alex Zhavoronkov,** Biogerontology Research Foundation, United Kingdom

*Topic editor Alex Zhavoronkov is the founder of Insilico Medicine, a company specializing in AI research. He is also a professor at the Buck Institute for Research on Aging. All other Topic Editors declare no competing interests with regards to the Research Topic subject.*

**Citation:** Pei, J., Zhavoronkov, A., eds. (2021). Artificial Intelligence for Drug Discovery and Development. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-649-4

# Table of Contents

ORIGINAL RESEARCH
published: 12 July 2019
doi: 10.3389/fphar.2019.00780

# Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease

*Salma Jamal*[1], *Abhinav Grover*[2] *and Sonam Grover*[1]*

[1] *JH-Institute of Molecular Medicine, Jamia Hamdard, New Delhi, India,* [2] *School of Biotechnology, Jawaharlal Nehru University, New Delhi, India*

Alzheimer's disease (AD) is a neurodegenerative disorder in which the death of brain cells takes place leading to loss of memory and decreased cognitive ability. AD is a leading cause of death worldwide and is progressive in nature with symptoms worsening over time. Machine learning–based computational predictive models based on 2D and 3D descriptors have been effective in identifying potential active compounds. However, the use of data from molecular dynamics (MD) trajectories for training machine learning models still needs to be explored. In the present study, descriptors have been extracted from the MD trajectories of caspase-8 ligand complexes to train models using artificial neural networks and random forest algorithms. Caspase-8 plays a key role in causing AD by cleaving amyloid precursor proteins during apoptosis leading to increased formation of the amyloid-beta peptide. A total of 43 ligands were docked using the glide module of Schrodinger software, and short MD simulations of 10 ns were performed for the calculation of MD descriptors. The MD descriptors were also combined with the 2D and 3D descriptors of chemical compounds, and individual descriptor based as well as combination models were generated. This study demonstrated that MD descriptors could be effectively used for the characterization of bioactive compounds along with lead prioritization and optimization.

**Keywords: Alzheimer's, caspase-8, machine learning, molecular dynamics trajectories, descriptors**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University,*
*China*

***Reviewed by:***
*Linfeng Zhang,*
*Princeton University,*
*United States*
*Denis Fourches,*
*North Carolina State University,*
*United States*

***Correspondence:***
*Sonam Grover*
*sonamgbt@gmail.com*
*sonamgrover@jamiahamdard.ac.in*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *25 March 2019*
***Accepted:*** *17 June 2019*
***Published:*** *12 July 2019*

***Citation:***
*Jamal S, Grover A and Grover S (2019) Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease.*
*Front. Pharmacol. 10:780.*
*doi: 10.3389/fphar.2019.00780*

## INTRODUCTION

Neurological disorders affect millions of people globally with Alzheimer's disease being the most common type of disease. Alzheimer's disease (AD) is the sixth prominent cause of death in the United States and, as per the data from the National Center for Health Statistics of the center for disease control (CDC), AD was responsible for approximately 110,561 deaths in 2015 (Alzheimer's Association, 2018). AD is the only disease among the top 10 causes of death with no means of prevention, treatment, or delay in progression. The disease is pathologically defined by protein aggregation and its impact on the function of neurons; therefore, studies have been primarily focusing on reducing protein aggregation and promoting clearance from the brain (Small et al., 2001). However, these therapies have been unsuccessful in clinical trials, which suggests targeting protein aggregation and clearance alone may not be sufficient to treat AD. Among the many factors responsible for AD such as amyloid hypothesis, cholinergic hypothesis, tau hypothesis, environmental risks, and genetic factors, it has been well established that approximately 70% of risk for the disease is attributable to genetics (Ballard et al., 2011). The previously discovered genes presenilin 1 (*PSEN1*), presenilin 1 (*PSEN2*), and amyloid precursor protein (*APP*) are accountable for the pathogenesis of AD in only about 5% of patients (Van Cauwenberghe et al., 2016). Considering the complex physiology of AD and multiple causes responsible for the disease, drug development against AD must consider all events related to the pathophysiology for more effective treatment strategies, which cannot be accomplished concentrating on one cause alone. The development of effective treatment options for AD has been of great interest considering the global burden of the disease, and thus, identification of more potent and selective inhibitors from the large pool of chemical compounds is imperative. Caspases have been reported to play an important role in AD due to the increase in β-amyloid levels by the cleavage of APP during apoptosis (Rohn et al., 2001). Multiple evidences are there which suggest that APP is a substrate for caspase-lead cleavage which is a crucial step in the AD process that may result in amyloid-beta formation, loss of synaptic activity, and behavior changes related with AD (Gervais et al., 1999; Cotman et al., 2005; Galvan et al., 2006). Recent studies have reported that activation of caspases leads to the formation of neurofibrillary tangles (NFT) (Gamblin et al., 2003; Rissman et al., 2004). Another study has confirmed the cleavage of tau by caspases in the early state of AD (Guillozet-Bongaarts et al., 2005). It has also been put forward that caspase-mediated truncation of tau is interrelated with the development of NFTs and beta-amyloids in AD (Dickson, 2004). In addition, all the caspases -1, -2, -3, -5, -6, -7, -8, and -9 have been identified to be transcriptionally elevated in AD (Castro et al., 2010). Caspase-8 has been labeled an originator caspase that further activates other downstream caspases, making this enzyme an attractive target for the identification and development of inhibitors. This could prevent unwanted cell death related to various neurodegenerative disorders (Watt et al., 1999). Caspase-8 has also been associated with synaptic plasticity as well as associated neurotoxicity through its downstream effector caspase-3, which points toward other supplementary mechanisms that might lead to AD (Rehker et al., 2017). Caspases play an important role in disease mechanisms associated with AD that include formation of beta-amyloids as well as NFTs and thus inhibiting caspases may lead to prevention of formation of plaques and tangles and also reducing disease progression. Computational predictive models have been of great use to researchers doing studies on drug discovery. Machine learning approaches have been used extensively for the identification of potential active compounds based on 2D and 3D molecular descriptors (Jamal et al., 2015; Wahi et al., 2015; Jamal et al., 2017). Although the previously developed models were successful for screening lakhs to millions of compounds, a high degree of reliability is required for prioritizing the top five or 10 compounds from a set of hundreds of possibilities. This necessitates the generation of more accurate hyper-predictive target-specific models utilizing the descriptors extracted from molecular dynamics (MD) trajectories and consideration of protein-ligand interactions (Ash and Fourches, 2017). Various quantitative structure activity relationship studies for the development of caspase-3 inhibitors have already been reported in the literature (Legewie et al., 2006; Wang et al., 2009; Firoozpour et al., 2012; Sharma et al., 2013). The present study was carried out to utilize the potential of MD-derived descriptors in predictive modeling of potent caspase inhibitors. Thus, the present study is based on the hypothesis that MD-based machine learning models could be extremely useful for lead optimization and chemical compound prioritization. Potential inhibitors of caspase-8 have been used for the calculation of 2D, 3D, and MD descriptors. The ligands were docked into caspase-8 protein, and the protein-ligand complexes were subjected to MD simulations to generate descriptors from MD trajectories. Further, artificial neural network and random forest machine learning algorithms were used to generate the models using an individual set of descriptors with two and three level combinations. The conformational dynamics of caspase-8 upon binding with the compound predicted to be active against the protein using 100 ns MD simulation was also explored. Moreover, pharmacophore model was developed using the ligands associated with caspase-8 which was further used for virtual screening to identify the new potential caspase inhibitors.

# METHODOLOGY

## Caspase-8 Data Set

In the present study, we used the caspase-8 data set comprised of ligands associated with caspase-8 retrieved from the ChEMBL (Gaulton et al., 2012) database (ChEMBL46860, ChEMBL46862, ChEMBL399983, ChEMBL304686, ChEMBL430105, ChEMBL 46849, and ChEMBL741342). A total of 81 compounds were obtained and preprocessed (Fourches et al., 2010), during which duplicates and compounds with approximate IC50 values were removed. Post-processing the data resulted in 43 compounds with pIC50 values ranging from 4.3 to 8.1, among which compounds with a pIC50 value above 6.5, were considered as active compounds while those with a pIC50 below 6.5 were considered as inactive compounds. The final data set including the molecule identifiers, SMILES, and pIC50 values has been provided in the Supporting Information.

## Molecular Docking

The X-ray crystal structure of human caspase-8 (PDB ID: 1qtn) in complex with acetyl-ile-glu-thr-asp-aldehyde peptide at a resolution of 1.2 Å was obtained from the protein data bank (PDB) (Parasuraman, 2012). The protein-ligand complex was preprocessed using Accelrys ViewerLite (Accelrys Inc., San Diego, CA, USA) during which ligands, water molecules, and heteroatoms were removed. Further, the protein was prepared with Preparation Wizard available from Schrodinger Suite (http://www.schrodinger.com/). Hydrogen bonds were added, and bond orders were assigned during protein preparation. The protonation states of residues were predicted using the PROPKA (Olsson et al., 2011) program at pH 7 followed by minimization of the protein using the OPLS3 force field (Sastry et al., 2013). The ligands associated with the caspase-8 protein were prepared using the LigPrep (Schrödinger, Inc., www.schrodinger.com) module of Schrodinger before molecular docking. The ligands were also minimized using the OPLS3 force field, and the possible ionization states were created at pH 7.0 ± 2.0. The tautomers were generated, specific chiralities of the ligands were retained, and 32 conformations per ligand were generated in case of indefinite chiralities. Next, using the Receptor Grid Generation section of the Glide (Halgren et al., 2004) module of Schrodinger, the binding site in protein was defined using the centroid of selected residues option in which the catalytic triad Cys360, His317, and Arg258 were chosen. A scaling factor of 1.0 was used to scale the van der Waals radii of receptor atoms having a partial atomic charge less than the specified cut-off, which was equal to 0.25. All other parameters were default. The prepared ligands were then docked into the active site of the receptor using an extra precision algorithm of Glide. The top-ranked pose for each ligand was selected and subjected to MD simulation studies.

## Molecular Dynamics Simulation Details

The top scoring protein-ligand complexes were subjected to 10-ns MD simulations to evaluate their structural and thermodynamic stability in the presence of explicit salt and solvents. All the MD simulation studies were performed using the GROMACS (Abraham et al., 2015) software version 5.0 and GROMOS96 force field. Prior to the MD simulation, each protein-ligand complex was prepared by the removal of the water molecules, addition of hydrogen atoms, capping of termini, treating disulphides, and finding overlaps. After the initial preparation, the model was solvated with a simple point charge (SPC) water model and Na+ and Cl- ions were added to maintain the neutrality of the system. The solvated system was then subjected to energy minimization for 50,000 steps using the steepest descent method until a maximum force of 10.0 kJ/mol was attained. An equilibration run was performed in two sequential steps, NVT (number of particles, volume, and temperature) equilibration, and NPT (number of particles, pressure, and temperature) equilibration during which pressure and temperature were kept to 1 bar and 300°C, respectively, for a maximum of 50,000 steps in both the types of equilibration. Further, a 10-ns MD simulation run was carried out to obtain a stable structure and time *versus* RMSD (root-mean square deviation) plot to ensure the stability of the system for the entire simulation run.

## Descriptors Computation

Molecular descriptors represent the chemical information of the ligands using numeric values. Three types of descriptors were used for modeling in the present study, 2D, 3D, and MD descriptors. The 2D descriptors included atom count, bond count, carbon types, hydrogen bond donor and acceptor count, Lipinski's rule of five, rotatable bonds count, topological surface area, van der Waals volume, and many more. The 3D descriptors included gravitational index descriptor, charged partial surface area, and length over breadth and moment of inertia descriptors, among others. The 3D-WHIM descriptors involved descriptors weighted by unit weights, van der Waals volumes, atomic masses, atomic polarizabilities, and Mulliken atomic electronegativites. A total of 770 2D descriptors and 115 3D descriptors were generated for each ligand conformation using PaDEL (Yap, 2011) software.

For MD descriptors, the trajectory of each protein-ligand complex was analyzed for three properties, radius of gyration (Rg), potential energy and total energy, and solvent accessible surface area (SASA). Each of the three MD descriptors was represented using the mean and standard deviation as described in other studies (Ash and Fourches, 2017; Riniker, 2017), resulting in a total of eight descriptors.

## Model Building

Machine learning (ML)–based modeling is learning from known properties and using the learned model systems to make predictions for unseen data. Using an in-house Perl script, the molecular descriptor files were split with 70% for a training set and 30% for a testing data set. The training set was used for generation of the models, and the test set was used for the assessment of model performance. An internal validation of the models generated using the training set was performed using *k*-fold cross validation, with *k* equal to 10 in the present work. Cross validation is a technique in which the data is divided into *k* subsets, with *k*-1 subsets used for model generation and the remaining subset used for testing purposes. This process is repeated until all the *k* folds have been used as a testing set at least once. The models were generated using individual 2D, 3D, and MD descriptors and their two level 2D+3D, 2D+MD, and 3D+MD and three level 2D+3D+MD combinations. The 2D, 3D, MD, 2D+3D, 2D+MD, 3D+MD, and 2D+3D+MD artificial neural network (ANN) and random forest (RF) models were generated using different parameters and finding the best combination of parameters.

## Machine Learning Algorithms

In the present study, two ML algorithms, ANN and RF, were used for building the models using Weka which is an ML software. ANN is a computational model that attempts to mimic the structure and function of neural networks in the human brain. It comprises a group of connected artificial neurons that process information and generate output. The ANN model used in the present study is multilayer perceptron (MLP), which is a feedforward ANN model using three or more layers including input and output layers along with hidden layers, to map input data and produce the correct output (Cheng et al., 2008).

RF is a decision tree based classifier that creates an assembly of decision trees and outputs the class that is the mode of the output of all the individual decision trees. The decision tree for each attribute is created by sampling the attributes, then using random selection. Next, the information gain criterion is used to select the best feature from the data which is used as the origin node of the tree. The origin node is then divided into sub-nodes, and the process is repeated until the sub-node becomes an output class. The final prediction is the class chosen by the majority of the trees (Breiman, 2001).

## Feature Selection

Feature, attribute, or descriptor selection is the procedure of identifying a subgroup of features that are relevant to the modeling and prediction task. Feature selection is performed to decrease the dimensionality of the data by eliminating insignificant features and thus reducing training time, removing redundant descriptors, simplifying models, and lessening overfitting of the models. Feature selection was performed at two levels, initially using the Remove Useless filter of the Weka (Bouckaert et al., 2010) ML tool followed by the selection of significant features. The Weka Remove Useless filter removed the descriptors having the same value for all compounds, as those descriptors did not contribute toward classification.

Two feature selection techniques were used, correlation-based feature selection (CFS) and relief attribute evaluation. CFS ranks features using a correlation based heuristic function which outputs a subset of features having a high correlation with the class but uncorrelated with each other (Hall, 1999). The following correlation based heuristic function is used for calculating the merits of a feature subset:

$$Ms = \frac{k\overline{rcf}}{k + k(k-1)\overline{rff}}$$

where *Ms* is the merit of feature subset *S* consisting of *k* features, *rff* is the mean of feature-feature correlation, and *rcf* is the mean of feature-class correlation.

The relief-based attribute selection algorithm calculates a feature score, ranks the features, and chooses the top ranked. The feature score is calculated using the Euclidean distances between features and their nearest neighboring instances. The training data set was used for feature selection, and the test set used to rid the data of any biasness (Kira and Rendell, 1992).

## Model Performance Evaluation

A total of 14 ML models were generated using ANN and RF algorithms, which were evaluated using accuracy, balanced accuracy, training error, generalization error, and a receiver-operating characteristic (ROC) plot. Accuracy ([{TP+TN/(TP+TN+FP+FN}]) is the proportion of correctly classified active and inactive compounds by the classification models. An ROC plot is a graph plotted as true positive rate (TPR) *vs* false positive rate (FPR, 1-specificity). TPR ([TP/{TP+FN}]) is the percentage of correctly classified actives while FPR (1-[TN/{TN+FP}]) is the proportion of correctly identified negatives.

## Pharmacophore Search and Virtual Screening

The 43 compounds used for the generation of ML models were used for ligand-based pharmacophore modeling using PharmaGist tool (Schneidman-Duhovny et al., 2008). A pharmacophore is a theoretical representation of features of ligand necessary for the recognition of ligand by the macromolecule and can be used to identify ligands that can bind to a common receptor through virtual screening. PharmaGist tools search for probable pharmacophores by multiple flexible alignment of input ligands and report the top scoring ones. The pharmacophore model developed in the present study was used for virtual screening to search through a total of 1,798 and 16 natural compounds from ZINCPharmer (Koes and Camacho, 2012) to get the similar hits from ZINC database. The top 10 most similar hits were subjected to Glide's XP docking with the caspase-8 protein used for the docking study with 43 ligands used in the present study.

# RESULTS

## Glide-Docking Analysis

A total of 43 active and inactive caspase-8-associated ligands were docked in the active site of the receptor protein, human caspase-8, using the extra precision (XP) docking approach. The XP docking scores of ligands ranged from −12.70 to −4.22 kcal/mol. The compounds having a pIC50 value above 6.5 categorized as actives corresponding to compound IDs 50267423, 50215849, 50215847, 50215835, 50297218, 50267430, and 50215896 had docking scores of −9.1 kcal/mol, −6.22 kcal/mol, −12.06 kcal/mol, −5.34 kcal/mol, −12.38 kcal/mol, −8.53 kcal/mol, and −7.16 kcal/mol, respectively. Additionally, we have generated the correlation plot between docking scores and pIC50. As is evident from the plot, the compounds having high pIC50 values had higher docking scores and *vice versa* (**Figure 1**).

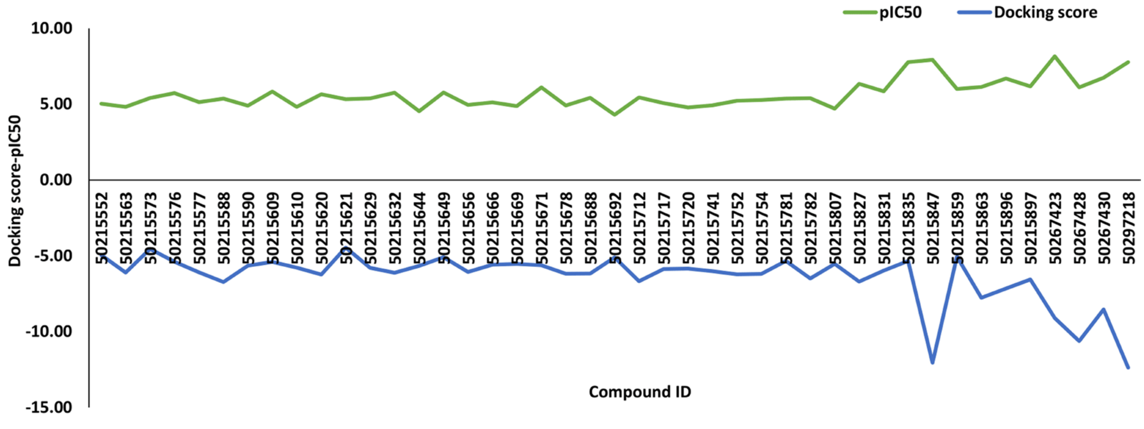


**FIGURE 1 |** The correlation plot between docking scores and pIC50.

## Feature Analysis

In the present study, feature analysis was performed using the Remove Useless filter, CFS, and relief-based attribute selection using Weka. The number of 2D descriptors was reduced from 770 to 387 after the Remove Useless filter was applied. The numbers of 3D and MD descriptors remained the same, 115 and 8, respectively.

E-state indices represent the electronic and topological character of an atom where the electronic state of an atom is encoded as perturbed by the electronic impact of other atoms in the molecule in context of the topological character of the molecule. The top ranked 2D features selected using CFS and relief-based selection included nwHBd, SwHBd, SHCHnX, minHCHnX, minwHBd, maxwHBd, maxHCHnX, and nHCHnX. The selected 2D features included count, sum, and minimum and maximum of E-states for weak hydrogen bond (H-bond) donors and atom type, H (estate: CHnX where nX corresponds to a halogen atom). Various studies have explained the importance of weak H-bonds in chemical and biological systems (Steiner, 1999).

The top ranked 3D features included FPSA-3, WK.unity, Wnu2.unity, WK.mass, Wnu2.mass, Weta3.volume, Wlambda3.mass, and TPSA. The selected 3D descriptors included WHIM descriptors which capture significant molecular 3D information that include shape, molecular size, atom distribution, and symmetry. These indices are computed using *x*, *y*, and *z* coordinates of a molecule using different weighing schemes like atomic mass, van der Waals volume, electronegativity, and atomic polarizabilities and have been used for QSAR modeling (Gramatica, 1997).

In the case of MD descriptors, all the eight descriptors that included mean and standard deviation of potential and total energy, Rg, and SASA were used for model generation. The MD descriptors included total and potential energy, Rg, and SASA where total and potential energy are mathematical forms of representations of protein-ligand interactions; Rg indicates the compactness of the protein, and that is how the secondary structures are compactly folded in to 3D structure of the protein. SASA is a measure of accessible surface of a molecule which further helps in secondary structure prediction. The number and description of features used in the present study have been provided in **Table 1**.

In addition to this, we also calculated importance of MD descriptors. This was carried out by computing average merit and average rank using CFS, relief-based attribute selection, and classifier attribute evaluation using ANN and RF classifiers. Average merit indicates the average accuracy loss when a particular feature is removed whereas average rank denotes the rank of the feature determined using 10-fold cross validation. The results indicated that potential energy of the protein-ligand complex was the most significant contributor toward classification followed by Rg, SASA, and total energy (**Table 2**).

## Model Predictions and Performances

A total of 14 ML models were generated in the present study using ANN and RF ML algorithms. These ANN and RF models (2D, 3D, MD, 2D+3D, 2D+MD, 3D+MD, and 2D+3D+MD) were generated using best combination of different parameters. Initially, we tried to the models using default parameters for ANN and RF algorithms. However, these did not perform well in terms of the statistical parameters used for model performance evaluation (**Table 3**). The training set consisted of 29 compounds, and the test data set consisted of 14 compounds. **Table 4** provides the performance metrics of all the generated ANN and RF models using the best combination of parameters. **Figure 2** illustrates the ROC plots generated for ANN and RF models using 2D, 3D, MD, 2D+3D, 2D+MD, 3D+MD, and 2D+3D+MD descriptors. The training and testing sets used for generating the models and the models build in the present study have been provided as **Supplementary Material**.

**TABLE 1 |** The number and description of features used in the present study.

| Type of descriptor | Initial number of descripted | Remove useless filter | Relief-based selection | Selected descriptors | Description as provided by PaDEL |
|---|---|---|---|---|---|
| 2D | 770 | 387 | 8 | nwHBd,<br>SwHBd,<br>SHCHnX,<br>minHCHnX<br>minwHBd,<br>maxwHBd,<br>maxHCHnX<br>nHCHnX | Atom type electrotopological state<br>Count of E-state for weak H-bond donors<br>Sum of E-state for weak H-bond donors<br>Sum of atom type H E-state: CHnX<br>Minimum atom type H E-state: CHnX<br>Minimum of E-state for weak H-bond donors<br>Maximum of E-state for weak H-bond donors<br>Maximum atom type H E-state: CHnX<br>Count of atom type H E-state: CHnX |
| 3D | 115 | 115 | 8 | FPSA-3,<br>WK.unity,<br>Wnu2.unity,<br>WK.mass,<br>Wnu2.mass,<br>Weta3.volume,<br>Wlambda3.mass<br>TPSA | Charged partial surface area<br>Non-directional WHIM weighted by unit weights<br>Directional WHIM weighted by unit weights<br>Non-directional WHIM weighted by atomic masses<br>Directional WHIM weighted by atomic masses<br>Directional WHIM weighted by van der Waals volumes<br>Directional WHIM weighted by atomic masses<br>Topological polar surface area |
| MD | 8 | 8 | 8 | Potential energy<br>Total energy<br>Radius of gyration<br>Solvent accessible surface area | |

**TABLE 2 |** Importance of molecular dynamics (MD) descriptors using correlation-based feature selection (CFS), relief-based attribute selection, and classifier attribute evaluation using artificial neural network (ANN) and random forest (RF) classifiers.

| MD descriptor | CFS | | Relief-based | | Classifier attribute evaluator (ANN) | | Classifier attribute evaluator (RF) | |
|---|---|---|---|---|---|---|---|---|
| | Average merit | Average rank | Average merit | Average rank | Average merit | Average rank | Average merit | Average rank |
| Total energy | 0.062 ± 0.032 | 2.6 ± 0.92 | 0.021 ± 0.028 | 3.3 ± 0.64 | −0.015 ± 0.023 | 2.2 ± 1.47 | −0.133 ± 0.032 | 3.8 ± 0.4 |
| Potential energy | 0.062 ± 0.032 | 1.8 ± 0.75 | 0.021 ± 0.028 | 2.3 ± 0.64 | −0.015 ± 0.023 | 2.4 ± 0.49 | −0.068 ± 0.046 | 2.5 ± 0.81 |
| Gyration | 0.048 ± 0.033 | 2.9 ± 1.14 | 0.014 ± 0.011 | 3.2 ± 0.98 | 0 ± 0 | 2.6 ± 0.49 | −0.036 ± 0.024 | 1.8 ± 0.75 |
| SASA | 0.057 ± 0.041 | 2.7 ± 1.27 | 0.092 ± 0.013 | 1.2 ± 0.6 | 0 ± 0 | 2.8 ± 1.47 | −0.028 ± 0.036 | 1.9 ± 1.04 |

**TABLE 3 |** The performance metrics of all the generated machine learning (ML) models using ANN and RF algorithms using default parameters.

| Machine learning algorithm | Descriptor type | Cross-validation accuracy (%) | Accuracy (%) | AUC | Balanced accuracy (%) | Training error | Generalization error | Training error | Generalization error |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MSE | RMSE | MSE | RMSE |
| Artificial neural network | 2D | 86.20 | 85.71 | 0.50 | 50.00 | 0.21 | 0.35 | 0.20 | 0.35 |
| | 3D | 82.75 | 85.71 | 0.91 | 70.50 | 0.16 | 0.37 | 0.16 | 0.37 |
| | MD | 82.75 | 85.71 | 0.66 | 50.00 | 0.27 | 0.39 | 0.22 | 0.34 |
| | 2D+3D | 89.65 | 85.71 | 0.91 | 70.50 | 0.10 | 0.28 | 0.16 | 0.38 |
| | 2D+MD | 75.86 | 85.71 | 0.58 | 50.00 | 0.29 | 0.46 | 0.20 | 0.35 |
| | 3D+MD | 89.65 | 78.57 | 0.87 | 66.50 | 0.10 | 0.25 | 0.20 | 0.42 |
| | 2D+3D+MD | 89.65 | 78.57 | 0.87 | 66.50 | 0.13 | 0.32 | 0.20 | 0.42 |
| Random forest | 2D | 86.20 | 85.71 | 0.50 | 50.00 | 0.21 | 0.35 | 0.21 | 0.35 |
| | 3D | 89.65 | 85.71 | 0.50 | 62.00 | 0.15 | 0.26 | 0.18 | 0.34 |
| | MD | 82.75 | 85.71 | 0.68 | 50.00 | 0.26 | 0.38 | 0.19 | 0.34 |
| | 2D+3D | 86.20 | 85.71 | 0.52 | 50.00 | 0.15 | 0.27 | 0.18 | 0.33 |
| | 2D+MD | 82.75 | 85.71 | 0.62 | 50.00 | 0.26 | 0.40 | 0.19 | 0.35 |
| | 3D+MD | 86.20 | 85.71 | 0.89 | 50.00 | 0.16 | 0.26 | 0.17 | 0.31 |
| | 2D+3D+MD | 86.20 | 85.71 | 0.87 | 50.00 | 0.17 | 0.28 | 0.16 | 0.30 |

**TABLE 4 |** The performance metrics of all the generated ANN and RF models using the best combination of parameters.

| Machine learning algorithm | Descriptor type | Cross-validation accuracy (%) | Accuracy (%) | AUC | Balanced accuracy (%) | Training error | Generalization error | Training error | Generalization error |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MSE | RMSE | MSE | RMSE |
| Artificial neural network | 2D | 86.20 | 85.71 | 0.50 | 50.00 | 0.21 | 0.25 | 0.2 | 0.35 |
| | 3D | 44.82 | 64.28 | 0.91 | 79.15 | 0.52 | 0.38 | 0.44 | 0.51 |
| | MD | 82.75 | 85.71 | 0.70 | 50.00 | 0.25 | 0.4 | 0.21 | 0.34 |
| | 2D+3D | 13.79 | 85.71 | 0.37 | 50.00 | 0.63 | 0.64 | 0.47 | 0.47 |
| | 2D+MD | 51.72 | 85.71 | 0.75 | 70.85 | 0.52 | 0.59 | 0.42 | 0.43 |
| | 3D+MD | 75.86 | 78.57 | 0.83 | 87.50 | 0.35 | 0.47 | 0.33 | 0.47 |
| | 2D+3D+MD | 62.06 | 78.57 | 0.91 | 87.50 | 0.48 | 0.55 | 0.39 | 0.47 |
| Random forest | 2D | 86.20 | 85.71 | 0.50 | 50.00 | 0.47 | 0.47 | 0.47 | 0.47 |
| | 3D | 82.75 | 78.57 | 0.79 | 66.65 | 0.24 | 0.35 | 0.32 | 0.37 |
| | MD | 55.17 | 71.42 | 1.00 | 83.35 | 0.44 | 0.53 | 0.32 | 0.38 |
| | 2D+3D | 65.51 | 57.14 | 0.77 | 54.15 | 0.37 | 0.43 | 0.43 | 0.47 |
| | 2D+MD | 86.20 | 85.71 | 0.75 | 50.00 | 0.21 | 0.38 | 0.17 | 0.35 |
| | 3D+MD | 89.65 | 92.85 | 0.79 | 75.00 | 0.17 | 0.26 | 0.19 | 0.31 |
| | 2D+3D+MD | 72.41 | 85.71 | 0.91 | 91.50 | 0.38 | 0.42 | 0.36 | 0.39 |

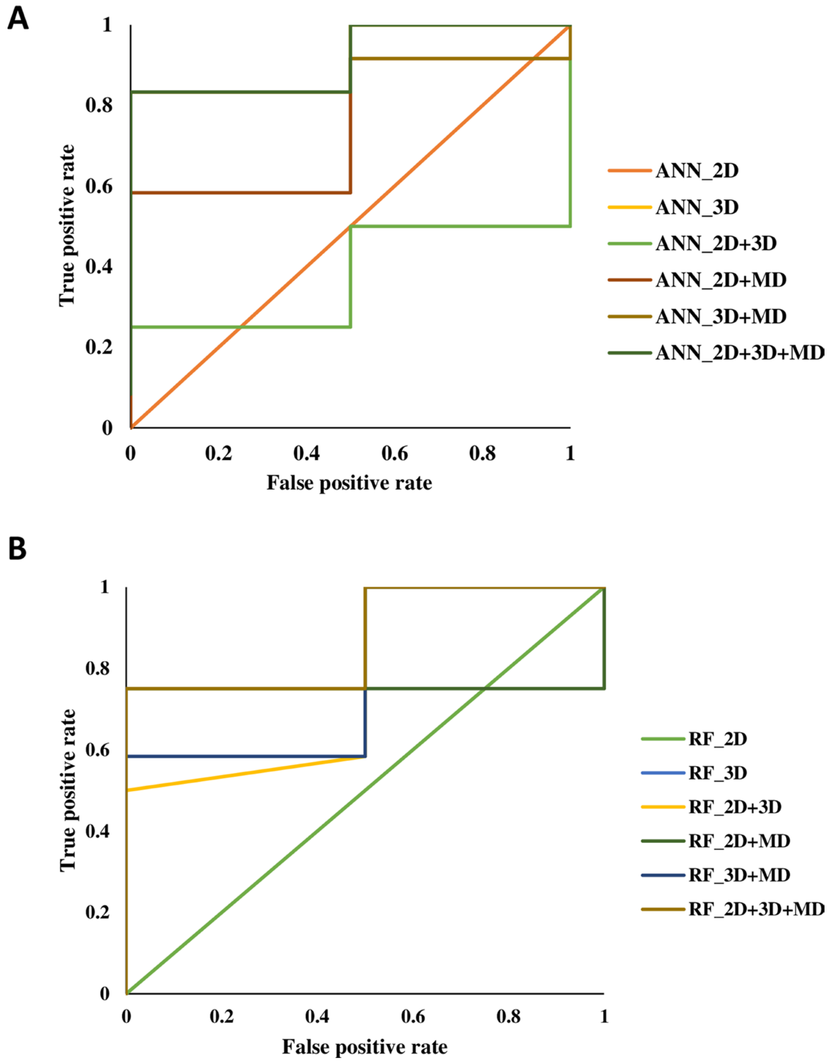


**FIGURE 2 |** The receiver-operating characteristic (ROC) plots generated for artificial neural network (ANN) and random forest (RF) models using 2D, 3D, MD, 2D+3D, 2D+MD, 3D+MD, and 2D+3D+MD descriptors.

## Modeling Using 2D Descriptors

The 2D ANN and RF models had an accuracy of 85.71%, balanced accuracy of 50.0%, and an AUC value of 0.50. The AUC value indicated these models were random predictors and thus were not considered for further predictions.

## Modeling Using 3D Descriptors

The 3D descriptor models had an accuracy, balanced accuracy, and AUC value of 64.28%, 79.15%, and 0.91, respectively, for the ANN model. In case of RF model, the accuracy, balanced accuracy, and AUC values corresponded to 78.57%, 66.65%, and 0.79. These results indicated that 3D compound descriptors play a vital role in the classification of compounds. The ANN model correctly predicted the two active compounds 50267423 and 50215896 and the other inactive compounds predicted as active by ANN had compound IDs 50215590, 50215632, 50215692, 50215782, and 50215859. The RF model gave the correct prediction for only one active compound, 50267423. The other inactive compounds predicted as active included 50215590 and 50215632.

## Modeling Using MD Descriptors

The models generated using MD descriptors had an accuracy of 85.71% and 71.42%, balanced accuracy of 50.0% and 83.35%, and an AUC value of 0.70 and 1.00 for ANN and RF models, respectively. The MD models had the most balanced accuracies and AUC values compared to the 2D and 3D descriptor models, in which either the accuracy was high and AUC value was low or vice versa. This clearly indicates the descriptors extracted from MD trajectories play a significant role in lead prioritization, resulting in most active compounds. The reduction in generalization error as compared to training error indicated that MD descriptors can perform well on new data. The ANN model correctly predicted the inactive compounds, but misclassified the compounds categorized as active. However, the RF model predicted the active compounds, 50267423 and 50215896, as active. The other compounds predicted as active included compounds corresponding to IDs 50215632, 50215720, 50215782, and 50267428.

## Modeling Using the Two Level Combination of 1D, 2D, and MD Descriptors

The models were generated by combining 1D, 2D, and 3D descriptors as 2D, 2D+MD, and 3D+MD. The 2D+3D descriptor models had an accuracy of 85.71% and 57.14%, balanced accuracy of 50.0% and 54.15%, and an AUC value of 0.37 and 0.77 for ANN and RF models, respectively. The 2D+3D RF model predicted one active compound, 50267423, accurately. In the case of RF models, the accuracy and balanced accuracy of the models remained the same when 2D descriptors were combined with MD descriptors. Although the accuracy was same (85.71%) in case of ANN models (2D and 3D), there was a significant increase in the balanced accuracy (from 50% to 70.85%) and AUC (from 0.37 to 0.75) upon addition of MD descriptors.

When the 3D descriptors were combined with MD descriptors, an increase in accuracy (from 64.28% to 78.57%) and balanced accuracy (from 79.15% to 87.50%) was observed in case of ANN models; however, there was a slight reduction (from 0.91 to 0.83) in the AUC value. In the case of models generated using the RF algorithm, the accuracy (from 78.57% to 92.85%) and balanced accuracy (from 66.65 to 75%) values improved while AUC (0.79) value remained the same in case of addition of MD+3D descriptors. The results clearly indicate the combination of models resulted in greater accuracy with the 3D+MD combination models being the most informative. As the 3D+MD combination models had the best performance, the compounds predicted as active by these models were corresponding to IDs 50267423, 50215590, and 50215720.

It was also observed that the models generated using 2D and 3D descriptors in combination with MD descriptors had low mean absolute error (MSE) and root mean squared error (RMSE) in comparison to models generated using 2D, 3D, and 2D+3D.

## Modeling Using the Combined 1D, 2D, and MD Descriptors

The models generated using the combination of all the three descriptors—2D, 3D, and MD—had high accuracy (ANN 78.57%; RF 92.85%) values, balanced accuracies (ANN 87.50%; RF 91.50%), and AUC (ANN 0.91; RF 0.87) values. The compounds predicted as active by both the models included 50267423, 50215590, and 50215720. The MD descriptors alone and in combination with 2D and 3D descriptors performed better in terms of generalization performance.

We also calculated the accuracy of ANN/RF model vis-a-vis the accuracy due to the different input. The accuracy obtained using different input dataset was higher in comparison to the ANN/RF model accuracies, indicating that the ML models generated in the present study would be able to predict outcomes for new unseen data.

## Molecular Dynamics Simulation Analysis of the Most Active Compound

Since most of the ANN and RF models were able to accurately predict this compound 50267423 as active among all the other predicted active compounds, the same was chosen for carrying out long MD simulations. The compound, 50267423, having a docking score of −9.10 kcal/mol was subjected to a 100ns MD simulation for an in depth study of its structural characteristics. As apparent from **Figure 3A**, the unbound caspase-8 protein was unstable, but became stable upon binding with compound 50267423. In both cases, the simulation reached convergence between 10–30ns with RMSD around 0.45 and 0.35 nm for the unbound caspase-8 and caspase-8_50267423 complex, respectively. Next, Rg was calculated to demonstrate the impact of compound 50267423 on the compactness of the protein. The protein had a compact packing in both unbound and bound forms (**Figure 3B**). Root mean square fluctuation (RMSF) analysis was performed to study the fluctuation on residues in the presence of the ligand. **Figure 3C** illustrates the RMSF in free caspase-8 and caspase-8_50267423 complex. The residues had enormous fluctuations in unbound caspase-8 while RMSF values were reasonably low, and the protein was very much stable in the presence of the 50267423 compound. Further, SASA was calculated, which was higher in the case of unbound caspase-8 protein in comparison to the SASA in the ligand-bound protein (**Figure 3D**). Thus, it is evident from the aforementioned results that the caspase-8 protein was highly stable upon binding with compound 50267423. The hydrogen bonding and hydrophobic interaction analyses were carried out for the caspase8-50267423 complex. The ligand formed five hydrogen bonds, which included two bonds with Trp420, two hydrogen bonds with Gln423, and one bond with Ser424, as demonstrated in **Figure 4**. The residues having hydrophobic interactions included Asp266, Leu315, Gln358, Ala404, Thr405, Ser411, Glu417, Gly418, Thr419, Tyr421, and Ile422 (**Figure 5**). The residues having hydrophobic interactions Gln358 and Ser411 have been shown to line the binding pocket in caspase-8 whereas the aromatic group of Ty420 which in the present study is forming two hydrogen bonds with the inhibitor has been shown to help to form the part of the pocket (Watt et al., 1999).

## Identification of Common Pharmacophore and Virtual Screening

Pharmacophore search using PharmaGist provided us a high-scoring pharmacophore containing compound corresponding to IDs 50267423 (most active compound) and other active compounds, 50215632, 50215590, 50215720, and 50215896. The pharmacophore model had a total of nine features which included one aromatic ring, one hydrophobic group, two hydrogen-bond donors, and three hydrogen-bond acceptors. This model will be of substantial help in design and development of novel caspase inhibitors. **Figures 6A, B** shows the pharmacophoric features of the most active

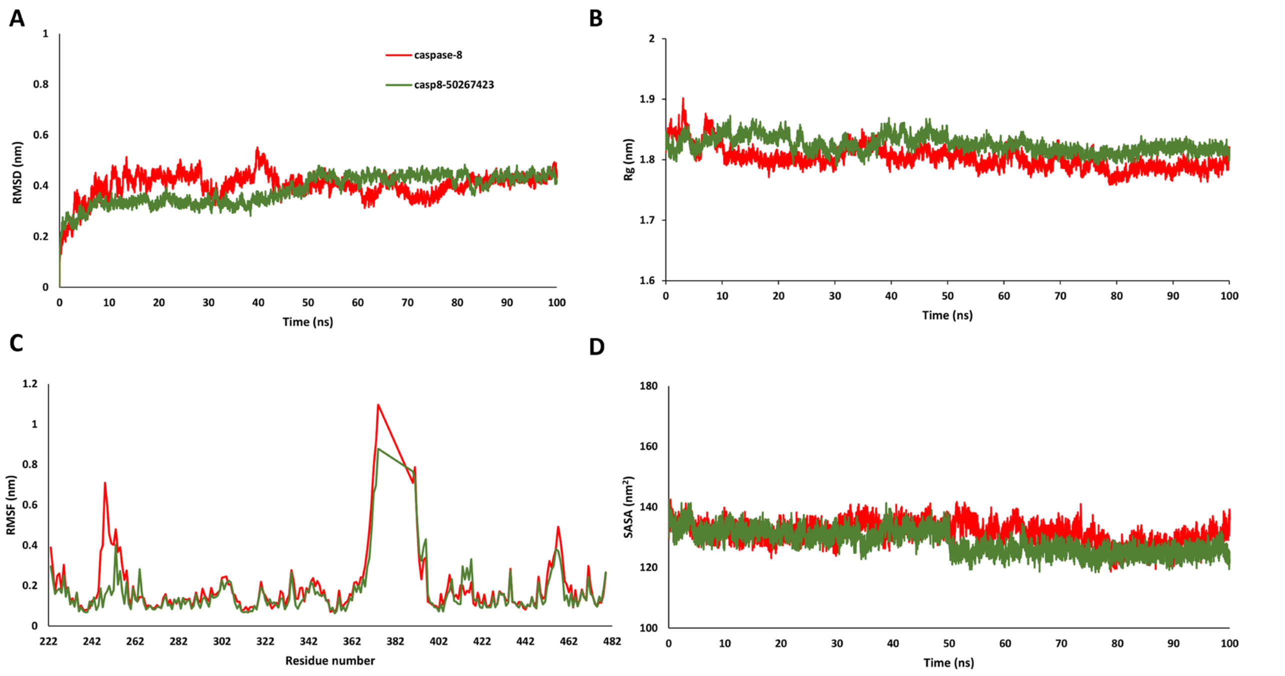


**FIGURE 3 | (A)** Root mean square deviation, **(B)** radius of gyration, **(C)** root mean square fluctuation, and **(D)** solvent accessible surface area plots for caspase8-50267423 complex.

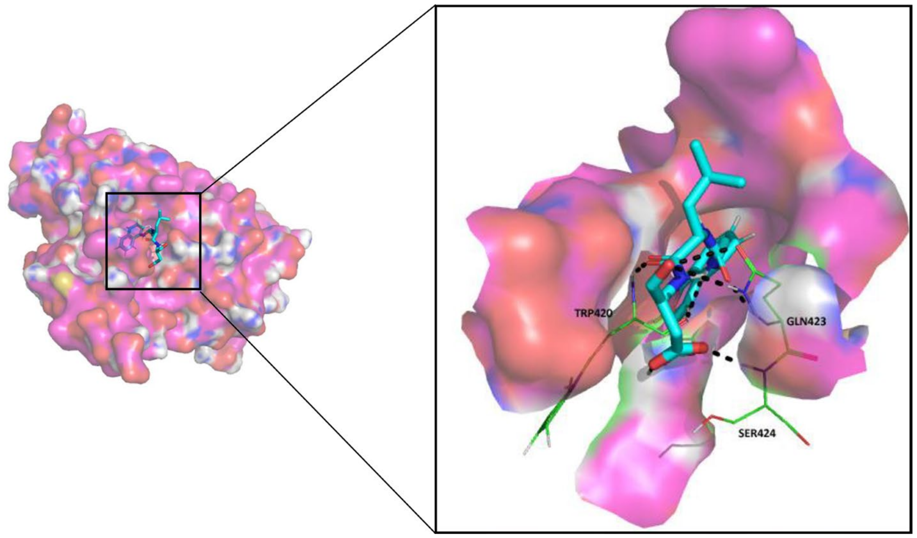


**FIGURE 4 |** The hydrogen bonding in caspase8-50267423 complex.

ligand, 50267423, and alignment of other active ligands to the pharmacophore model. A total of 129 hits were obtained which matched the pharmacophoric features of the most active compound 50267423. The ZINC IDs of the 129 hits have been provided in supporting information. The molecular docking analysis of the top five leads revealed that the XP scores of the compounds ranged between −10.775 and −9.423 (**Table 5**).

## DISCUSSION

AD is a chronic progressive long-term neurodegenerative disorder that affects millions of people worldwide and thus needs immediate attention. The current drugs available in the market can only temporarily improve upon the symptoms and delay the progression of the disease but could not stop it from progressing

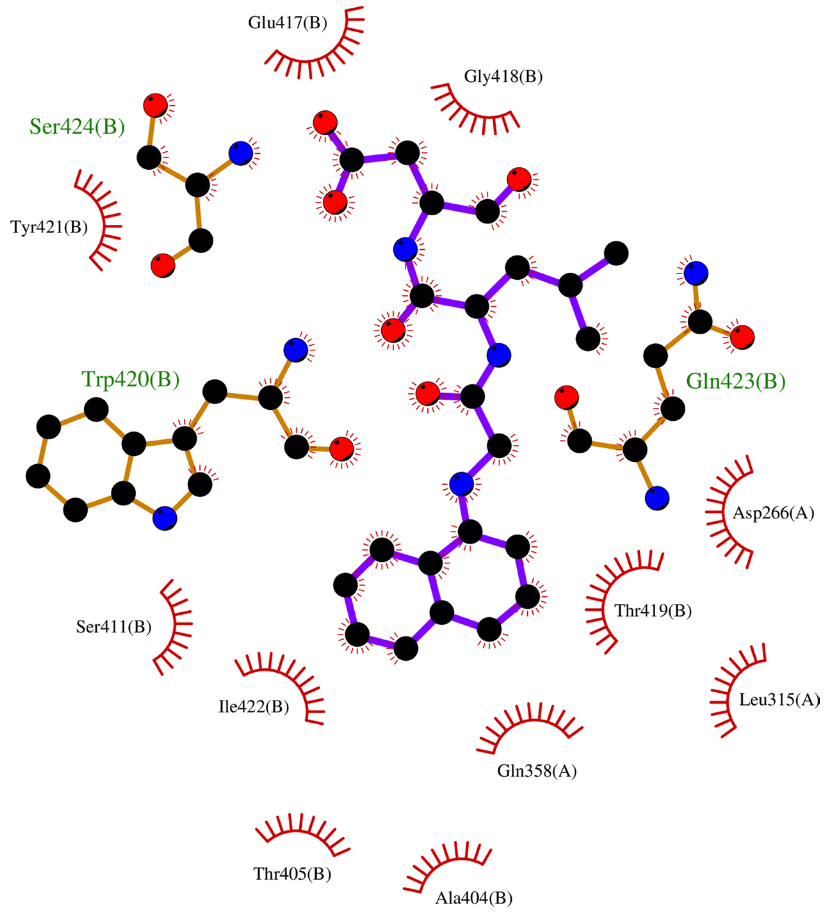


**FIGURE 5 |** The hydrophobic interactions in caspase8-50267423 complex.

and deteriorating the cognitive functions further. This study is based on the hypothesis that incorporating protein-ligand interactions for lead prioritization could lead to identification of compounds with highest binding affinities. In our previous studies, we had used molecular descriptors of chemical compounds to generate ML models for the classification of biologically active compounds (Jamal et al., 2015; Jamal et al., 2017). The properties extracted from MD trajectories have not been yet used for the classification of active compounds. The present work involved generation of ML models based on MD trajectories for prioritization of chemical compounds and lead optimization. Using Glide, we performed molecular docking of caspase-8-associated compounds and performed 10-ns MD simulations of top scoring conformation of each ligand and caspase-8 protein-ligand complex. Several 2D and 3D descriptors were generated, and MD descriptors were obtained from MD simulation trajectories. Various feature selection, Remove Useless filter, CFS, and relief-based attribute selection techniques were used to identify a subset of features having high contribution toward classification. The predictive models were generated using 2D, 3D, and MD descriptors and their combinations, 2D+3D, 2D+MD, 3D+MD, and 2D+3D+MD. Two ML algorithms, ANN and RF, were used for model building. The results obtained indicated that the MD descriptors performed better than 2D and 3D descriptors individually as well as in combinations. The MD descriptors clearly improved the classification performance of the models thus suggesting that the longer simulations as well as the MD descriptors in combination with 2D and 3D descriptors could lead to accurate and efficient lead optimization and prioritization. Another study conducted by Ash and Fourches in 2017 also confirmed the hypothesis that the descriptors extracted from MD trajectories are highly informative descriptors and could be effectively used not only for screening chemical libraries but for drug candidate design and prioritization (Ash and Fourches, 2017). Additionally, we also used a nine-point pharmacophore model consisting of three hydrogen-bond acceptor, two hydrogen-bond donors, one hydrophobic group, and one aromatic ring. This pharmacophore model was used for virtual screening of ZINC library of chemical compounds which led to the identification of 129 hits. The five lead compounds were subjected to molecular docking analysis which resulted in compounds having docking scores between −10.775 and −9.423 indicating that these compounds could be used as potential caspase-8 inhibitors.

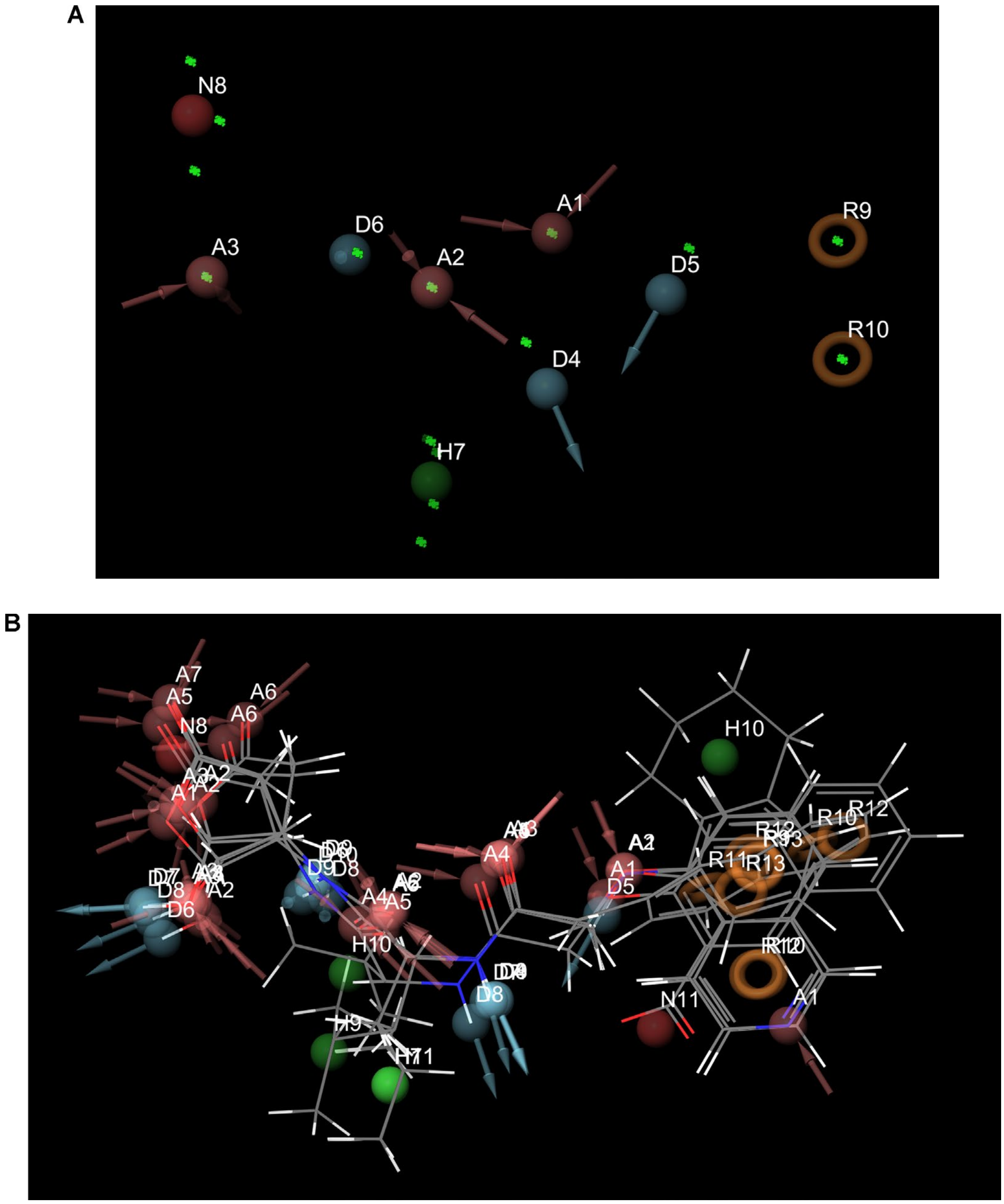


**FIGURE 6 | (A)** and **(B)** The pharmacophoric features of the most active ligand, 50267423, and alignment of other active ligands to the pharmacophore model. The color classification of the features is hydrogen bond acceptor (red), hydrogen bond donor (blue), hydrophobic (green), and aromatic ring (orange).

**TABLE 5 |** The molecular docking analysis of the top five ZINC compounds obtained after virtual screening using pharmacophore.

| ZINC database ID | Glide XP score | Interacting residues (hydrogen bond) |
|---|---|---|
| ZINC38200481 | −10.775 | Arg260 (2), Gln358 (1), Arg413 (3) |
| ZINC01576107 | −10.775 | Arg260 (2), Gln358 (1), Arg413 (3) |
| ZINC02384806 | −10.729 | Arg260 (1), Gln358 (1), Arg413 (2) |
| ZINC38570006 | −9.702 | Arg260 (2), Gln358 (1), Ser411(1), Arg413 (4) |
| ZINC38569951 | −9.423 | Arg260 (2), Gln358 (1), Ser411(1), Arg413 (4) |

## CONTRIBUTION TO THE FIELD STATEMENT

Dementia is a syndrome, usually chronic or progressive in nature, which leads to decline in cognitive function resulting in loss of ability of thinking and performing routine activities and majorly effects elderly population. Alzheimer's is a progressive disease during which the symptoms of dementia get worse over time. The current treatment regimen can only improve upon the systems for short term causing a temporary relief though cannot stop the disease from progression. Thus, there is a need of better treatment options which can stop the development of the disease. The high throughput screening studies have resulted in large number of compounds among which many compounds are in clinical trials and can be potential drugs against AD. However, selection of compounds with huge potential activity against Alzheimer's remains a problem to be addressed. The present study involves generation of predictive classification models using molecular dynamics descriptors which could lead to the identification of bioactive compounds and aid lead optimization and prioritization.

## SUPPORTING INFORMATION

Supporting information includes final data set including the molecule identifiers, SMILES and pIC50 values and the training and testing files and the models generated in the present study.

## DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

## AUTHOR CONTRIBUTIONS

SJ conceived, designed and performed the experiments. SJ, SG, and AG analyzed the data. SG contributed reagents/materials/analysis tools. All authors contributed to the writing of the manuscript.

## ACKNOWLEDGMENTS

SJ acknowledges a Young Scientist Fellowship from the Department of Health Research (DHR), India. AG and SG are grateful to University Grants Commission, India for the Faculty Recharge Position. SG is grateful to Jamia Hamdard for DST Purse grant.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00780/full#supplementary-material

## REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001

Ash, J., and Fourches, D. (2017). Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *J. Chem. Inf. Mode* 57, 1286–1299. doi: 10.1021/acs.jcim.7b00048

Alzheimer's Association (2018). 2018 Alzheimer's disease facts and figures. *Alzheimers Dement*. 14, 367–429. doi: 10.1016/j.jalz.2018.02.001

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., and Jones, E. (2011). Alzheimer's disease. *Lancet* 377, 1019–1031. doi: 10.1016/S0140-6736(10)61349-9

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., et al. (2010). WEKA—experiences with a java open-source project. *J. Mach. Learn. Res.* 11, 2533–2541.

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Castro, R. E., Santos, M. M., Gloria, P. M., Ribeiro, C. J., Ferreira, D. M., Xavier, J. M., et al. (2010). Cell death targets and potential modulators in Alzheimer's disease. *Curr. Pharm. Des.* 16, 2851–2864. doi: 10.2174/138161210793176563

Cheng, J., Tegge, A. N., and Baldi, P. (2008). Machine learning methods for protein structure prediction. *IEEE Rev. Biomed. Eng.* 1, 41–49. doi: 10.1109/RBME.2008.2008239

Cotman, C. W., Poon, W. W., Rissman, R. A., and Blurton-Jones, M. (2005). The role of caspase cleavage of tau in Alzheimer disease neuropathology. *J. Neuropathol. Exp. Neurol.* 64, 104–112. doi: 10.1093/jnen/64.2.104

Dickson, D. W. (2004). Apoptotic mechanisms in Alzheimer neurofibrillary degeneration: cause or effect? *J. Clin. Invest.* 114, 23–27. doi: 10.1172/JCI22317

Firoozpour, L., Sadatnezhad, K., Dehghani, S., Pourbasheer, E., Foroumadi, A., Shafiee, A., et al. (2012). An efficient piecewise linear model for predicting activity of caspase-3 inhibitors. *Daru* 20, 31. doi: 10.1186/2008-2231-20-31

Fourches, D., Muratov, E., and Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Mode* 50, 1189–1204. doi: 10.1021/ci100176x

Galvan, V., Gorostiza, O. F., Banwait, S., Ataie, M., Logvinova, A. V., Sitaraman, S., et al. (2006). Reversal of Alzheimer's-like pathology and behavior in human APP transgenic mice by mutation of Asp664. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7130–7135. doi: 10.1073/pnas.0509695103

Gamblin, T. C., Chen, F., Zambrano, A., Abraha, A., Lagalwar, S., Guillozet, A. L., et al. (2003). Caspase cleavage of tau: linking amyloid and neurofibrillary tangles in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10032–10037. doi: 10.1073/pnas.1630428100

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–1107. doi: 10.1093/nar/gkr777

Gervais, F. G., Xu, D., Robertson, G. S., Vaillancourt, J. P., Zhu, Y., Huang, J., et al. (1999). Involvement of caspases in proteolytic cleavage of Alzheimer's amyloid-beta precursor protein and amyloidogenic a beta peptide formation. *Cell* 97, 395–406. doi: 10.1016/S0092-8674(00)80748-5

Gramatica, R. T. P. (1997). The Whim Theory: new 3D molecular descriptors for Qsar in environmental modelling. *SAR. QSAR. Environ. Res.* 7, 89–115. doi: 10.1080/10629369708039126

Guillozet-Bongaarts, A. L., Garcia-Sierra, F., Reynolds, M. R., Horowitz, P. M., Fu, Y., Wang, T., et al. (2005). Tau truncation during neurofibrillary tangle evolution in Alzheimer's disease. *Neurobiol. Aging* 26, 1015–1022. doi: 10.1016/j.neurobiolaging.2004.09.019

Halgren, T. A., Muphy, R., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. New Zealand, The University of Waikato: Department of Computer Science. Hamilton.

Jamal, S., Goyal, S., Shanker, A., and Grover, A. (2015). Checking the STEP-associated trafficking and internalization of glutamate receptors for reduced cognitive deficits: a machine learning approach-based cheminformatics study and its application for drug repurposing. *PLoS One* 10, e0129370. doi: 10.1371/journal.pone.0129370

Jamal, S., Goyal, S., Shanker, A., and Grover, A. (2017). Machine learning and molecular dynamics based insights into mode of actions of insulin degrading enzyme modulators. *Comb Chem. High Throughput Screen* 20, 279–291. doi: 10.2174/1386207320666170130144443

Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection." in *Machine Learning Proceedings 1992*, eds. D. Sleeman and P. Edwards

(San Francisco, CA: Morgan Kaufmann), 249–256. doi: 10.1016/ B978-1-55860-247-2.50037-1

Koes, D. R., and Camacho, C. J. (2012). ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* 40, W409–414. doi: 10.1093/nar/gks378

Legewie, S., Bluthgen, N., and Herzel, H. (2006). Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. *PLoS Comput. Biol.* 2, e120. doi: 10.1371/journal.pcbi.0020120

Olsson, M. H., Sondergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: consistent treatment of internal and surface residues in Empirical pKa predictions. *J. Chem. Theory Comput.* 7, 525–537. doi: 10.1021/ct100578z

Parasuraman, S. (2012). Protein data bank. *J. Pharmacol. Pharmacother.* 3, 351–352. doi: 10.4103/0976-500X.103704

Rehker, J., Rodhe, J., Nesbitt, R. R., Boyle, E. A., Martin, B. K., Lord, J., et al. (2017). Caspase-8, association with Alzheimer's disease and functional analysis of rare variants. *PLoS One.* 12, e0185777. doi: 10.1371/journal.pone.0185777

Riniker, S. (2017). Molecular dynamics fingerprints (MDFP): machine learning from MD Data to predict free-energy differences. *J. Chem. Inf. Mode* 57, 726–741. doi: 10.1021/acs.jcim.6b00778

Rissman, R. A., Poon, W. W., Blurton-Jones, M., Oddo, S., Torp, R., Vitek, M. P., et al. (2004). Caspase-cleavage of tau is an early event in Alzheimer disease tangle pathology. *J. Clin. Invest.* 114, 121–130. doi: 10.1172/JCI200420640

Rohn, T. T., Head, E., Nesse, W. H., Cotman, C. W., and Cribbs, D. H. (2001). Activation of caspase-8 in the Alzheimer's disease brain. *Neurobiol. Dis.* 8, 1006–1016. doi: 10.1006/nbdi.2001.0449

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi: 10.1007/s10822-013-9644-8

Schneidman-Duhovny, D., Dror, O., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2008). PharmaGist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Res.* 36, W223–228. doi: 10.1093/nar/gkn187

Sharma, S., Basu, A., and Agrawal, R. K. (2013). Pharmacophore modeling and docking studies on some nonpeptide-based caspase-3 inhibitors. *Biomed. Res. Int.* 2013, 306081. doi: 10.1155/2013/306081

Small, D. H., Mok, S. S., and Bornstein, J. C. (2001). Alzheimer's disease and Abeta toxicity: from top to bottom. *Nat. Rev. Neurosci.* 2, 595–598. doi: 10.1038/35086072

Steiner, T. (1999). "Weak Hydrogen Bonds," in *Implications of Molecular and Materials Structure for New Technologies*. Eds. J. A. K. Howard, F. H. Allen, and G. P. Shields (Dordrecht: Springer), 360. doi: 10.1007/978-94-011-4653-1_13

Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med* 18, 421–430. doi: 10.1038/gim.2015.117

Wahi, D., Jamal, S., Goyal, S., Singh, A., Jain, R., Rana, P., et al. (2015). Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents. *Syst. Synth. Biol.* 9, 33–43. doi: 10.1007/s11693-015-9162-1

Wang, Q., Mach, R. H., and Reichert, D. E. (2009). Docking and 3D-QSAR studies on isatin sulfonamide analogues as caspase-3 inhibitors. *J. Chem. Inf. Mode* 49, 1963–1973. doi: 10.1021/ci900144x

Watt, W., Koeplinger, K., Mildner, A. M., Heinrikson, R. L., Tomasselli, A. G., and Watenpaugh, K. D. (1999). The atomic-resolution structure of human caspase-8, a key activator of apoptosis. *Structure* 7, 1135–1143. doi: 10.1016/S0969-2126(99)80180-4

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466–1474. doi: 10.1002/jcc.21707

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 22 August 2019
doi: 10.3389/fphar.2019.00924

# Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods

*Dingyan Wang[1,2], Chen Cui[1,2], Xiaoyu Ding[1,2], Zhaoping Xiong[3], Mingyue Zheng[1]*, Xiaomin Luo[1]*, Hualiang Jiang[1,3] and Kaixian Chen[1,3]*

*[1] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China, [2] College of Pharmacy, University of Chinese Academy of Sciences, Beijing, China, [3] School of Life Science and Technology, ShanghaiTech University, Shanghai, China*

Scoring functions play an important role in structure-based virtual screening. It has been widely accepted that target-specific scoring functions (TSSFs) may achieve better performance compared with universal scoring functions in actual drug research and development processes. A method that can effectively construct TSSFs will be of great value to drug design and discovery. In this work, we proposed a deep learning–based model named DeepScore to achieve this goal. DeepScore adopted the form of PMF scoring function to calculate protein–ligand binding affinity. However, different from PMF scoring function, in DeepScore, the score for each protein–ligand atom pair was calculated using a feedforward neural network. Our model significantly outperformed Glide Gscore on validation data set DUD-E. The average ROC-AUC on 102 targets was 0.98. We also combined Gscore and DeepScore together using a consensus method and put forward a consensus model named DeepScoreCS. The comparison results showed that DeepScore outperformed other machine learning–based TSSFs building methods. Furthermore, we presented a strategy to visualize the prediction of DeepScore. All of these results clearly demonstrated that DeepScore would be a useful model in constructing TSSFs and represented a novel way incorporating deep learning and drug design.

**Keywords: virtual screening, target-specific scoring function, deep learning, drug discovery, DUD-E**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Cao Dongsheng,*
*Central South University, China*
*Hongming Chen,*
*AstraZeneca, Sweden*

****Correspondence:***
*Mingyue Zheng*
*myzheng@simm.ac.cn*
*Xiaomin Luo*
*xmluo@simm.ac.cn*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *20 May 2019*
***Accepted:*** *22 July 2019*
***Published:*** *22 August 2019*

***Citation:***
*Wang D, Cui C, Ding X, Xiong Z, Zheng M, Luo X, Jiang H and Chen K (2019) Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods. Front. Pharmacol. 10:924. doi: 10.3389/fphar.2019.00924*

## INTRODUCTION

Structure-based drug design (SBDD) has been widely used in industry and academia (Andricopulo et al., 2009; Morrow et al., 2012). There are three main categories of tasks for SBDD methods: virtual screening, *de novo* drug design, and ligand optimization. Virtual screening generally refers to the process of identifying active compounds among molecules selected from a virtual compound library. By utilizing the three-dimensional information of proteins, structure-based virtual screening is believed to be more efficient than traditional virtual screening methods. The key factor for guaranteeing the success of structure-based virtual screening is the quality of scoring functions. Theoretically, a scoring function is capable of predicting the binding affinity of a protein–ligand complex structure, and thus can be used for predicting the binding pose of a ligand or screening a virtual compound library to find potential active compounds.

Classic scoring functions can be divided into three categories: force field–based, knowledge-based, and empirical (Liu et al., 2017). For a long time, researchers have found that machine learning and deep learning methods had an excellent performance in helping constructing different kinds of scoring functions. Especially recently, convolutional neural network (CNN) utilizing the structural information of protein–ligand complexes has shown promise in predicting binding affinity and virtual screening (Ragoza et al., 2017; Stepniewska-Dziubinska et al., 2018). A deep learning model constructed using CNN by Imrie et al. represented the state-of-the-art on several virtual screening benchmarks (Imrie et al., 2018). However, the authors also found that fine-tuning a general model on subsets of a specific protein family resulted in a significant improvement. This reflects the fact that no single scoring function is suitable for every target. Moreover, in practice, a medicinal chemist is usually concerned about only one target at a time and hope that the scoring function he uses has the best performance on this target. The most common and direct way to address this issue is to build a target-specific scoring function (TSSF) for the specific target. TSSFs have been widely used in virtual screening campaign and proved to be useful in variable kinds of important drug targets including kinases (Xu et al., 2017; Berishvili et al., 2018) and GPCRs (Kooistra et al., 2016).

Based on the fact mentioned above, it is of great value to design a method that can effectively construct TSSFs. Several methods have been proposed to address this problem. In 2005, Antes et al. presented a model called Parameter Optimization using Ensemble Methods (POEM) which applied the design of experiments (DOE) approach and ensemble methods to the optimization of TSSFs in molecular docking (Antes et al., 2005). They fitted FlexX and ScreenScore to the kinase and ATPase protein classes and got a promising result. In 2010, Xue et al. developed a kinase-specific scoring function named kinase-PMF in order to score ATP-competitive inhibitors (Xue et al., 2010). Their work showed that TSSFs achieved better performance compared with general scorings. In 2011, Li et al. proposed a scoring function building strategy named SVM-SP based on support vector machine (SVM) (Li et al., 2011). They tailored SVM-SP to each target in the test set and found that it outperformed many other scoring functions including Glide. In 2015, Wang et al. introduced a strategy named TS-Chemscore to build TSSFs based on a known universal scoring function by a regression process on energy contributions (Wang et al., 2015). In 2017, Yan et al. used a residue-based interaction decomposition method with SVM to develop a target-specific discrimination model called protein–ligand empirical interaction components-SVM (PLEIC-SVM) (Yan et al., 2017). Their results showed that PLEIC-SVM was a useful tool in filtering the docking poses.

Here, we introduce a deep learning–based method named DeepScore used for constructing TSSFs. The purpose of DeepScore is rescoring the docking poses generated from docking software like Glide. DeepScore uses the scoring model of PMF scoring function, where the score for a protein–ligand complex is derived from the sum of protein–ligand atom pair-wise interactions within a distance range. The score for a single protein–ligand atom pair is calculated using a fully connected neural network. Since consensus scoring methods have shown to be useful in improving the performance considering the results from several different models (Teramoto and Fukunishi, 2008; Ericksen et al., 2017), we further proposed DeepScoreCS by combining the results of DeepScore and Glide Gscore together. The directory of useful decoys–enhanced (DUD-E) was used as the benchmark to quantitatively assess the model. 12 metrics were calculated and used for making comparison between Gscore, DeepScore, DeepScoreCS, and some other TSSF models reported by previous studies.

# MATERIALS AND METHODS

## Data Preparation

The directory of useful decoys–enhanced (DUD-E) benchmarking set (Mysinger et al., 2012) was used for training and evaluating the model. DUD-E is a data set designed for helping benchmark docking software and scoring functions. There are 102 targets in DUD-E. Each target is provided with 224 active ligands and 13,835 decoys on average. DUD-E has been widely used for evaluating the virtual screen ability of scoring functions (Chaput et al., 2016; Ericksen et al., 2017; Ragoza et al., 2017; Yan et al., 2017; Imrie et al., 2018). Although it has been reported by some literature that there exists noncausal bias in DUD-E (Sieg et al., 2019), we still use it to evaluate our model since there is no better data set so far.

The first step is to generate docking poses for actives and decoys. We noticed that, in other similar work, a variety of docking methods were used in this step, including Glide (Yan et al., 2017), AutoDock Vina (Imrie et al., 2018), DOCK (Pereira et al., 2016), PLANTS (Kurkinen et al., 2018), and so on. Even using the same docking program, sometimes different docking protocols were adopted (Chaput et al., 2016; Yan et al., 2017). It should be emphasized that, strictly speaking, only the rescoring results from the same docking poses are comparable.

Since the ligands in DUD-E have been assigned appropriate protonation states, we followed the approach in (Chaput et al., 2016) that ligands were used without any modified. Receptors were prepared with protein preparation wizard from Schrodinger suit (Schrödinger, LLC, New York, NY, 2015-2). Ligands were docked using Glide (Friesner et al., 2006) in SP mode with default options.

## Descriptors and Model

Through data preparation step, the best poses ranked by Gscore were selected for actives and decoys. To rescore the docking poses from Glide, we utilized the form of the potential of mean-force (PMF) scoring function (Muegge and Martin, 1999) to calculate the score for each protein–ligand complex. In PMF scoring function, the score for a complex is defined as the sum of overall protein–ligand atom pair-wise interactions within a specific cutoff radius:

$$PMFScore_{complex} = \sum_i \sum_j A\left(i, j, distance_{ij}\right) \text{ for } distance_{ij} < cutoff\ distance \quad (1)$$

where i is the ligand atom, j is the receptor atom, $distance_{ij}$ is the distance between atom i and atom j, and A is the function used for calculating the PMF between atom i and atom j.

In Pafnucy (Stepniewska-Dziubinska et al., 2018), a structure-based CNN model, 19 features were used for describing an atom. In DeepScore, almost same features but with minor modifications were used (see **Table 1**). The features included the information of atom type, hybridization state, heavy valence, hetero valence, partial charge, and whether the atom was aromatic/hydrophobic/hydrogen-bond donor/hydrogen-bond acceptor/in a ring. Heavy valence and hetero valence were represented as one-hot vectors in DeepScore instead of integers in Pafnucy.

**TABLE 1 |** Atom features used in DeepScore.

| Atom feature name | Feature length | Features |
|---|---|---|
| Type | 9 | B, C, N, O, P, S, Se, halogen, and metal |
| Hybridization | 4 | 1, 2, 3, other |
| Heavy valence[a] | 4 | 1, 2, 3, other |
| Hetero valence[b] | 5 | 0, 1, 2, 3, other |
| Partial charge | 1 | Value |
| Hydrophobic | 1 | 1 (True) or 0 (false) |
| Aromatic | 1 | 1 (True) or 0 (false) |
| Hydrogen-bond donor | 1 | 1 (True) or 0 (false) |
| Hydrogen-bond acceptor | 1 | 1 (True) or 0 (false) |
| Ring | 1 | 1 (True) or 0 (false) |

[a] *The number of bonds with other heavy atoms.*
[b] *The number of bonds with other heteroatoms.*

Cutoff distance was changed to an accepted distance range in DeepScore. For each complex, atom pairs between 2 and 8 Å were sorted in the ascending order of length, and only 500 shortest pairs were taken into consideration. Distance was also discretized with bins equally distanced by 0.025 Å between 2 and 8 Å. The feature for a protein–ligand atom pair was comprised of the concatenation of the ligand atom feature vector, the protein atom feature vector, and the one-hot-encoded distance, which made the length of an atom pair feature 80 bins long (Eq. 2-1). The score for an atom pair (i-j) was calculated as Eq. 2-2 using a 2-hidden layer fully connected network. The sizes of weight matrix $W_1$, $W_2$, and $W_3$ were 80×128,128×64,64×1, respectively. $b_1$, $b_2$, and $b_3$ were biases. Rectified linear unit (ReLU) was used as activation function. Final score, or DeepScore, for a protein–ligand complex was calculated as Eq. 2-3. In Eq. 2-3, i and j refer to the ligand atom and the receptor atom respectively. All calculated scores of selected protein–ligand atom pairs were summed up to determine the final score. Overview of the workflow is also shown in **Figure 1**.

$$Feature_{ij} = concatenate\left(Feature_i,\ Distance_{ij}, Feature_j\right) \quad (2\text{-}1)$$

$$DeepScore_{ij} = W_3\left(\mathrm{ReLU}\left(W_2\left(ReLU\left(W_1 Feature_{ij} + b_1\right)\right) + b_2\right)\right) + b_3 \quad (2\text{-}2)$$

$$DeepScore_{complex} = \sum_{i-j} DeepScore_{ij} \text{ for selected atom pair } i - j \quad (2\text{-}3)$$

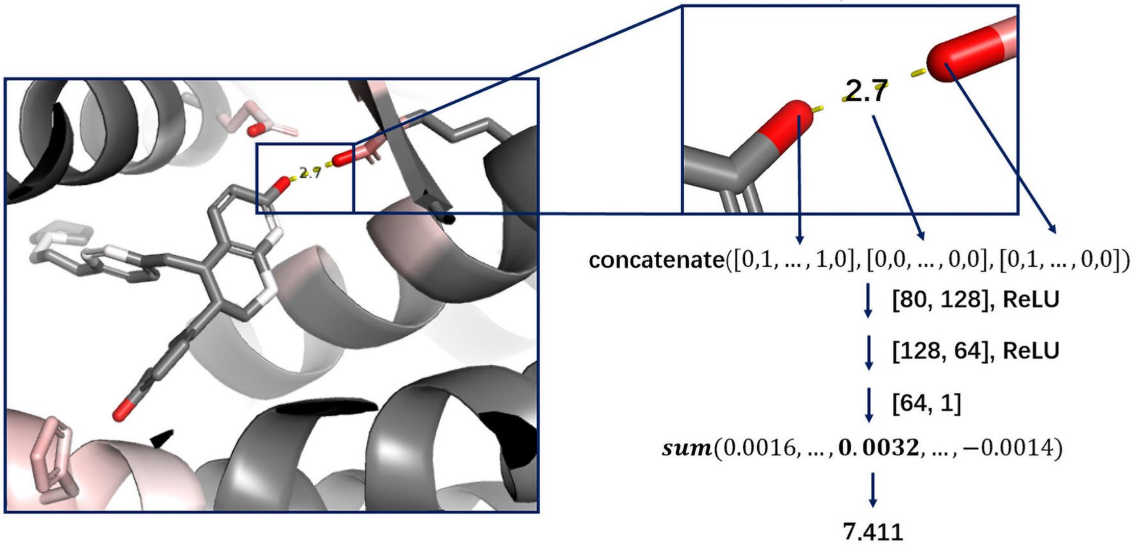

2.7
2.7
concatenate([0,1, … , 1,0], [0,0, … , 0,0], [0,1, … , 0,0])
[80, 128], ReLU
[128, 64], ReLU
[64, 1]
sum(0.0016, … , 0.0032, … , −0.0014)
7.411

**FIGURE 1 |** Workflow of DeepScore model construction.

## Loss Function

In deep learning processes, the usual practice while dealing with a two possible classification problem is to put two units in the output layer and transform the outputs using softmax function. The outputs, which represent the probability of classes 0 and 1, respectively, are then used for calculating the loss with cross entropy loss or other loss functions. However, in practice, we found that the cross entropy loss function did not apply to our model very well. We tried some other loss functions and found that modified Huber loss (Eq. 3) (Zhang, 2004) was more suitable. The formula of modified Huber loss is shown in Eq. 3, where f(x) refers to the output of the model and y refers to the label (1 for actives and -1 for decoys). It should be noted that, unlike general scoring functions, the possible scoring range of DeepScore is the entire real number filed. A score greater than zero indicates that the model considers the compound to be active, whereas a score less than zero is inactive. Another important point is that scores between different targets are not comparable.

$$\mathrm{L}\left(\mathrm{y},\ \mathrm{f(x)}\right)=\begin{cases}\left[\max\left(0,\ 1-yf(x)\right)\right]^2 & for\ \ yf(x)\geq -1,\\ -4yf(x) & otherwise.\end{cases} \tag{3}$$

## Training

Five-fold cross validation test was performed on each target in DUD-E. For each target, the whole data set was split into five parts at first. Within each fold, three parts were used as training set, one part as validation set, and one part as test set. The order of [(training set)/validation set/test set], we used during cross validation was ([1,2,3]/4/5), ([2,3,4]/5/1), ([3,4,5]/1/2), ([4,5,1]/2/3), and ([5,1,2]/3/4). Early stopping strategy was used for avoiding overfitting. For each training epoch, the area under the curve of precision recall curve (PRC-AUC) on validation set was calculated. If the performance did not improve within eight epochs, training was stopped, and the best model was saved and evaluated on test set. Mean value of the metrics of five folds on test set was calculated and used as the performance of the model. To make it fair, the performance of Gscore was also calculated in the same way.

It should be noticed that there existed a dramatic class imbalance in our data sets as the number of decoys was almost 50 times of that of actives. To overcome this problem, we adopted the random undersampling strategy. Over an epoch, we did not use the whole training set to train the model. Instead, parts of decoys were randomly selected out to make sure that the number of actives and decoys was the same in an epoch. The reason why we chose undersampling was that, compared with other methods like oversampling, the training procedure using this strategy was significantly faster.

Our model was implemented using PyTorch 1.0 (https://pytorch.org/) in python. Each model was trained using Adam optimizer with a batch size of 32, a learning rate of 0.001, and a weight decay of 0.001.

## Evaluation Metrics

The area under the curve of receiver operating characteristic curve (ROC-AUC), the PRC-AUC, enrichment factor (EF), and ROC Enrichment Factor (ROC-EF) were calculated for each fold in order to evaluate the performance of the model. ROC-AUC is a traditional metric for assessing the performance of a classification model. However, under the circumstance that the number of negative samples is obviously larger than the number of positive, like our mission, PRC-AUC is usually a more appropriate choice to replace ROC-AUC since ROC-AUC may not reflect the early enrichment ability of the model (Truchon and Bayly, 2007). EF is the fraction of actives within a certain percentage of ranking list divided by the fraction in whole data set. Because the way of calculating EF simulates actual virtual screening scenarios where only a small fraction of ligands are picked out to carry out biological test, EF is one of the gold standards used for evaluating ranking ability of scoring functions. ROC-EF is another metric recommended by Jain et al. (Jain and Nicholls, 2008) to quantify early enrichment. It refers to the rate of true-positive rate (TPR) to false-positive rate (FPR) at certain FPR. Both EF and ROC-EF were calculated at five different levels of percentage: 0.5%, 1%, 2%, 5%, and 10%. Thus, there were in all 12 metrics for evaluating the models.

## Consensus Scoring

When the correlation between the statistical errors of multiple models is low, combining the predicted values of these models in a certain way usually performs better than any single one model. This is the basic idea of ensemble learning (Dietterich, 2000). We adopted this strategy and used Eq. 4 to calculate DeepScoreCS for a complex. In Eq. 4, c is a coefficient that can be adjusted. More details will be showed and discussed in Results and Discussion part.

$$\mathrm{DeepScoreCS}=\mathrm{DeepScore}\times\mathrm{c}+\mathrm{Gscore}\times\left(1-\mathrm{c}\right),\ 0\leq\mathrm{c}\leq 1 \tag{4}$$

# RESULTS AND DISCUSSION

## Model Architecture

Deep learning models are usually regarded as black boxes since the information of which features that are important can hardly be interpreted from the model. Although CNN based scoring functions, like Pafnucy from which the atom features of DeepScore were borrowed, have achieved state-of-the-art performance in benchmark test, and become the representative of deep learning–based scoring functions, treating the whole protein–ligand complex as a 3D picture is still counterintuitive. Thus, in consideration of interpretation, we chose to reform the classic PMF scoring function. The neural network in DeepScore is only used to facilitate the learning of atom-pair potentials; meanwhile, the overall framework of PMF scoring function is preserved. DeepScore is able to directly give the score of each atom pair, which makes the model's output easy to explain. To the best of our knowledge, DeepScore is the first model to use this framework.

## Glide Screening

Glide docking results are provided in **Table S1**. For DUD-E data set, the mean value of ROC-AUC gained from Glide was 0.82, which showed a significant better screening ability compared with other docking software, like AutoDock Vina (0.703) (Imrie et al., 2018). To ensure the reliability of docking poses, we compared Boltzmann-enhanced discrimination ROC (BEDROC, α=80.5) of our results with (Chaput et al., 2016) on each target, since we used the same docking software and similar docking protocol with them. The scatter plotting is shown in **Figure 2**. Our results showed a high correlation with (Chaput et al., 2016), which ensured that the docking poses are credible.

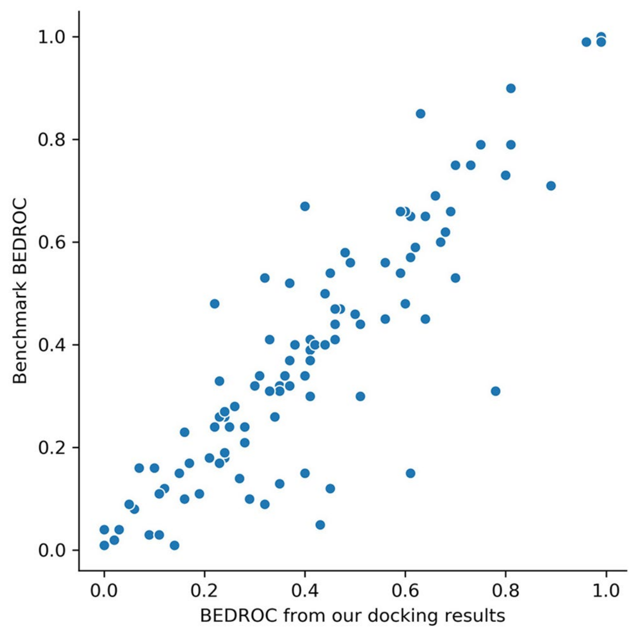


**FIGURE 2 |** BEDROC scores (α=80.5) on 102 targets of our screening results *versus* the results from benchmark (Chaput et al., 2016). Each dot represents a target.

## DeepScore

ROC-AUC, PRC-AUC, EF (0.5%, 1%, 2%, 5%, and 10%), ROC-EF (0.5%, 1%, 2%, 5%, and 10%) of Gscore, and DeepScore on all 102 targets were calculated (see **Figure 3**, **Table S2** and **Table S3**). **Figure 3** shows that DeepScore performs better than Gscore significantly. DeepScore had an excellent performance on ROC-AUC where all the targets showed an improvement *versus* Gscore. The mean values of 12 metrics were all increased by using DeepScore, as shown in **Table 2**.

The improvement of performance on some targets was obvious. For example, for target FPPS (farnesyl diphosphate synthase), the ROC-AUC of Gscore was 0.54, which indicated that Gscore just randomly scored actives and decoys on FPPS. On the other side, ROC-AUC of DeepScore was 1.00 which demonstrated that DeepScore could almost perfectly separate actives and decoys. Similar situation also arose in (Ragoza et al., 2017). In this study, authors found that AutoDock Vina got a worse-than-random ROC-AUC of 0.29 on FPPS, while the

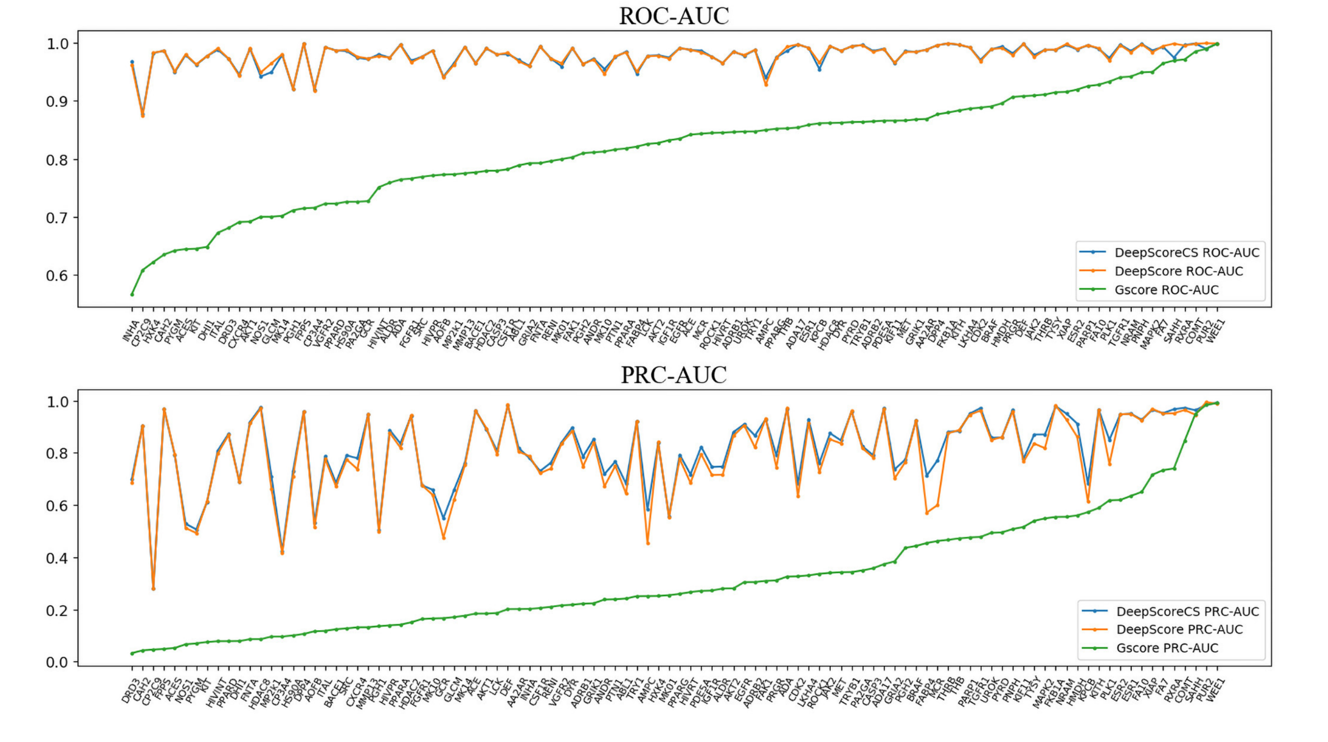


**FIGURE 3 |** ROC-AUC (upper panel) and PRC-AUC (lower panel) of cross validation performance on each target. Targets are sorted by the performance of Gscore.

**TABLE 2 |** Average performance of Gscore, DeepScore, and DeepScoreCS on DUD-E data set.

| | Gscore | DeepScore | | DeepScoreCS | |
|---|---|---|---|---|---|
| | Value | Value | Better-1[a] | Value | Better-2[b] |
| ROC-AUC | 0.817 | 0.979 | 102 | 0.978 | 49 |
| PRC-AUC | 0.317 | 0.796 | 100 | 0.814 | 81 |
| EF0.5% | 30.625 | 55.275 | 94 | 57.149 | 51 |
| EF1% | 24.335 | 52.218 | 98 | 53.658 | 65 |
| EF2% | 17.203 | 39.716 | 100 | 40.075 | 60 |
| EF5% | 9.122 | 18.200 | 102 | 18.200 | 40 |
| EF10% | 5.573 | 9.472 | 101 | 9.448 | 20 |
| ROC-EF0.5% | 51.522 | 148.948 | 100 | 151.986 | 66 |
| ROC-EF1% | 31.239 | 81.614 | 102 | 82.164 | 47 |
| ROC-EF2% | 18.689 | 43.320 | 102 | 43.498 | 42 |
| ROC-EF5% | 9.423 | 18.417 | 101 | 18.365 | 35 |
| ROC-EF10% | 5.680 | 9.500 | 101 | 9.484 | 28 |

[a] *Better-1 column refers to the number of targets where DeepScore outperforms Gscore.*

[b] *Better-2 column refers to the number of targets where DeepScoreCS outperforms DeepScore.*

"DUD-E only model" they trained also performed excellently with a ROC-AUC of 0.98. The authors supposed that the reason why AutoDock Vina performed so poorly was that the docking poses of actives were incorrect. However, we found that the wrong docking poses may not be the main reason. As is shown in **Figures 4A**, **B**, more than half of actives were docked correctly by Glide, where the bisphosphonate group chelated with the magnesium ions, but the performance of Gscore was still very poor. Despite this, we agree with (Ragoza et al., 2017) that the perfect performance of no matter their model or DeepScore was because of simply recognizing the biphosphate group or polarity of molecules since very few decoys possessed phosphorus. It is an extreme example but still highlights two facts. First, DUD-E data set exists the problem of obvious structure differences between decoys and actives, which may result in artificial enrichment during the evaluation of scoring functions and virtual screening methods. Second, TSSFs are more useful than universal scoring functions in the case where the subject is only a single target, because the factors that play a leading role in protein–ligand binding modes in different kinds of targets are different.

## DeepScoreCS

As has been mentioned in Methods part, we further investigated if consensus methods could improve the performance of the model in our mission. Eq. 4 was used for calculating the mixture model consensus scores of Gscore and DeepScore. It was important to set an appropriate coefficient c for Eq. 4, and obviously, the best c on each target should be different from each other. Grid searching was used for settling this problem. For each training fold, after the training had stopped, the scores on

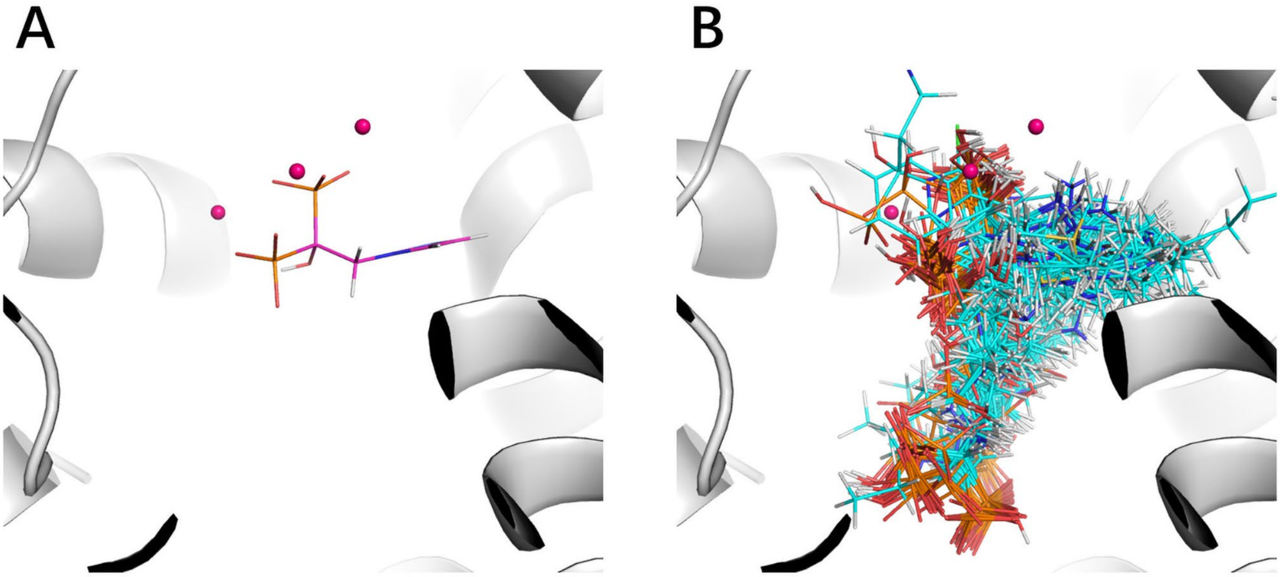


**FIGURE 4 |** The binding site of FPPS (PDB ID 1zw5). **(A)** Crystal structure ligand. **(B)** Superposition of all the docking poses of actives.

validation set were determined by the best DeepScore model. Then, different coefficient c from 0 to 1 with step 0.01 was chosen to calculate DeepScoreCS scores on validation set according to Eq. 4. The coefficient c with best PRC-AUC on validation set was used on test set to evaluate the performance of DeepScoreCS. The results are shown in **Table 2**. It can be seen that the improvement of performance by conducting consensus experiment is not obvious. The mean values of PRC-AUC, EF0.5%, EF1%, EF2%, ROC-EF0.5%, ROC-EF1%, and ROC-EF2% increased slightly, while the rest metrics decreased. Most of targets (81/102) got higher PRC-AUC. To investigate whether the performance of the model may actually benefit from consensus methods, we quantitatively examined the improvement of PRC-AUC on each target. The results are presented in **Figure 5**. In **Figure 5**, each point represents a target, X-axis represents the best coefficient c (mean value of five folds) on this target, and Y-axis represents the improvement on PRC-AUC, which is calculated by the PRC-AUC of DeepScoreCS minus that of DeepScore. Targets with higher PRC-AUC are painted blue, and targets with lower PRC-AUC are painted red. It can be noticed that, although on most targets, the impact of consensus strategy was just random perturbation ($|\Delta AUC| < 0.025$), no target got a significant decrease on AUC ($\Delta AUC < -0.025$). On the other hand, for more than 20 targets, $\Delta$AUC was larger than 0.025. Especially for three targets (AMPC, MCR, and FABP4), the increase of AUC was significant ($\Delta$AUC > 0.1). These results demonstrated that the consensus method was worthy of trying since it would not weaken the performance of the model, and for few targets, the performance would be significantly improved.

## Comparing With Previous Studies

We compared our results with two previous similar studies to check if our model showed better performance.

First, we compared the performance of DeepScore with PLEIC-SVM constructed by Yan et al. (2017). They used 36 targets to train and test their model, so we selected the scores of overlapped targets to make comparison. The results are shown in **Table 3** and **Figure 6**. **Table 3** clearly indicates that DeepScore performed better than PLEIC-SVM. The average ROC-AUC, ROC0.5%, ROC%1, ROC2%, and ROC5% (ROC10% of PLEIC-SVM was not provided) for all 36 targets increased from 0.93, 0.58, 0.64, 0.69, and 0.77 to 0.98, 0.78, 0.85, 0.89, and 0.94, respectively, by using DeepScore. Among these metrics, ROC0.5% is the most important one since the early enrichment ability of scoring functions is paid more attention in the context of virtual screening. **Figure 6** shows that DeepScore outperforms on most of the targets on ROC0.5%. On some targets, such as FNTA, the improvement was dramatic (for FNTA, ROC0.5% increased from 0.31 to 0.92 by using DeepScore). However, for GCR, CDK2, BACE1, and PRGR, DeepScore only got a similar or slightly worse performance.

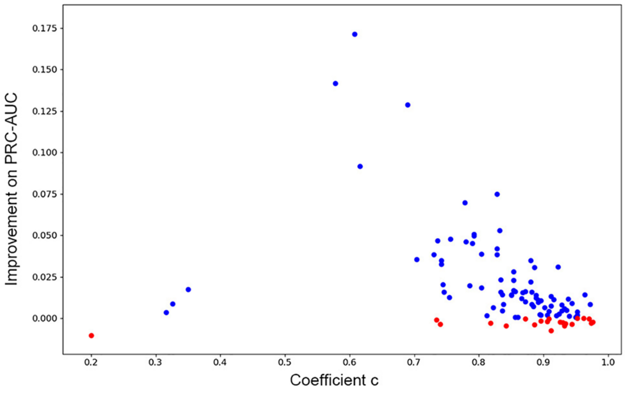


**FIGURE 5 |** The improvement of PRC-AUC on each target using consensus method. Each point represents a target. Y-axis represents the value of PRC-AUC of DeepScoreCS minus that of DeepScore. Blue dot means that the improvement is positive while red means negative (the performance became worse through consensus method). X-axis represents the mean value of the coefficient c DeepScoreCS used.

The workflow of PLEIC-SVM included a process of tuning parameters for SVM model. It should be noticed that, limited by the huge number of targets, we did not perform hyperparameter optimization for every model. In another word, all models were trained under the same set of hyperparameters (learning rate, network structure, etc.). Considering the fact that hyperparameters may significantly affect the performance of machine learning models (also pointed out by (Yan et al., 2017)), it is reasonable to infer that the performance of DeepScore will be further improved by hyperparameter optimization.

We also compared our model with RF-Score. Wójcikowski et al. adopted the same protocol (DUD-E, single target, five-fold cross validations) to evaluate the target-specific virtual screening ability of RF-Score (Wójcikowski et al., 2017). Descriptors from three versions of RF-Score and ligand binding conformations from three docking programs (AutoDock Vina, DOCK 3.6, and DOCK 6.6) were used for training the model. In all, nine RF-Score models were obtained for testing in their study. The comparison results are presented in **Table 4**. It shows that DeepScore outperforms the nine RF-Score models on all of the metrics.

## Sensitivity to Docking Program

Above results have shown that DeepScore works well with the docking poses generated from Glide. To examine whether

**TABLE 3 |** Performance comparison between PLEIC-SVM and DeepScore.

| | PLEIC-SVM | DeepScore |
|---|---|---|
| ROC-AUC | 0.93 | **0.98** |
| ROC0.5%$^a$ | 0.58 | **0.78** |
| ROC1%$^b$ | 0.64 | **0.85** |
| ROC2%$^c$ | 0.69 | **0.89** |
| ROC5%$^d$ | 0.77 | **0.94** |

*Performance values of PLEIC-SVM are collected from (Yan et al., 2017). Better results are highlighted in bold.*

$^a$ *ROC0.5% = ROC-EF0.5% / 200.*

$^b$ *ROC1% = ROC-EF1% / 100.*

$^c$ *ROC2% = ROC-EF2% / 50.*

$^d$ *ROC5% = ROC-EF5% / 20.*

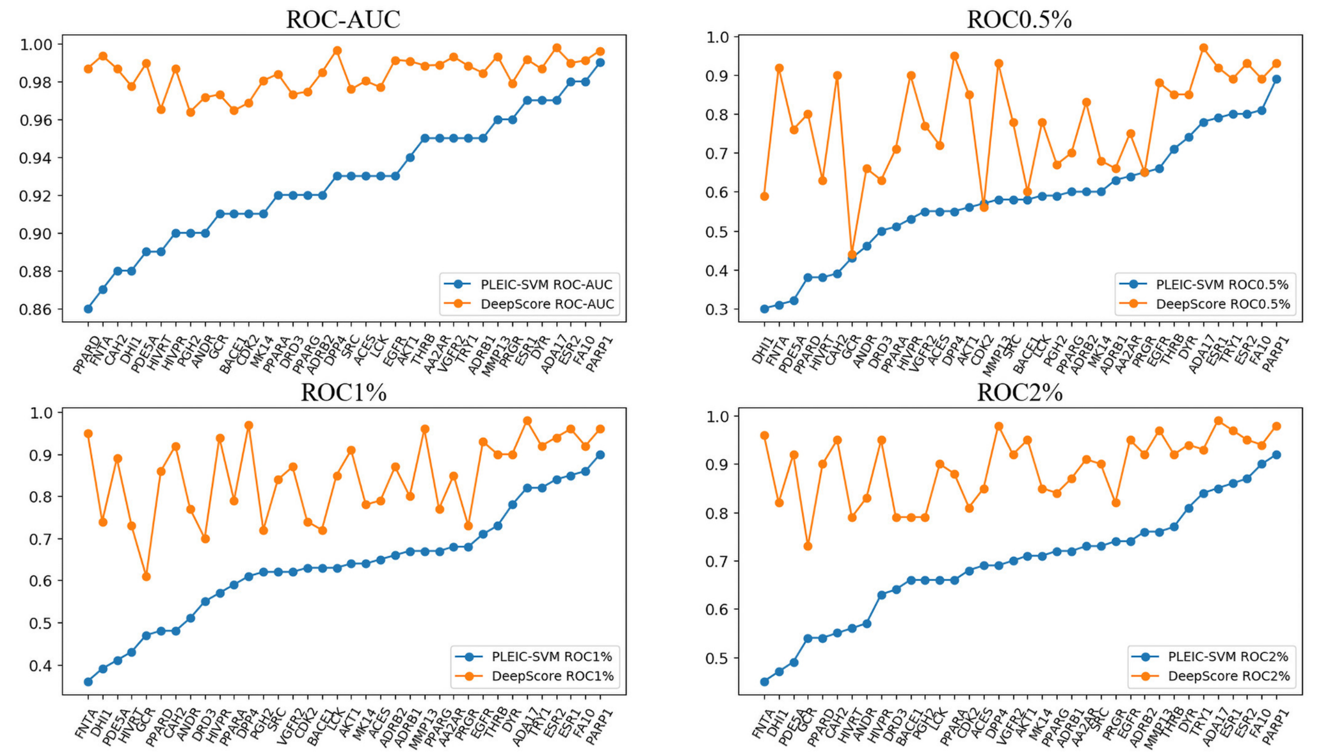


**FIGURE 6 |** The performance of PLEIC-SVM and DeepScore. Targets are sorted by the performance of PLEIC-SVM.

**TABLE 4 |** Performance comparison between RF-Score and DeepScore.

| Model name | ROC-AUC | EF1% | EF2% | EF5% | EF10% |
|---|---|---|---|---|---|
| DeepScore | **0.98** | **52.22** | **39.72** | **18.20** | **9.47** |
| AV-RF-V1 | 0.82 | 29.69 | 21.07 | 11.74 | 7.1 |
| AV-RF-V2 | 0.84 | 34.75 | 24.37 | 12.99 | 7.55 |
| AV-RF-V3 | 0.84 | 32.72 | 23.04 | 12.6 | 7.47 |
| D3.6-RF-V1 | 0.84 | 36.28 | 25.3 | 13.3 | 7.71 |
| D3.6-RF-V2 | 0.87 | 43.43 | 29.72 | 14.76 | 8.25 |
| D3.6-RF-V3 | 0.87 | 41.1 | 28.27 | 14.61 | 8.2 |
| D6.6-RF-V1 | 0.77 | 27.42 | 18.65 | 10.37 | 6.42 |
| D6.6-RF-V2 | 0.80 | 34.3 | 22.07 | 11.73 | 6.96 |
| D6.6-RF-V3 | 0.79 | 32.05 | 21.56 | 11.47 | 6.88 |

*Performance values of RF-Score are collected from the Supporting Information of (Wójcikowski et al., 2017). Best results are highlighted in bold.*

DeepScore is sensitive to docking program, we regenerated all ligand poses using AutoDock Vina (Trott and Olson, 2010) and repeated the above process. ROC-EFs of test results were calculated and shown in **Tables S4** and **S5** to quantitatively assess the influence of changing docking program on the virtual screening ability of DeepScore. Obvious differences can be observed on some targets in **Table S5**. For example, DeepScore-ADV (AutoDock Vina) achieved a ROC-EF0.5% of 160.65 on HS90A which represented an improvement of 37.01% over the ROC-EF0.5% achieved by DeepScore-Glide (117.25). But on PLK1, ROC-EF0.5% dropped by 60.61 (DeepScore-ADV 84.76 *vs*. DeepScore-Glide 145.37). Generally speaking, DeepScore-ADV got a similar performance with DeepScore-Glide in terms of mean values (see **Table S4**). It can be concluded that the screening ability of DeepScore is robust and insensitive to the docking program used, on the premise that the docking program can provide reliable docking poses.

## Case Study and Visualization

An appropriate visualization method will be beneficial for lead optimization. Some deep learning–based scoring functions, like DenseFS that uses 3D CNN (Hochuli et al., 2018; Imrie et al., 2018), are rather cumbersome in explaining the results of the model. The form of DeepScore makes the interpretation and visualization of the model much more intuitive. Here, we used four targets, AA2AR, CDK2, ESR1, and DPP4, as examples to show how to visually analyze the prediction results of DeepScore. These four targets were randomly selected and belong to four different protein families: AA2AR (adenosine A2a receptor, GPCR), CDK2 (cyclin-dependent kinase 2, kinase), ESR1 (estrogen receptor alpha, nuclear receptor), and DPP4 (dipeptidyl peptidase IV, protease).

We showed the contribution of every ligand (or protein) atom to binding by coloring each atom different shades of red. Given a protein–ligand complex, the score for each atom pair could be calculated through Eq. 2-2 under a certain model. The contribution of an atom was equivalent to the sum of the scores of all atom pairs involving this atom. All of the ligand and protein atoms were initially painted dark gray. Then, atoms that contributed positively would be painted different shades of red, and the color of atoms with negative contributions would not change. The atom with the highest positive score in ligand/protein would be painted in the deepest red. The shades of the red of other atoms indicated the relative magnitude of the contribution of the atom to the contribution of the atom colored deepest red. We randomly selected a positive ligand for each target and analyzed the binding mode of the ligand to the target using above coloring strategy.

**AA2AR** A2A adenosine receptors (AA2ARs) belong to G protein–coupled receptors (GPCRs). From the pharmacophore model, we have known that for AA2AR antagonists, basic structures include a hydrogen-bond donor, an N-containing aromatic ring, a large lipophilic region, and a smaller lipophilic region (Mantri et al., 2008). In **Figure 7**, the binding mode of an active obeying these pharmacophore rules is presented, and different regions are labeled. It can be seen that DeepScore highlighted the importance of the N-containing aromatic ring and the smaller lipophilic region by painting them red. The rest structures were taken as less important.

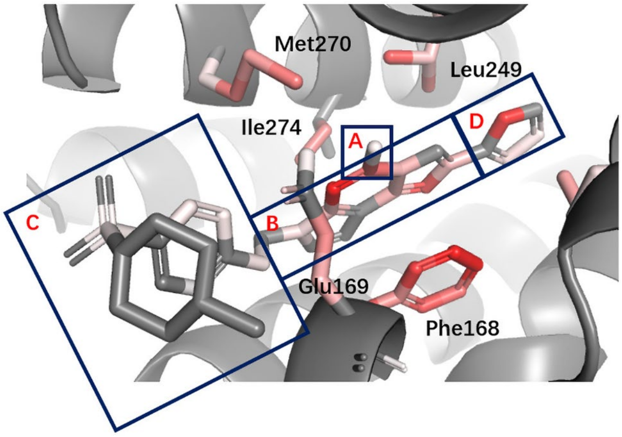


**FIGURE 7 |** Binding mode analysis of CHEMBL418564 with AA2AR receptor (DeepScore =1.875, PDB ID 3eml). A to D refer to the four different parts in pharmacophore model of AA2AR antagonists. A, hydrogen-bond donor. B, N-containing aromatic ring. C, large lipophilic region. D, smaller lipophilic region.

**CDK2** Cyclin-dependent kinases (CDKs) belong to serine/threonine family protein kinases. CDK2 is an ideal clinical target used for the treatment of breast cancer. Previous studies have shown that Leu83 residue is involved in the hydrogen bond formed with ligand (Wang et al., 2018). DeepScore also gave a high score to Leu83 and the nearest aromatic group (**Figure 8**).

**ESR1** Estrogen receptor alpha (ER alpha, ESR1) is a target for the treatment of breast cancer. Yan et al. used the information extracted by their model (PLEIC-SVM) to statistically analyze the average hydrophobic and hydrogen-bond interactions between residues of binding pocket and ligands for ESR1 (Yan et al., 2017). They found that the hydrogen bonds formed between the ligand and three residues, Glu353, Arg394, and His524, were the decisive factors in distinguishing between actives and decoy. As shown in **Figure 9**, DeepScore also ranked exact these residues as the most important three ones.

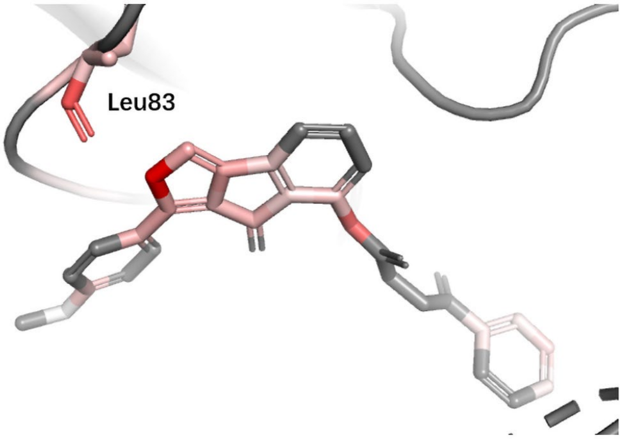


**FIGURE 8 |** Binding mode analysis of CHEMBL363077 with CDK2 receptor (DeepScore = 1.805, PDB ID 1h00).

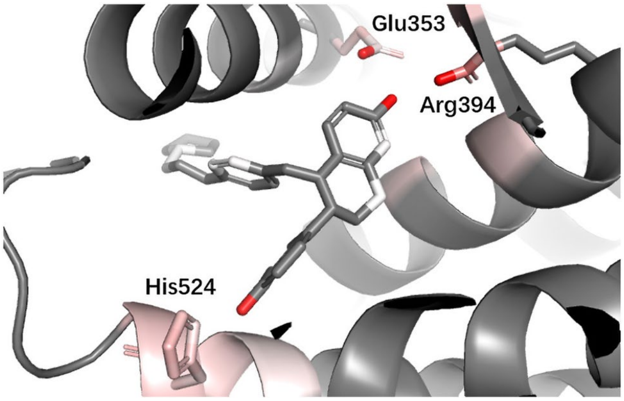


**FIGURE 9 |** Binding mode analysis of CHEMBL56306 with ESR1 receptor (DeepScore = 7.411, PDB ID 1sj0).

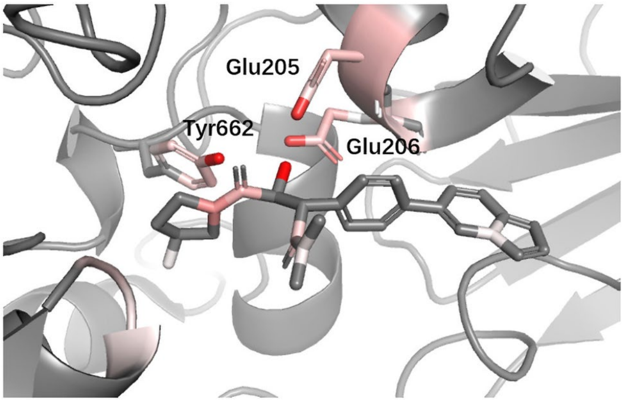


**FIGURE 10 |** Binding mode analysis of CHEMBL378637 with DPP4 receptor (DeepScore = 3.549, PDB ID 2i78).

**DPP4** Dipeptidyl peptidase-IV (DPP4) inhibitors are used for treating diabetes mellitus. According to a recent review about DPP4 inhibitors, Glu205, Glu206, and Tyr662 in DPP4 are believed to be the most import anchor points helping inhibitors recognize DPP IV. Since we used different protein with (Ojeda-Montes et al., 2018), for the convenience of comparison, we performed sequence alignment and renumbered all residues so that the residue number we used could match (Ojeda-Montes et al., 2018). In **Figure 10**, it can be seen that DeepScore also favored these three residues and gave them fairly high scores.

## CONCLUSION

In this work, we introduced a novel strategy for training target-specific protein–ligand scoring functions used for structure-based virtual screening. The model outperformed Glide Gscore significantly and made progress with respect to some metrics compared with traditional machine learning–based models. These results demonstrate that our model is able to further improve the screening effect by rescoring docking poses generated from docking software. There still remains more space for improving DeepScore. Like PMF scoring function, energy terms were treated implicitly in DeepScore, which made the model more difficult to capture important protein–ligand interactions. The cutoff distance we chose may be too short, causing long-range interactions not to be captured. However, on the other side, during the experiment, we found that a larger cutoff distance would significantly increase the noise and calculation cost. The most valuable aspect of DeepScore is that it represents a novel atom-pair-based machine learning scoring strategy. With the deeper integration of deep learning and chemical informatics, we believe that deep learning–based scoring functions will further develop in the future.

## DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://dude.docking.org/.

## AUTHOR CONTRIBUTIONS

XL and MZ designed the study and are responsible for the integrity of the manuscript. DW, XD, and CC performed the analysis and all calculations. DW mainly wrote the manuscript. ZX contributed to data processing. HJ and KC gave conceptual advice. All authors discussed and commented on the manuscript.

## FUNDING

This work was supported by the National Science & Technology Major Project "Key New Drug Creation and Manufacturing Program" of China (Number:2018ZX09711002), National Natural Science Foundation of China (81573351), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12020372), grants from Science and Technology Commission of Shanghai Municipality (18431907100), and Fudan-SIMM Joint Research Fund (FU-SIMM20174007).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00924/full#supplementary-material

## REFERENCES

Andricopulo, A., Salum, L., and Abraham, D. (2009). Structure-based drug design strategies in medicinal chemistry. *Curr. Top. Med. Chem.* 9, 771–790. doi: 10.2174/156802609789207127

Antes, I., Merkwirth, C., and Lengauer, T. (2005). POEM: parameter optimization using ensemble methods: application to target specific scoring functions. *J. Chem. Inf. Model.* 45, 1291–1302. doi: 10.1021/ci050036g

Berishvili, V. P., Voronkov, A. E., Radchenko, E. V., and Palyulin, V. A. (2018). Machine learning classification models to improve the docking-based screening: a case of PI3K-tankyrase inhibitors. *Mol. Inform.* 37, e1800030. doi: 10.1002/minf.201800030

Chaput, L., Martinez-Sanz, J., Saettel, N., and Mouawad, L. (2016). Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminform.* 8, 1–17. doi: 10.1186/s13321-016-0167-x

Dietterich, T. G. (2000). "Ensemble methods in machine learning", in *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, vol. 1857, pp. 1–5. doi: 10.1007/3-540-45014-9_1

Ericksen, S. S., Wu, H., Zhang, H., Michael, L. A., Newton, M. A., Hoffmann, F. M., et al. (2017). Machine learning consensus scoring improves performance across targets in structure-based virtual screening. *J. Chem. Inf. Model.* 57, 1579–1590. doi: 10.1021/acs.jcim.7b00153

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* 49, 6177–6196. doi: 10.1021/jm051256o

Hochuli, J., Helbling, A., Skaist, T., Ragoza, M., and Koes, D. R. (2018). Visualizing convolutional neural network protein-ligand scoring. *J. Mol. Graph. Model.* 84, 96–108. doi: 10.1016/j.jmgm.2018.06.005

Imrie, F., Bradley, A. R., Van Der Schaar, M., and Deane, C. M. (2018). Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J. Chem. Inf. Model.* 58, 2319–2330. doi: 10.1021/acs.jcim.8b00350

Jain, A. N., and Nicholls, A. (2008). Recommendations for evaluation of computational methods. *J. Comput. Aided. Mol. Des.* 22, 133–139. doi: 10.1007/s10822-008-9196-5

Kooistra, A. J., Vischer, H. F., McNaught-Flores, D., Leurs, R., De Esch, I. J. P., and De Graaf, C. (2016). Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* 6, 1–21. doi: 10.1038/srep28288

Kurkinen, S. T., Niinivehmas, S., Ahinko, M., Lätti, S., Pentikäinen, O. T., and Postila, P. A. (2018). Improving docking performance using negative image-based rescoring. *Front. Pharmacol.* 9, 1–15. doi: 10.3389/fphar.2018.00260

Li, L., Khanna, M., Jo, I., Wang, F., Ashpole, N. M., Hudmon, A., et al. (2011). Target-specific support vector machine scoring in structure-based virtual screening: computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation. *J. Chem. Inf. Model.* 51, 755–759. doi: 10.1021/ci100490w

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., et al. (2017). Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.* 50, 302–309. doi: 10.1021/acs.accounts.6b00491

Mantri, M., de Graaf, O., van Veldhoven, J., Goblyos, A., von Frijtag Drabbe Kunzel, J. K., Mulder-Krieger, T., et al. (2008). 2-Amino-6-furan-2-yl-4-substituted nicotinonitriles as A 2A adenosine receptor antagonists. *J. Med. Chem.* 51, 4449–4455. doi: 10.1021/jm701594y

Morrow, J. K., Chen, L., Phatak, S. S., Zhang, S., Du-Cuny, L., and Tran, H. T. (2012). From laptop to benchtop to bedside: structure-based drug design on protein targets. *Curr. Drug Metab.* 18, 1217–1239. doi: 10.2174/138920012799362837

Muegge, I., and Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* 42, 791–804. doi: 10.1021/jm980536j

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Ojeda-Montes, M. J., Gimeno, A., Tomas-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Valls, C., et al. (2018). Activity and selectivity cliffs for DPP-IV inhibitors: lessons we can learn from SAR studies and their application to virtual screening. *Med. Res. Rev.* 38, 1874–1915. doi: 10.1002/med.21499

Pereira, J. C., Caffarena, E. R., and Dos Santos, C. N. (2016). Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506. doi: 10.1021/acs.jcim.6b00355

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* 57, 942–957. doi: 10.1021/acs.jcim.6b00740

Sieg, J., Flachsenberg, F., and Rarey, M. (2019). In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* 59, 947–961. doi: 10.1021/acs.jcim.8b00712

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 34, 3666–3674. doi: 10.1093/bioinformatics/bty374

Teramoto, R., and Fukunishi, H. (2008). Consensus scoring with feature selection for structure-based virtual screening. *J. Chem. Inf. Model.* 48, 288–295. doi: 10.1021/ci700239t

Trott, O., and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Truchon, J. F., and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* 47, 488–508. doi: 10.1021/ci600426e

Wang, W. J., Huang, Q., Zou, J., Li, L. L., and Yang, S. Y. (2015). TS-Chemscore, a target-specific scoring function, significantly improves the performance of scoring in virtual screening. *Chem. Biol. Drug Des.* 86, 1–8. doi: 10.1111/cbdd.12470

Wang, Y., Chen, Y., Cheng, X., Zhang, K., Wang, H., Liu, B., et al. (2018). Design, synthesis and biological evaluation of pyrimidine derivatives as novel CDK2 inhibitors that induce apoptosis and cell cycle arrest in breast cancer cells. *Bioorganic Med. Chem.* 26, 3491–3501. doi: 10.1016/j.bmc.2018.05.024

Wójcikowski, M., Ballester, P. J., and Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* 7, 1–10. doi: 10.1038/srep46710

Xu, D., Li, L., Zhou, D., Liu, D., Hudmon, A., and Meroueh, S. O. (2017). Structure-based target-specific screening leads to small-molecule CaMKII Inhibitors. *ChemMedChem* 12, 660–677. doi: 10.1002/cmdc.201600636

Xue, M., Zheng, M., Xiong, B., Li, Y., Jiang, H., and Shen, J. (2010). Knowledge-based scoring functions in drug design. 1. Developing a target-specific method for kinase-ligand interactions. *J. Chem. Inf. Model.* 50, 1378–1386. doi: 10.1021/ci100182c

Yan, Y., Wang, W., Sun, Z., Zhang, J. Z. H., and Ji, C. (2017). Protein-ligand empirical interaction components for virtual screening. *J. Chem. Inf. Model.* 57, 1793–1806. doi: 10.1021/acs.jcim.7b00017

Zhang, T. (2004). "Solving large scale linear prediction problems using stochastic gradient descent algorithms" in *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. New York, NY, USA: ACM, vol. 116. doi: 10.1145/1015330.1015332

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fphar.2019.00913

# Identification of Novel Antibacterials Using Machine Learning Techniques

*Yan A. Ivanenkov*[1,2,3,4]*, *Alex Zhavoronkov*[4], *Renat S. Yamidanov*[1,4], *Ilya A. Osterman*[3,5], *Petr V. Sergiev*[5,6], *Vladimir A. Aladinskiy*[2,4], *Anastasia V. Aladinskaya*[2,4], *Victor A. Terentiev*[1,2,4], *Mark S. Veselov*[1,2,4], *Andrey A. Ayginin*[1,2], *Victor G. Kartsev*[7], *Dmitry A. Skvortsov*[3,8], *Alexey V. Chemeris*[1], *Alexey Kh. Baimiev*[1], *Alina A. Sofronova*[9], *Alexander S. Malyshev*[10], *Gleb I. Filkov*[2], *Dmitry S. Bezrukov*[3,5], *Bogdan A. Zagribelnyy*[3], *Evgeny O. Putin*[11], *Maria M. Puchinina*[2] *and Olga A. Dontsova*[3,5,6]

*[1] Institute of Biochemistry and Genetics Russian Academy of Science (IBG RAS) Ufa Scientific Centre, Ufa, Russia, [2] Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia, [3] Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia, [4] Insilico Medicine, Inc. Johns Hopkins University, Rockville, MD, United States, [5] Skolkovo Institute of Science and Technology, Skolkovo, Russia, [6] Department of Chemistry and A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia, [7] InterBioScreen ltd, Chernogolovka, Russia, [8] Faculty of Biology and Biotechnologies, Higher School of Economics, Moscow, Russia, [9] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [10] Faculty of Medicine, Lomonosov Moscow State University, Moscow, Russia,[11] Computer Technologies Lab, ITMO University, St. Petersburg, Russia*

OPEN ACCESS

***Edited by:***
*Wolff Mayer Kirsch,*
*Loma Linda University Medical Center (LLUMC),*
*United States*

***Reviewed by:***
*Pio Maria Furneri,*
*University of Catania, Italy*
*Rosanna Papa,*
*Sapienza University of Rome, Italy*
*Paola Rondón-Villarreal,*
*Universidad de Santander,*
*Colombia*

***Correspondence:***
*Yan A. Ivanenkov*
*yan@insilicomedicine.com*

***Specialty section:***
*This article was submitted to Experimental Pharmacology and Drug Discovery,*
*a section of the journal Frontiers in Pharmacology*

***Received:*** *15 February 2019*
***Accepted:*** *19 July 2019*
***Published:*** *27 August 2019*

***Citation:***
*Ivanenkov YA, Zhavoronkov A, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Veselov MS, Ayginin AA, Kartsev VG, Skvortsov DA, Chemeris AV, Baimiev AK, Sofronova AA, Malyshev AS, Filkov GI, Bezrukov DS, Zagribelnyy BA, Putin EO, Puchinina MM and Dontsova OA (2019) Identification of Novel Antibacterials Using Machine Learning Techniques.*
*Front. Pharmacol. 10:913.*
*doi: 10.3389/fphar.2019.00913*

Many pharmaceutical companies are avoiding the development of novel antibacterials due to a range of rational reasons and the high risk of failure. However, there is an urgent need for novel antibiotics especially against resistant bacterial strains. Available *in silico* models suffer from many drawbacks and, therefore, are not applicable for scoring novel molecules with high structural diversity by their antibacterial potency. Considering this, the overall aim of this study was to develop an efficient *in silico* model able to find compounds that have plenty of chances to exhibit antibacterial activity. Based on a proprietary screening campaign, we have accumulated a representative dataset of more than 140,000 molecules with antibacterial activity against *Escherichia coli* assessed in the same assay and under the same conditions. This intriguing set has no analogue in the scientific literature. We applied six *in silico* techniques to mine these data. For external validation, we used 5,000 compounds with low similarity towards training samples. The antibacterial activity of the selected molecules against *E. coli* was assessed using a comprehensive biological study. Kohonen-based nonlinear mapping was used for the first time and provided the best predictive power (av. 75.5%). Several compounds showed an outstanding antibacterial potency and were identified as translation machinery inhibitors *in vitro* and *in vivo*. For the best compounds, MIC and $CC_{50}$ values were determined to allow us to estimate a selectivity index (SI). Many active compounds have a robust IP position.

**Keywords: novel antibacterials, machine learning techniques, translation inhibitors, virtual screening, Kohonen-based SOM**

## INTRODUCTION

To the current date, a huge number of diverse small-molecule compounds have been reported as having antibacterial activity against different bacterial strains (Kohanski et al., 2010; Mohr, 2016; Naeem et al., 2016; Kaczor et al., 2017). However, almost all of them were discovered more than a half-century ago, and they are of natural origin, for example, penicillins (Fleming, 2001), cephalosporins (Brotzu, 1948), tetracyclines (Bryer et al., 1948), aminoglycosides (Schatz et al., 1944), and macrolides (McGuire et al., 1952). Some trivial structural modifications were introduced into their structure to improve pharmacokinetic features, reduce off-target side effects, and overcome bacterial resistance, which resulted in a broader range of next-in-class analogues, which were brought to market as well (Abouelhassan et al., 2019; Guan et al., 2019). On the contrary, fluoroquinolones [FQs, e.g., ciprofloxacin (Bauernfeind and Petermuller, 1983)] and linezolid (Spangler et al., 1996) are classified as synthetic antibiotics bearing a structure suitable for modification, and it is not surprising that more than 40 FQs were launched. For instance, lascufloxacin (Kishii et al., 2017), a broad-spectrum antibacterial drug, by Kyorin, is currently undergoing registration procedure in Japan as an oral formulation, while tedizolid, a linezolid analogue, developed by Merck & Co., was approved in 2014 (USA) against acute bacterial skin and skin structure infection (ABSSSI). According to Thomson Integrity Database, more than 4,000 molecules have been claimed as antibacterials during the past 5 years, including the most recent nontrivial 2-oxo-1,3-oxazolidines (2017 US 463908) by Johns Hopkins University, 1*H*-imidazo[4,5-*c*]quinolines by Pfizer (2018 US 629152), and 2-oxo-1,2-dihydrospiro-indoles by Shaanxi University of Science Technology (2018 CN 10285257). Twenty new antibacterial chemotypes have been discussed in the *Journal of Medicinal Chemistry* for the last 2 years (see *Supporting Information*). Many pharmaceutical companies, including big pharma alliances, have recently focused on antibacterial vaccines in their pre-clinical and clinical pipelines, for instance, VLA-1701 (Phase II) (Clinialtrialsgov, NIH, 2018c), ETEC (Phase I) (Clinicaltrialsgov, NIH, 2019b), GC-3107 (Phase I) (Clinicaltrialsgov, NIH, 2017a), PF-06842433 (Phase II) (Clinicaltrialsgov, NIH, 2018a), and PF-06886992 (Phase I), Vi-TCV (Phase III) (Clinicaltrialsgov, NIH, 2018b), rhGM-CSF (Phase II/III) (Clinicaltrialsgov, NIH, 2019d), and LEP-F1/GLA-SE (Phase I) (Clinicaltrialsgov, NIH, 2019c). Several small-molecule antibacterial compounds are currently evaluated in different clinical trials, including *N*-thiadiazolo-substituted piperidine (DS-2969; Phase I, Daiichi Sankyo), two boron-containing molecules [(GSK-070 (Clinicaltrialsgov, NIH, 2019a) and VNRX-5133 (Clinicaltrialsgov, NIH, 2017b); Phase I, GSK, and Phase I, VenatoRx Pharmaceuticals, respectively], benzimidazole-substituted 2*H*-chromen (tegoprazan; registered in 2018 for the treatment of gastroesophageal reflux disease in Korea, RaQualia), novel monobactam (BOS-228; Phase II, Novartis), 2-oxo-3,4-dihydro-1,8-naphthyridine (afabicin bis; Phase II, GSK), substituted 3-phenyl-1*H*-pyrrol-olorofim (Phase II, F2G Ltd.), original 1,6-diazabicyclo[3.2.1]octane-2-carboxamide (nacubactam, a β-lactamase inhibitor; Phase I, Roche), and 1*H*-pyrrolo[3,2-*b*]pyridine (TBA-7371; against tuberculosis, Phase I, AstraZeneca). At first glance, there are no principal barriers in this field; however, this speculative conclusion is rather illusory. *De facto*, biological evaluation of many molecules was discontinued due to the lack of efficiency and resistance barriers. The rate of failure outcomes within this sector is close to that observed in anticancer indication. Anyhow, a relatively high risk of failure makes this area much less attractive for the drug design and development in contrast to other easy-to-use therapeutic niches. Indeed, in recent years, global pharmaceutical players have shied away from this field and have shifted focus to more lucrative long-term treatments to manage generally chronic conditions (Projan, 2003). Considering the industry's reluctance to invest and support the development of new small-molecule antibiotics, academia and minor pharmaceutical companies are strategically positioned to play a key role in the initial stages of lead identification and optimization. Therefore, the improvement of primarily hit identification programs can dramatically extend a pool of promising lead candidates. Under these conditions, machine learning techniques can be reasonably regarded as one of the most appropriate and effective tools to perform rational selection of the most attractive compounds and to achieve significant success during initial rounds of HTS, thereby providing many diverse starting points for subsequent optimization and development.

Although many QSAR models for describing and predicting the antibacterial activity of small-molecule compounds have been published to date, they are mostly focused on an individual class of compounds or on a pre-defined scaffold (Morjan et al., 2015; Leemans et al., 2016). As a rule, such models are not applicable for diverse compound libraries at all, because input parameters, for example, molecular descriptors, are mainly selected to properly describe the chemical space around a chemotype studied. There are some examples of generalized *in silico* models for the prediction of antibacterial potency of heterogeneous series of molecules (**Table 1**). Most of them were trained with small- to moderate-sized training sets (Garcia-Domenech and de Julian-Ortiz, 1998; Tomas-Vert et al., 2000; Mishra et al., 2001; Cronin et al., 2002; Aptula et al., 2003; Molina et al., 2004; Murcia-Soler et al., 2004; Cherkasov, 2005; Gonzalez-Diaz et al., 2005; Marrero-Ponce et al., 2005; Yang et al., 2009) collected using three data sources of antibiotics (Glasby, 1978; Negwer, 1987; Maynard, 1996). As a result, they contain activity values determined in different assays and conditions with no information about their effective concentration. However, recently published models have utilized more comprehensive and qualitative databases (Karakoc et al., 2006; Yang et al., 2009; Wang et al., 2014; Masalha et al., 2018). For instance, in 2006, Karakoc and colleagues used a complete small-molecule collection that included 4,346 compounds bearing "*vecchio*" scaffolds, particularly 520 antibiotics, 562 bacterial metabolites, 958 drugs, 1,220 drug-like compounds, and 1,104 human metabolites (Karakoc et al., 2006). In 2018, Masalha et al. built a predictive model based on 3,500 molecules, but this dataset was collected using different sources that could provide a great bit of false-positive results (Masalha et al., 2018). Although the database contained compounds with high diversity in structure, most of them were well-known chemical entities and natural products (e.g., caffeine and ricinine), representing the active and inactive domains, respectively. In contrast, in this work, we utilized our large proprietary dataset of highly diverse molecules

**TABLE 1 |** *In silico* models for the development of novel antibacterial compounds.

| No. | $N_{total}$ | $N_{antibiotics}$ | Number of variables | Technique$^{a)}$ | Overall accuracy$^{b)}$ (%) | Ref. |
|---|---|---|---|---|---|---|
| **1** | 111 | 60 | 7 | LDA | 93.8/91.5$^{**}$ | (Garcia-Domenech and |
| | | | | ANN | 89.0/97.9$^{**}$ | de Julian-Ortiz, 1998) |
| **2** | 664 | 249 | 62 | ANN | 94.8$^{**}$ | (Tomas-Vert et al., 2000) |
| **3** | 59 | 24 | 17 | LDA | 85.0/84.0$^{***}$ | (Mishra et al., 2001) |
| **4** | 661 | 249 | 6 | LDA | 92.6/93.6$^{*}$ | (Cronin et al., 2002) |
| | | | | BLR | 94.7/94.3$^{*}$ | |
| **5** | 664 | 249 | 3 | LDA | 90.1$^{**}$ | (Aptula et al., 2003) |
| | | | | BLR | 92.1$^{**}$ | |
| **6** | 351 | 213 | 7 | LDA | 91.0/89.0$^{***}$ | (Molina et al., 2004) |
| **7** | 433 | 217 | 6 | LDA | 85.7/87.5$^{**}$ | (Murcia-Soler et al., 2004) |
| | | | 62 | ANN | 98.7/91.4$^{**}$ | |
| **8** | 667 | 363 | 7 | LDA | 92.9/94.0$^{**}$ | (Gonzalez-Diaz et al., 2005) |
| **9** | 657 | 249 | 34 | ANN | 92.9$^{**}$/100.0$^{***}$ | (Cherkasov, 2005) |
| **10** | 2,030 | 1,006 | 8 | LDA$^{c)}$ | 90.4/89.3$^{**}$/93.1$^{***}$ | (Marrero-Ponce et al., 2005) |
| **11** | 4,346 | 520 | 62 | kNN | 95.0/95.0$^{*}$/84.4$^{***}$ | (Karakoc et al., 2006) |
| | 611 | 230 | 36 | SVC | 100.0$^{*}$/100.0$^{**}$/98.1$^{***}$ | (Yang et al., 2009) |
| **12** | | | | kNN | 97.7$^{**}$/96.1$^{***}$ | |
| | | | | DT | 98.6$^{*}$/92.3$^{**}$/91.0$^{***}$ | |
| **13** | 7,517 | 2,066 | 21 | kNN$^{c)}$ | 99.2$^{*}$/81.8$^{**}$/78.3$^{***}$ | (Wang et al., 2014) |
| **14** | 2,230 | 1,051 | 3 | LDA$^{c)}$ | 85.6/87.2$^{**}$/86.2$^{***}$ | (Castillo-Garit et al., 2015) |
| **15** | 3,500 | 628 | 4 | ISE | 94.6/72.0$^{***}$ | (Masalha et al., 2018) |

*$^{a)}$LDA, linear discriminant analysis; ANN, artificial neural network; BLR, binary logistic regression; kNN, k-nearest neighbors; MLR, multiple linear regression; SVC, support vector classification; DT, decision tree; ISE, iterative stochastic elimination. $^{b)}$*Cross-validation; **internal test set; ***external test set. $^{c)}$Models that demonstrated the highest quality with external test set*

(*vide infra*) with low structural similarity towards the reported antibacterial compounds. This set was improved by antibacterial compounds obtained from Thomson Integrity Database.

Furthermore, the predictive power of many published models was not verified by cross-validation or by using an external validation set of fairly diverse compounds (Garcia-Domenech and de Julian-Ortiz, 1998; Tomas-Vert et al., 2000; Aptula et al., 2003; Murcia-Soler et al., 2004; Gonzalez-Diaz et al., 2005). Nevertheless, only a small part of these models was employed in a routine virtual screening practice (Marrero-Ponce et al., 2005; Wang et al., 2014; Castillo-Garit et al., 2015; Masalha et al., 2018) and resulted in the discovery of novel hit compounds with a remarkable antibacterial activity (Gonzalez-Diaz et al., 2005; Wang et al., 2014; Masalha et al., 2018). In 2015, Castillo-Garit and co-workers performed a ligand-based virtual screening study of 116 molecules with reported antibacterial activity using the developed QSAR model (Castillo-Garit et al., 2015). The model demonstrated good predictive ability in differentiation between active and inactive molecules. In 2014, an *in silico* study was carried out by Wang et al. using Guangdong Small Molecule Tangible Library (7,500 compounds) to search for new anti-MRSA agents and led to the identification of 56 primarily hits (Wang et al., 2014). Among them, 12 compounds were not reported previously as anti-MRSA agents and exhibited good activity against three MRSA strains. However, for the best compounds, only MIC values against bacterial cell lines were measured with no information about, for example, cytotoxicity towards eukaryotic cells. Therefore, it is hard to assess the SI of these molecules and further perspectives. In contrast, in this study, $CC_{50}$ values against the selected eukaryotic cell lines were determined to estimate this index for the most promising compounds.

For a long time, linear discriminant analysis (LDA) (Garcia-Domenech and de Julian-Ortiz, 1998; Mishra et al., 2001; Cronin et al., 2002; Aptula et al., 2003; Molina et al., 2004; Murcia-Soler et al., 2004; Gonzalez-Diaz et al., 2005; Marrero-Ponce et al., 2005; Karakoc et al., 2006; Castillo-Garit et al., 2015) and ANN (Garcia-Domenech and de Julian-Ortiz, 1998; Tomas-Vert et al., 2000; Murcia-Soler et al., 2004; Cherkasov, 2005; Karakoc et al., 2006) were the most popular machine learning methods that were used for prediction of antibacterial activity. On the contrary, few studies successfully implemented other *in silico* techniques, for example, binary logistic regression (BLR) (Cronin et al., 2002; Aptula et al., 2003), SVM (Yang et al., 2009; Wang et al., 2014), kNN (Karakoc et al., 2006; Yang et al., 2009; Wang et al., 2014), and decision tree (DT) (Yang et al., 2009). Therefore, herein, we placed particular focus on powerful and high-performance machine learning techniques that were not applied for antibacterials before, including Kohonen-based SOMs.

# MATERIALS AND METHODS

## Biological Evaluation

### High-Throughput Screening

Primary antibacterial activity of small-molecule compounds was assessed using our unique HTS platform described previously (Osterman et al., 2016). This approach allows us to estimate the mechanism of action of hit molecules based on the specific double-reporter system. Briefly, the red fluorescent protein gene rfp was placed under the control of a sulA promoter that was induced by SOS response. The gene of the fluorescent protein, katushka2S, was inserted downstream of the tryptophan attenuator. Two tryptophan codons were replaced by alanine codons, with simultaneous replacement of the complementary part of the attenuator to prevent the formation of a secondary structure that influences transcription termination. Thereby, the expression of katushka2S is observed only upon exposure to ribosome-stalling compounds. *E. coli* strains BW25113 or JW5503 were transfected with the designed plasmid called pDualrep2. As a result, it was possible to differentiate between three mechanisms of antibacterial action in "one-pot" format: DNA damage (expression of rfp), translation inhibition (expression of katushka2S), and others (inhibition of bacterial growth without expression of any reporter gene). The described assay was successfully validated using well-known antibacterial molecules and antibiotics (**Supplementary Figure 1**). Molecules were purchased from vendor collections and dissolved in DMSO at a concentration of 17 mg/ml (for the first round of HTS). Then, solutions of the compounds were spotted on agar plates with the reporter strain by a 96-channel pipetting head of a Janus liquid handling station (PerkinElmer) in a volume of 2 µl of each sample. Erythromycin (ERY, 1 µl) and levofloxacin (LVX, 1 µl) were added in each plate as control samples. After 16 h of incubation at 37°C, the Petri plates were scanned by a ChemiDoc system (Bio-Rad). Antibacterial activity was preliminary estimated by a thorough visual analysis, measurement of growth inhibition zone and MIC values: 0–4 mm ("–"), 4–7 mm ("+/–"), 7–11 mm ("+"), 11–16 mm ("++"; 25 µg/ml < MIC < 200 µg/ml), 16–20 mm ("+++"; 6.25 < MIC < 25), and 20–25 mm ("++++"; MIC < 6.25). Compounds with an insignificant growth inhibition area ("–," "+/–," and "+"; MIC > 200 µg/ml) were defined as inactive because of a relatively high concentration of compounds was used during this step. Molecules that caused strong inhibition of bacterial growth ("++," "+++," and "++++") were classified as active.

## *In Vitro* Translation Inhibition

### *$^{14}$C-Test*

*E. coli* ΔtolC strain was used to assess translation inhibition *in vivo*. Bacterial cells were cultivated in M9 medium to OD600 0.3–0.5. Then, the tested molecule was added at a concentration of 10 times higher than the determined MIC value to the 200 µl of the cells. After 2-min incubation, 1 µl of $^{14}$C-labeled valine (256 mCi/mmol) was added to the sample. Cells were incubated further for 2 min at 37°C. After incubation was completed, the sample was centrifuged, culture medium was separated, and lysis was performed with 20 µl HU buffer. The resulting mixture (5–10 µl) was subjected to polyacrylamide gel electrophoresis. The 10% SDS–PAGE gel was run for 60 min at 120 V and stained with Coomassie Brilliant Blue dye. The detection of $^{14}$C-labeled valine was carried out after 48 h by means of Typhoon GE Phosphorimager.

### *In Vitro Luciferase Assay*

*In vitro* transcribed firefly luciferase mRNA was translated in a cell-free system based on S30 cellular extract from *E. coli*. The samples were tested at a final concentration of 100 times lower than that used in the cell-based assay (*vide supra*). To investigate the effect of the selected molecules on the prokaryotic ribosome, a mixture of isolated ribosomes with a compound was kept at 37°C for 5 min without mRNA. Then, mRNA (200 ng) was added to the reaction mixture, and translation was initiated in a 10-ml reaction volume at 37°C for 30 min (Osterman et al., 2017). The translation of mRNA encoding luciferase was evaluated by measurement of enzyme activity using 0.1 mM D-luciferin and a spectrophotometer (PerkinElmer). Two control samples were used: negative (1% DMSO as an indicator that no translation inhibition occurred) and positive (ERY at a final concentration of 0.01 mg/ml as a translation inhibitor). All the measured values were normalized using the positive control baseline and expressed as a percentage.

## MTT Test

Cytotoxicity was assessed using the MTT (3-(4,5-dimethylthiazol-2-yl)2,5-diphenyl tetrazolium bromide) assay following the standard protocol with some modifications. Four thousand cells per well for VA13 cell line and 2,500 cells per well for MCF7, HEK293T, and A549 cell lines were plated out in 135 µl of DMEM/F12 media in a 96-well plate and incubated at 37°C, 5% $CO_2$ for 18 h before treatment. Then, the tested compound (15 µl, media/DMSO solution, the final DMSO concentration in the media was 0.5% or less) was added, and the cell samples were incubated for 72 h. The tested molecule in final concentrations of 50 nM–100 µM (eight dilutions), in triplicate, was applied. Doxorubicin (2 nM–6 µM) was used as a positive control. At the end of the incubation, MTT was added into the media (up to 0.5 mg/ml), and cells were incubated for 2 h followed by removal of the media and addition of DMSO (100 µl). The amount of MTT reduced by cells to its blue formazan derivative was measured spectrophotometrically at 565 nM using a plate reader and normalized to the values for cells treated with the media/DMSO only. $CC_{50}$ value was calculated with "GraphPad Prism 5" software (GraphPad Software, Inc., San Diego, CA). Cytotoxicity of some compounds was also assessed by an independent biological team. Compounds were tested at a single concentration of 10 µM, and the survival rate was obtained.

## Minimum Inhibitory Concentration

MICs in LB and M9 medium were determined using broth microdilution assay (Wiegand et al., 2008). The cell concentration was adjusted to approximately 5 × $10^5$ cells/ml. The tested compound was serially diluted twofold in a 96-well microplate (100 µl per well). Microplates were covered and incubated at 37°C with shaking. The OD600 of each well was measured, and the lowest concentration of the tested compound that resulted in no growth after 16–20 h was assigned to MIC value.

# Reference Database and Pre-Processing

The crude reference database for *in silico* modeling contained a total of 145,000 small-molecule compounds. Most of them (132,641 molecules) were outputted from our HTS campaign: 1,786 active and 130,855 inactive compounds (a hit rate for a random HTS was typical, 1.35). It should be especially noted that these compounds were highly dissimilar in structure to known antibiotics and antibacterial compounds because the prime goal of our previous work was to identify novel antibacterial scaffolds. The database was improved by the known antibacterial compounds obtained from Thomson Reuters Integrity database in order to increase the number of active samples and to cover the entire chemical space. In total, 12,347 molecules were added. Duplicate structures were removed using ChemoSoft software. Antibacterial molecules frequently contain specific substructures that are rather unusual in other therapeutic indications. Therefore, in this case, several medicinal chemistry filters (MCFs) cannot be properly applied to exclude undesired molecules. Thus, only "absolutely" nondrug-like molecules (e.g., metal-, silicon- and phospho-organic compounds, extensive linear aliphatic moieties, and sugars) as well as compounds containing highly toxic or unstable/reactive groups (e.g., strained heterocycles, isatines, hydroxamic acids, acylated imidazoles, and disulfides) were eliminated. Charged items were redrawn and presented in their neutral form, salt parts were deleted, and errors in structures were manually corrected. Then, the database was clustered using ChemoSoft software with the following parameters: Tanimoto similarity threshold ≥ 0.5 and the number of structures per cluster ≥ 5. In order to increase the common diversity of the dataset and to decrease the number of overrepresented structures, only 30 members with upper diversity coefficients per each cluster, as well as singletons, were retained. As a result, the final database contained 74,567 compounds (8,724 active and 65,843 inactive). The main parameters of the training dataset are listed in **Table 2**.

## Molecular Descriptors

Molecular descriptors (total of 1,749) were calculated for the whole training dataset using Dragon, ChemoSoft, MOE, and SmartMining (Pletnev et al., 2009) software tools. The number of descriptors was reduced to 1,243 by the omission of constant, near-constant, and highly correlated ($R > 0.9$) descriptors. *A priori*, we excluded a series of ordinary descriptors (e.g., number of exact and query substructures as well as fingerprints) to overcome overfitting, like in the case of β-lactams, fluoroquinolones, linezolid analogues, and other structure-biased antibacterials, and to objectively describe the input chemical space by a comprehensive set of key physicochemical molecular properties related with antibacterial potency. Then, the *t*-statistic was calculated for the remaining descriptors. Those with the best *t*-values were selected accounting their theoretical impact on the studied phenomenon (**Supplementary Table 1**) followed by PCA analysis (*Supporting Information*). As a result, we selected 40 molecular descriptors to perform the learning procedure. These include topological and electrotopological descriptors, lipophilicity and polarity indexes, the number of potential H-bond donors and H-bond acceptors, number of free-rotatable bonds and drug-likeness violation, atomic contribution to molar refractivity and autocorrelation, partial van der Waals surface area, and symmetry indexes (**Supplementary Table 2**).

## *In Silico* Modeling

### SOM

SOM (Kohonen, 1990) is one of the most powerful machine learning techniques that map multidimensional data onto lower-dimensional subspaces where geometric relationships between points indicate their similarity. Considering this fact, the output may be easily interpreted. However, this method requires a large amount of input data in order to achieve an appropriate predictive power. Kohonen-based SOM was constructed in SmartMining Software. The map size was 30 × 30 nodes (2D representation, of total 900 nodes, random distribution threshold was 82 samples per neuron), tetragonal cell, learning epochs: 2,000, initial learning rate: 0.3 (*linear decay*), initial learning radius: 15 (*linear decay*), activation function: Gaussian, winning neuron was determined using the standard Euclidean metrics, initial weight coefficients: random distribution, input vector: 40 descriptors (*not normalized*). Three independent randomizations were used to assess the reproducibility and stability of the model. After the unsupervised training process was completed, neurons were prioritized based on the following privileged factor (PF): $N_i^{ab}$ (%)/$N_i^{nab}$ (%), where $N_i^{ab}$ is the percent of antibacterials located in the *i*th neuron and while $N_i^{nab}$ is the percent of nonantibacterials located in the same neuron and vice versa. PF value greater than 1 was used as a threshold to assign neurons to one of these two classes.

### kNN

kNN (Zhang, 2016) is one of the simplest machine learning algorithms. However, its predictive performance and low computational costs make it one of the most used machine learning methods. This algorithm is based on feature similarity: the test sample is classified according to the nearest neighbors from the training dataset. However, the simplicity of kNN is associated with its inability to achieve an appropriate classification performance in case of complex data. In order to achieve the best predictive power, the following parameters of the classifier were varied: a number of neighbors (3–9, default 5), weights ("uniform" or "distance"), power parameter for the Minkowski metric ($p = 1$ for Manhattan distance and $p = 2$ for Euclidean distance).

**TABLE 2 |** Key features of the training dataset.

| Number of compounds | Active | Inactive | Diversity* | Unique heterocycles | | | Clusters** | Av. cluster size | Singletons |
|---|---|---|---|---|---|---|---|---|---|
| | | | | All | Active | Inactive | | | |
| 74,567 | 8,724 | 65,843 | 0.86 | 3,961 | 1,146 | 3,370 | 2,021 | 15 | 22,521 |

*Reverse Tanimoto metrics; **min. cluster size, 5; max. cluster size, 30; similarity threshold, 0.5.*

### Training Dataset Segmentation

Considering that the following *in silico* techniques use a supervised learning procedure, the randomized training datasets were subdivided into three categories in order to correctly estimate their classification accuracy: training set, cross-validation set, and internal test set (Balakin et al., 2004). The cross-validation subset was used to avoid model overfitting during the learning procedure, and the internal testing subset was used for pre-validation of the developed models. The learning settings were varied in order to reach the best classification accuracy. All the algorithms below were realized using scikit-learn library for Python 3.6.

### GB

Gradient boosting (Friedman, 2001) is one of the most powerful machine learning methods. It is an ensemble technique, in which new models (*decision trees*) are added to correct the errors by the existing models. Models are added sequentially until no further improvements can be made. GB is relatively resistant to an increase in the number of decision trees, so this usually leads to greater performance. Increasing the maximum depth does not always improve the prediction quality and may lead to overfitting and an increase in training time. The learning parameters were varied in order to reach the best classification power. The default values were the following: The number of trees was 100, and maximum tree depth was 3.

### RF

In contrast to GB, random forest (Breiman, 2001) is based on "fully grown" decision trees that are trained independently using a random sample of data. It should be noted that both GB and RF may be trained without preparation of the input data (scaling or normalization). One of the main advantages of RF compared with GB is the simplicity of model tuning. However, it is less resistant to an increase in the number of basic classifiers that also leads to a dramatic increase in computational costs. The default parameters of the model were the following: the number of trees was 10, and the maximum depth was not limited (building a tree until all leaves were empty or all leaves contained less than two elements).

### FFN (Feedforward Neural Network)

Artificial neural networks (Sazli, 2006) usually perform slightly better than the classifiers described above. However, overfitting is the main problem during the training procedure. Thus, different regularization techniques, parameters tuning, and accurate feature selection are required to achieve an appropriate classification accuracy. Moreover, FFN training procedure requires intense computational cost than do the other classifiers. One of the main disadvantages of neural networks is their "black box" nature. It is hard to understand how the prediction has been made. The three-layer neural network was constructed as follows: 30 neurons in the input layer, 100–150 neurons in the second layer, 30–80 neurons in the third layer, and 1 neuron in the output layer. The number of learning epochs varied from 1000 to 2000; initial learning rate was 0.1 (linear decay coefficient 0.01); weights were initialized randomly; dropout technique was used to prevent overfitting.

### SVM

SVM (Cortes and Vapnik, 1995) is a supervised machine learning algorithm that can be used for both classification and regression tasks. In this algorithm, each data item is plotted as a point in $n$-dimensional space with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the best hyperplane that differentiates two predefined classes. The main advantage of SVM is the possibility of using different kernels. Kernels are functions that transform low-dimensional input space to a higher-dimensional space where the classes can be separated. However, it is usually hard to choose hyperparameters of the SVM for sufficient generalization performance. The following parameters of SVM were used: penalty parameter ($1.0 \leq C \leq 10.0$) and kernel (linear, RBF, polynomial, and sigmoid).

## Experimental Validation

All the models described above were used to predict the antibacterial activity of novel molecules (5,000) randomly selected from the available vendor`s collections. These testing samples were selected using a threshold Tanimoto-based similarity value < 0.5 towards the training samples. All the compounds obtained were investigated for their antibacterial potency using the assays listed above. Biological results were then used to assess the prediction power of the models.

# RESULTS

## High-Throughput Screening

We used extensive proprietary data on the antibacterial activity of small-molecule compounds obtained during our HTS campaign. Screening molecules were selected from the stocks based on the following core principles: a) a relatively low structural similarity towards the reported antibacterial compounds and antibiotics, b) maximum diversity in structure per each cluster, c) all the remaining singletons (molecules that were not fitted in any cluster) were included as well, and d) maximal covering of the common chemical space provided by suppliers. In general, we used two collections of commercially available compounds by ChemDiv and IBS. While ChemDiv stock mainly contains organic compounds of synthetic origin, IBS basically focuses on nature-based molecules and their close analogues. To our satisfaction, to the current date, these obstacles have been overcome, and we have recently initiated a follow-up HTS round with EA compounds.

## Reference Database and Pre-Processing

To estimate the quality of covering the whole chemical space by the pool of the selected compounds, we constructed Sammon-based nonlinear map using the descriptors listed above (Sammon, 1969). Prior to mapping, we performed clustering analysis and rejected molecules with high similarity in structure per each cluster. We also applied soft MCFs to the final database for the exclusion of marginal nondrug-like structures. During Sammon mapping, we observed that about

70% of the collections used were normally covered by the remaining molecules. Therefore, we can speculate that they reflect the key features of the collections used more reliably and objectively versus random selection. In other words, we were trying to reach maximal covering and diversity with a minimal number of compounds. The pre-processed database was then used as a training set for *in silico* modeling.

## Molecular Descriptor Feature Selection

It should be especially noted that several of the selected molecular descriptors were described previously as important for statistically significant separation between antibacterial and nonantibacterial compounds (Araya-Cloutier et al., 2018). Distributions for the representative descriptors used for learning herein are depicted in **Figure 1**. As shown in **Figure 1**, Hy (F2 = 0.75) and HBD (F2 = 0.72) were among the best scored variables with *t*-coefficient higher than 40. Of the total, 25 molecular descriptors were classified as core inputs on the basis of *t*-stat analysis ($t > 30$). Several descriptors were less significant providing lower *t*-values, for instance, S(═N–) ($t = 8.7$, F4 = 0.48), GVWAI-80 ($t = 24$, F12 = 0.62), and M1 ($t = 29$, F1 = 0.83); however, in contrast to Hy and HBD, they were disposed in distinct areas of the common PCA plot (*Supporting Information*) and, therefore, contributed well to the exposition of the input chemical space, as well as the classification. Indeed, the exclusion of inputs with low-rate *t*-value led to the reduction of classification accuracy. Moreover, we performed non-parametric Mann–Whitney *U* test to prove the correctness of the description selection. The results of *U* test correlated with *t*-stat values (**Supplementary Table 6**). Based on the performed PCA analysis, 18 molecular descriptors were found to reflect 90% of the entire variability.

## *In Silico* Modeling

Because the dataset was highly imbalanced (eight times more inactive molecules vs. active), the models constructed using the remaining algorithms classified the majority of the samples as inactive, achieving a relatively high overall accuracy. Manipulations with the class weights parameters did not lead to any improvement in the classification ability. As a result, we constantly observed overfitting passages that are highly undesirable in machine learning. In order to balance the numbers of the compounds of both classes, the inactive molecules were clustered (Tanimoto similarity > 0.7), and the singletons were added to the nearest clusters. Each cluster was equally split into four different subsets (**Supplementary Table 3**), which resulted in four independent training sets (15,961 inactive and 7,724 active molecules in each "echelon"). The remaining 2,000 inactive and 1,000 active compounds were merged to form the internal test set. The antibacterial activity of the molecules was predicted using all these models and assigned for each sample based on the consensus score value. RF classifier demonstrated a relatively low classification accuracy. The best results were obtained using 100 classification trees (other parameters were kept as default). The average accuracy with the internal test set was 79.5% (90.2% for inactive and 68.8% for active compounds). GB provided almost the same results (average accuracy was 79.9%). More advanced algorithms (SVM and FFN) performed slightly better than decision tree-based classifiers. It should be noted that data scaling is strongly required to achieve better performance for these techniques. Thus, each descriptor vector was standardized using the scaling tool of scikit-learn library. The best classification results with SVM were obtained using the following parameters: penalty parameter of the error term ($C = 10.0$, kernel = rbf because other kernels, such as sigmoid, polynomial, and linear, demonstrated worse results). The

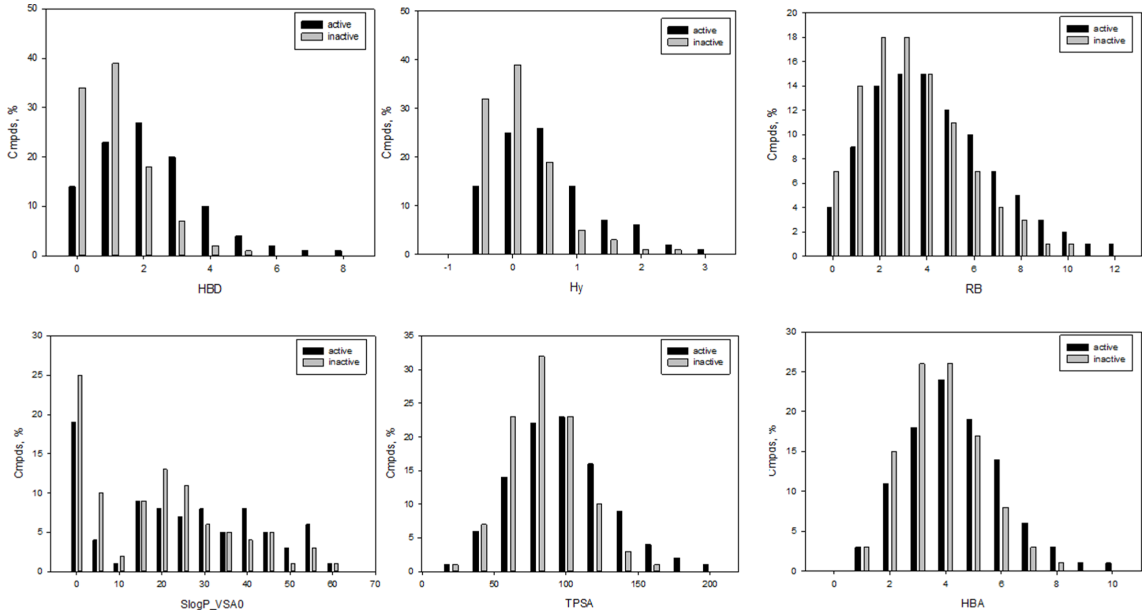


**FIGURE 1 |** Representative examples of molecular descriptors included in the final set of input variables; HBD, number of potential H-bond donors, Hy, hydrophilic factor, RB, number of free-rotatable bonds; LogP_VSA, reflects hydrophobic and hydrophilic effects; TPSA, polar surface area; HBA, number of potential H-bond acceptors.

average accuracy of the constructed models was 82.7% (91.5% for inactive and 73.9% for active compounds). FFN was implemented using Keras library for Python 3.6. Three-layer FFN showed the best average classification accuracy 81.3% (90.5% for inactive and 73.1% for active compounds). Other parameters of the neural network were the following: number of training epochs = 1,000; regularization, dropout (0.3 for each layer); activation function, sigmoid; and initial learning rate = 0.01.

The resulting Kohonen map is presented in **Figure 2**. As shown in **Figure 2A**, antibacterial compounds are located predominantly within a tight area of the whole map distinct from the nodes abundantly populated by nonantibacterial molecules (**Figure 2B**). The arrows in **Figure 2A** denote the location of different antibacterial drug classes in the map. At the final iteration, learning vector quantization error (LVQE) was relatively low, reaching a maximum value of 0.012 (**Supplementary Figure 2**). More than 90% of samples provided LVQE of less than 0.002. Therefore, we can conclude that the constructed model has very good generalization ability and learning outcome. The stability of the model was verified using three independent randomizations. Moreover, the addition of a fuzzy input with stochastic variables did not strongly affect the quality of classification. Upon examination, there were few "dead" neurons within the constructed map. The average classification accuracy was 77.5% and 69.8% without and with a random threshold, respectively.

The best predictive power with the internal test set was obtained using kNN. The most significant parameters that affected the classification accuracy were the number of neighbors (the best value = 3) and the weight function used in prediction (the best one was "distance" that weighted points by the inverse of their distance, so closer neighbors of a query point had a greater influence than neighbors that were further away). Euclidean metric was used for distance calculation. The prediction accuracy was 83.2% (88.7% for inactive and 77.7% for active compounds).

In addition, we performed a comprehensive statistical analysis of various nonheterocyclic (**Figure 3A**) as well as heterocyclic (**Figure 3B**) fragments presented in both classes. Among the first category, the methoxy (30.5% and 35% for active and inactive compounds, respectively) and carbonyl groups (39% and 25%) are the most represented. Nonantibacterial compounds contain 1.56 times greater number of carbonyl fragments in contrast to antibacterials, while the methoxy group does not provide a statistically significant separation between two classes studied. The accuracy of propanoyl moiety among inactive compounds is 3 times higher than in active samples. Carboxylic, α,β-unsaturated carbonyl, and allyl are the most characteristic moieties for antibacterial compounds: respectively 3.75, 6, and 9 times higher rate than the inactive class. With respect to heterocycles, indole is the most represented (12%) heterocyclic fragment among antibacterials. The rate of indole, imidazole, quinoline, and benzimidazole fragments is greatly biased towards antibacterial compounds, while furan and piperazine (~7%) are 2.3 times more abundant in nonantibacterials. In addition, 1,3-benzodioxole fragment is privileged for inactive molecules, while isoxazole is equally found in both classes. It should be especially noted that several molecular descriptors included in the final set for performing learning procedure are closely related with the statistical observations below. For instance, the common polarity encoded by S(–OH), S(–O–), S(═N–), S(> N–), HB2, a_acc, O-057/061, PEOE_VSA_FPOS, and TPSA corresponds to methoxy, carbonyl, propanoyl, carboxylic, and α,β-unsaturated carbonyl groups and heterocycles, while Hy and SlogP_VSA0 contribute to lipophilicity, particularly taking into account linear and branched alkyl moieties as well as aromatic fragments. Topology of a structure relates, for example, with M1, SPI, EEig07x, Q′, VEA2, and GATS1m.

## Experimental Validation

To investigate the prediction power of the constructed models, we used an external test set of 5K small-molecule compounds with similarity in structure of less than 0.5 to the whole training set. These molecules were randomly selected and purchased from ChemDiv and IBS collections. Antibacterial activity of the

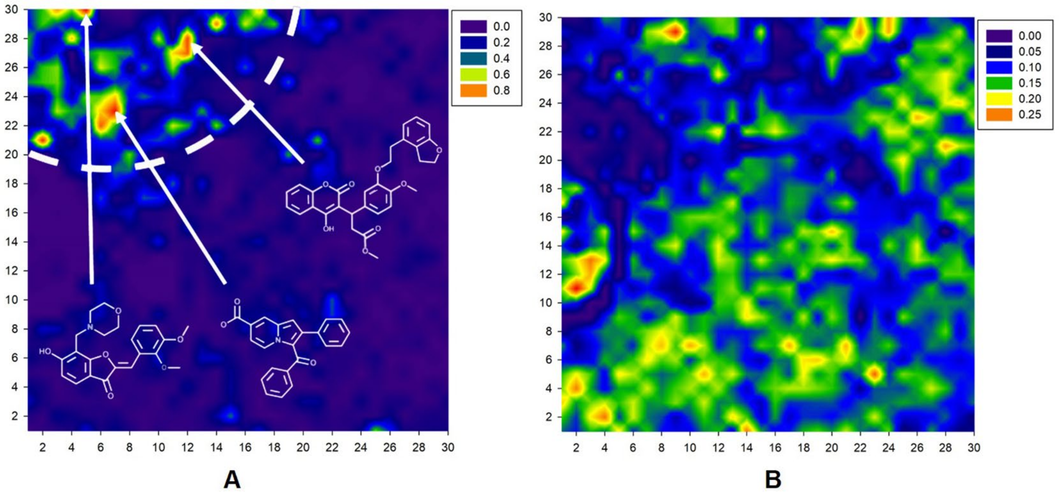


**FIGURE 2 |** A 30 × 30 2D Kohonen SOM for discrimination between antibacterial **(A)** and nonantibacterial **(B)** compounds within the same map. Color gradient corresponds to the percentage of molecules. Basic contours of the map were smoothed for a convenient visual inspection.

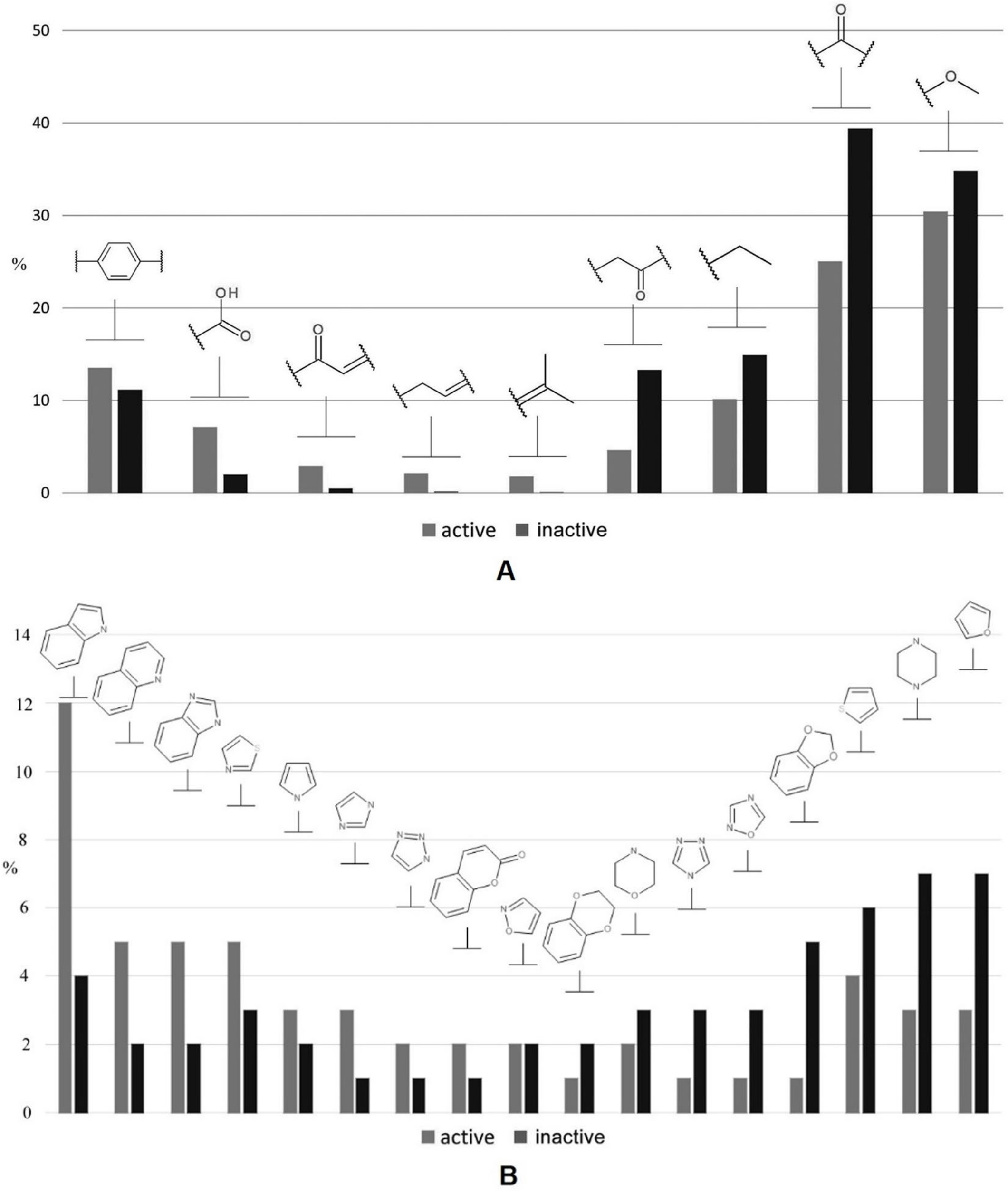
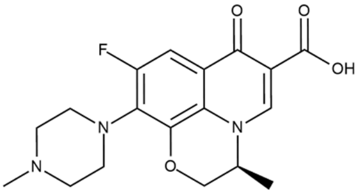
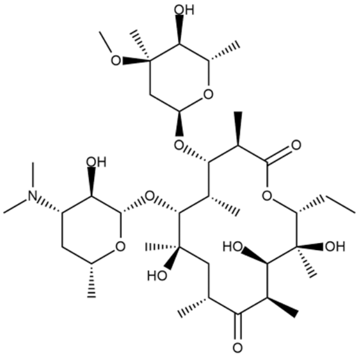
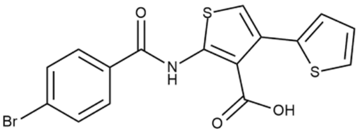
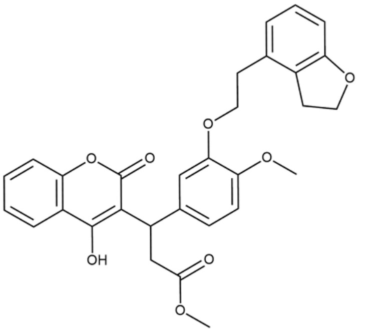
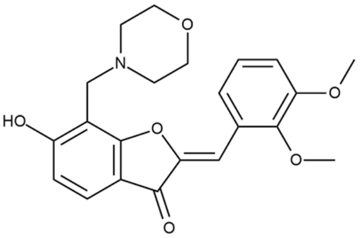
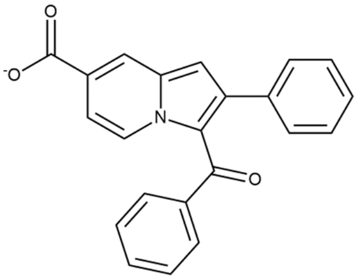


**FIGURE 3 |** A brief statistical analysis on basic nonheterocyclic **(A)** and heterocyclic **(B)** fragments presented in antibacterial and nonantibacterial compounds.

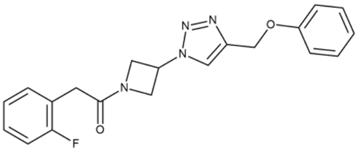
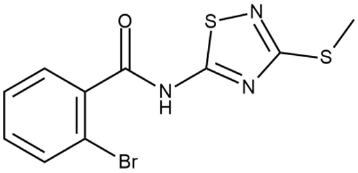
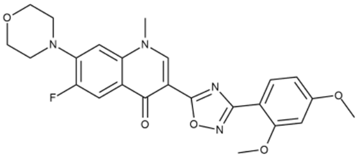
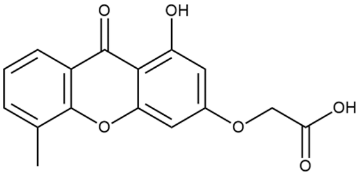
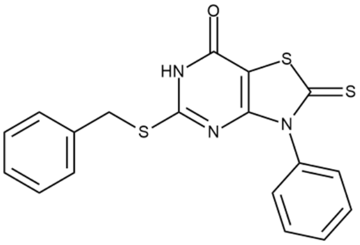
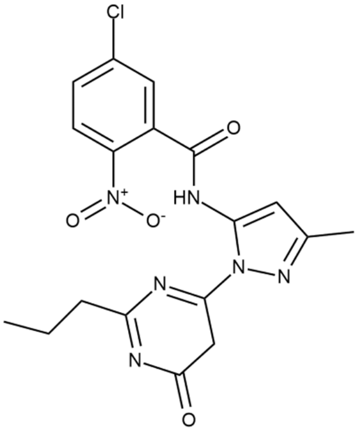
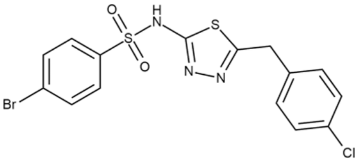
compounds was predicted using the developed models and then evaluated following the biological protocols described above. We did not use a consensus score value per each sample and retained compounds, which were predicted as inactive to estimate the prediction "resolution" of the models towards both classes used. This allowed us to get a valuable feedback on a possible overfitting or bias during the learning procedures. The first round of HTS has resulted in 371 active compounds (hit rate = 7.4%) followed by rescreen procedure performed at lower concentration. Rescreen confirmed moderate-to-high antibacterial activity for 65% of molecules. It should be especially noted that among all the active molecules from the initial HTS (molecules included in the training set) and from the external test set, only a few compounds showed a considerable inhibition activity against *E. coli*$^{wt}$. Several compounds demonstrated a robust SOS response or inhibition of translation machinery. A few compounds showed both signals but with a relatively low intensity. A brief summary of the performed biological evaluation is presented in **Table 3**. Due to confidentiality reasons, we cannot disclose the structures of the lead compounds. As shown in **Table 3**, among the listed molecules, the highest antibacterial potency was revealed for FQ analogue **7** (MIC = 0.8 μg/ml), 6*H*-thiazolo[4,5-*d*]pyrimidinone **9** (MIC < 0.2 μg/ml), (6-oxo-1*H*-pyrimidin-2-yl)pyrazole **10** (MIC < 0.2 μg/ml), substituted thiadiazoles **11** (MIC = 0.8 μg/ml), hydroxy-pyrazole

**TABLE 3 |** Representative examples of active compounds that were correctly predicted as antibacterials. The detailed biological results are presented in **Supplementary Figure 3**.

| ID | Structure | Activity | ID (from database) | MIC (μg/ml, ΔtolC) | Mechanism of action | In vitro translation | $^{14}$C-test | SI* | IP** |
|---|---|---|---|---|---|---|---|---|---|
| **LVX** |  | ++++ | - | 0.016 ± 0.009 | SOS | – | – | H | - |
| **ERY** |  | ++++ | - | 2.5 ± 0.5 | T | + | + | M | - |
| **1** |  | +++ | STOCK1S-88700 | 1.8 ± 0.8 | T | + | + | M | M |
| **2** |  | +++ | STOCK1N-86948 | 2 ± 0.4 | T | + | ± | M | M |
| **3** |  | ++++ | STOCK1N-55723 | 3.9 ± 1.4 | S + T | + | – | L | H |
| **4** |  | +++ | D090-0093 | 6.25 ± 1.3 | T | + | + | H | H |

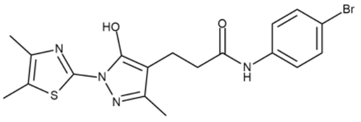
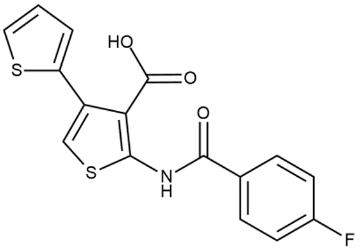
*(Continued)*

**TABLE 3 |** Continued

| ID | Structure | Activity | ID (from database) | MIC (µg/ml, ΔtolC) | Mechanism of action | In vitro translation | $^{14}$C-test | SI* | IP** |
|---|---|---|---|---|---|---|---|---|---|
| **5** |  | ++ | P991-0387 | 12.5 ± 1.9 | T | ± | – | H | H |
| **6** |  | + | F333-0013 | 42 ± 5 | T | + | + | H | M |
| **7** |  | +++ | F418-0205 | 0.8 | SOS | – | – | H | M |
| **8** |  | +++ | STOCK1N-64226 | 20.8 | SOS | – | – | L | H |
| **9** |  | +++ | F092-0369 | <0.2 | O | – | – | H | M |
| **10** |  | +++ | F269-0279 | <0.2 | O | – | – | H | H |
| **11** |  | +++ | Y030-6952 | 0.8 | O | – | – | H | H |

*(Continued)*

**TABLE 3 |** Continued

| ID | Structure | Activity | ID (from database) | MIC (μg/ml, ΔtolC) | Mechanism of action | In vitro translation | $^{14}$C-test | SI* | IP** |
|---|---|---|---|---|---|---|---|---|---|
| 12 |  | +++ | D475-2799 | 0.8 | O | – | – | M | H |
| 13 |  | +++ | STOCK2S-91453 | 1.8 ± 0.8 | T | + | + | M | M |

*SI, selectivity index = $CC_{50}$ (μg/ml or %)/MIC (μg/ml); H, high, SI > 100; M, moderate, 20 < SI < 100; L, low, SI < 20; T, translation inhibition; SOS, SOS response, O, other mechanism of action; **IP, intellectual property; L (low), match antibacterial Markush structure; M (moderate), match non-antibacterial Markush structure (but not listed among examples); H (high), clear IP status.

**12** (MIC = 0.8 μg/ml), and bithiophene **13** (MIC = 1.8 μg/ml). Compounds **1** and **2** strongly inhibited translation at 16 μg/ml and provided good SI. Furthermore, compound **2** showed a comparative antibacterial potency against several mutant strains (*these results will be published shortly*). Two compounds **7** and **8** induced a considerable SOS response, MIC = 0.8 and 20.8 μg/ml, respectively; however, compound **8** has lower SI. Among molecules acting *via* other mechanisms, compound **11** can be attributed to a wide class of sulfanilamide-based inhibitors (PABA analogues) of dihydropteroate synthase. Cytotoxicity against a panel of eukaryotic cells is summarized in the Supporting Information (**Supplementary Table 4**). The IP position of the molecules was assessed using SciFinder and Integrity Databases.

In order to make the study more informative, two hit compounds, **11** (4-bromo-*N*-{5-[(4-chlorophenyl)methyl]-1,3,4-thiadiazol-2-yl}benzene-1-sulfonamide) and **13** (5′-(4-fluorobenzamido)-[2,3′-bithiophene]-4′-carboxylic acid), were studied on antimicrobial activity against selected archival strains: *E. coli* ATCC 25922, *Klebsiella pneumoniae* 181210171-2, *Pseudomonas aeruginosa* ATCC 27853, *Staphylococcus aureus* ATCC USA 206, and *Candida albicans* 181210169-1 (**Table 4**). These substances exhibited modest activity against gram-negative bacteria *K. pneumoniae*. A similar pattern was observed on *C. albicans* multi-resistant strain. Compound **13** only slightly inhibited the growth of *E. coli*, while compound **11** did not demonstrate activity against this strain. No activity against *P. aeruginosa* was detected. The outstanding antimicrobial activity of compounds **11** and **13** was revealed in the tests on gram-positive cocci of *S. aureus*. The growth inhibition zone during the tests exceeded 20 mm.

## DISCUSSION

In contrast to numerous focused QSAR studies with recruiting small- or medium-sized reference databases of small-molecule compounds having a common scaffold or high similarity in structure, generalized *in silico* approaches for solving nontrivial classification problems cannot be adequately applied without a representative and comprehensive training dataset harmonically populated with a sufficient number of appropriate samples. These samples should almost ideally cover a whole input space providing maximum diversity. Any pattern within this space should contain bits of information important for learning procedure to achieve theoretically valid and interpreted results. This issue becomes one of the most crucial limitations, especially in the area of

**TABLE 4 |** Antibacterial activity of compounds **11** (4-bromo-*N*-{5-[(4-chlorophenyl)methyl]-1,3,4-thiadiazol-2-yl}benzene-1-sulfonamide) and **13** (5′-(4-fluorobenzamido)-[2,3′-bithiophene]-4′-carboxylic acid) against selected archival strains.

| Species | Strain ID | Source | Activity | |
|---|---|---|---|---|
| | | | Compound 11 | Compound 13 |
| *Escherichia coli* | ATCC 25922 | ATCC* | – | ± |
| *Klebsiella pneumoniae* | 181210171-2 | Clinic of the Bashkir State Medical University | ± | + |
| *Pseudomonas aeruginosa* | ATCC 27853 | ATCC | – | – |
| *Staphylococcus aureus* | ATCC USA 206 | Clinic of the Bashkir State Medical University | ++++ | ++++ |
| *Candida albicans* | 181210169-1 | ATCC | + | ± |

*ATCC, American Type Culture Collection.

**TABLE 5** | Overall performance of *in silico* modeling.

| Model | Classification accuracy (%)* | | | | | | | | | Prediction accuracy active/inactive (%)** |
|---|---|---|---|---|---|---|---|---|---|---|
| Set | Training | | | Cross-validation | | | Internal test | | | |
| Subset | Active | Inactive | Average | Active | Inactive | Average | Active | Inactive | Average | |
| **SOM** | 75 | 80 | 77.5 | – | – | – | – | – | – | 73/78 |
| **FFN** | 83.2 | 93.4 | 88.3 | 74.2 | 90.5 | 82.4 | 73.1 | 90.5 | 81.3 | 72/77 |
| **RF** | 100 | 100 | 100 | 70.7 | 92.7 | 81.7 | 68.8 | 90.2 | 79.5 | 69/80 |
| **GB** | 79.2 | 95.7 | 87.5 | 68.5 | 91.1 | 79.3 | 68 | 87 | 77.5 | 68/81 |
| **SVM** | 84.5 | 97.6 | 91.0 | 73.5 | 91.7 | 82.6 | 73.9 | 91.5 | 82.7 | 73/78 |
| **kNN** | 100 | 100 | 100 | 77.9 | 87.9 | 82.9 | 77.7 | 88.7 | 83.2 | 63/76 |

**Values for the best randomization; ** external test set.*

computer-aided modeling and the prediction of antibacterial activity. Considering a long-term period of a permanent stagnation in the field of development of novel antibacterials, a revolutionary breakthrough can be achieved using more powerful *in silico* approaches with an improved mining ability and prediction quality. These models are likely to achieve success in searching for principally new antibacterial chemotypes and to possibly overcome an overwhelming bacterial resistance.

In our work, the best results of *in silico* modeling were obtained using more advanced machine learning methods: Kohonen SOM, FFN, and SVM (**Table 5**). As it was expected, the predictive power of kNN was insufficient for this task. RF and GB performed slightly better. However, they failed to predict active molecules correctly in consequence of their tendency to overfit. Thus, Kohonen SOM, FFN, and SVM can be used for the *in silico* assessment of antibacterial activity. However, as it was discussed earlier, FFN and SVM did not perform well on the highly imbalanced dataset, and it was decided to split it in a different manner. Despite the achieved predictive power, the manipulations with input data may result in loss of information and require additional time-consuming data preparation steps. Thereby, Kohonen SOM is likely to be a more preferable and effective tool due to the following reasons: a) the resulting maps are very convenient for visual inspection of patterns occupied by compounds from different classes and for the distribution of molecular descriptor values within the map; b) in contrast to other machine learning techniques described above, overfitting was not observed for SOM using the training set of 73,000 samples, thereby providing more appropriate and reliable generalization; and c) in addition to a range of implemented settings, there are some advanced modifications of the algorithm (e.g., neural gas, convex combination, Grossberg-layer hybrid SOM, and Duane-Desieno method), which can be effectively used for improving the learning procedure and discrimination ability.

Experimental *in vitro* validation of developed models during first round of HTS and following rescreen procedure demonstrated relatively high hit rate considering the random compound selection for the external test set. Most of the most active molecules (**Table 3**) have moderate-to-high selectivity index and IP status. Two compounds (**11** and **13**) demonstrated satisfactory activities against several archival strains of microorganisms.

In summary, for the first time, we used a very large database of our proprietary HTS results to construct a highly discriminative and robust *in silico* model able to score molecules by their antibacterial potency against *E. coli*. The main focus was placed on compounds with low similarity in structure to the reported antibacterials, as well as maximum diversity. Forty of the most reliable molecular descriptors were rationally selected from a whole pool of more than 1,700 calculated features. The final set of descriptors reflects several key aspects in privileged structures presented in antibacterial or nonantibacterial compounds and significant patterns hidden in the input chemical space. Cumulative *in silico* modeling with recruiting several machine learning techniques showed that, using this dataset, two polar categories of compounds could be successfully separated providing good classification index. These models were then used to predict the antibacterial and nonantibacterial potency of novel compounds, which were not included in the parent database. Molecules from this external pool bore a relatively low structural similarity towards the training samples. Subsequent biological evaluation confirmed an attractive predictive power of the developed models. In particular, Kohonen-based SOM has not been used previously for solving the title task and demonstrated very promising results. Although we cannot disclose the structures of the best hits because of confidentiality reasons, the presented active molecules showed good antibacterial activity and can be reasonably regarded as convenient starting points for further optimization and morphing. Some compounds effectively inhibited translation in prokaryotes and showed no or weak cytotoxicity against a small panel of eukaryotic cell lines, thereby providing a benefit SI. With the use of the specific professional databases, the IP position of the molecules was preliminary assessed. The developed model can be effectively applied especially in academic organizations or small- to moderate-sized pharmaceutical companies to perform the rational selection of compounds for the primary HTS campaigns, thereby reducing the total costs of the entire R&D process.

## DATA AVAILABILITY

The datasets generated for this study are included in the manuscript and/or the **Supplementary Files**.

## AUTHOR CONTRIBUTIONS

GF and BZ prepared the database. VA, AVA, MV, AAA, DB, and EP performed the computational experiments. VT performed substructure and data analysis. RY, IO, PS, DS, AC, AB, and AS performed the biological experiments. MP performed the IP analysis. YI, AZ, VK, and OD conceived and supervised the study. AM prepared the manuscript.

## FUNDING

The authors would like to kindly acknowledge the Ministry of Education and Science of the Russian Federation, government grant 20.9907.2017/VU (expert opinion, discussion, manuscript preparation, and substructure analysis) and Russian Science Foundation No. 17-74-30012, IBG RAS Ufa (biological evaluation, compound selection, and purchasing as well as *in silico* modeling).

## ACKNOWLEDGMENTS

This study was performed in the context of the collaboration between Lomonosov Moscow State University (Russia), Moscow Institute of Physics and Technology (Russia), Institute of Biochemistry and Genetics Russian Academy of Science (Ufa, Russia), IBS (Russia), Enamine (USA), Information Technologies, and Mechanics and Optics (St. Petersburg). We kindly acknowledge all the scientists implicated in this multidisciplinary study.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00913/full#supplementary-material

## REFERENCES

Abouelhassan, Y., Garrison, A. T., Yang, H., Chavez-Riveros, A., Burch, G. M., and Huigens, R. W., 3rd (2019). Recent progress in natural-product-inspired programs aimed to address antibiotic resistance and tolerance. *J. Med. Chem.* doi: 10.1021/acs.jmedchem.9b00370

Aptula, A. O., Kühne, R., Ebert, R.-U., Cronin, M. T. D., Netzeva, T. I., and Schüürmann, G. (2003). Modeling discrimination between antibacterial and non-antibacterial activity based on 3D molecular descriptors. *QSAR Comb. Sci.* 22 (1), 113–128. doi: 10.1002/qsar.200390001

Araya-Cloutier, C., Vincken, J. P., van de Schans, M. G. M., Hageman, J., Schaftenaar, G., den Besten, H. M. W., et al. (2018). QSAR-based molecular signatures of prenylated (iso) flavonoids underlying antimicrobial potency against and membrane-disruption in Gram positive and Gram negative bacteria. *Sci. Rep.* 8 (1), 9267. doi: 10.1038/s41598-018-27545-4

Balakin, K. V., Ivanenkov, Y. A., Skorenko, A. V., Nikolsky, Y. V., Savchuk, N. P., and Ivashchenko, A. A. (2004). In silico estimation of DMSO solubility of organic compounds for bioscreening. *J. Biomol. Screen.* 9 (1), 22–31. doi: 10.1177/1087057103260006

Bauernfeind, A., and Petermuller, C. (1983). In vitro activity of ciprofloxacin, norfloxacin and nalidixic acid. *Eur. J. Clin. Microbiol.* 2 (2), 111–115. doi: 10.1007/BF02001575

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi: 10.1023/A:1010933404324

Brotzu, G. (1948). *Ricerche su di un nuovo antibiotico*. Cagliari: Lavori dell'Istituto d'Igiene di Cagliari.

Bryer, M. S., Schoenbach, E. B., Chandler, C. A., Bliss, E. A., and Long, P. H. (1948). Aureomycin: experimental and clinical investigations. *JAMA* 138 (2), 117–119. doi: 10.1001/jama.1948.02900020013004

Castillo-Garit, J. A., Marrero-Ponce, Y., Barigye, S. J., Medina-Marrero, R., Bernal, M. G., de la Vega, J. M. G., et al. (2015). In silico antibacterial activity modeling based on the TOMOCOMD-CARDD approach. *J. Braz. Chem. Soc.* 26 (6), 1218–1226. doi: 10.5935/0103-5053.20150087

ChemoSoft [Online]. Chemical Diversity Labs, Inc. Available: http://chemosoft.com/modules/db/ [Accessed 02/14/2019].

Cherkasov, A. (2005). Inductive QSAR descriptors. distinguishing compounds with antibacterial activity by artificial neural networks. *Int. J. Mol. Sci.* 6 (1), 63–86. doi: 10.3390/i6010063

Clinicaltrialsgov, NIH. (2017a). Study to investigate the safety and efficacy of GC3107 (BCG vaccine) in healthy adults. https://clinicaltrials.gov/show/NCT03363178. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2017b). VNRX-5133 SAD/MAD safety and PK in healthy adult volunteers. https://clinicaltrials.gov/show/NCT02955459. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2018a). A trial to evaluate a multivalent pneumococcal conjugate vaccine in healthy adults 50-85 years of age. https://clinicaltrials.gov/show/NCT03313050. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2018b). Clinical efficacy of typhoid conjugate vaccine (Vi-TCV) among children age 9 months through 12 years in Blantyre, Malawi. https://clinicaltrials.gov/show/NCT03299426. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2018c). Study confirming a human challenge model and investigating the safety of VLA1701. https://clinicaltrials.gov/ct2/show/NCT03576183. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2019a). An early bactericidal activity, safety and tolerability of GSK3036656 in subjects with drug-sensitive pulmonary tuberculosis. https://clinicaltrials.gov/show/NCT03557281. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2019b). Dose escalating study of a prototype CS6 subunit vaccine with a modified heat-labile enterotoxin from enterotoxigenic Escherichia coli (ETEC). https://clinicaltrials.gov/ct2/show/NCT03404674. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2019c). Phase 1 LEP-F1 + GLA-SE vaccine trial in healthy adult volunteers. https://clinicaltrials.gov/show/NCT03302897. [Accessed 08/22/2019].

Clinicaltrialsgov, NIH. (2019d). Pilot trial of inhaled molgramostim in nontuberculous mycobacterial (NTM) infection. https://clinicaltrials.gov/show/NCT03421743. [Accessed 08/22/2019].

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007/BF00994018

Cronin, M. T., Aptula, A. O., Dearden, J. C., Duffy, J. C., Netzeva, T. I., Patel, H., et al. (2002). Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.* 42 (4), 869–878. doi: 10.1021/ci025501d

Dragon [Online]. Milan (Italy): Talete s.r.l. Available: http://www.talete.mi.it/about/about.htm [Accessed 02/14/2019].

Fleming, A. (2001). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. *Bull. World Health Organ.* 79 (8), 780–790.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29 (5), 1189–1232. doi: 10.1214/aos/1013203451

Garcia-Domenech, R., and de Julian-Ortiz, J. V. (1998). Antimicrobial activity characterization in a heterogeneous group of compounds. *J. Chem. Inf. Comput. Sci.* 38 (3), 445–449. doi: 10.1021/ci9702454

Glasby, J. S. (1978). *Encyclopedia of antibiotics*. Manchester: Woodhouse.

Gonzalez-Diaz, H., Torres-Gomez, L. A., Guevara, Y., Almeida, M. S., Molina, R., Castanedo, N., et al. (2005). Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J. Mol. Model.* 11 (2), 116–123. doi: 10.1007/s00894-004-0228-3

Guan, Q., Huang, S., Jin, Y., Campagne, R., Alezra, V., and Wan, Y. (2019). Recent advances in the exploration of therapeutic analogues of gramicidin S, an old but still potent antimicrobial peptide. *J. Med. Chem*. doi: 10.1021/acs.jmedchem.9b00156

Kaczor, A. A., Polski, A., Sobotka-Polska, K., Pachuta-Stec, A., Makarska-Bialokoz, M., and Pitucha, M. (2017). Novel antibacterial compounds and their drug targets—successes and challenges. *Curr. Med. Chem.* 24 (18), 1948–1982. doi: 10.2174/0929867323666161213102127

Karakoc, E., Cherkasov, A., and Sahinalp, S. C. (2006). Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics* 22 (14), e243–e251. doi: 10.1093/bioinformatics/btl259

Kishii, R., Yamaguchi, Y., and Takei, M. (2017). In vitro activities and spectrum of the novel fluoroquinolone lascufloxacin (KRP-AM1977). *Antimicrob. Agents Chemother.* 61 (6), e00120–17. doi: 10.1128/AAC.00120-17

Kohanski, M. A., Dwyer, D. J., and Collins, J. J. (2010). How antibiotics kill bacteria: from targets to networks. *Nat. Rev. Microbiol.* 8 (6), 423–435. doi: 10.1038/nrmicro2333

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE Inst. Electr. Electron. Eng.* 78 (9), 1464–1480. doi: 10.1109/5.58325

Leemans, E., Mahasenan, K. V., Kumarasiri, M., Spink, E., Ding, D., O'Daniel, P. I., et al. (2016). Three-dimensional QSAR analysis and design of new 1,2,4-oxadiazole antibacterials. *Bioorg. Med. Chem. Lett.* 26 (3), 1011–1015. doi: 10.1016/j.bmcl.2015.12.041

Marrero-Ponce, Y., Medina-Marrero, R., Torrens, F., Martinez, Y., Romero-Zaldivar, V., and Castro, E. A. (2005). Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.* 13 (8), 2881–2899. doi: 10.1016/j.bmc.2005.02.015

Masalha, M., Rayan, M., Adawi, A., Abdallah, Z., and Rayan, A. (2018). Capturing antibacterial natural products with in silico techniques. *Mol. Med. Rep.* 18 (1), 763–770. doi: 10.3892/mmr.2018.9027

Maynard, R. L. (1996). *The Merck index: 12th edition*. New York: Merck.

McGuire, J. M., Bunch, R. L., Anderson, R. C., Boaz, H. E., Flynn, E. H., Powell, H. M., et al. (1952). Ilotycin, a new antibiotic. *Antibiot. Chemother. (Northfield)* 2 (6), 281–283.

Mishra, R. K., Garcia-Domenech, R., and Galvez, J. (2001). Getting discriminant functions of antibacterial activity from physicochemical and topological parameters. *J. Chem. Inf. Comput. Sci.* 41 (2), 387–393. doi: 10.1021/ci000303c

Mohr, K. I. (2016). History of antibiotics research. *Curr. Top. Microbiol. Immunol.* 398, 237–272. doi: 10.1007/82_2016_499

Molecular Operating Environment [Online]. Chemical Computing Group. Available: http://www.chemcomp.com/software.html [Accessed 02/14/2019].

Molina, E., Diaz, H. G., Gonzalez, M. P., Rodriguez, E., and Uriarte, E. (2004). Designing antibacterial compounds through a topological substructural approach. *J. Chem. Inf. Comput. Sci.* 44 (2), 515–521. doi: 10.1021/ci0342019

Morjan, R. Y., Al-Attar, N. H., Abu-Teim, O. S., Ulrich, M., Awadallah, A. M., Mkadmh, A. M., et al. (2015). Synthesis, antibacterial and QSAR evaluation of 5-oxo and 5-thio derivatives of 1,4-disubstituted tetrazoles. *Bioorg. Med. Chem. Lett.* 25 (18), 4024–4028. doi: 10.1016/j.bmcl.2015.04.070

Murcia-Soler, M., Perez-Gimenez, F., Garcia-March, F. J., Salabert-Salvador, M. T., Diaz-Villanueva, W., Castro-Bleda, M. J., et al. (2004). Artificial neural networks and linear discriminant analysis: a valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Comput. Sci.* 44 (3), 1031–1041. doi: 10.1021/ci030340e

Naeem, A., Badshah, S. L., Muska, M., Ahmad, N., and Khan, K. (2016). The current case of quinolones: synthetic approaches and antibacterial activity. *Molecules* 21 (4), 268. doi: 10.3390/molecules21040268

Negwer, M. (1987). *Organic–chemical drugs and their synonyms*. Berlin: Akademie.

Osterman, I. A., Khabibullina, N. F., Komarova, E. S., Kasatsky, P., Kartsev, V. G., Bogdanov, A. A., et al. (2017). Madumycin II inhibits peptide bond formation by forcing the peptidyl transferase center into an inactive state. *Nucleic Acids Res.* 45 (12), 7507–7514. doi: 10.1093/nar/gkx413

Osterman, I. A., Komarova, E. S., Shiryaev, D. I., Korniltsev, I. A., Khven, I. M., Lukyanov, D. A., et al. (2016). Sorting out antibiotics' mechanisms of action: a double fluorescent protein reporter for high-throughput screening of ribosome and DNA biosynthesis inhibitors. *Antimicrob. Agents Chemother.* 60 (12), 7481–7489. doi: 10.1128/AAC.02117-16

Pfizer Pipeline [Online]. Pfizer Web Site. Available: https://www.pfizer.com/sites/default/files/product-pipeline/01302018_PipelineUpdate.pdf [Accessed 02/14/2019].

Pletnev, I. V., Ivanenkov, Y. A., and Tarasov, A. V., (2009). "Dimensionality reduction techniques for pharmaceutical data mining," in *Pharmaceutical data mining: approaches and applications for drug discovery*. Editor. K. V. Balakin (Hoboken, NJ: John Wiley & Sons, Inc), 420–455.

Projan, S. J. (2003). Why is big pharma getting out of antibacterial drug discovery? *Curr. Opin. Microbiol.* 6 (5), 427–430. doi: 10.1016/j.mib.2003.08.003

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 18 (5), 401–409. doi: 10.1109/T-C.1969.222678

Sazli, M. H. (2006). A brief review of feed-forward neural networks. *Commun. Fac. Sci. Univ. Ank. Ser.* 50 (1), 11–17. doi: 10.1501/0003168

Schatz, A., Bugie, E., and Waksman, S. A. (1944). Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. *Proc. Soc. Exper. Biol. Med.* 55, 66–69. doi: 10.3181/00379727-55-14461

Spangler, S. K., Jacobs, M. R., and Appelbaum, P. C. (1996). Activities of RPR 106972 (a new oral streptogramin), cefditoren (a new oral cephalosporin), two new oxazolidinones (U-100592 and U-100766), and other oral and parenteral agents against 203 penicillin-susceptible and -resistant pneumococci. *Antimicrob. Agents Chemother.* 40 (2), 481–484. doi: 10.1128/AAC.40.2.481

Thomson Integrity Database [Online]. Thomson Integrity. Available: https://integrity.thomson-pharma.com/integrity/xmlxsl [Accessed 02/14/2019].

Tomas-Vert, F., Pérez-Giménez, F., Salabert-Salvador, M. T., García-March, F. J., and Jaén-Oltra, J. (2000). Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *Theochem* 504, 249–259. doi: 10.1016/S0166-1280(00)00366-3

Wang, L., Le, X., Li, L., Ju, Y., Lin, Z., Gu, Q., et al. (2014). Discovering new agents active against methicillin-resistant *Staphylococcus aureus* with ligand-based approaches. *J. Chem. Inf. Model.* 54 (11), 3186–3197. doi: 10.1021/ci500253q

Wiegand, I., Hilpert, K., and Hancock, R. E. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* 3 (2), 163–175. doi: 10.1038/nprot.2007.521

Yang, X. G., Chen, D., Wang, M., Xue, Y., and Chen, Y. Z. (2009). Prediction of antibacterial compounds by machine learning approaches. *J. Comput. Chem.* 30 (8), 1202–1211. doi: 10.1002/jcc.21148

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* 4 (11), 218–224. doi: 10.21037/atm.2016.03.37

**Conflict of Interest Statement:** Authors YI, AZ, RY, VA, AA, VT, and MV were employed by Insilico Medicine, Inc. Author VK was employed by InterBioScreen ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

TECHNOLOGY AND CODE
published: 05 September 2019
doi: 10.3389/fphar.2019.00971

# ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method

*Xiangeng Wang†, Yanjing Wang†, Zhenyu Xu, Yi Xiong* and Dong-Qing Wei**

*State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China*

OPEN ACCESS

**Edited by:**
*Jianfeng Pei,*
*Peking University,*
*China*

**Reviewed by:**
*Cao Dongsheng,*
*Central South University,*
*China*
*Quan Zou,*
*University of Electronic Science and*
*Technology of China, China*

***Correspondence:**
*Yi Xiong*
*xiongyi@sjtu.edu.cn*
*Dong-Qing Wei*
*dqwei@sjtu.edu.cn*

*†These authors have contributed equally to this work*

**Specialty section:**
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

**Received:** *10 June 2019*
**Accepted:** *29 July 2019*
**Published:** *05 September 2019*

**Citation:**
*Wang X, Wang Y, Xu Z, Xiong Y and Wei D-Q (2019) ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method. Front. Pharmacol. 10:971. doi: 10.3389/fphar.2019.00971*

Anatomical Therapeutic Chemical (ATC) classification system proposed by the World Health Organization is a widely accepted drug classification scheme in both academic and industrial realm. It is a multilabeling system which categorizes drugs into multiple classes according to their therapeutic, pharmacological, and chemical attributes. In this study, we adopted a data-driven network-based label space partition (NLSP) method for prediction of ATC classes of a given compound within the multilabel learning framework. The proposed method ATC-NLSP is trained on the similarity-based features such as chemical–chemical interaction and structural and fingerprint similarities of a compound to other compounds belonging to the different ATC categories. The NLSP method trains predictors for each label cluster (possibly intersecting) detected by community detection algorithms and takes the ensemble labels for a compound as final prediction. Experimental evaluation based on the jackknife test on the benchmark dataset demonstrated that our method has boosted the absolute true rate, which is the most stringent evaluation metrics in this study, from 0.6330 to 0.7497, in comparison to the state-of-the-art approaches. Moreover, the community structures of the label relation graph were detected through the label propagation method. The advantage of multilabel learning over the single-label models was shown by label-wise analysis. Our study indicated that the proposed method ATC-NLSP, which adopts ideas from network research community and captures the correlation of labels in a data driven manner, is the top-performing model in the ATC prediction task. We believed that the power of NLSP remains to be unleashed for the multilabel learning tasks in drug discovery. The source codes are freely available at https://github.com/dqwei-lab/ATC.

**Keywords: drug classification, multilabel classification, label correlation, label space partition, label propagation**

## INTRODUCTION

The Anatomical Therapeutic Chemical (ATC) Classification System (MacDonald and Potvin, 2004), maintained by the World Health Organization Collaborating Centre for Drug Statistics Methodology, is the most widely accepted and canonical scheme for drug categorization. This system assigns different group labels for drugs based on the organ or systems where they take effect and/

or their therapeutic, pharmacological, and chemical attributes. The ATC system is a strict hierarchy, including five levels of classification, and for the first level, there are 14 main groups: 1) alimentary tract and metabolism (coded by **A**); 2) blood and blood-forming organs (coded by **B**); 3) cardiovascular system (coded by **C**); 4) dermatologicals (coded by **D**); 5) genitourinary system and sex hormones (coded by **G**); 6) systemic hormonal preparations, excluding sex hormones and insulins (coded by **H**); 7) anti-infectives for systemic use (coded by **J**); 8) antineoplastic and immunomodulating agents (coded by **L**); 9) musculoskeletal system (coded by **M**); 10) nervous system (coded by **N**); 11) antiparasitic products, insecticides, and repellents (coded by **P**); 12) respiratory system (coded by **R**); 13) sensory organs (coded by **S**); and 14) various (coded by **V**). Given a new compound, prediction of its ATC classes can provide us with deeper insights into its therapeutic indications and side effects, thus accelerating both basic research and drug development (Hutchinson et al., 2004; Dunkel et al., 2008).

Traditionally, identification of ATC classes for a new drug using experimental methods is both time- and resource-consuming. Therefore, *in silico* prediction of ATC classes of a compound by machine learning techniques is a hot field in drug discovery and development. Previous studies (Dunkel et al., 2008; Wu et al., 2013) formulate the prediction of ATC classes as a single-label learning task, which is suggested to be inappropriate due to the multilabel nature of this biological system (Chou, 2013). Within the multilabel learning framework, Cheng et al. (2017b) proposed a multilabel predictor iATC-mISF, which utilized multilabel Gaussian kernel regression and three types of features (chemical–chemical interaction, structural similarity, and fingerprint similarity). The iATC-mISF has been upgraded as iATC-mHyb (Cheng et al., 2017a) by further incorporating drug ontological information. Besides one-dimensional representation of features, inspired by the histograms of oriented gradients (HoG) method proposed by the computer vison community (Dalal and Triggs, 2005), Nanni and Brahnam (2017) reshaped the features into two-dimensional matrix and performed slightly better than iATC-mISF. Continuing in this direction, the same group (Lumini and Nanni, 2018) applied pretrained convolutional neural networks models on the two-dimensional feature matrix as a featurizer and achieved best performance among the previously published methods on this task.

Typically, multilabel (ML) classification algorithms are classified into three major groups: algorithm adaptation, problem transformation, and ensembles of multilabel classifier (EMLC) (Wan et al., 2017). Algorithm adaptation methods incorporate specific tricks that modify traditional single-label learning algorithms into multilabel ones. The representative algorithm of this group is ML-*k*NN (Zhang and Zhou, 2005). For the problem transformation method, it converts multilabel learning problem into one or more single-label problems. The common strategies for such a transformation include binary relevance, classifier chains, label ranking, and label powerset (LP) (Read et al., 2011). LP trains models on each possible subset of label sets (Gibaja and Ventura, 2014). For a dataset with high cardinality in the large label set, LP is prone to be overfitting because of the exponentially increased number of subsets. To tackle the overfitting nature of label powerset, (Tsoumakas et al., 2011) proposed the RA*k*EL*d* method, which divides the label set into *k* disjoint subsets and use label powerset in these subsets. One major drawback of RA*k*EL*d* is that the *k* is arbitrarily chosen without incorporating the label correlations, which can be possibly learnt from the training data. The **n**etwork-based **l**abel **s**pace **p**artition (NLSP) (Szymański et al., 2016) is an EMLC built upon ML. This NLSP method divides the label set into *k* small-sized label sets (possibly intersecting) by a community detection method, which can incorporate the label correlation structures in the training set, such that it finally learns *k* representative ML classifiers. As a result, NLSP tackles much less subsets compared to LP on the original label set and selects *k* in a data-driven manner. For more detailed explanation of multilabel learning, refer to (Zhang and Zhou, 2014; Moyano et al., 2018).

In this study, we adopted an NLSP method to explore the correlation among labels. Our NLSP method was evaluated on a benchmark dataset (Chen et al., 2012) by the jackknife test. The proposed method demonstrates its superiority over other state-of-the-art approaches by our experimental results. The main strength of our method hinges on two aspects. On the one hand, the NLSP clusters the label space into subspaces and utilizes the correlation among labels. On the other hand, the ensemble learning nature of NLSP on the overlapping subspace could further improve model performance. Interesting patterns on the label relation graph were also detected by NLSP. In addition, the label-wise analysis of the best NLSP model was performed to provide experimental biologists with more insights.

# MATERIALS AND METHODS

## Benchmark Dataset and Sample Formulation

We utilized the same dataset as the previous study (Cheng et al., 2017b) to facilitate model comparison. This dataset consists of 3,883 drugs, and each drug is labeled with at least one or more of 14 main ATC classes. It is a tidy dataset where no missing value and contradictory record. The UpSet visualization technique (Lex et al., 2014) was used for quantitative analysis of interactions of label sets.

Then, we adopted the same method provided by (Cheng et al., 2017b) to represent the drug samples. The dataset can be formulated in set notation as the union of elements in each class: $\mathbb{S} = \mathbb{S}_1 \bigcup \mathbb{S}_2 \ldots \bigcup \mathbb{S}_{14}$ (1), and a sample $D$ can be represented by concatenating the following three types of features.

1. A 14-dimentional vector, $D^{\mathbf{Int}} = [\Phi_1\Phi_2\Phi_3 \ldots \Phi_{14}]^T$ (2), which represents its maximum interaction score $\Phi_i$ (Kotera et al., 2012) with the drugs in each of the 14 $\mathbb{S}_i$.
2. A 14-dimentional vector, $D^{\mathbf{StrSim}} = [\Psi_1\Psi_2\Psi_3 \ldots \Psi_{14}]^T$ (3) which represents its maximum structural similarity score $\Psi_i$ (Kotera et al., 2012) with the drugs in each of the 14 $\mathbb{S}_i$.
3. A 14-dimentional vector, $D^{\mathbf{FigSim}} = [\mathrm{T}_1\mathrm{T}_2\mathrm{T}_3 \ldots \mathrm{T}_{14}]^T$ (4), which represents its molecular fingerprint similarity score $\mathrm{T}_i$ (Xiao et al., 2013) with the drugs in each of the 14 $\mathbb{S}_i$.

Therefore, a given drug $D$ is formulated by:

$$D = D^{\mathbf{Int}} \oplus D^{\mathbf{StrSim}} \oplus D^{\mathbf{FigSim}} = [@_1 @_2 @_3 \ldots @_{42}]^T \tag{5}$$

Where $\oplus$ represents the symbol for orthogonal sum and where

$$@_{\mathrm{u}} = \begin{cases} \Phi_u (1 \le u \le 14) \\ \Psi_u (15 \le u \le 28) \\ \mathrm{T}_u (29 \le u \le 42) \end{cases} \tag{6}$$

For more details, refer to Cheng et al. (2017b).

## Measuring Label Correlation

In order to evaluate the correlation between two labels, we calculated the bias corrected Cramér's V statistic for all the label pairs (Bergsma, 2013). Cramér's V (sometimes referred to as Cramér's phi and denoted as φc) statistic is a measure of association between two nominal variables, ranging from 0 to 1 (inclusive). The bias corrected Cramér's V statistic is given by (here $n$ denotes sample size and $\chi^2$ stands for the chi-square statistic without a continuity correction for a contingency table with $r$ rows and $c$ columns)

$$\tilde{V} = \sqrt{\frac{\tilde{\varphi}^2}{\tilde{m}}} \tag{7}$$

where

$$\tilde{\varphi}^2 = \max(0, \varphi^2 - \frac{(r-1)(c-1)}{n-1})) \tag{8,}$$

$$\varphi^2 = \frac{\chi^2}{n} \tag{9}$$

and

$$\tilde{m} = \min(\tilde{r} - 1, \tilde{c} - 1) \tag{10,}$$

$$\tilde{r} = r - \left( \frac{(r-1)^2}{n-1} \right) \tag{11,}$$

$$\tilde{c} = c - \left( \frac{(\mathrm{c}-1)^2}{n-1} \right) \tag{12.}$$

## Network-Based Label Space Partition

The NLSP is a newly proposed multilabel learning method and has achieved top performance in some predictive tasks (Szymański et al., 2016). In this study, we adopted the data-driven NLSP method for prediction of ATC classes of a compound. NLSP divides the predictive modeling task into the training and classification phase.

In the training phase, four steps are preformed:

1. Establishing a label co-occurrence graph on the training set. The label co-occurrence graph $G$ has the label set $L$ as the vertex set and the edge between two vertices (labels) exists if at least one sample $S$ in training set $D_{train}$ is assigned by these two labels $l_i$ and $l_j$ together (here $l_i$, $l_j$ denote labels of the set $L_s$, which stands for the assigned label set of a sample $S$; $\|\ \|$ stands for the cardinality of a given set):

$$E = \left\{ \{l_i, l_j\} : (\exists (S, L_s) \in \mathrm{D}_{train}) \left( l_i \in L_s \wedge l_j \in L_s \right) \right\} \tag{13}$$

We can also easily assign weights to $G$ by defining a counting function $w\colon L \to \mathbb{N}$:

$$\begin{aligned} w\left(l_i, l_j\right) &= \text{number of sample } S \text{ that have both labels assigned} \\ &= \left\| \left\{ S : (S, L_s) \in \mathrm{D}_{train} \wedge l_i \in L_s \wedge l_j \in L_s \right\} \right\| \end{aligned} \tag{14}$$

2. Detecting community on the label co-occurrence graph. There are various community detection algorithms. In this study, we utilized the following two methods to identify communities because both of the two methods have linear time complexity:

   a) **Largest modularity using incremental greedy search (Louvain method)** (Blondel et al., 2008): This method is based on greedy aggregation of communities, beginning with communities with single convex and merging the communities iteratively. In each step, two communities are merged when the merging makes the highest contribution to modularity. The algorithm halts when there is no merge that could increase current modularity. This method is frequently referred as "Louvain method" in the network research community. The detailed explanation of this method is described in **Supplementary Method S1**.

   b) **Multiple async label propagation (LPA)** (Raghavan et al., 2007): This method assigns unique tags to every vertex in a graph and then iteratively updates the tags of every vertex. This update reassigns the tag of the majority of neighbors to the central vertex. The updating order of vertices shuffled at each iteration. The algorithm is stopped when all vertices have tags identical to the dominant tag in proximity. The detailed description of LPA is appended in **Supplementary Method S2**.

3. For each community $C_i$, corresponding training set $D_i$ is created by taking the original dataset with label columns presented in $L_i$.
4. For each community, a base predictor $b_i$ is learnt on the training set $D_i$. In this study, we compared the performance of five types of base predictors:

   (a) **Extremely randomized trees (ERT)** (Geurts et al., 2006; Li et al., 2019) is an ensemble method that adds more randomness compared to random forests by the random top–down splitting of trees instead of computing the locally optimal cut-point for each feature under consideration. This increase in randomness allows to reduce the variance of the model a bit, at the expense of a slightly greater increase in bias.

(b) **Random forests (RF)** (Breiman, 2001) is an ensemble method that combines the probabilistic predictions of a number of decision tree-based classifiers to improve the generalization ability over a single estimator.

(c) **Support vector machine (SVM)** (Cortes and Vapnik, 1995) is a widely used classification algorithm which tries to find the maximum margin hyperplane to divide samples into different classes. Incorporated by kernel trick, this method could handle both linear and no-linear decision boundary.

(d) **Extreme gradient boosting (XGB)** (Chen and Guestrin, 2016) is a newly proposed boosting method, which has achieved state-of-the-art performance on many tasks with tabular training data (Chen et al., 2018). Traditional gradient boosting machine is a meta algorithm to build an ensemble strong learner from weak learners such as decision trees, while XGB is an efficient and distributed implementation of gradient boosting machine.

(e) **Multilayer perceptron (MLP)** (Ruck et al., 1990) is a supervised learning algorithm which could learn nonlinear models. It has one or more nonlinear hidden layers between the input and output. For each hidden layer, different numbers of hidden neurons can be assigned. Each hidden neuron yields a weighted linear summation of the values from the previous layer, and the nonlinear activation function is followed. The weights are learnt through backpropagation algorithm or variations upon it.

In the classification phase, we just perform predication on all communities detected in the training phase and fetch the union of assigned labels:

$$b(S) = \bigcup_{j=1}^{k} b_i(S) \tag{15}$$

## Parameter Tuning

There are two layers of hyperparameters tunable for NLSP:

1. The base learner: we chose five types of base learners.

   (a) Extremely randomized trees: we tuned the hyperparameter of number of trees at [500, 1000], other hyperparameters are at the default values.

   (b) Random forests: we tuned hyperparameter of number of trees at [500, 1,000], other hyperparameters are at the default values.

   (c) Support vector machine: we tuned the hyperparameter of $C$ (penalty) at [0.01, 0.1, 1, 10, 100], we chose the radial basis function with gamma value of $\frac{1}{N_{features}} = \frac{1}{42}$, other hyperparameters are at the default values.

   (d) Extreme gradient boosting: we tuned the hyperparameter of number of trees at [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], other hyperparameters are at the default values.

   (e) Multilayer perceptron: We tuned the hyperparameter of hidden layer sizes at [50, 100, 200, 500, 1,000], other hyperparameters are at the default values.

2. The cluster: for each type of base learner, we try to compare two community detection methods.

   (a) Largest modularity using incremental greedy search (Blondel et al., 2008).

   (b) Multiple async label propagation (Raghavan et al., 2007).

## Performance Measures of Multilabel Learning

Evaluation of a multilabel learning model is not a trivial task (Zhang et al., 2015; Yuan et al., 2016; Zhang et al., 2017; You et al., 2018; Xiong et al., 2019; You et al., 2019). Inspired by the definition of Chou *et al.* (Chou, 2013) and practice of Madjarov et al. (2012), we utilized the following five metrics to evaluate the multilabel learning models throughout this work.

$$\begin{cases} Aiming = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k^*\right\|}\right) \\ Coverage = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k\right\|}\right) \\ Accuracy = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}\right) \\ Absolute\ True = \frac{1}{N}\sum_{k=1}^{N}\ (\mathbb{L}_k, \mathbb{L}_k^*) \\ Hamming\ loss = \frac{1}{N}\sum_{k=1}^{N}\left\|\mathbb{L}_k \ominus \mathbb{L}_k^*\right\| \end{cases} \tag{16}$$

where $N$ is the total number of samples, $M$ is the total number of labels, $\bigcup$ represents union in set theory and $\bigcap$ represents intersection in set theory, $\mathbb{L}_k$ denotes the true label set of $k$-th sample, $\mathbb{L}_k^*$ means the predicted label vector of $k$-th sample, $\ominus$ stands for the symmetric difference between two sets, and

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, if\ all\ the\ labels\ in\ \mathbb{L}_k\ equal\ \mathbb{L}_k^* \\ 0, otherwise \end{cases} \tag{17}$$

In order to avoid the zero-divisor problem generated by all negative predictions, we add a pseudo-number 1 to 0 divisors in the calculation of the aiming metric. These above metrics have been used in a series of studies (Cheng et al., 2017a; Cheng et al., 2017b; Nanni and Brahnam, 2017).

## Performance Measures of Single-Label Learning

Apart from the metrics in the multilabel framework, we also utilized the following metrics to assess the single-label classification models.

$$\begin{cases} Accuracy = \dfrac{TP + TN}{TP + TN + FN + FP} \\ Specificity = \dfrac{TN}{TN + FP} \\ Recall = \dfrac{TP}{TP + FN} \\ F1 = \dfrac{2TP}{2TP + FP + FN} \end{cases} \tag{18}$$

where $TP$, $TN$, $FN$, and $TN$ are true positives, true negatives, false positives, and false negatives for the prediction of each label, respectively. These metrics have widely been used in a large number of bioinformatics applications recently (Feng et al., 2017; Niu and Zhang, 2017; Sun et al., 2017; Wang et al., 2017; Xu et al., 2017; He et al., 2018; Li et al., 2018; Pan et al., 2018; Qiao et al., 2018; Xiong et al., 2018; Xu et al., 2018; Zhang et al., 2018; Bian et al., 2019; Wei et al., 2019a; Wei et al., 2019b; Zou et al., 2019). In addition, we also calculated the area under the receive operating characteristic curve (AUC) by the trapezoidal rule.

## Model Validation Method

There are mainly three methods to evaluate the generalization ability of a classification model, such as the independent testing method, $k$-fold cross validation, and the jackknife method. In order to fairly compare our proposed model with previous works on the same benchmark dataset, we utilized the jackknife method for the model validation in the multilabel learning framework. Jackknife is a resampling method for parameter estimation. The jackknife estimation of a parameter is constructed by calculating the parameter for each subsample omitting the $i$-th observation and then takes the mean value of these parameters as final estimation.

In the model validation of single-label analysis, we utilized 10 times repeated 10-fold cross validation (10 × 10-fold CV) method. In $k$-fold cross validation (CV), the sample set is randomly partitioned into $k$ subsets with equal size. Of the $k$ subsets, one subset is selected as the validation data for testing the model, and the remaining $k - 1$ subsets are used for training. The cross-validation process is then repeated $k$ times (the folds), with each of the $k$ subsets used exactly once as the validation data. The 10-fold cross-validation is proven to be a better alternative of jackknife method in terms of bias, variance, and computation complexity (Kohavi, 1995). We also repeated 10-fold CV 10 times in shuffled benchmark dataset to further reduce the estimation variance.

# RESULTS AND DISCUSSION

## Label Correlation Analysis

One major advantage of multilabel learning framework is the explicit exploitation of label correlations (Zhang and Zhou, 2014). We calculated bias corrected Cramér's V statistics for all the label pairs and depicted them in a heatmap manner (**Figure 1A**), and the UpSet visualization of label intersections is depicted in **Figure 1B**. The results indicated that 46 drugs are both labeled as ATC category 4 (dermatologicals) and ATC category 12 (respiratory system), 43 drugs are both labeled as ATC category 13 (sensory organs) and ATC category 7 (anti-infectives for systemic use), which can be explained by the fact that many widely applied corticosteroids, such as dexamethasone, betamethasone, and fluocortolone, can be used both in dermatology and respirology medicine. We also found that several label sets are correlated, especially for ATC category 4 (dermatologicals) and ATC category 13 (sensory organs), of which the Cramér's V statistic

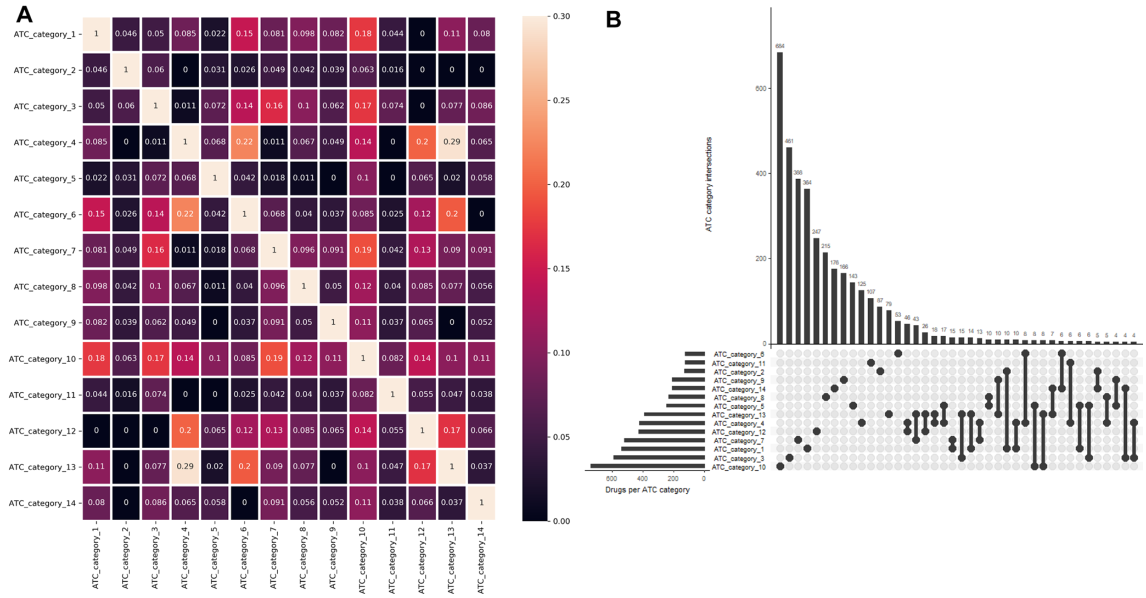


**FIGURE 1 |** Label correlation landscape. **(A)** The pair wise visualization of Cramér's V statistics for all the labels in a heatmap manner. **(B)** The UpSet visualization of label intersections. The horizontal bar shows the number of drugs per ATC category, and the vertical bar shows the number of drugs per ATC category intersection.

is 0.29. Details about the pairwise intersection numbers of drugs and the pairwise Cramér's V statistics between all the labels are shown in **Table S1** and **Table S2**.

## Multilabel Performance Comparison

**Table 1** shows the prediction performances based on the jackknife test among different methods on the benchmark dataset. We found the absolute true value of almost all our NLSP-based methods performed better than that of other methods, which is the most stringent metric for multilabel learning. Among all the NLSP-based models, the NLSP-XGB-LPA performs the best, consistently better than all the other methods trained on benchmark dataset, in terms of aiming, coverage, accuracy, and absolute true. As for the value of absolute true, our NLSP-XGB-LPA has boosted ~11.67% compared to the best deep learning model trained on the same benchmark dataset (Lumini and Nanni, 2018). As for the clusterer, we found that the LPA method performs consistently better than the Louvain method in all the NLSP-based models (**Figure S1**), so we append the suffix of "-LPA" to all the NLSP-based models. We then trained the final NLSP-XGB-LPA model on the full benchmark dataset using previous optimized hyperparameters. This model can be accessed through https://github.com/dqwei-lab/ATC.

## Label Community Analysis

One major innovation of NLSP method is the construction of label relation graph, which is built on the concept of label co-occurrence (Szymanski and Kajdanowicz, 2019). The communities detected in the label relation graph will not only help to improve the classification performance but also provide us with deeper insights of the intrinsic label structure. We extracted the community membership information from the final model of NLSP-XGB-LPA (shown in **Figure 2**). We found that there are two communities detected, in which ATC category 8 (anti-infectives for systemic use) lies in a unique community. In terms of medicinal chemistry and clinical pharmacotherapeutics, anti-infectives for systemic use are structure variant and usage limited compared to other 16 types of drugs. For example, daptomycin (DB00080) is one of the anti-infectives for systemic use, which is composed of an unusual molecular structure of lipopeptide with limited indications for skin and skin structure infections caused by Gram-positive infections, *S. aureus* bacteremia, and right-sided *S. aureus* endocarditis (Henken et al., 2010). The community membership learnt from benchmark dataset is surprising but intuitive. This result suggests the potential pattern extraction power of network-based machine learning models in terms of pharmacology.

## Single-Label Analysis

Apart from multilabel learning metrics, it is often useful to evaluate multilabel learning models in a label-wise manner (Michielan et al., 2009; Mayr et al., 2016). We utilized the parameters of the best-performing model of NLSP-XGB-LPA and conducted 10 times repeated 10-fold cross-validation ($10 \times 10\text{-fold CV}$) because the jackknife test is rather time consuming. The details are listed in **Table 2**. We found that our NLSP-XGB-LPA performs well in all the single-label subtasks of ATC prediction, especially for the label of "anti-infectives for systemic use," reaching an AUC at 0.9946. Compared to a dedicated single-label classification system for cardiovascular system (Gurulingappa et al., 2009), our best-performing multilabel model boosted the value of accuracy from 0.8947 into 0.9490.

**TABLE 1 |** Comparison with other state-of-the-art multilabel predictors.

| Method | DL[a] | Aiming | Coverage | Accuracy | Absolute true | Hamming loss |
|---|---|---|---|---|---|---|
| EnsANet_LR ⊕ DO[c] (τ = 0.25)<br>(Lumini and Nanni, 2018) | Yes | 0.7957 | 0.8335 | 0.7778 | 0.7090 | Not available |
| EnsANet_LR ⊕DO[c] (τ = 0.5)<br>(Lumini and Nanni, 2018) | Yes | 0.9011 | 0.7162 | 0.7232 | 0.6871 | |
| EnsLIFT<br>(Nanni and Brahnam, 2017) | No | 0.7818 | 0.7577 | 0.7121 | 0.6330 | |
| iATC-mHyb[c]<br>(Cheng et al., 2017a) | No | 0.7191 | 0.7146 | 0.7132 | 0.6675 | |
| Chen et al.<br>(Chen et al., 2012) | No | 0.5076 | 0.7579 | 0.4938 | 0.1383 | |
| iATC-mISF<br>(Cheng et al., 2017b) | No | 0.6783 | 0.6710 | 0.6641 | 0.6098 | |
| NLSP-ERT-LPA | No | 0.7948 | 0.7691 | 0.7578 | 0.7213 | 0.03817 |
| NLSP-RF-LPA | No | 0.8072 | 0.7889 | 0.7778 | 0.7489 | **0.03427** |
| NLSP-SVM-LPA | No | 0.7844 | 0.7529 | 0.7370 | 0.6925 | 0.04322 |
| NLSP-XGB-LPA | No | **0.8135**[b] | **0.7950** | **0.7828** | **0.7497** | 0.03429 |
| NLSP-MLP-LPA | No | 0.7958 | 0.7858 | 0.7591 | 0.7090 | 0.04032 |

[a] *DL denotes whether this model is a deep learning-based method.*

[b] *The bold value stands for the best value of specific metrics.*

[c] *These models are trained on a modified benchmark dataset, whose metrics are not comparable to our model.*

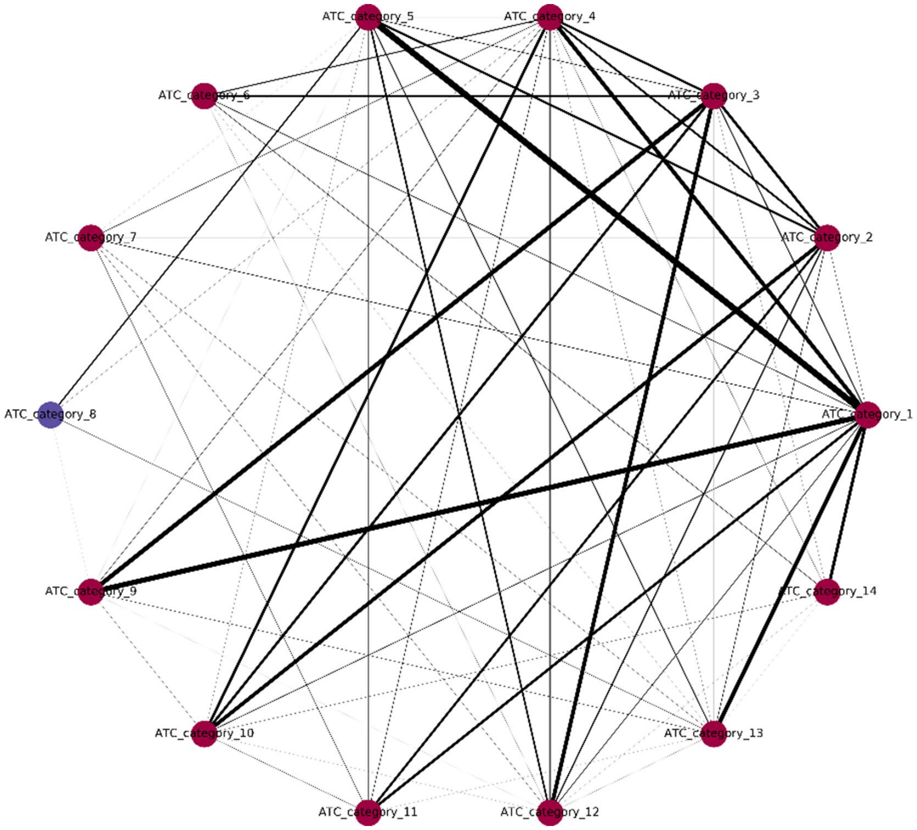


**FIGURE 2 |** Label relation graph. Different colors stand for different communities. The line width represents the weight between two labels. Communities are detected by multiple async label propagation method, while the weight represents the frequency of label co-occurrence.

**TABLE 2 |** Label-wise analysis of best-performing multilabel learning model.

| Predictive label | Accuracy | Specificity | Recall | F1 score | AUC | Evaluation method |
|---|---|---|---|---|---|---|
| Alimentary tract and metabolism | 0.9269 | 0.7312 | 0.7549 | 0.7406 | 0.9550 | 10 × 10-fold CV |
| Blood and blood forming organs | 0.9793 | 0.7754 | 0.5644 | 0.6430 | 0.9493 | 10 × 10-fold CV |
| Cardiovascular system | 0.9490 | 0.8371 | 0.8274 | 0.8306 | 0.9752 | 10 × 10-fold CV |
| Dermatologicals | 0.9403 | 0.7966 | 0.6038 | 0.6845 | 0.9472 | 10 × 10-fold CV |
| Genitourinary system and sex hormones | 0.9691 | 0.8148 | 0.6682 | 0.7294 | 0.9539 | 10 × 10-fold CV |
| Systemic hormonal preparations, excluding sex hormones and insulins | **0.9867**[a] | 0.8227 | 0.7605 | 0.7816 | 0.9940 | 10 × 10-fold CV |
| Anti-infectives for systemic use | 0.9793 | **0.9276** | **0.9170** | **0.9215** | **0.9946** | 10 × 10-fold CV |
| Antineoplastic and immunomodulating agents | 0.9792 | 0.8683 | 0.7724 | 0.8126 | 0.9804 | 10 × 10-fold CV |
| Musculoskeletal system | 0.9820 | 0.8707 | 0.7836 | 0.8209 | 0.9842 | 10 × 10-fold CV |
| Nervous system | 0.9511 | 0.8581 | 0.8913 | 0.8733 | 0.9825 | 10 × 10-fold CV |
| Antiparasitic products, insecticides and repellents | 0.9863 | 0.8312 | 0.7358 | 0.7714 | 0.9803 | 10 × 10-fold CV |
| Respiratory system | 0.9573 | 0.8432 | 0.7516 | 0.7923 | 0.9720 | 10 × 10-fold CV |
| Sensory organs | 0.9492 | 0.8206 | 0.6367 | 0.7140 | 0.9487 | 10 × 10-fold CV |
| Various | 0.9717 | 0.7681 | 0.6997 | 0.7241 | 0.9703 | 10 × 10-fold CV |
| Cardiovascular system (Gurulingappa et al., 2009) | 0.8947 | | Not available | | | 100 × bootstrapping |
| Cardiovascular system (Gurulingappa et al., 2009) | 0.7712 | | | | | Test set |
| SuperPred (Dunkel et al., 2008) | 0.676[b] | | | | | Jackknife |

[a] *The bold value stands for the best value of specific metrics.*

[b] *The mean accuracy of flattened 850 ATC classes.*

## CONCLUSION

Based upon the NLSP method, we have achieved the state-of-the-art performance on the benchmark dataset using the similarity-based features such as chemical–chemical interaction and structural and fingerprint similarities of a compound to other compounds belonging to the different ATC categories. Label community and single-label analysis were also performed on the benchmark dataset. There are three major conclusions can be reached. First, compared to dedicated single-label models (Dunkel et al., 2008; Gurulingappa et al., 2009), multilabel learning framework could improve the performance on single-label metrics by incorporating label correlation information. Second, compared to feature engineering tricks (Nanni and Brahnam, 2017; Lumini and Nanni, 2018), the introduction of new method such as NLSP could generate more performance improvement. Third, at least in the ATC prediction task, the NLSP method, which adopts ideas from network research community and captures the correlation of labels in a data-driven manner, can perform better than the models based on deep learning techniques, especially in the absolute true rate metric. The idea behind NLSP method is fascinating, and the power of NLSP remains to be unleashed for the multilabel learning tasks in drug discovery.

Although the NLSP method was the first time to be applied to the multilabel classification task in pharmacology and achieved good performance in the preliminary results, there are shortcomings in several aspects in this study. First, the similarity-based features are not recalculated for the specific communities detected by the NLSP methods. Second, the rigidity of the model validation can be improved by the independent external dataset. Last but not the least, the number of communities detected by NLSP on this drug classification problem is too low, which may be not an ideal dataset for proving the predictive power of the NLSP-based method. These problems can be addressed in the further studies.

## DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or **Supplementary Files**.

## AUTHOR CONTRIBUTIONS

YX, D-QW and XW contributed conception and design of the study; XW and YW organized the database; XW, YW and ZX performed the statistical analysis; XW wrote the first draft of the manuscript; XW and YW wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

## FUNDING

This work was supported by the funding from National Key Research Program (contract no. 2016YFA0501703), National Natural Science Foundation of China (grant no. 31601074, 61872094, 61832019), and Shanghai Jiao Tong University School of Medicine (contract no. YG2017ZD14).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00971/full#supplementary-material

## REFERENCES

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* 42 (3), 323–328. doi: 10.1016/j.jkss.2012.10.002

Bian, Y., Jing, Y., Wang, L., Ma, S., Jun, J. J., and Xie, X. Q. (2019). Prediction of orthosteric and allosteric regulations on cannabinoid receptors using supervised machine learning classifiers. *Mol. Pharm.* 16 (6), 2605–2615. doi: 10.1021/acs.molpharmaceut.9b00182

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory E.* 2008 (10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi: 10.1023/A:1010933404324

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA: ACM). doi: 10.1145/2939672.2939785

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23 (6), 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Chen, L., Zeng, W. M., Cai, Y. D., Feng, K. Y., and Chou, K. C. (2012). Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. *PLoS One* 7 (4), e35254. doi: 10.1371/journal.pone.0035254

Cheng, X., Zhao, S. G., Xiao, X., and Chou, K. C. (2017a). iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget* 8 (35), 58494–58503. doi: 10.18632/oncotarget.17028

Cheng, X., Zhao, S. G., Xiao, X., and Chou, K. C. (2017b). iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* 33 (3), 341–346. doi: 10.1093/bioinformatics/btw644

Chou, K. C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* 9 (6), 1092–1100. doi: 10.1039/c3mb25555g

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007/BF00994018

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings*, 886–893. doi: 10.1109/CVPR.2005.177

Dunkel, M., Gunther, S., Ahmed, J., Wittig, B., and Preissner, R. (2008). SuperPred: drug classification and target prediction. *Nucleic Acids Res.* 36, W55–W59. doi: 10.1093/nar/gkn307

Feng, P., Zhang, J., Tang, H., Chen, W., and Lin, H. (2017). Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. *Interdiscip. Sci.* 9 (4), 540–544. doi: 10.1007/s12539-016-0193-4

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63 (1), 3–42. doi: 10.1007/s10994-006-6226-1

Gibaja, E., and Ventura, S. (2014). Multi-label learning: a review of the state of the art and ongoing research. *WIREs Data Mining Knowl. Discov.* 4 (6), 411–444. doi: 10.1002/widm.1139

Gurulingappa, H., Kolarik, C., Hofmann-Apitius, M., and Fluck, J. (2009). Concept-based semi-automatic classification of drugs. *J. Chem. Inf. Model.* 49 (8), 1986–1992. doi: 10.1021/ci9000844

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

Henken, S., Bohling, J., Martens-Lobenhoffer, J., Paton, J. C., Ogunniyi, A. D., Briles, D. E., et al. (2010). Efficacy profiles of daptomycin for treatment of invasive and noninvasive pulmonary infections with *Streptococcus pneumoniae. Antimicrob. Agents Chemother.* 54 (2), 707–717. doi: 10.1128/AAC.00943-09

Hutchinson, J. M., Patrick, D. M., Marra, F., Ng, H., Bowie, W. R., Heule, L., et al. (2004). Measurement of antibiotic consumption: a practical guide to the use of the anatomical thgerapeutic chemical classification and definied daily dose system methodology in Canada. *Can. J. Infect. Dis.* 15 (1), 29–35. doi: 10.1155/2004/389092

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," (Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.).

Kotera, M., Hirakawa, M., Tokimatsu, T., Goto, S., and Kanehisa, M. (2012). The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.* 802, 19–39. doi: 10.1007/978-1-61779-400-1_2

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20 (12), 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, X., Xu, Y., Lai, L., and Pei, J. (2018). Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Mol. Pharm.* 15 (10), 4336–4345. doi: 10.1021/acs.molpharmaceut.8b00110

Li, Y., Niu, M., and Zou, Q. (2019). ELM-MHC: an improved MHC identification method with extreme learning machine algorithm. *J. Proteome Res.* 18 (3), 1392–1401. doi: 10.1021/acs.jproteome.9b00012

Lumini, A., and Nanni, L. (2018). Convolutional neural networks for ATC classification. *Curr. Pharm. Des.* 24 (34), 4007–4012. doi: 10.2174/138161282 4666181112113438

MacDonald, K., and Potvin, K. (2004). Interprovincial variation in access to publicly funded pharmaceuticals: a review based on the WHO Anatomical Therapeutic Chemical Classification System. *Can. Pharm. J.* 137 (7), 29–34. doi: 10.1177/171516350413700703

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit.* 45 (9), 3084–3104. doi: 10.1016/j.patcog.2012.03.004

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3, 80. doi: 10.3389/fenvs.2015.00080

Michielan, L., Terfloth, L., Gasteiger, J., and Moro, S. (2009). Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J. Chem. Inf. Model.* 49 (11), 2588–2605. doi: 10.1021/ci900299a

Moyano, J. M., Gibaja, E. L., Cios, K. J., and Ventura, S. (2018). Review of ensembles of multi-label classifiers: models, experimental study and prospects. *Inf. Fusion* 44, 33–45. doi: 10.1016/j.inffus.2017.12.001

Nanni, L., and Brahnam, S. (2017). Multi-label classifier based on histogram of gradients for predicting the anatomical therapeutic chemical class/classes of a given compound. *Bioinformatics* 33 (18), 2837–2841. doi: 10.1093/bioinformatics/btx278

Niu, Y., and Zhang, W. (2017). Quantitative prediction of drug side effects based on drug-related features. *Interdiscip. Sci.* 9 (3), 434–444. doi: 10.1007/s12539-017-0236-5

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34 (9), 1473–1480. doi: 10.1093/bioinformatics/btx822

Qiao, Y., Xiong, Y., Gao, H., Zhu, X., and Chen, P. (2018). Protein–protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinformatics* 19 (1), 14. doi: 10.1186/s12859-018-2009-5

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76 (3), 036106. doi: 10.1103/PhysRevE.76.036106

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Mach. Learn.* 85 (3), 333–359. doi: 10.1007/s10994-011-5256-5

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., and Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans. Neural Netw.* 1 (4), 296–298. doi: 10.1109/72.80266

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 18 (1), 277. doi: 10.1186/s12859-017-1700-2

Szymanski, P., and Kajdanowicz, T. (2019). Scikit-multilearn: a scikit-based Python environment for performing multi-label classification. *J. Mach. Learn. Res.* 20 (1), 209–230.

Szymański, P., Kajdanowicz, T., and Kersting, K. (2016). How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy* 18 (8), 282. doi: 10.3390/e18080282

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2011). Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* 23 (7), 1079–1089. doi: 10.1109/TKDE.2010.164

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17 (17–18), 1700262. doi: 10.1002/pmic.201700262

Wang, N. N., Huang, C., Dong, J., Yao, Z. J., Zhu, M. F., Deng, Z. K., et al. (2017). Predicting human intestinal absorption with modified random forest approach: a comprehensive evaluation of molecular representation, unbalanced data, and applicability domain issues. *RSC Adv.* 7 (31), 19007–19018. doi: 10.1039/C6RA28442F

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2019a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35 (8), 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019b). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. doi: 10.1093/bioinformatics/btz408

Wu, L., Ai, N., Liu, Y., Wang, Y., and Fan, X. (2013). Relating anatomical therapeutic indications by the ensemble similarity of drug sets. *J. Chem. Inf. Model.* 53 (8), 2154–2160. doi: 10.1021/ci400155x

Xiao, X., Min, J. L., Wang, P., and Chou, K. C. (2013). iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.* 337, 71–79. doi: 10.1016/j.jtbi.2013.08.013

Xiong, Y., Qiao, Y., Kihara, D., Zhang, H. Y., Zhu, X., and Wei, D. Q. (2019). Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates. *Curr. Drug Metab.* 20 (3), 229–235. doi: 10.2174/1389200219666181019094526

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Xu, Y., Chen, P., Lin, X., Yao, H., and Lin, K. (2018). Discovery of CDK4 inhibitors by convolutional neural networks. *Future Med. Chem.* 11, 165–177 doi: 10.4155/fmc-2018-0478

You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., et al. (2019). NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.* 47 (W1), W379–W387. doi: 10.1093/nar/gkz388

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 34 (14), 2465–2473. doi: 10.1093/bioinformatics/bty130

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32 (12), i18–i27. doi: 10.1093/bioinformatics/btw244

Zhang, M. L., and Zhou, Z. H. (2005). A k-nearest neighbor based algorithm for multi-label classification. *2005 IEEE International Conference on Granular Computing, Vols 1 and 2*, 718–721. doi: 10.1109/GRC.2005.1547385

Zhang, M. L., and Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* 26 (8), 1819–1837. doi: 10.1109/TKDE.2013.39

Zhang, W., Liu, F., Luo, L., and Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinformatics* 16, 365. doi: 10.1186/s12859-015-0774-y

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA–protein interactions. *PLoS Comput. Biol.* 14 (12), e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinformatics* 18 (Suppl 13), 464. doi: 10.1186/s12859-017-1875-6

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 25 (2), 205–218. doi: 10.1261/rna.069112.118

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

REVIEW
published: 05 November 2019
doi: 10.3389/fphar.2019.01303

# Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research

*Laurianne David*[1,2]*, *Josep Arús-Pous*[1,3], *Johan Karlsson*[4], *Ola Engkvist*[1], *Esben Jannik Bjerrum*[1], *Thierry Kogej*[1], *Jan M. Kriegl*[5], *Bernd Beck*[5] *and Hongming Chen*[1,6]*

*[1] Hit Discovery, Discovery Sciences, Biopharmaceutical R&D, AstraZeneca, Gothenburg, Sweden, [2] Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany, [3] Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland, [4] Quantitative Biology, Discovery Sciences, Biopharmaceutical R&D, AstraZeneca, Gothenburg, Sweden, [5] Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany, [6] Chemistry and Chemical Biology Centre, Guangzhou Regenerative Medicine and Health – Guangdong Laboratory, Guangzhou, China*

OPEN ACCESS

**Edited by:**
*Jianfeng Pei,*
*Peking University, China*

**Reviewed by:**
*Alexander Sedykh,*
*Sciome LLC, United States*
*Maxim Kuznetsov,*
*Insilico Medicine, Inc., United States*

***Correspondence:**
*Laurianne David*
*Laurianne.david1@gmail.com*
*Hongming Chen*
*Hongming.Chen71@hotmail.com*

**Specialty section:**
*This article was submitted to Experimental Pharmacology and Drug Discovery, a section of the journal Frontiers in Pharmacology*

**Received:** *07 August 2019*
**Accepted:** *14 October 2019*
**Published:** *05 November 2019*

**Citation:**
*David L, Arús-Pous J, Karlsson J, Engkvist O, Bjerrum EJ, Kogej T, Kriegl JM, Beck B and Chen H (2019) Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research. Front. Pharmacol. 10:1303. doi: 10.3389/fphar.2019.01303*

In recent years, the development of high-throughput screening (HTS) technologies and their establishment in an industrialized environment have given scientists the possibility to test millions of molecules and profile them against a multitude of biological targets in a short period of time, generating data in a much faster pace and with a higher quality than before. Besides the structure activity data from traditional bioassays, more complex assays such as transcriptomics profiling or imaging have also been established as routine profiling experiments thanks to the advancement of Next Generation Sequencing or automated microscopy technologies. In industrial pharmaceutical research, these technologies are typically established in conjunction with automated platforms in order to enable efficient handling of screening collections of thousands to millions of compounds. To exploit the ever-growing amount of data that are generated by these approaches, computational techniques are constantly evolving. In this regard, artificial intelligence technologies such as deep learning and machine learning methods play a key role in cheminformatics and bio-image analytics fields to address activity prediction, scaffold hopping, *de novo* molecule design, reaction/retrosynthesis predictions, or high content screening analysis. Herein we summarize the current state of analyzing large-scale compound data in industrial pharmaceutical research and describe the impact it has had on the drug discovery process over the last two decades, with a specific focus on deep-learning technologies.

**Keywords: Artificial intelligence, deep learning, Chemogenomics, Large-scale data, pharmaceutical industry**

## INTRODUCTION

Digital data, in all shapes and sizes, are growing exponentially. According to the National Security Agency of the United States, the Internet is processing around 1.8 billion GB of data per day (Macarron et al., 2011). In 2011, digital information has grown nine times in volume in just 5 years (Mayr and Bojanic, 2009) and by 2020, its amount in the world is expected to reach 35 trillion GB (Borman, 1999). The recent development of deep learning and other artificial intelligence methods is fuelled by the desire to seek greater insight among the ever-increasing amount of data in several key industries and powered by technological advancements as in, for example, computer vision, natural language processing, internet of things (IoT), or computer hardware.

Over the past decade, there has been a remarkable increase in the amount of available compound activity, biomedical (Borman, 1999; Mayr and Bojanic, 2009; Schamberger et al., 2011), and genomics data (Guyer and Collins, 1995; Human Genome Project Results; Wilson and Nicholls, 2015) thanks to the rapid development of high-throughput screening (HTS) and gene sequencing technologies. Typically, databases in pharma companies contain around 1–4 million compounds with biological data for several thousands of biological end-points such as targets or activities in cellular assays. Furthermore, due to the increasing level of automation and standardization, larger data sets of consistent conditions have become available. All chemical compounds synthesized and/or extracted from publications represent around 96 million compounds (Kim et al., 2019). Even though only a small fraction of them have associated biological information (Wang et al., 2014; Kim, 2016), these chemogenomics data sets alone already represent a formidable task for predictive modelling work.

The usage of new automation technologies resulted in a large volume of data, which has promoted the usage of machine learning (ML) methods. ML methods such as support vector machine (SVM), random forest (RF), or neural networks (NNs) have been used for data modelling in cheminformatics and bioinformatics for a long time. Only recently, various deep learning methods have become more popular due to the availability of large-scale training sets and high-performance computer hardware. An important difference between deep learning and previous ML methods is the flexibility of NN architectures and input/output data structures in deep learning methods and the automated extraction of features from raw data representations. This flexibility allows to design models that fit to the characteristics of the prediction problem (Wu et al., 2018; Xiong et al., 2019; Yang et al., 2019). Some of the popular NN architectures include convolutional NNs, recurrent NNs, autoencoders, and fully connected deep NNs. These deep learning methods have been applied (Ramsundar et al., 2017; Chen et al., 2018) on aspects of compound activity prediction (Dahl et al., 2014; Ma et al., 2015; Koutsoukas et al., 2017), *de novo* molecular design (Brown et al., 2019), protein–ligand interaction prediction (Lenselink et al., 2017; Feinberg et al., 2018), predictive toxicity (Mayr et al., 2016), and reaction prediction (Segler and Waller, 2017b). In this review, we will provide an overview on various types of large-scale data sets that are available in pharmaceutical industry. Such data sets offer a wealth of information that are unavailable in the public domain and give rise to a broad range of applications. Furthermore, we will exemplify the applications of artificial intelligence, in particular deep-learning technologies, that are powered through these large data sets on various problems in drug discovery.

## LARGE-SCALE COMPOUND DATA IN PHARMACEUTICAL INDUSTRY

The past two decades have seen an acceleration of compound data generation in pharmaceutical industry driven by the technical advancement of HTS (Mayr and Bojanic, 2009; Macarron et al., 2011), parallel chemical synthesis (Borman, 1999), as well as the by the introduction of automation in sequencing and imaging. The various types of large-scale compound data in pharmaceutical research are illustrated in **Figure 1**. A small molecule database belongs to the core infrastructure of industrial pharma R&D in order to store the results of lead identification and optimization campaigns, which are used for, e.g., structure–activity–relationship (SAR) analyses. The typical size of a compound collection at major pharma companies ranges from 1 to 4 million compounds (Schamberger et al., 2011; Kogej et al., 2013). Compound activity data (including Administration Distribution Metabolism Excretion Toxicology (ADMET) end points) are the major part of the “Compound Data Estate” in pharmaceutical industry. Most of the SAR data come from the HTS campaigns carried out during the drug discovery projects, which typically comprise crude readouts generated from *in vitro* assays at single compound concentration—so called single-shot-potency—in the primary screening stage, and more accurate concentration response data (IC50s, EC50s, etc.) derived from multiple compound concentration experiments. Pharmaceutical databases allow for in-depth studies that may not be achievable with public data. Indeed, structuration and curation of private databases are done with the inclusion of concepts such as screening campaigns or lead optimization programs, which make possible a faster and easier analysis of high-quality data. Occasionally, the overall number of SAR data points in pharmaceutical companies was disclosed in the past; some numbers reported in literature are listed in **Table 1**. Although this information is not up-to-date, it can still give a sense of the scale of experimental compound data in pharmaceutical industry.

Comparing with conventional HTS screening with a limited number of data readouts per compound, high-content screening (HCS) (Bickle, 2010) using automated microscopy generates images with multi-parameter readouts that provide an information-rich characterization of cellular phenotypic responses to small molecules. It has become an important tool for compound profiling and has led to a substantial increase in the amount of compound profiling data. For example, 460,800 images were produced through a screen comprising 100 384-well plates imaged with three fluorescent channels at four independent sites per well (Boutros et al., 2015). Hundreds of parameters can be extracted from each cell in the image quantifying information of morphological, geometric, intensity, and texture-based features. Recently Janssen reported (Simm et al., 2018) an image dataset for 524,371 compounds originally used for the detection of glucocorticoid receptor (GCR) nuclear translocation. For each cell in the image, 842 features were extracted, corresponding to roughly 440 million data points. The usage of image-based compound profiling data will be discussed in a subsequent section.

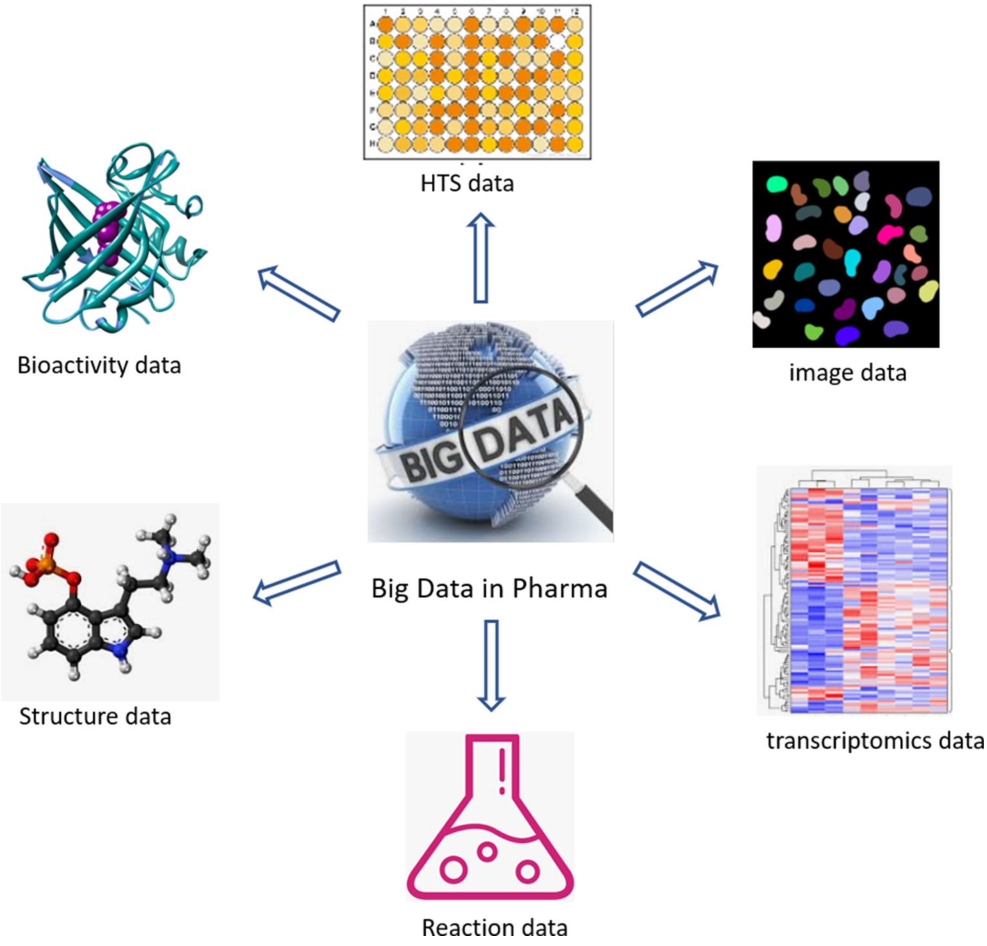


**FIGURE 1 |** Different categories of large-scale compound data in industrial pharmaceutical research.

**TABLE 1 |** Number of SAR data point in large pharmaceutical companies reported in literatures.

| Company | # of SAR point | Date | Reference |
|---|---|---|---|
| AstraZeneca | 150 million single-shot SAR points, 14 million[a] CR SAR points | Up to 2008 | (Proffitt, 2008; Muresan et al., 2011) |
| Boehringer Ingelheim | 260 million single-shot SAR points, 7 million CR SAR points | Up to 2011 | (Beck, 2012) |
| Pfizer | 0.6 million CR SAR points | Up to 2005 | (Paolini et al., 2006) |
| Johnson & Johnson | 30 million SAR points | Up to 2006 | (Agrafiotis et al., 2007) |

*a) This number includes external sources, up to 2012.*

High throughput mRNA expression profiling can be used to characterize the response of cell culture models to perturbations such as small molecules acting as pharmacologic modulators (Lamb et al., 2006; Iorio et al., 2013). These compounds induce transcriptional effects that can be used as gene signatures to discover new connections among compounds, pathways, and diseases. With one of these technologies, known as L1000™ Expression Profiling (profiling for 978 gene expressions) (De Wolf et al., 2016; Genometry), thousands of compounds can be screened per day at lower costs than conventional microarray techniques (Subramanian et al., 2017). Merck reported the screening of a set of 3,699 compounds using the Genometry L1000 platform to unveil a new target for compounds (Filzen et al., 2017). Janssen announced (How library-scale gene-expression profiling is changing drug discovery; Pascale, 2015) that they will use Genometry's L1000 platform to generate gene-expression profiles for 250,000 compounds from Janssen's small-molecule screening library. It is expected that more pharmaceutical companies will adopt similar technologies and approaches to generate large-scale transcriptomics data for compound profiling.

With the continuous increase in the amount and heterogeneity of data that are generated and stored in large repositories, the question of how to ensure and sustain data integrity gained more and more attention. The generation and storage of large amounts of data require significant investments in IT infrastructure. These investments are justified not only by efficiency gains for ongoing projects through elimination of manual steps to compile and analyze project-relevant data that ultimately lead to decisions on whether or not to pursue a certain molecule or compound class, but also perhaps even more so by the prospect to discover knowledge across projects as described for example in recent publications by Novartis (Wassermann et al., 2015a) or Boehringer Ingelheim (BI) (Beck, 2012). All this is only possible if the data context is provided alongside the data itself, and when there is a profound understanding of the data quality. One important aspect for consideration is the assay technology that is applied for compound testing. The direct interference of compounds with an assay technology is a source for systematic errors, which should be considered when analyzing the respective data sets. In a recent example at BI (Beck et al., 2015), the screening deck was assayed against an ion channel target for neuroprotection by means of a fluorometric imaging plate reader (FLIPR) assay (Sullivan et al., 1999). The screen yielded a high hit rate, and using a systematic overlap analysis with results from previous FLIPR campaigns, a large number of compounds most likely to be false positives were excluded from labor-intensive follow-up activities. Other important aspects regarding data quality are, for instance, compound purity, autofluorescence, or physicochemical properties such as aggregation propensity (Jadhav et al., 2010), which can have a significant influence on assay results and need therefore to be taken into account as decision-relevant context. This can be accomplished by computational surrogate parameters or auxiliary experiments such as high-throughput solubility determination *via* nephelometry (Fligge and Schuler, 2006).

Typically, data repositories within pharmaceutical companies evolve over years, and the best practices as to which data to store in such systems do so as well. This leads to situations in which legacy data are hardly comparable with present results, thereby limiting the chances to add value from mining data, which were generated at significantly different points in time. Efforts to set up data governance structures and to employ modern technologies around meta data management and central nomenclatures aim to address this issue and are currently underway in many companies (Proffitt, 2008).

## BIOLOGICAL PROFILING DESCRIPTORS FOR HIT EXPANSION

Traditionally, cheminformatic approaches focused on the use of molecular descriptors that are related to structure in order to describe the biological activities of compounds. Among them, structural fingerprints have been intensively used in similarity search, clustering, as well as in building SAR models (Willett, 2011). This is largely based on the hypothesis that structurally similar molecules are likely to bind to the same group of protein and then—as a consequence—share similar biological profiles (Martin et al., 2002; Keiser et al., 2007; Willett, 2011). In the late 1980s, NCI pioneered the implementation of a biological fingerprint to access the similarity of compounds (Paul et al., 1989). In contrast to structural fingerprints, biological activity data are utilized to describe a compound, neglecting structural features. Furthermore, with the recent advent of phenotypic screening, we observe an increasing awareness that the cellular effects of a compound can be described by its interaction with the proteome, without requiring the knowledge of the molecular structure.

Efforts have been devoted to transpose various types of biological responses into fingerprint format that could be used to access biological similarity of ligands (Kauvar et al., 1995; Fliri et al., 2005a; Fliri et al., 2005b; Plouffe et al., 2008; Dixon and Villar, 2010). Recently, researchers of Novartis reported the use of the huge amount of in-house HTS data for this purpose (Petrone et al., 2012). The aggregated data from 195 biochemical and cell-based assays for around 1.5 million of compounds have been employed to generate biological fingerprints, so called *HTS-FP*. They stressed the usefulness in mixing biochemical and cell-based data in detecting molecules that can produce similar phenotype without necessarily presenting the same mode of action (Petrone et al., 2012). They demonstrated the complementarity between the *HTS-FP* and a state-of-the-art molecular fingerprint [e.g., ECFP4 (Rogers and Hahn, 2010)] in similarity searches, especially in relation to the scaffold hopping potential of *HTS-FP* to identify structurally diverse hits. On the other hand, biological fingerprints were found to be more efficient in a study related to screening plate selection and hit expansion (Petrone et al., 2012). Additionally, it was observed that biological fingerprint-based clusters contain compounds that interact with targets that operate jointly in the cell. In further work, the combination of *HTS-FP* with structural fingerprints *via* the use of various machine-learning approaches has showed promising results in HTS hit expansion (Riniker et al., 2014). Other studies showed the usefulness of *HTS-FP* for iterative screening purpose (Paricharak et al., 2016). *HTS-FP* has one major drawback though, which is that predictions cannot be made for compounds that have not been previously tested in any HTS assays. In addition, HTS predominantly produces much more *inactive* than *active*, which consequently leads to quite sparse *HTS-FP*. To tackle these issues, Laufkötter et al. (2019) have developed a method where missing bioactivity data were compensated by considering structural data in a so-called combined fingerprint (CESFP) (**Figure 2**). They reported a significant improvement when using CESFP compared to the use of *HTS-FP* and Extended Circular Fingerprints (ECFP) alone in random-forest based activity prediction models. This indicates a clear synergistic effect between structural and biological fingerprints. *HTS-FP* have also been employed for multitask ML. In a recent study, it was observed that *HTS-FP* and ECFP based activity predictions, while comparable in performance, could return hits containing different chemotypes, suggesting that combining these approaches can be an efficient way to explore the bioactive chemical space (Sturm et al., 2019).

Leveraging the transcriptional data such as gene expression profile (gene signature) in a cell could be another way to construct a biological profile descriptor. The publicly funded CMap database (Connectivity Map; Lamb et al., 2006) initially contained profiles of 164 drugs and later expanded to 1,309 FDA-approved small molecules. These small molecules were tested in five human cell lines, generating over 7,000 gene expression profiles in the database (Lamb et al., 2006). Compound induced gene signature profiles have been used for finding diverse hits (Lamb et al., 2006) and drug repositioning (Ishimatsu-Tsuji et al., 2010; Sirota et al., 2011). Although

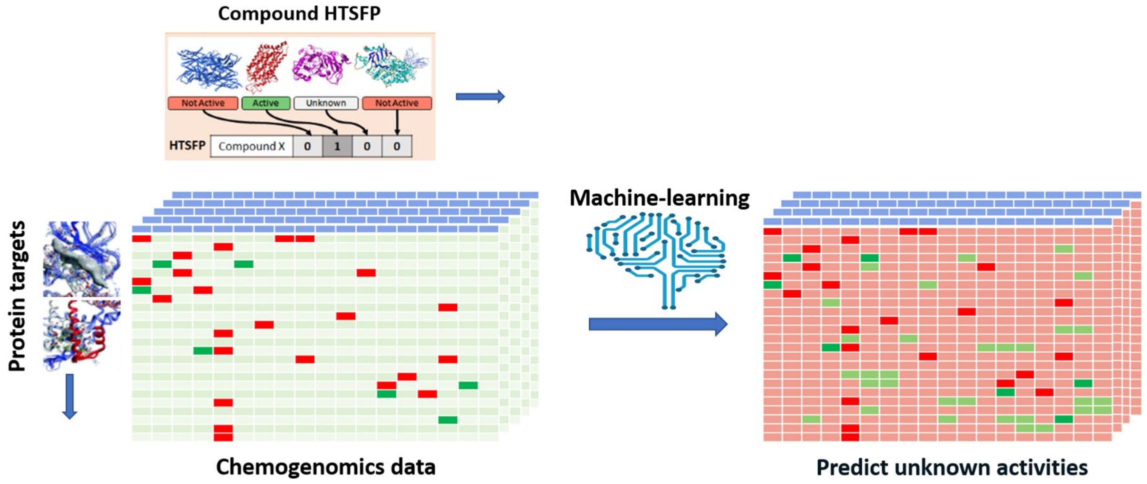


**FIGURE 2 |** Illustration of applying HTS-FP for building multi-task learning models. A chemogenomic matrix represents the interactions between the compound collection and a panel of biological target. Such a matrix is very often sparsely filled activities and missing cells represent unknown activity for the compound/target pair. Employing machine learning and HTSFP is an example of how unknown activities can be predicted.

generating this kind of compound related cell perturbation data is still quite expensive, several pharmaceutical companies, as mentioned earlier, are moving in the direction of generating such data in a large scale. It can be expected that transcriptomics-based biological descriptors will be explored for hit identification in the future. Other biological descriptors derived from multiplexed image data have been reported and successfully used for several tasks, which will be discussed in the subsequent imaging section.

## ANALYSIS OF IMAGE-BASED PROFILING DATA WITH MACHINE LEARNING

In the drug discovery process, biological imaging and image analysis are widely used at various stages ranging from preclinical research to clinical trials. Imaging techniques enable the visualization of phenotype and behavior at multiple levels, including full body of humans or animals, organs, tissues, cells, subcellular compartments, and single molecules. A wide range of available imaging techniques can help to reveal the distribution of a drug in the body, organ, and cell as well as its mechanism of action. Such techniques rely on image datasets obtained through automated microscopy. An example of a large-scale image dataset is given by The Cell Image Library (Bray et al., 2017), which contains 919,265 five-channel fields of view related to 30,616 compounds. The most common imaging techniques are automated microscopy using several fluorescent markers as well as label free microscopy such as brightfield and digital phase contrast. These imaging techniques and the downstream data analysis produce a large amount of data and associated extracted features. For several decades, automatic analysis methods (Boutros et al., 2015) have been successfully applied to identify objects such as organs, tissue types, cells, and subcellular compartments. Effects of diseases and drugs could be quantified by applying statistics and ML methods on the features that were extracted from the images in post-processing efforts. However, recent developments in deep NNs and specifically convolutional NNs (CNNs) are revolutionizing the field and setting new gold standards for key tasks such as segmentation and classification (Kraus et al., 2016; Chen et al., 2016; Dürr and Sick, 2016; Kraus et al., 2017). These new methods not only achieve better results but also avoid the time-consuming manual work of designing features and searching analysis methods for specific tasks. To achieve this, relatively large annotated data sets and substantial computational resources as provided in modern GPU clusters are required for training.

Deep neural nets (typically CNNs) have now been successfully applied for most tasks occurring in automated cell and tissue microscopy image analysis, including denoising (Su et al., 2015), super resolution (Nehme et al., 2018; Ouyang et al., 2018; Rivenson et al., 2018; Wang et al., 2019), stain normalization (Janowczyk et al., 2017), hit identification (Simm et al., 2018), protein localization (Pärnamaa and Parts, 2017), cell cycle phase classification (Eulenberg et al., 2017), mechanism of action classification (Kensert et al., 2019), focus quality check (Yang et al., 2018), segmentation both in 2D and 3D (often using some version of a U-net architecture (Ronneberger et al., 2015)), and modality estimation (Christiansen et al., 2018). Many tasks fall in the area of classification, including tasks such as quality control (Yang et al., 2018), object detection (Ren et al., 2017; Hung et al., 2018), or outcome classification (Cireşan et al., 2013). Classification can be performed either on the image level or on the object level. In the latter case, it is linked to a localization or detection task to identify objects in a given image. One common two-step approach used is to first select candidate regions and then classify them. Alternatively, the network output consists of a probability map, which is analyzed in a postprocessing step to identify the objects. A typical architecture for classification is shown in **Figure 3**.

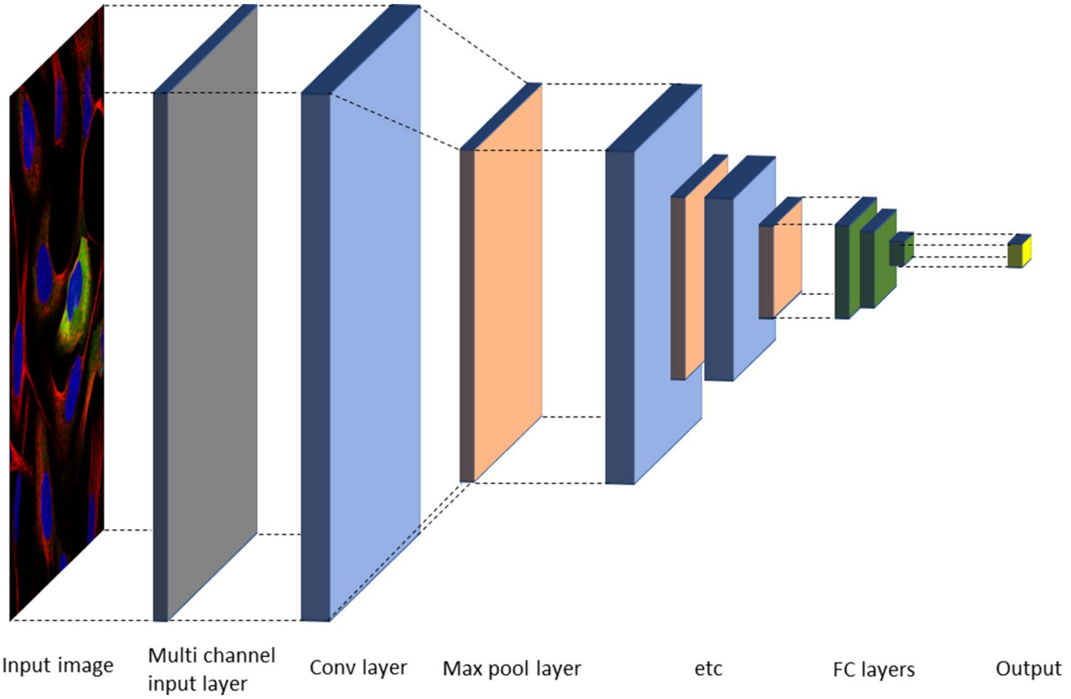


**FIGURE 3 |** Typical neural network architecture for image classification. Alternating convolutional and max pool layers are followed by a number of fully connected layers, and finally an output layer with either sigmoid or softmax functions, depending on the task (Gawehn et al., 2016).

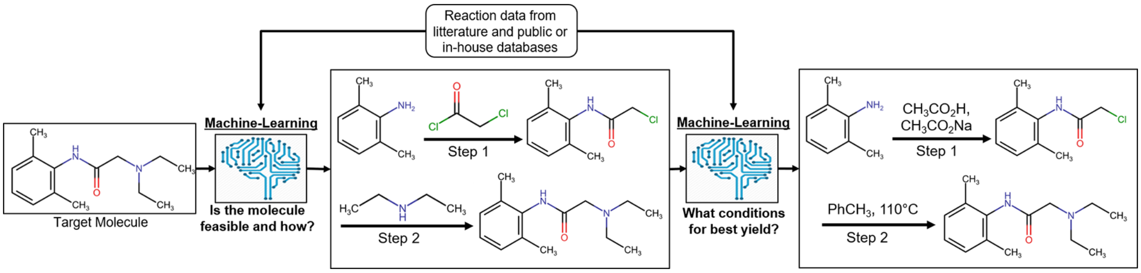

Reaction data from litterature and public or in-house databases
Target Molecule
Machine-Learning
Is the molecule feasible and how?
Step 1
Step 2
Machine-Learning
What conditions for best yield?
$CH_3CO_2H$, $CH_3CO_2Na$
Step 1
$PhCH_3$, 110°C
Step 2

**FIGURE 4 |** Process of reaction prediction on an exemplary target molecule [lidocaine (Reilly, 2009)]. Machine-learning methods are applied to, first, predict the synthetic feasibility of the molecule and, second, predict the chemical context leading to the best yield possible for the reaction.

Since large amounts of annotated data are often not available for a specific task, strategies such as transfer learning are often applied, e.g., for classification tasks (Kensert et al., 2019; Zhang et al.). This starts with a pretrained neural net from a different task where a large data set is available. The model is then used as an initialization for the new task and fine-tuned for the task at hand. The last output layers of the original network are often not reused but trained for the new task from scratch.

As mentioned above, HCS where cells are exposed to different compounds followed by automated multichannel microscopy and subsequent automatic feature extraction is producing much richer data for screening than traditional HTS. More advanced analysis of cells exposed to chemical perturbations allows to identify related spatial and temporal information. Different biological descriptors derived from multiplexed image data have been reported (Loo et al., 2007; Young et al., 2008; Feng et al., 2009; Caicedo et al., 2017). Reisen et al. (2015) derived a biological fingerprint from HCS. Their HCS fingerprints are based on an automatic analysis of a panel of imaging assays that recorded morphological changes within six different cellular compartments upon testing of 2,725 compounds with well-characterized mode of actions. These fingerprints were then used in classifying the compounds into clusters, which were subsequently annotated with target activities from bioactive molecules from different databases such as ChEMBL, Gostar (Gostardb), Drug bank (Knox et al., 2011), Integrity (Thomson Reuters), or Metabase (Thomson Reuters). Phenotypic responses were successfully classified for 52% of the tested compounds, and different phenotypes were identified that could be linked to the modulation of individual targets, cellular pathways, or disease genes (Reisen et al., 2015). Later, Simm et al. (2018) built a supervised machine-learning model based on fingerprints obtained from morphological features extracted from high-throughput (cell) imaging (HTI) screening data. Their method enabled the identification of additional hits that were diverse from those obtained in a primary screen. More recently, end-to-end convolutional NNs (Hofmarcher et al., 2019) were used on cell-painting images to predict assay activity as a multitask prediction problem. A number of common architectures were compared to each other as well as to the baseline model constructed with CellProfiler (Carpenter et al., 2006) extracted features. End-to-end models were shown to be able to deliver better results without first extracting features from the images.

## PREDICTING COMPOUND ACTIVITY USING LARGE CHEMOGENOMICS MODELS

One of the main purposes of chemogenomics (Caron et al., 2001) is to obtain a matrix containing all the possible and impossible interactions between compounds covering the entire chemical space and biological proteins. Despite the advances in HTS (Hertzberg and Pope, 2000) techniques, which made it possible to test hundreds of thousands of compounds against a biological target in very little time, it seems quite unlikely that we will ever obtain a full chemogenomic matrix due to the complexity of the chemical space (Reymond, 2015) and the cost and time such a task would require due to the sheer size of the chemical space. It is, however, possible to computationally predict interactions between chemical compounds and panels of biological targets. The generation of such chemogenomic models is enabled by large databases that contain compounds with annotated biological activities. An applied example of activity predictions relying on chemogenomic models is shown in **Figure 2**. As previously mentioned, a large amount of SAR datapoints from assays with constant conditions and well-characterized quality can be found in private pharmaceutical companies' databases. In the public domain, the most known databases are ChEMBL (Davies et al., 2015; Gaulton et al., 2016), PubChem (Kim et al., 2019), and BindingDB (Gilson et al., 2015). ChEMBL is a manually curated database of bioactive molecules with drug-like properties. PubChem is a repository for screening data and BindingDB contains affinity measurements data. ChEMBL and BindingDB data were manually extracted from peer-reviewed journal articles. Furthermore, large amounts data from publications and patents are available in commercial databases such as Reaxys (Reaxys Database) and SciFinder.

A major topic that has been briefly addressed previously is the necessity of data standardization and curation prior to building a predictive model. Chemical structures can be represented by different types of notations (SMILES, InChI, etc.) (InChI and InChIKeys for chemical structures; Weininger, 1988; Weininger et al., 1989; Heller et al., 2015), and bioactivity data typically originate from different assay formats and are reported in a variety of units. One recent example of such a standardization exercise was reported by Sun et al. (2017) and resulted in the creation of a unified dataset, ExCAPE-DB, covering over 70 million SAR data points coming from PubChem and ChEMBL. In another study, Mervin et al. (2015) mined ChEMBL active compounds and PubChem inactive compounds to construct a dataset of 195 million bioactivity data points and investigated the impact of inactive data on the performance of a predictive model.

Several models (Wang et al., 2013; Sushko et al., 2014; Hughes et al., 2016) employing various ML methods or virtual screening are available for target predictions and compound reactivity prediction, but only a few were derived from larger datasets. Studies on small-scale datasets (i.e., on very few assays or targets) can lead to misinterpretation of results or incorrect generalization as their applicability domain is limited. When using small dataset, there is a risk of investigating compounds that do not cover a wide range of the chemical space. In such a scenario, predictive models would show excellent performance when applied on structurally similar compounds but would fail to predict the activity of compounds pertaining to other series. Most compound-target profiles are sparsely filled. One method to compensate missing data is to combine bioactivity data with structural data as we have discussed in the previous section. Applying ML methods on large chemogenomic datasets has been reported in literature. Mervin et al. (2015) constructed a dataset of over 195 million bioactive data points and demonstrated that the inclusion of inactivity data improves the accuracy of predictive models. Another example for modelling large-scale chemogenomic data was reported by Martin et al. (2019) and produced activity predictions as accurate as an experimental 4-concentration $IC_{50}$s. A profile-QSAR (pQSAR) model based on 11,805 Novartis assays was applied on 5.5 million Novartis compounds, leading to a total of 50 billion predictions. This model is updated monthly. Recently, deep learning methods were also applied to build multi-task models. A study by Mayr et al. (2018) applied a variety of ML methods on a dataset of 45,000 compounds contained in more than 1,000 assays extracted from ChEMBL. It was shown that deep-learning outperforms all the other tested methods [i.e., RF (Breiman, 2001), SVM (Cortes and Vapnik, 1995), K-Nearest-Neighbors (Silverman and Jones, 1989), Similarity Ensemble Approach (Keiser et al., 2007), Naïve Bayes (Zhang, 2004) statistics] for target predictions. The strength of this analysis relies on the fact that it was not biased by specific chemical structures or a particular structure representation of the compounds, as the dataset covered a wide range of target families, and various types of fingerprints were employed. This analysis showed that the performance of the predictive model increases with the training set size, confirming that effort should be put into creating large dataset for ML methods. Efforts for estimating prediction uncertainty of ML models have also been reported, for example, conformal prediction framework-based methods (Bosc et al., 2019; Cortés-Ciriano and Bender, 2019) and Bayesian-based approaches (Zhang and Lee, 2019). A study (Tsubaki et al., 2019) employed GNN and CNN to infer protein–compound interaction predictions and determine the importance of each subsequences of the proteins in the interaction. In **Table 2**, we summarized some studies in which DNN has been shown to outperform traditional ML approaches.

**TABLE 2 |** Performances comparison of traditional ML and DL in Drug Discovery.

| Ref. | Performance traditional ML | Performance deep-learning |
|---|---|---|
| (Koutsoukas et al., 2017) (1) | RF: MCC = 0.89 | DNN: MCC = 0.91 |
| (Dahl et al., 2014) (2) | RF: AUC = 0.78 | MT NN: AUC = 0.82 |
| (Lenselink et al., 2017) | SVM: MCC = 0.50, BEDROC = 0.88<br>RF: MCC = 0.56, BEDROC = 0.82 | DNN_MC: MCC = 0.57, BEDROC = 0.92 |
| (Mayr et al., 2016) | SVM: AUC = 0.71 | ST: AUC = 0.72<br>MT: AUC = 0.75 |
| (Feinberg et al., 2018) | RF: Pearson = 0.783 | GNN: Pearson = 0.822 |
| (Segler and Waller, 2017b) | LR: Acc = 0.86 (reaction prediction)<br>LR: Acc = 0.64 (retrosynthesis) | NN: Acc = 0.92 (reaction prediction)<br>NN: Acc = 0.78 (retrosynthesis) |
| (Wu et al., 2018) (3) | SVM: AUC = 0.822 | GC: AUC = 0.829 |
| (Xiong et al., 2019) (4) | SVM: AUC = 0.792 | Attentive FP: AUC = 0.832 |
| (Yang et al., 2019) (5) | RF: AUC = 0.619 | FFN: AUC = 0.788 |
| (Ma et al., 2015) (6) | RF: $R^2$ = 0.42 | DNN: $R^2$ = 0.49 |
| (Ramsundar et al., 2017) (7) | RF: $R^2$ = 0.428 | ST: $R^2$ = 0.448<br>MT: $R^2$ = 0.468 |

*LR, ST, MT, GC, GNN, and FFN refer to Linear Regression, Single- and Multi-Task, Graph Convolution, Graph, and Feedforward Neural Network, respectively. (1) Averaged performance on validation sets over 7 datasets. (2) Averaged performance on test sets over 19 datasets. (3) Performance on a test subset of the Tox21 dataset. (4) Performance on the HIV dataset. (5) Performance on the Tox21 dataset. (6) Averaged performance over 15 datasets. (7) Model performance on a test set.*

Although it is crucial to have a sufficient amount of training data to infer target predictions, having high-quality data is also necessary. Indeed, available activity data can be erroneous due to the problematic nature of the compounds (Dahlin et al., 2015) (e.g., reactivity, impurity, aggregation, technology hitters, etc.) or the experimental conditions in which they were tested (concentration, assay technology, plate type, etc.). The integration of such erroneous and heterogenous data can have an impact on predictive models. Various methods have been developed to detect such problematic compound behaviors, the most popular one being the Pan-Assay Interference Substructure (PAINS) filters (Baell and Holloway, 2010). A significant number of compounds that were initially considered as potential leads were found to be false positives. PAINS filters are substructures that were frequently observed among these compounds. It has now become usual to apply these filters when selecting compounds for follow-up studies. However, the PAINS filters were derived from compounds tested in only one specific HTS technology (namely, AlphaScreen) and do not cover the entire chemical space. Thus, these filters should be applied with care (Baell and Nissink, 2018). Stork et al. (2018, 2019) developed the Hit Dexter model to predict frequent-hitter, aggregator, PAINS, dark chemical matter (Wassermann et al., 2015b), and other potential nuisance compounds. The Hit Dexter model is based on a set of extensively tested compounds from PubChem represented by their 2D molecular fingerprints. The Badapple model (Yang et al., 2016) was developed to filter out promiscuous compounds based on a scaffold promiscuity analysis. Such predictive models and substructure filters are crucial for compounds triaging and data accuracy; however, the characteristics of the data under investigation and the aim of the screening project have to be taken into consideration when applying those filters. Promiscuous compounds, while giving rise to possible negative side effects due to their potential interactions with multiple targets, can still be of great interest because of their polypharmacology. In a similar manner, compounds interfering with an assay technology should not be discarded from a drug discovery process but should, however, be tested in a different technology based on dissimilar mechanisms. Sample impurity is another factor to consider regarding promiscuity. If the purity of each sample tested is known, it is easy to filter out everything that did not match the requested quality criterion. If this is not the case, one can use in-house data to detect promiscuous samples in the screening deck (Beck, 2012).

Other criterion to consider in HTS the druglikeness of a compound, which is determined by the compound's physicochemical (PC) and toxicological properties. Various quality control pipelines created to filter out compounds employ straightforward filtering rules (Hsieh et al., 2015; Zhai et al., 2016), while some other employ ML techniques such as deep-learning (Liu et al., 2019) methods. In pharmaceutical companies and academic institutes, PC filters are tuned depending on the type of compounds found in the chemical libraries (Brenk et al., 2008; Pearce et al., 2006; Cumming et al., 2013). PC properties-based rules ensure that compounds have similar properties to other drugs based on historical data and have a good probability to be synthesizable and non-toxic. Furthermore, structural alerts have been created (Sushko et al., 2012) to flag potential toxic compounds in terms, for example, of mutagenicity (Tennant and Ashby, 1991) or skin sensitization (Barratt et al., 1994).

Very recently, a new consortium of pharmaceutical, technology, and academic partners has launched the "MELLODDY" (Machin**e** Learning Ledger Orchestration for Drug Discover**y**) project (MELLODDY Consortium| Twitter; Pharma Companies Join Forces to Train AI for Drug Discovery Collectively). The project involves 17 partners from across Europe and receives funding from the EU Innovative Medicines Initiative (IMI) as a public–private partnership. MELLODDY aims to train chemogenomics models across multi-partner (10 pharma companies) datasets while ensuring privacy preservation of both the data and the models by developing a platform using federated learning. It will be interesting to see their efforts regarding data standardization and generation of a large high-quality data set and the results of such an approach.

## MODELLING CHEMICAL REACTIONS FROM LARGE-SCALE SYNTHESIS DATA

It is of crucial importance in drug discovery to be able to predict the feasibility of chemical reactions (Engkvist et al., 2018). It ranges from predicting synthetic feasibility for compounds identified in virtual screening in early drug discovery as well as for hit expansion in the lead generation phase to late stage modifications during lead optimization and to predict possible synthetic routes for upscaling of the synthesis of clinical candidates (**Figure 4**).

Synthetic predictions have a long history dating back to rule-based programs in the 1960s (Corey and Todd Wipke, 1969). Several aspects have made reaction informatics a field for active research during recent years. Besides established commercial products with reactions extracted from literature, reaction data have been extracted from electronic laboratory notebooks (ELNs) (Christ et al., 2012) and patents. Schneider et al. (2016) used text-mining to extract 1.15 million unique whole reaction schemes, including reaction roles and yields, from pharmaceutical patents. The reactions were assigned to well-known reaction types such as Wittig olefination or Buchwald–Hartwig amination using an expert system. Also, large-scale reaction data can be generated from high-throughput experimentation. Schematically reaction informatics can be divided into two subfields, retrosynthetic analysis, where a molecule is analyzed and a set of reactions and building blocks are proposed to synthesize the molecule, and forward reaction prediction, where it is predicted if a set of building blocks will react or not and at which conditions a reaction will occur. In recent years, there has been a paradigm shift on how retrosynthesis routes can be predicted. While historically rule-based systems were the most popular method, more recently several studies using ML have shown superior results. One advantage of ML algorithms is that they are generalized methods and not dependent on rigid predefined rules for describing the exact reaction.

In the following, we will focus on recent examples of predicting how to synthesize molecules by mining large corpora of experimental synthesis data. For more general reviews, we refer to recent publications (Warr, 2014; Coley et al., 2018). Segler and Waller (2017b) used reaction fingerprint descriptors to classify reactions. Both hand-coded and automatically extracted reaction rules were used to classify reactions from literature. Three million reactions were classified with the hand-coded rules, while almost 5 million reactions were classified with the automatically extracted reaction rules. Reaction classification models were built with artificial NNs (ANNs). ANNs were found to be superior in predicting reactions than a rule-based system. In another article, they showed that reaction graphs with reactions extracted from literature can be used to predict novel reactions (Segler and Waller, 2017a). A knowledge graph consisting of 14 million molecules was generated, and 8 million reactions and probable novel reactions could be inferenced from. Studies were also published for predicting the reactivity of protecting groups (Lin et al., 2016); 142,000 catalytic hydrogenation reactions were extracted from literature. The reactions were described with condensed graphs of reaction fingerprints. The models showed high accuracy (90%) for predicting optimal conditions for deprotection of protecting groups. The models were also used to identify contradictions in reactivity charts created manually by experts. Coley et al. (2017) developed predictive ML models using 15,000 reactions extracted from US patents. They created a set of candidate reactions based on enumeration of a set of reactants and reaction templates. In a second step, the candidate reactions were described by a set of reaction descriptors, and a NN model was trained to prioritize the candidate reactions. The model predicted the correct reaction in 72% of the cases, the correct reaction was found in 87% of the cases among the top three predicted reactions, and it was found to be among the top five predicted reactions in 91% of the cases. A recent example of predicting reaction conditions with a large data set was published by Gao et al. (2018). They developed a NN model to predict the chemical context [catalyst(s), solvent(s), reagent(s)] and the most suitable temperature for any particular organic reaction. Reactions were extracted from Reaxys and filtered according to various criteria, resulting in ~10 million example reactions. The models were trained on these reactions and were able to propose conditions where a close match to the recorded catalyst, solvent, and reagent was found within the top 10 predictions in 69.6% of the cases. Another noteworthy development in the reaction prediction field is the construction development of a retrosynthesis system using deep learning technologies. Segler et al. (2018b) reported such a system, in which the system reaction DNN models derived from literature reaction data were combined with Monte Carlo Tree Search (MCTS) to identify a set of reactions and building blocks that could be used to synthesize the desired molecule. While most studies have used a reaction template to describe the reaction, it has been shown recently that a template free seq-2-seq approach (i.e., directly translate product SMILES to the predicted reactants in reaction SMILES format) also can give promising results for synthesis prediction (Schwaller et al., 2018a; 2018b). An alternative way of predicting the synthetic pathway exploiting through learned policies has just been published (Schreck et al., 2019).

## DATA DRIVEN *DE NOVO* MOLECULE DESIGN THROUGH GENERATIVE MODELS AND DATA AUGMENTATION

Even though industrial compound-bioactivity datasets have millions of data points, many assay results for specific compound series (typical for the lead optimization stage of a drug discovery project) have much less SAR data. However, these datasets can still be augmented and be further exploited with deep learning approaches, such as QSAR and generative modelling. Data augmentation is the process of adding noise or artificial perturbation to the samples in the dataset before training the model in order to make the final models more robust to overfitting (Arús-Pous et al., 2019b). Moreover, in some cases, data augmentation can give additional information to the model. A simple analogy can be found in building image classification models. For instance, a single image with a “dog” will still be recognizable even if it is rotated, cropped slightly, changed in terms of contrast or lightness, etc. Therefore, a single labelled image can be multiplied into multiple training set entries, thus expanding the dataset.

Similar approaches have also been used in areas relevant to pharmaceutical research such as predicting concentrations of chemical compounds from spectroscopy data (Bjerrum et al., 2017) and building QSAR models from chemical images (Goh et al., 2017). In molecular deep learning models, many architectures use the SMILES as molecular representation

(Bjerrum, 2017), which is obtained by assigning a unique number to each atom in the molecule and then traversing the molecular graph using that order. Commonly, a canonical SMILES representation of each molecule is used, which is obtained by calculating a unique numbering for molecules (Weininger et al., 1989). This representation is served as a way of uniquely identifying molecules. Nevertheless, most molecules can have more than one SMILES representation obtained by only changing the numbering of the atoms, meaning that different SMILES start in different atoms of the molecule and traverse it in different ways (**Figure 5**). Randomized SMILES for the same compound can thus be used for data augmentation.

A great surge of interest in cheminformatics applications of deep learning has happened in recent years when NNs were used to generate molecules represented by SMILES strings (Olivecrona et al., 2017; Gómez-Bombarelli et al., 2018; Segler et al., 2018a). Recurrent NN (RNN) trained with a set of SMILES strings can generate molecules that are not present in the training set but that have similar properties as the training samples. These

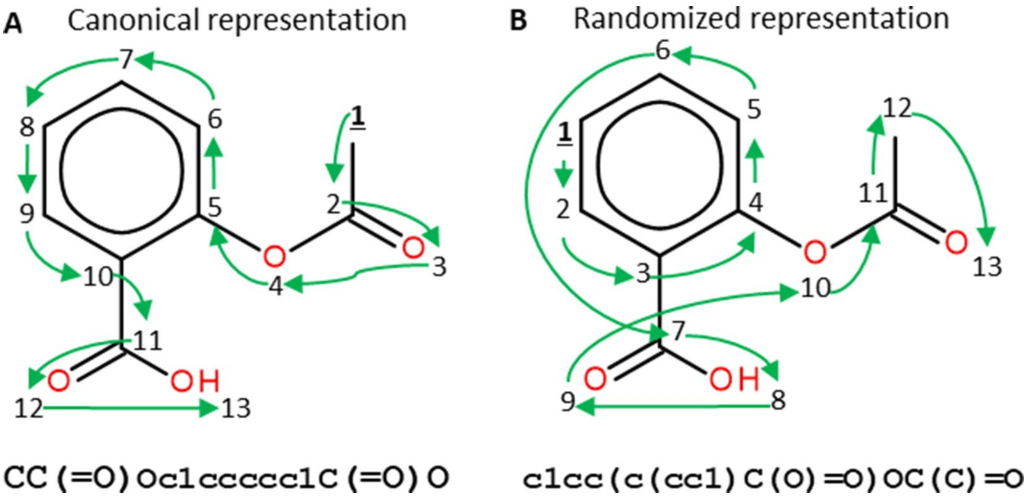


**FIGURE 5 |** Canonical **(A)** and randomized **(B)** SMILES representations of Aspirin. Numbers represent the atom numberings assigned by the canonicalization algorithm **(A)** or randomized **(B)**. Green arrows indicate how the molecular graph is traversed. Both SMILES strings represent the same molecule but, as the atom numbering changes, the generated SMILES strings do too. Figure extracted with permission from Arús-Pous et al. (2019b).

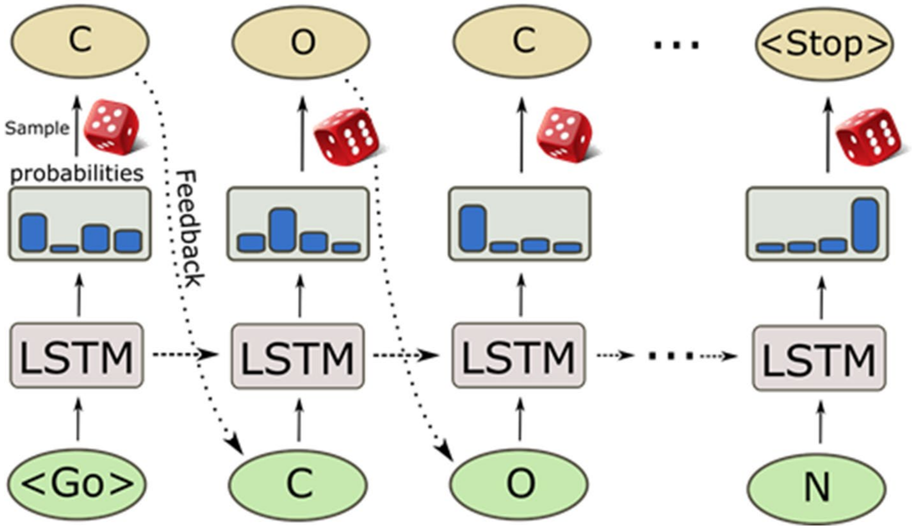


**FIGURE 6 |** Sampling process of a pre-trained recurrent neural network. The generation process starts with a GO token, and at each step, the model computes a probability distribution of all possible characters. Then, the next character is sampled from it and fed back to predict the next character. The internal memory in the long short-term memory (LSTM) cells enables the predictions to take previous characters into account when generating the next character.

deep learning-based generative models are entirely data driven and do not rely on any predefined reaction/transformation rules, in contrast to the traditional library enumeration methods for generating chemical structures (Schneider and Fechner, 2005). Molecules are generated character by character as SMILES strings by randomly sampling the probability distribution of the next character to sample (**Figure 6**). This process generates a very high ratio of valid SMILES, especially thanks to the use of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU) (Cho et al., 2014) cells that capture long-range relationships such as ring closures and branches. Additionally, pre-training on a large set of chemical structures [such as ChEMBL, ZINC (Sterling and Irwin, 2015), etc.] and the subsequent application of transfer learning to smaller datasets can be used to generate focused datasets with an enrichment of active compounds (Segler et al., 2018a). The pre-trained RNNs can also be used to directly optimize toward desirable properties (Olivecrona et al., 2017). This triggered the development of a plethora of novel architectures and techniques in the last years, such as Variational AutoEncoders (VAEs) (Kingma and Welling, 2013; Polykovskiy et al., 2018b; Zhavoronkov et al., 2019), Differentiable Neural Computers (DNCs) (Putin et al., 2018), Generative Adversarial Networks (GANs) (Guimaraes et al., 2017; Prykhodko et al., 2019), and Bayesian optimization method for structure optimization (Pyzer-Knapp, 2018). Besides the SMILES string based *de novo* structure generation methods, algorithms of generating molecules based on molecular graphs have also been proposed and, by using them, methods molecules can be directly generated step-by-step as molecular graphs (Jin et al., 2018; You et al., 2018; Elton et al., 2019; Xu et al., 2019).

Data augmentation techniques have also been applied in molecular generative models. For example, they have shown to improve the quality of the chemical space generated in VAEs (Bjerrum and Sattarov, 2018) and RNNs (Arús-Pous et al., 2019b) in terms of performance of latent vector-based QSAR models (Bjerrum and Sattarov, 2018) and coverage of targeted chemical space (Arús-Pous et al., 2019b). However, there is no consensus on how to measure and compare the performances of generative models. Some approaches have been published, such as MOSES (Polykovskiy et al., 2018a) and Guacamol (Brown et al., 2019), but they are not able to fully characterize the complete chemical space generated. To solve this problem, an approach using the negative log-likelihood (NLL) of generated molecules was recently described (Arús-Pous et al., 2019a). It is able to characterize the models by their completeness, i.e., how many molecules from the target chemical space are sampled, uniformity, i.e., how uniform are those being sampled, and closedness, i.e., how many molecules outside of the target chemical space are being sampled. More specifically, it was found that models trained with 1 million molecules sampled randomly from GDB-13 (Blum and Reymond, 2009), an enumerated database containing 970 million drug-like compounds with up to 13 heavy atoms, are able to generate up to 68% of the entire database when the canonical SMILES representation is used for model training, while the coverage increases to 83%, when non-canonical randomized SMILES are used. It indicates that data augmentation based on randomized SMILES generation has an impact on what models can learn. Moreover, models trained with randomized SMILES generate a much more uniform and closed chemical space than those trained with canonical SMILES.

Deep-learning-based generative model has been applied successfully for prospective design of new druglike molecules with desired activities (Merk et al., 2018). Compounds were generated using a recurrent NN trained on a large set of bioactive compounds. By transfer learning, this general model was fine-tuned on recognizing retinoid X and peroxisome proliferator-activated receptor agonists. The five top-ranking compounds were synthesized and investigated in cell-based assays. Four of these compounds showed a strong affinity toward the targets, with nanomolar to low-micromolar receptor modulatory activity. Generative modelling can also be applied to other chemical entities, such as peptides (Grisoni et al., 2018; Müller et al., 2018), but no method for data augmentation has been described up to now. A potential challenge might be that it is not possible to simply permute the amino acid sequence of peptides as it is done with the arbitrary atom order in SMILES strings, although it may be possible to integrate data from larger unlabelled datasets. PSI-BLAST similarity searching has been used to expand the prior dataset of known active compounds before generation and selection in iterative optimization rounds (Yoshida et al., 2018). This suggests that bioinformatics approaches area a viable way to find the natural variation for the amino acid substitutions and thus enable data set expansion. The drug-like chemical space is estimated to have at least $10^{24}$ molecules (Bohacek et al., 2010), and it is not feasible to fully enumerate. Nevertheless, deep-learning-based generative models combined with data augmentation techniques have the potential to provide a way to sample large regions of the drug-like chemical space. In combination with synthesis routes prediction, this would deliver a tremendous boost for compound design in pharmaceutical research.

## CONCLUSION

Over the past years, large amounts of heterogeneous data characterizing the biological action of small molecules have been accumulated in pharmaceutical R&D, stored in both proprietary and publicly available data bases. The origin of these data ranges from biochemical or cellular assays to experiments that investigate the impact of compounds on transcriptomics signatures and assays with imaging readouts. These fast-growing data have fuelled the application of data-savvy ML methods, and in particular deep learning, in order to detect patterns that allow to derive hypotheses for compound-mediated effects on biological (model) systems or to generate predictive models that can be employed at various stages during identification and optimization of new drug candidates. Together with deep-learning-based approaches to sample the drug-like chemical space that—depending on the use case—can be applied with or without predictions of synthetic accessibility, a plethora of potential high-impact applications is emerging. It offers the opportunity to accelerate early drug discovery and to enable a much more comprehensive exploration of the chemical space and the biological effects of its members than traditional wet lab and virtual screening approaches.

## AUTHOR CONTRIBUTIONS

JMK, BB, and HC wrote the section Large-Scale Compound Data in Pharmaceutical Industry. TK wrote the section Biological Profiling Descriptors for Hit Expansion. JK wrote the section Analysis of Image-Based Profiling Data With Machine Learning. LD wrote the section Predicting Compound Activity Using Large Chemogenomics Models. OE wrote the section Modelling Chemical Reactions From Large-Scale Synthesis Data. JA-P and EB wrote the section Data Driven de Novo Molecule Design Through Generative Models and Data Augmentation. LD and HC co-supervised the manuscript.

## FUNDING

LD and JA-P have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska Curie grant agreement No 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu). The article reflects only the authors view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

## REFERENCES

Agrafiotis, D. K., Alex, S., Dai, H., Derkinderen, A., Farnum, M., Gates, P., et al. (2007). Advanced Biological and Chemical Discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model.* 47, 1999–2014. doi: 10.1021/ci700267w

Arús-Pous, J., Blaschke, T., Ulander, S., Reymond, J. L., Chen, H., and Engkvist, O. (2019a). Exploring the GDB-13 chemical space using deep generative models. *J. Cheminform.* 11, 20. doi: 10.1186/s13321-019-0341-z

Arús-Pous, J., Johansson, S., Ptykhodko, O., Bjerrum, E. J., Tyrchan, C., and Reymond, J.-L. (2019b). Randomized SMILES strings improve the quality of molecular generative models. *ChemRxiv Prepr.* Available at: https://chemrxiv.org/articles/Randomized_SMILES_Strings_Improve_the_Quality_of_Molecular_Generative_Models/8639942/1 [Accessed July 5, 2019]. doi: 10.26434/chemrxiv.8639942.v2

Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi: 10.1021/jm901137j

Baell, J. B., and Nissink, J. W. M. (2018). Seven year itch: pan-assay interference compounds (PAINS) in 2017 - utility and limitations. *ACS Chem. Biol.* 13, 36–44. doi: 10.1021/acschembio.7b00903

Barratt, M. D., Basketter, D. A., and Roberts, D. W. (1994). Skin sensitization structure-activity relationships for phenyl benzoates. *Toxicol. Vitr.* 8, 823–826. doi: 10.1016/0887-2333(94)90077-9

Beck, B. (2012). BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds. *Bioorg. Med. Chem.* 20, 5428–5435. doi: 10.1016/j.bmc.2012.04.023

Beck, B., Seeliger, D., and Kriegl, J. M. (2015). The impact of data integrity on decision making in early lead discovery. *J. Comput. Aided Mol. Des.* 29, 911–921. doi: 10.1007/s10822-015-9871-2

Bickle, M. (2010). The beautiful cell: high-content screening in drug discovery. *Anal. Bioanal. Chem.* 398, 219–226. doi: 10.1007/s00216-010-3788-3

Bjerrum, E. J. (2017). SMILES enumeration as data augmentation for neural network modeling of molecules. *ArXiv*.

Bjerrum, E. J., Glahder, M., and Skov, T., (2017). Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics 1–10.

Bjerrum, E. J., and Sattarov, B. (2018). Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* 8, 131. doi: 10.3390/biom8040131

Blum, L. C., and Reymond, J. L. (2009). 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc* 131, 8732–8733. doi: 10.1021/ja902302h

Bohacek, R. S., McMartin, C., and Guida, W. C. (2010). ChemInform abstract: the art and practice of structure-based drug design: a molecular modeling perspective. *ChemInform* 27, no–no. doi: 10.1002/chin.199617316

Borman, S. (1999). Reducing time to drug discovery. *Chem. Eng. News* 77, 33–48. doi: 10.1021/cen-v077n010.p033

Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., and Leach, A. R. (2019). Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J. Cheminform.* 11, 4. doi: 10.1186/s13321-018-0325-4

Boutros, M., Heigwer, F., and Laufer, C. (2015). Microscopy-based high-content screening. *Cell* 163, 1314–1325. doi: 10.1016/J.CELL.2015.11.007

Bray, M. A., Gustafsdottir, S. M., Rohban, M. H., Singh, S., Ljosa, V., Sokolnicki, K. L., et al. (2017). A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* 6, 1–5. doi: 10.1093/gigascience/giw014

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I. H., Frearson, J., et al. (2008). Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* 3, 435–444. doi: 10.1002/cmdc.200700139

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C., (2019). GuacaMol: benchmarking models for de novo molecular design. doi: 10.1021/acs.jcim.8b00839

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., et al. (2017). Data-analysis strategies for image-based cell profiling. *Nat. Methods* 14, 849–863. doi: 10.1038/nmeth.4397

Caron, P. R., Mullican, M. D., Mashal, R. D., Wilson, K. P., Su, M. S., and Murcko, M. A. (2001). Chemogenomic approaches to drug discovery. *Chem. Biol.* 5, 464–470. Available at: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html. [Accessed May 27, 2019]. doi: 10.1016/S1367-5931(00)00229-5

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I., Friman, O., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7, R100. doi: 10.1186/gb-2006-7-10-r100

Chen, C. L., Mahjoubfar, A., Tai, L.-C., Blaby, I. K., Huang, A., Niazi, K. R., et al. (2016). Deep learning in label-free cell classification. *Sci. Rep.* 6, 21471. doi: 10.1038/srep21471

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1724–1734 doi: 10.3115/v1/D14-1179

Christ, C. D., Zentgraf, M., and Kriegl, J. M. (2012). Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* 52, 1745–1756. doi: 10.1021/ci300116p

Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., Skibinski, G., Lipnick, S., et al. (2018). In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* 173, 792–803.e19. doi: 10.1016/j.cell.2018.03.040

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J., (2013). *Mitosis detection in breast cancer histology images with deep neural networks*. Berlin, Heidelberg: Springer, 411–418. doi: 10.1007/978-3-642-40763-5_51

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., and Jensen, K. F. (2017). Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443. doi: 10.1021/acscentsci.7b00064

Coley, C. W., Green, W. H., and Jensen, K. F. (2018). Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289. doi: 10.1021/acs.accounts.8b00087

Connectivity Map Available at: https://www.broadinstitute.org/connectivity-map-cmap [Accessed October 24, 2019].

Corey, E. J., and Todd Wipke, W. (1969). Computer-assisted design of complex organic syntheses. *Science* (80-.) 166, 178–192. doi: 10.1126/science.166.3902.178

Cortés-Ciriano, I., and Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *J. Chem. Inf. Model.* 59, 3330–3339. doi: 10.1021/acs.jcim.9b00297

Cortes, C., and Vapnik, V. (1995). Support vector networks machine active learning with applications to text classification. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M., and Chen, H. (2013). Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discovery* 12, 948–962. doi: 10.1038/nrd4128

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR Predictions. *ArXiv*. Available at: http://arxiv.org/abs/1406.1231 [Accessed September 25, 2019].

Dahlin, J. L., Nissink, J. W. M., Strasser, J. M., Francis, S., Higgins, L., Zhou, H., et al. (2015). PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *J. Med. Chem.* 58, 2091–2113. doi: 10.1021/jm5019093

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., et al. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Web Serv. Issue Publ. Online* 43, W612–W620. doi: 10.1093/nar/gkv352

De Wolf, H., De Bondt, A., Turner, H., and Göhlmann, H. W. (2016). Transcriptional characterization of compounds: lessons learned from the public LINCS data. *Assay Drug Dev. Technol.* 14, 252–260. doi: 10.1089/adt.2016.715

Dixon, S. L., and Villar, H. O. (2010). ChemInform abstract: bioactive diversity and screening library selection *via* Affinity fingerprinting. *ChemInform* 30, no–no. doi: 10.1002/chin.199916265

Dürr, O., and Sick, B. (2016). Single-cell phenotype classification using deep convolutional neural networks. *J. Biomol. Screen.* 21, 998–1003. doi: 10.1177/1087057116631284

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* 4, 828–849. doi: 10.1039/c9me00039a

Engkvist, O., Norrby, P.-O., Selmi, N., Lam, Y., Peng, Z., Sherer, E. C., et al. (2018). Computational prediction of chemical reactions: current status and outlook. *Drug Discovery Today* 23, 1203–1218. doi: 10.1016/J.DRUDIS.2018.02.014

Eulenberg, P., Köhler, N., Blasi, T., Filby, A., Carpenter, A. E., Rees, P., et al. (2017). Reconstructing cell cycle and disease progression using deep learning. *Nat. Commun.* 8, 463. doi: 10.1038/s41467-017-00623-3

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). PotentialNet for molecular property prediction. *ACS Cent. Sci.* 4, 1520–1530. doi: 10.1021/acscentsci.8b00507

Feng, Y., Mitchison, T. J., Bender, A., Young, D. W., and Tallarico, J. A. (2009). Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discovery* 8, 567–578. doi: 10.1038/nrd2876

Filzen, T. M., Kutchukian, P. S., Hermes, J. D., Li, J., and Tudor, M. (2017). Representing high throughput expression profiles *via* perturbation barcodes reveals compound targets. *PloS Comput. Biol.* 13, e1005335. doi: 10.1371/journal.pcbi.1005335

Fligge, T. A., and Schuler, A. (2006). Integration of a rapid automated solubility classification into early validation of hits obtained by high throughput screening. *J. Pharm. Biomed. Anal.* 42, 449–454. doi: 10.1016/j.jpba.2006.05.004

Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005a). Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U. S. A.* 102, 261–266. doi: 10.1073/pnas.0407790101

Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005b). Biospectra analysis: Model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* 48, 6918–6925. doi: 10.1021/jm050494g

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., and Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* 4, 1465–1476. doi: 10.1021/acscentsci.8b00357

Gaulton, A., Hersey, A., -l Nowotka, M., Patrícia Bento, A., Chambers, J., Mendez, D., et al. (2016). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, 945–954. doi: 10.1093/nar/gkw1074

Gawehn, E., Hiss, J. A., and Schneider, G. (2016). Deep learning in drug discovery. *Mol. Inform.* 35, 3–14. doi: 10.1002/minf.201501008

Genometry Available at: https://www.linkedin.com/company/genometry-inc/about/ [Accessed October 24, 2019].

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2015). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, 1045–1053. doi: 10.1093/nar/gkv1072

Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O., and Baker, N., (2017). Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR Models.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Gostardb. Available at: www.gostardb.com/gostar/.

Grisoni, F., Neuhaus, C. S., Gabernet, G., Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Designing anticancer peptides by constructive machine learning. *ChemMedChem* 13, 1300–1302. doi: 10.1002/cmdc.201800204

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A., (2017). Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. doi: arXiv:1705.10843v3

Guyer, M. S., and Collins, F. S. (1995). How is the Human Genome Project doing, and what have we learned so far? *Proc. Natl. Acad. Sci. U. S. A.* 92, 10841–10848. doi: 10.1073/pnas.92.24.10841

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. *J. Cheminform.* 7, 23. doi: 10.1186/s13321-015-0068-4

Hertzberg, R. P., and Pope, A. J. (2000). High-throughput screening: new technology for the 21st century. *Curr. Opin. Chem. Biol.* 4, 445–451. doi: 10.1016/S1367-5931(00)00110-1

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S., and Klambauer, G. (2019). Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *J. Chem. Inf. Model.* 59, 1163–1171. doi: 10.1021/acs.jcim.8b00670

How library-scale gene-expression profiling is changing drug discovery Available at: https://www.statnews.com/sponsor/2017/02/17/library-scale-gene-expression-profiling-changing-drug-discovery/ [Accessed October 24, 2019].

Hsieh, J.-H., Sedykh, A., Huang, R., Xia, M., and Tice, R. R. (2015). A data analysis pipeline accounting for artifacts in Tox21 quantitative high-throughput screening assays. *J. Biomol. Screen.* 20, 887–897. doi: 10.1177/1087057115581317

Hughes, T. B., Dang, N., Miller, G. P., and Swamidass, S. J. (2016). Modeling reactivity to biological macromolecules with a deep multitask network. *ACS Cent. Sci.* 2, 529–537. doi: 10.1021/acscentsci.6b00162

Human Genome Project Results Available at: https://www.genome.gov/human-genome-project/results [Accessed October 24, 2019].

Hung, J., Ravel, D., Lopes, S. C. P., Rangel, G., Nery, O. A., Malleret, B., et al. (2018). Applying faster R-CNN for object detection on malaria images. Available at: http://arxiv.org/abs/1804.09548 [Accessed June 20, 2019].

InChI and InChIKeys for chemical structures Available at: https://www.inchi-trust.org/ [Accessed October 24, 2019].

Iorio, F., Rittman, T., Ge, H., Menden, M., and Saez-Rodriguez, J. (2013). Transcriptional data: a new gateway to drug repositioning? *Drug Discovery Today* 18, 350–357. doi: 10.1016/j.drudis.2012.07.014

Ishimatsu-Tsuji, Y., Soma, T., and Kishimoto, J. (2010). Identification of novel hair-growth inducers by means of connectivity mapping. *FASEB J.* 24, 1489–1496. doi: 10.1096/fj.09-145292

Jadhav, A., Ferreira, R. S., Klumpp, C., Mott, B. T., Austin, C. P., Inglese, J., et al. (2010). Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* 53, 37–51. doi: 10.1021/jm901070c

Janowczyk, A., Basavanhally, A., and Madabhushi, A. (2017). Stain normalization using sparse autoEncoders (StaNoSA): application to digital pathology. *Comput. Med. Imaging Graph.* 57, 50–61. doi: 10.1016/j.compmedimag.2016.05.003

Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. Available at: http://arxiv.org/abs/1802.04364 [Accessed September 26, 2019].

Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, Å., Bukar, R., et al. (1995). Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118. doi: 10.1016/1074-5521(95)90283-X

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206. doi: 10.1038/nbt1284

Kensert, A., Harrison, P. J., and Spjuth, O. (2019). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS Discovery Adv. Life Sci. R&D* 24, 466–475. doi: 10.1177/2472555218818756

Kim, S. (2016). Getting the most out of PubChem for virtual screening. *Expert Opin. Drug Discovery* 11, 843–855. doi: 10.1080/17460441.2016.1216967

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019a). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi: 10.1093/nar/gky1033

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. Available at: http://arxiv.org/abs/1312.6114 [Accessed September 26, 2019].

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* 39, D1035–D1041. doi: 10.1093/nar/gkq1126

Kogej, T., Blomberg, N., Greasley, P. J., Mundt, S., Vainio, M. J., Schamberger, J., et al. (2013). Big pharma screening collections: more of the same or unique libraries? the AstraZeneca–Bayer Pharma AG case. *Drug Discovery Today* 18, 1014–1024. doi: 10.1016/J.DRUDIS.2012.10.011

Koutsoukas, A., Monaghan, K. J., Li, X., and Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* 9, 42. doi: 10.1186/s13321-017-0226-y

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 32, i52–i59. doi: 10.1093/bioinformatics/btw252

Kraus, O. Z., Grys, B. T., Ba, J., Chong, Y., Frey, B. J., Boone, C., et al. (2017). Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* 13, 924. doi: 10.15252/msb.20177551

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* (80-. ) 313, 1929–1935. doi: 10.1126/science.1132939

Laufkötter, O., Sturm, N., Bajorath, J., Chen, H., and Engkvist, O. (2019). Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold-hopping capability. *chemRxiv*. 11, 54. doi: 10.26434/chemrxiv.7725209.v1

Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W. T., Kowalczyk, W., et al. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* 9, 45. doi: 10.1186/s13321-017-0232-0

Lin, A. I., Madzhidov, T. I., Klimchuk, O., Nugmanov, R. I., Antipin, I. S., and Varnek, A. (2016). Automatized assessment of protective group reactivity: a step toward big reaction data analysis. *J. Chem. Inf. Model.* 56, 2140–2148. doi: 10.1021/acs.jcim.6b00319

Liu, K., Sun, X., Jia, L., Ma, J., Xing, H., Wu, J., et al. (2019). Chemi-net: a molecular graph convolutional network for accurate drug property prediction. *Int. J. Mol. Sci.* 20, 3389. doi: 10.3390/ijms20143389

Loo, L.-H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nat. Methods* 4, 445–453. doi: 10.1038/nmeth1032

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* 55, 263–274. doi: 10.1021/ci500747n

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., et al. (2011). Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* 10, 188–195. doi: 10.1038/nrd3368

Martin, E. J., Polyakov, V. R., Zhu, X.-W., Tian, L., Mukherjee, P., and Liu, X. (2019). All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration $IC_{50s}$ for 8558 Novartis Assays. *J. Chem. Inf. Model.* doi: 10.1021/acs.jcim.9b00375

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12213076 [Accessed June 20, 2019]. doi: 10.1021/jm020155c

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3, 80. doi: 10.3389/fenvs.2015.00080

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9, 5441–5451. doi: 10.1039/C8SC00148K

Mayr, L. M., and Bojanic, D. (2009). Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* 9, 580–588. doi: 10.1016/j.coph.2009.08.004

MELLODDY Consortium| Available at: https://cordis.europa.eu/project/rcn/223634/factsheet/en [Accessed October 24, 2019]

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* 37, 1700153. doi: 10.1002/minf.201700153

Mervin, L. H., Afzal, A. M., Drakakis, G., Lewis, R., Engkvist, O., and Bender, A. (2015). Target prediction utilising negative bioactivity data covering large chemical space. *J. Cheminform.* 7, 51. doi: 10.1186/s13321-015-0098-y

Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model.* 58, 472–479. doi: 10.1021/acs.jcim.7b00414

Muresan, S., Petrov, P., Southan, C., Kjellberg, M. J., Kogej, T., Tyrchan, C., et al. (2011). Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* 16, 1019–1030. doi: 10.1016/j.drudis.2011.10.005

Nehme, E., Weiss, L. E., Michaeli, T., and Shechtman, Y. (2018). Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica* 5, 458. doi: 10.1364/OPTICA.5.000458

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* 9, 48. doi: 10.1186/s13321-017-0235-x

Ouyang, W., Aristov, A., Lelek, M., Hao, X., and Zimmer, C. (2018). Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* 36, 460–468. doi: 10.1038/nbt.4106

Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815. doi: 10.1038/nbt1228

Paricharak, S., IJzerman, A. P., Bender, A., and Nigsch, F. (2016). Analysis of iterative screening with stepwise compound selection based on novartis in-house HTS data. *ACS Chem. Biol.* 11, 1255–1264. doi: 10.1021/acschembio.6b00029

Pärnamaa, T., and Parts, L. (2017). Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *Genes|Genomes|Genetics* 7, 1385–1392. doi: 10.1534/g3.116.033654

Pascale, C. (2015). Genometry Announces Deal with Janssen for Library-Scale Gene-Expression Profiling | Business Wire. Available at: https://www.businesswire.com/news/home/20151007006618/en#.VhZdNWTBzRZ [Accessed June 20, 2019].

Paul, K. D., Shoemaker, R. H., Hodes, L., Monks, A., Scudiero, D. A., Rubinstein, L., et al. (1989). Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* 81, 1088–1092. doi: 10.1093/jnci/81.14.1088

Pearce, B. C., Sofia, M. J., Good, A. C., Drexler, D. M., and Stock, D. A. (2006). An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* 46, 1060–1068. doi: 10.1021/ci050504m

Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., et al. (2012). Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* 7, 1399–1409. doi: 10.1021/cb3001028

Pharma Companies Join Forces to Train AI for Drug Discovery Collectively Available at: https://www.biopharmatrend.com/post/97-pharma-companies-join-forces-to-train-ai-for-drug-discovery-collectively/ [Accessed June 5, 2019].

Plouffe, D., Brinker, A., McNamara, C., Henson, K., Kato, N., Kuhen, K., et al. (2008). *In silico* activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci.* 105, 9059–9064. doi: 10.1073/pnas.0802982105

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al., (2018a). Molecular sets (MOSES): a benchmarking platform for molecular generation models.

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., et al. (2018b). Entangled conditional adversarial autoencoder for de novo drug discovery. *Mol. Pharm.* 15, 4398–4405. doi: 10.1021/acs.molpharmaceut.8b00839

Proffitt, A. (2008). AstraZeneca invests in data, discovery management - bio-IT World. Available at: http://www.bio-itworld.com/issues/2008/july-august/best-practices-astrazeneca.html [Accessed June 20, 2019].

Prykhodko, O., Johansson, S., Kotsias, P.-C., Bjerrum, E. J., Engkvist, O., and Chen, H., (2019). A de novo molecular generation method using latent vector based generative adversarial network. doi: 10.26434/chemrxiv.8299544.v1

Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sanchez-Lengeling, B., Aspuru-Guzik, A., et al. (2018). Reinforced adversarial neural computer for *de novo* molecular design. *J. Chem. Inf. Model.* 58, 1194–1204. doi: 10.1021/acs.jcim.7b00690

Pyzer-Knapp, E. O. (2018). Bayesian optimization for accelerated drug discovery. *IBM J. Res. Dev.* 62, 2, 1–2:7. doi: 10.1147/JRD.2018.2881731

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., et al. (2017). Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* 57, 2068–2076. doi: 10.1021/acs.jcim.7b00146

Reaxys Database. Available at: https://www.reaxys.com/#/login [Accessed October 24, 2019].

Reilly, T. J. (2009). The preparation of lidocaine. *J. Chem. Educ.* 76, 1557. doi: 10.1021/ed076p1557

Reisen, F., Sauty de Chalon, A., Pfeifer, M., Zhang, X., Gabriel, D., and Selzer, P. (2015). Linking phenotypes and modes of action through high-content screen fingerprints. *Assay Drug Dev. Technol.* 13, 415–427. doi: 10.1089/adt.2015.656

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Reymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48, 722–730. doi: 10.1021/ar500432k

Riniker, S., Wang, Y., Jenkins, J. L., and Landrum, G. A. (2014). Using information from historical high-throughput screens to predict active compounds. *J. Chem. Inf. Model.* 54, 1880–1891. doi: 10.1021/ci500190p

Rivenson, Y., Göröcs, Z., Günaydın, H., Zhang, Y., Wang, H., Ozcan, A., et al., (2018). "*Conference on lasers and electro-optics*," in *deep learning microscopy: enhancing resolution, field-of-view and depth-of-field of optical microscopy images using neural networks* (Washington, D.C: OSA), AM1J.5. doi: 10.1364/CLEO_AT.2018.AM1J.5

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Ronneberger, O., Fischer, P., and Brox, T., (2015). *U-Net: convolutional networks for biomedical image segmentation*. Cham: Springer, 234–241. doi: 10.1007/978-3-319-24574-4_28

Schamberger, J., Grimm, M., Steinmeyer, A., and Hillisch, A. (2011). Rendezvous in chemical space? Comparing the small molecule compound libraries of bayer and schering. *Drug Discovery Today* 16, 636–641. doi: 10.1016/j.drudis.2011.04.005

Schneider, G., and Fechner, U. (2005). Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discovery* 4, 649–663. doi: 10.1038/nrd1799

Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A., and Landrum, G. A. (2016). Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* 59, 4385–4402. doi: 10.1021/acs.jmedchem.6b00153

Schreck, J. S., Coley, C. W., and Bishop, K. J. M. (2019). Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci*. 5, 970–981. doi: 10.1021/acscentsci.9b00055

Schwaller, P., Gaudin, T., Lányi, D., Bekas, C., and Laino, T. (2018a). "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* 9, 6091–6098. doi: 10.1039/c8sc02339e

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2018b). Molecular Transformer - a model for uncertainty-calibrated chemical reaction prediction. Available at: http://arxiv.org/abs/1811.02633 [Accessed June 25, 2019].

SciFinder. Available at: https://scifinder.cas.org [Accessed October 24, 2019]

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018a). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi: 10.1021/acscentsci.7b00512

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018b). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610. doi: 10.1038/nature25978

Segler, M. H. S., and Waller, M. P. (2017a). Modelling chemical reasoning to predict and invent reactions. *Chem. A Eur. J.* 23, 6118–6128. doi: 10.1002/chem.201604556

Segler, M. H. S., and Waller, M. P. (2017b). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. A Eur. J.* 23, 5966–5971. doi: 10.1002/chem.201605499

Silverman, B. W., and Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important contribution to nonparametric discriminant analysis and density estimation: commentary on fix and hodges (1951). *Int. Stat. Rev./Rev. Int. Stat.* 57, 233. doi: 10.2307/1403796

Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., et al. (2018). Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.* 25, 611–618.e3. doi: 10.1016/j.chembiol.2018.01.015

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77–96ra77. doi: 10.1126/scitranslmed.3001318

Sterling, T., and Irwin, J. J. (2015). ZINC 15 – Ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

Stork, C., Chen, Y., Šícho, M., and Kirchmair, J. (2019). Hit Dexter 2.0: Machine-learning models for the prediction of frequent hitters. *J. Chem. Inf. Model.* 59, 1030–1043. doi: 10.1021/acs.jcim.8b00677

Stork, C., Wagner, J., Friedrich, N. O., de Bruyn Kops, C., Šícho, M., and Kirchmair, J. (2018). Hit dexter: a machine-learning model for the prediction of frequent hitters. *ChemMedChem* 13, 564–571. doi: 10.1002/cmdc.201700673

Sturm, N., Sun, J., Vandriessche, Y., Mayr, A., Klambauer, G., Carlsson, L., et al. (2019). Application of bioactivity profile-based fingerprints for building machine learning models. *J. Chem. Inf. Model.* 59, 962–972. doi: 10.1021/acs.jcim.8b00550

Su, H., Xing, F., Kong, X., Xie, Y., Zhang, S., and Yang, L. (2015). "Robust Cell Detection and Segmentation in Histopathological Images Using Sparse Reconstruction and Stacked Denoising Autoencoders," in Medical image computing and computer-assisted intervention: MICCAI. International Conference on Medical Image Computing and Computer-Assisted Intervention. 383–390. doi: 10.1007/978-3-319-24574-4_46

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 Profiles. *Cell* 171, 1437–1452.e17. doi: 10.1016/j.cell.2017.10.049

Sullivan, E., Tucker, E. M., and Dale, I. L., (1999). "*Calcium signaling protocols*," in *measurement of [Ca<sup<2+</sup>]; Using the fluorometric imaging plate reader (FLIPR)* (New Jersey: Humana Press), 125–134. doi: 10.1385/1-59259-250-3:125

Sun, J., Jeliazkova, N., Chupakin, V., Golib-Dzib, J. F., Engkvist, O., Carlsson, L., et al. (2017). ExCAPE-DB: An integrated large scale dataset facilitating big data analysis in chemogenomics. *J. Cheminform.* 9, 1–9. doi: 10.1186/s13321-017-0203-5

Sushko, I., Salmina, E., Potemkin, V. A., Poda, G., and Tetko, I. V. (2012). ToxAlerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* 52, 2310–2316. doi: 10.1021/ci300245q

Sushko, Y., Novotarskyi, S., Körner, R., Vogt, J., Abdelaziz, A., and Tetko, I. V. (2014). Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. *J. Cheminform.* 6, 1–18. doi: 10.1186/s13321-014-0048-0

Tennant, R. W., and Ashby, J. (1991). Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutat. Res. Genet. Toxicol.* 257, 209–227. doi: 10.1016/0165-1110(91)90002-D

Thomson Reuters. Available at: https://www.thomsonreuters.com/en.html [Accessed October 24, 2019].

Tsubaki, M., Tomii, K., and Sese, J. (2019). Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35, 309–318. doi: 10.1093/bioinformatics/bty535

Wang, H., Rivenson, Y., Jin, Y., Wei, Z., Gao, R., Günaydın, H., et al. (2019). Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nat. Methods* 16, 103–110. doi: 10.1038/s41592-018-0239-0

Wang, L., Ma, C., Wipf, P., Liu, H., Su, W., and Xie, X.-Q. (2013). TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J.* 15, 395–406. doi: 10.1208/s12248-012-9449-z

Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., et al. (2014). PubChem BioAssay: 2014 update. *Nucleic Acids Res.* 42, D1075–D1082. doi: 10.1093/nar/gkt978

Warr, W. A. (2014). A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Mol. Inform.* 33, 469–476. doi: 10.1002/minf.201400052

Wassermann, A. M., Lounkine, E., Davies, J. W., Glick, M., and Camargo, L. M. (2015a). The opportunities of mining historical and collective data in drug discovery. *Drug Discovery Today* 20, 422–434. doi: 10.1016/j.drudis.2014.11.004

Wassermann, A. M., Lounkine, E., Hoepfner, D., Le Goff, G., King, F. J., Studer, C., et al. (2015b). Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol.* 11, 958–966. doi: 10.1038/nchembio.1936

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36. doi: 10.1021/ci00057a005

Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* 29, 97–101. doi: 10.1021/ci00062a008

Willett, P. (2011). Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. *Wiley Interdiscip. Rev. Data Min. Knowl. Discovery* 1, 241–251. doi: 10.1002/widm.26

Wilson, B. J., and Nicholls, S. G. (2015). The human genome project, and recent advances in personalized genomics. *Risk Manage. Healthc. Policy* 8, 9–20. doi: 10.2147/RMHP.S58728

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi: 10.1039/c7sc02664a

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* acs.jmedchem.9b00959. doi: 10.1021/acs.jmedchem.9b00959

Xu, Y., Lin, K., Wang, S., Wang, L., Cai, C., Song, C., et al. (2019). Deep learning for molecular generation. *Future Med. Chem.* 11, 567–597. doi: 10.4155/fmc-2018-0358

Yang, J. J., Ursu, O., Lipinski, C. A., Sklar, L. A., Oprea, T. I., and Bologa, C. G. (2016). Badapple: promiscuity patterns from noisy evidence. *J. Cheminform.* 8, 29. doi: 10.1186/s13321-016-0137-3

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59, 3370–3388. doi: 10.1021/acs.jcim.9b00237

Yang, S. J., Berndl, M., Michael Ando, D., Barch, M., Narayanaswamy, A., Christiansen, E., et al. (2018). Assessing microscope image focus quality with deep learning. *BMC Bioinf.* 19, 77. doi: 10.1186/s12859-018-2087-4

Yoshida, M., Hinkley, T., Tsuda, S., Abul-Haija, Y. M., Mcburney, R. T., Kulikov, V., et al. (2018). Exploring sequence space for antimicrobial peptides using evolutionary algorithms and machine learning. available at: https://blogit.itu.dk/evoblissproject/wp-content/uploads/sites/19/2018/03/yoshida_2018_preprint_Using-Evolutionary-Algorithms-and-Machine-Learning-to-Explore-Sequence-Space-for-the-Discovery-of-Antimicrobial-Peptides_.pdf [Accessed August 2, 2019].

You, J., Liu, B., Ying, R., Pande, V., and Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation. Available at: http://arxiv.org/abs/1806.02473 [Accessed September 26, 2019].

Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G.-W., Tao, C. Y., et al. (2008). Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* 4, 59–68. doi: 10.1038/nchembio.2007.53

Zhai, Y., Chen, K., Zhong, Y., Zhou, B., Ainscow, E., Wu, Y.-T., et al. (2016). An automatic quality control pipeline for high-throughput screening hit identification. *J. Biomol. Screen.* 21, 832–841. doi: 10.1177/1087057116654274

Zhang, H. (2004). "*Proceedings of the seventeenth international florida artificial intelligence research society conference, FLAIRS 2004*," in *the optimality of Naive Bayes*, 562–567. Available at: https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf [Accessed September 25, 2019].

Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S., Ye, J., et al. (2015). Deep model based transfer and multi-task learning for biological image analysis in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1475–1484 doi: 10.1145/2783258.2783304

Zhang, Y., and Lee, A. A. (2019). Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* 10, 8154–8163. doi: 10.1039/c9sc00616h

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040. doi: 10.1038/s41587-019-0224-x

**Conflict of Interest:** Authors LD, JA-P, JK, OE, EB, TK and HC were employed by AstraZeneca. Authors JMK and BB were employed by Boehringer Ingelheim Pharma GmbH & Co. KG.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/fphar.2019.01301

# HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction *via* Chou's Five-Step Rule

*Ping Xuan[1], Hui Cui[2], Tonghui Shen[1]*, Nan Sheng[1] and Tiangang Zhang[3]**

*[1] School of Computer Science and Technology, Heilongjiang University, Harbin, China, [2] Department of Computer Science and Information Technology, La Trobe University, Bundoora, VIC, Australia, [3] School of Mathematical Science, Heilongjiang University, Harbin, China*

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University,*
*China*

***Reviewed by:***
*Feng Zhu,*
*Zhejiang University,*
*China*
*Kuo-Chen Chou,*
*The Gordon Life Science Institute,*
*United States*

***Correspondence:***
*Tonghui Shen*
*shen.tonghui@163.com*
*Tiangang Zhang*
*zhang@hlju.edu.cn*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *01 July 2019*
***Accepted:*** *11 October 2019*
***Published:*** *08 November 2019*

***Citation:***
*Xuan P, Cui H, Shen T, Sheng N and Zhang T (2019) HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction via Chou's Five-Step Rule. Front. Pharmacol. 10:1301. doi: 10.3389/fphar.2019.01301*

Identifying new treatments for existing drugs can help reduce drug development costs and explore novel indications of drugs. The prediction of associations between drugs and diseases is challenging because their similarities and relations are complicated and non-linear. We propose a HeteroDualNet model to address this issue. Firstly, three types of matrices are extracted to represent intra-drug similarities, intra-disease similarity and drug-disease associations. The intra-drug similarities consider three drug features and a newly introduced drug-related disease correlation. Secondly, an embedding mechanism is proposed to integrate these matrices in a heterogenous drug-disease association layer (hetero-layer). Further, a neighbouring heterogeneous layer (hetero-layer-N) is constructed to incorporate the biological premise that similar drugs can often treat related diseases. Finally, a dual convolutional neural network is built with hetero-layer and hetero-layer-N as two branches to learn from characteristics of drug-disease and the relations of their neighbours simultaneously. HeteroDualNet outperformed the other four methods in comparison over a public dataset of 763 drugs and 681 diseases in terms of Areas Under the Curves of Receiver Operating Characteristics and Precision-Recall, and recall rate at top *k*. Case study of five drugs further proved the capacity of HeteroDualNet in finding reliable disease candidates of drugs as validated by database records or literature. Our findings show that the embedded heterogenous layers of original and neighbouring drug-disease representations in a dual neural network improved the association prediction performance.

**Keywords: drug-disease association prediction, multiple kinds of similarities, neighbouring heterogeneous layer, deep learning, dual convolutional neural network**

## INTRODUCTION

The research and development (R&D) processes of new drugs are time-consuming and expensive. Stringent drug testing and approvals are required for an invented new drug to make it to market. For instance, it takes an average of 15 years from preliminary examination of compounds to clinical trials of drug candidates, and finally to drug marketing, while the estimated investment cost is about 800 million dollars (Adams and Brantner, 2006; Tamimi and Ellis, 2009; Pushpakom et al., 2018). However, even in the case of a significant amount of time and capital investment, the R&D of new drugs still faces high failure risks (Li et al., 2016). Meanwhile, the number of new drugs approved by major drug regulatory agencies around the world is decreasing year by year (Grabowski, 2004; Nosengo, 2016). According to the statistics of the US Food and Drug Administration (FDA), the average success rate of new drugs approved from 2003 to 2011 was less than 10% (Padhy and Gupta, 2011; Hay et al., 2014; Pritchard et al., 2017). Therefore, the conventional R&D productivity of new drugs has been stagnant in the last few decades (Paul et al., 2010).

Given the challenges faced by conventional drug R&D techniques, there are significant needs of innovative drug development strategies to increase R&D productivity, which is one of the essential priorities in the pharmaceutical industry. Drug repositioning techniques, or the so-called reuse of existing drugs, have been proved of its advantages over the conventional drug R&D strategies. (Hurle et al., 2013) Drug repositioning is the process to identify new indications for existing drugs and is playing an essential role in the state-of-the-art drug R&D process. Drug repositioning can be applied to drugs which have been approved to market. Because those drugs have passed the procedures of laboratory, pharmacokinetics, toxicology and safety testing, drug developers can use these drugs in clinical trials directly. In this way, drug repositioning skips those procedures and will significantly reduce the time and financial costs in drug development. At the same time, it also reduces the risks of drug development failure. Thus, drug repositioning has attracted great interests in the pharmaceutical industry and research community (Hurle et al., 2013).

Drug repositioning aims to find potential indications for existing drugs (Shim and Liu, 2014; Chen et al., 2016). Computational methods in biology are playing increasingly important roles in the stimulation, development and finding of new drugs (Chou, 2015). To develop useful predictors for biological systems *via* computing models, Chou's 5-steps (Chou, 2011; Chou, 2019b) are used by recent publications (Chou, 2019a; Awais et al., 2019; Ehsan et al., 2019; Hussain et al., 2019). These steps provide guidance in the development and validation of computerized methods, which include selection of a valid benchmark dataset for training and testing, representation of samples by effective formulation to reflect intrinsic correlations with the target, development of algorithms for prediction, objective performance evaluation by cross-validation, and consideration of public accessibility by web-server.

Several methods have been proposed to predict drug-disease associations. For example, Chiang and Butte proposed a technique based on the internal correlation of networks to predict the potential drug-disease associations (Chiang and Butte, 2009). Sirota et al. developed a prediction method by integrating the common gene expressions of drugs and diseases (Sirota et al., 2011). Besides, Yang and Agarwal et al. proposed to infer the new drug-disease associations by using the phenotypic information on drug side effects (Yang and Agarwal, 2011). Most of these methods are designed for early-stage drugs which have multiple uses and treatment plans. They cannot be used for association prediction when there are no common gene expressions and side effects information between drugs and diseases.

With the increasing amount and variety of drug-related data, recent research has been focusing on integrating multimodality information to investigate the potential uses of drugs. Gottlieb et al. proposed a classification model which used various associations of drug and disease as distinguish signatures. A logistical regression model was then used to predict the indications of drugs (Gottlieb et al., 2011). A kernel-based strategy was proposed to integrate molecular structure, molecular activity, and phenotypic information for drug repositioning (Wang et al., 2013). Heterogenous networks have also been investigated to predict drug indications. Heterogeneous networks are constructed by associating drugs, diseases, targets and genes. The prediction can be achieved by approaches such as network clustering (Wu et al., 2013), priority ranking (Martinez et al., 2015), network topology measurement (Chen et al., 2015), or iteration (Wang et al., 2014b). Given these heterogeneous networks, some other models integrated multiple chemical features such as chemical phenotype of drugs and molecular characteristics of diseases. Then the prediction of new drug indications can be achieved by proteochemometric models (Dakshanamurthy et al., 2012; Yu et al., 2015), statistical (Iwata et al., 2015) or sparse subspace learning (Liang et al., 2017; Xuan et al., 2019) models.

Most of the above existing methods for drug-disease association predictions are shallow models. The associations between drugs and diseases, however, are non-linear and complicated. It is challenging for these shallow models to dig out advanced level while hidden drug-disease relations. Thus, there are great necessities to develop models to learn the deep representations of drug-disease associations for improved drug indication prediction.

In this work, we propose a novel convolutional network with heterogeneous layers and dual branches, referred to as HeteroDualNet, for drug-disease association prediction. Our first unique contribution is the extraction of three types of matrices for the representation and indexing of intra-drug similarity, drug-disease similarity and drug-disease associations. When constructing intra-drug similarity matrices, we consider both regular drug features, including chemical substructures, domains and annotations of target proteins, and a newly introduced feature calculated by drug-related disease correlations. The second contribution is that we construct a new heterogenous drug-disease association layer (hetero-layer) to associate three types of matrices by a proposed embedding mechanism. Further, a drug-disease association layer with neighbouring information (hetero-layer-N) is constructed by the embedding mechanism to reflect the biological premise that similar drugs can often treat related diseases. Finally, HeteroDualNet is built to predict drug-disease associations with hetero-layer and hetero-layer-N as two branches to learn from both original and neighbouring characteristics of drugs and diseases simultaneously. We also investigate the prediction capacity of the proposed model in therapeutic drug indications by case studies of five drugs.

# MATERIALS AND METHODS

## Dataset

We obtained the data of drugs and diseases from a published work (Wang et al., 2014a). There are 763 drugs, 681 diseases and 3051 known drug-disease associations. The characteristics of each drug include 881 chemical substructures which were initially derived from the chemical fingerprints extracted from the PubChem database (Wang et al., 2009); 1,426 target protein domains from the InterPro database (Mitchell et al., 2015); and 4,447 target protein annotations obtained from the UniProt database (Uniprot, 2010). The similarities among diseases were calculated by (Wang et al., 2010) and provided in the dataset.

## Hypothesis and Framework

We hypothesize that a dual neural network which integrates features of drugs, drug-related disease correlations, and the biological premise of drugs and diseases will improve the performance of drug-disease association predictions. The overview of the proposed method is shown in **Figure 1**. Given the input dataset, the drugs and diseases information is firstly extracted and indexed by three types of similarity matrices in terms of intra-drug, intra-disease and drug-disease. Then, a heterogenous drug-disease association layer, referred by hetero-layer, is constructed by a proposed embedding mechanism to associate those matrices among drugs and diseases. Another heterogeneous layer with neighbouring information, denoted by hetero-layer-N, is built to represent the biological premise that similar drugs can often treat related diseases. Lastly, the dual convolutional neural network is constructed by integrating hetero-layer and hetero-layer-N using a fully connected layer.

## Drug and Disease Similarity and Association Representation

We define three types of matrices to represent and index the information of drugs and diseases in terms of intra-drug similarity, intra-disease similarity and drug-disease associations.

### Intra-Disease Similarity Matrix

Intra-disease similarities were calculated and provided by (Wang et al., 2010) based on semantic information of diseases (Wang et al., 2010). This information was also used in published work such as Liang et al. (2017) and Zhang et al. (2018). The similarity between disease $d_i$ and the disease $d_j$ is denoted by $\boldsymbol{D}(i,j) \in [0,1]$. where is the intra-disease similarity matrix and $N^{DI}$ is the number of diseases. The greater $D(i,j)$ is, the higher similarity between diseases $d_i$ and $d_j$.

### Intra-Drug Similarity Matrix

Four intra-drug similarity matrices are obtained by calculating the similarities between drugs from four perspectives, including the chemical substructures, target protein domain information, target protein annotations and the related disease information of drugs.

The first three intra-drug similarity matrices of chemical substructure, domain and annotation information of target proteins represent that if two drugs have more common chemical substructures, target protein domains or gene ontology information, the more similar they are. Thus, we calculate these three intra-drug similarity matrices by cosine similarity measurement (Liang et al., 2017).

To calculate the first three intra-drug similarity matrices, we firstly obtain matrices of features and drugs. The feature matrix of chemical feature and all the drugs is denoted by $\boldsymbol{F}_1 \in \mathbb{R}^{N_1^F \times N^{DR}}$ where $N_1^F$ is the number of chemical substructure features, and $N^{DR}$ is the number of drugs. Similarly, the feature matrix of protein domain and drugs is $\boldsymbol{F}_2 \in \mathbb{R}^{N_2^F \times N^{DR}}$ and that of protein annotation and drugs is $\boldsymbol{F}_3 \in \mathbb{R}^{N_3^F \times N^{DR}}$, where $N_2^F$ is the number of target protein domain feature and $N_3^F$ is the number of target protein annotation. Each element of the vectors is 1 or 0 according to whether the drug has such a feature. Given the dataset used in this paper, $N_1^F = 881$, $N_2^F = 1426$ and $N_3^F = 4{,}447$. Let $\boldsymbol{f}_{t,i}$ be the feature vector of $i$-th drug $r_i$ in the $t$-th feature matrix $F_t$ ($1 \le t \le 3$), the similarity $\boldsymbol{R}_t(i,j)$ between drugs $r_i$ and $r_j$ in terms of feature $t$ is calculated by cosine similarity measurement as

$$\boldsymbol{R}_t(i,j) = \frac{f_{t,i} \cdot f_{t,j}}{||f_{t,i}||\,||f_{t,j}||}. \tag{1}$$

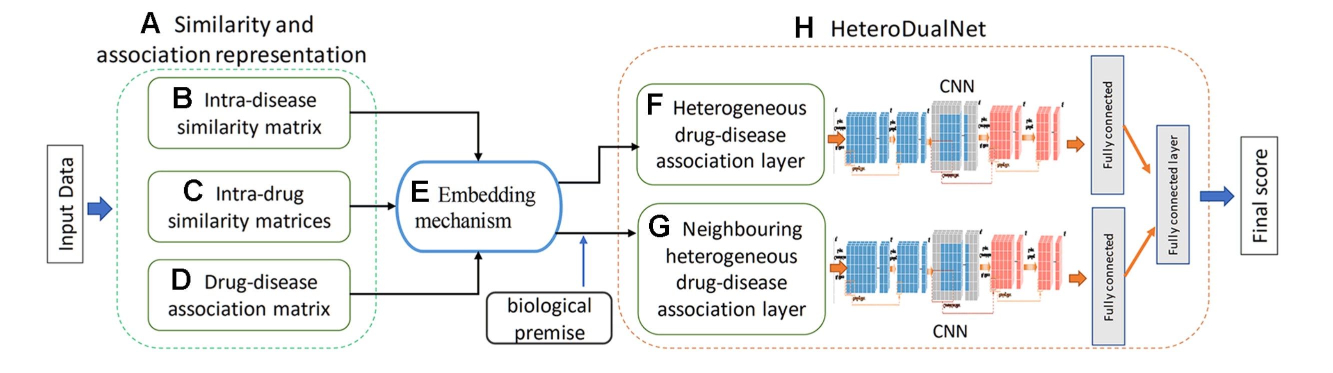

A Similarity and association representation
B Intra-disease similarity matrix
C Intra-drug similarity matrices
D Drug-disease association matrix
Input Data
E Embedding mechanism
biological premise
H HeteroDualNet
F Heterogeneous drug-disease association layer
G Neighbouring heterogeneous drug-disease association layer
CNN
CNN
Fully connected
Fully connected
Fully connected layer
Final score

**FIGURE 1 |** Overview of the proposed HeteroDualNet model for drug-disease association prediction. Given input data, **(A)** similarity and association representations are extracted including **(B)** intra-disease similarity, **(C)** intra-drug similarity, and **(D)** drug-disease association. Then **(E)** an embedding mechanism is proposed to embed these matrices. The final drug-disease association score is obtained by **(H)** HeteroDualNet with **(F)** heterogeneous and biological premise enhanced **(G)** neighboring heterogeneous drug-disease association layers.

where $R_t(i,j)\in[0,1]$ and higher values indicate higher similarity between a pair of drugs.

The fourth intra-drug similarities matrix $\boldsymbol{R}_4 \in \mathbb{R}^{N^{DR}\times N^{DR}}$ is obtained based on the idea that if two drugs are associated with similar diseases, the drugs are more likely to be correlated. Given the dataset of diseases $\boldsymbol{DI}=\{d_k|k\in[1,N^{DI}]\}$ and intra-disease similarity matrix $D$ if $i$-th drug $r_i$ is associated with a subset of diseases $\boldsymbol{DI}_m \subset \boldsymbol{DI}$, and drug $r_j$ is related to a disease subset $\boldsymbol{DI}_n$, the similarity $\boldsymbol{R}_4(i,j)$ between $i$-th and $j$-th drugs can be obtained by calculating the similarity between $\boldsymbol{DI}_m$ and $\boldsymbol{DI}_n$ as proposed in our previous work (Xuan et al., 2019) by

$$\boldsymbol{R}_4(i,j)=\frac{\sum_{k=1}^{num(DI_m)}\max\left(D\left(d_{i,k},d_{j,*}\right)\right)+\sum_{k=1}^{num(DI_n)}\max\left(D\left(d_{j,k},d_{i,*}\right)\right)}{num\left(DI_m\right)+num\left(DI_n\right)} \quad (2)$$

where $num(DI_m)$ denotes the number of elements in $DI_m$. $d_{i,k}$ represents the $k$th disease related with drug $r_i$, $d_{j,*}$ denotes all the related diseases of drug $r_j$, and $max(D(d_{i,k},d_{j,*}))$ is the maximum similarity between drug $r_j^{'}s$ $k$th related disease and all the related diseases of $r_j$. Similarly, $max(D(d_{i,k},d_{j,*}))$ denotes the maximum similarity between drug $r_j^{'}s$ $k$th related disease and all the associated diseases of $r_i$. The final similarity between $r_i$ and $r_j$ is obtained by the average maximum similarities between diseases in their relevant disease subsets $DI_m$ and $DI_n$.

### Drug-Disease Association Matrix

The drug-disease association matrix is denoted by $\boldsymbol{A}\in\mathbb{R}^{N^{DR}\times N^{DI}}$ where an element can be 0 or 1. 1 indicates that a drug and a disease are related, and the association is available; while 0 represents that the relation between a drug and a disease is unknown. Among all the 763 drugs and 681 diseases in the dataset, 3051 drug-disease associations are available. The remaining unknown associations are to be predicted.

## HeteroDualNet Architecture

The sparsity of drug-disease associations makes it challenging to dig out the hidden characteristics and relations between drugs and diseases. We construct HeteroDualNet, a dual convolutional neural network with heterogeneous layers, to predict drug-disease associations. One branch integrates the three matrices of drugs and diseases by a heterogeneous association layer (hetero-layer); the other branch incorporates the neighbouring information in a neighbouring heterogenous layer (hetero-layer-N). The two heterogeneous layers are learnt by passing through convolutional and pooling layers and joint by a connection module. The final association score is obtained by weighted voting of association scores from two branches.

### Embedding Mechanism for Heterogeneous Drug-Disease Association Matrix

The heterogenous drug-disease association layer is built upon an embedded matrix of afore-extracted matrices. An embedding mechanism is proposed based on the idea that if two drugs are more similar, the more likely they are associated with related diseases, whereas two similar diseases tend to be associated with similar drugs. Given intra-drug matrices $\boldsymbol{R}_t$, drug-disease association matrix $A$ and intra-disease matrix $D$**,** the heterogeneous matrix $\boldsymbol{X}_L$ of drug $r_i(i\in[1,N^{DR}])$ and disease $d_k(k\in[1,N^{DI}])$ is obtained by the following embedding procedures.

Firstly, row vectors $R_t(i,*)$ are combined sequentially as $\boldsymbol{X}_{\mathrm{L},11}=[\boldsymbol{R}_1(i,*); \boldsymbol{R}_2(i,*); \boldsymbol{R}_3(i,*); \boldsymbol{R}_4(i,*)]$ where $\boldsymbol{R}_t(i,*)$ denotes the $i$-th row in an intra-drug similarity matrix $R_t$ which records the $t$-th type of similarities between $r_i$ and all drugs, $t$ = 1,2,3,4 denotes chemical substructures, target protein domains, target protein annotations and related disease information respectively. Secondly, the transposed column vector $\boldsymbol{A}^{\mathrm{T}}(*,k)$ is concatenated under $\boldsymbol{R}_4(i,*)$ as $\boldsymbol{X}_{L,21}$ where $\boldsymbol{A}(*,k)$ is the $k$th column of $A$ which contains the associations between $d_k$ and all the drugs. Thirdly, $A(i,*)$ is repeated four times and spliced to the right of each row in $\boldsymbol{X}_{L,11}$ as $\boldsymbol{X}_{L,12}=[\boldsymbol{A}(i,*); \boldsymbol{A}(i,*); \boldsymbol{A}(i,*); \boldsymbol{A}(i,*)]$ where $\boldsymbol{A}(i,*)$ denotes the $i$th row of $A$ which includes the associations between $r_i$ and all the diseases. Lastly, $\boldsymbol{D}(k,*)$ is spliced under $X_{L,12}$ where $\boldsymbol{D}(k,*)$ is the $k$th row of $\boldsymbol{D}$ containing the similarities between $d_k$ and all the diseases. The final embedded matrix $\boldsymbol{X}_L\in\mathbb{R}^{5\times(N_r+N_d)}$ of drug $r_i$ and disease $d_k$ is formed as

$$\boldsymbol{X}_L=\begin{bmatrix}\boldsymbol{X}_{L,11} & \boldsymbol{X}_{L,12}\\ \boldsymbol{X}_{L,21} & \boldsymbol{X}_{L,22}\end{bmatrix}=\begin{bmatrix}\boldsymbol{R}_1(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_2(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_3(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_4(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{A}^{\boldsymbol{T}}(*,k) & \boldsymbol{D}(k,*)\end{bmatrix} \quad (3)$$

Given such a heterogeneous matrix $X_L$, the unknown drug-disease relations can be inferred *via* the correlations between diseases. In the meanwhile, the unavailable associations can be derived upon the similarities between drugs. In **Figure 2**, we illustrate the embedding procedure and use drug $r_2$ and disease $d_1$ whose association is unknown as an example. If $r_2$ is very similar to $r_3$ and $r_4$ (as shown in **Figure 2A**),$r_3$ and $r_4$ are closely associated with $d_1$(**Figure 2B**), it can be inferred that $r_2$ is more likely to be associated with $d_1$. Alternatively, if $d_1$ is similar to $d_4$ (shown in **Figure 2C**), and $d_4$ is related with $r_2$ (**Figure 2B**), a high possibility that $r_2$ is associated with $d_1$ can be derived.

### Neighbouring Heterogeneous Association Matrix

The neighbouring heterogeneous drug-disease association matrix $X_{L\text{-}N}$ embeds the neighbours of drug $r_i$ and disease $d_k$. The embedding is proposed based on the biological premise that if the neighbours of a drug are associated with the neighbours of a disease, there is a high probability that the drug and the disease are associated. The embedding procedures considering the neighbours of $r_i$ and $d_k$ is: Firstly, we find drugs $r_m$, $r_n$, $r_p$,

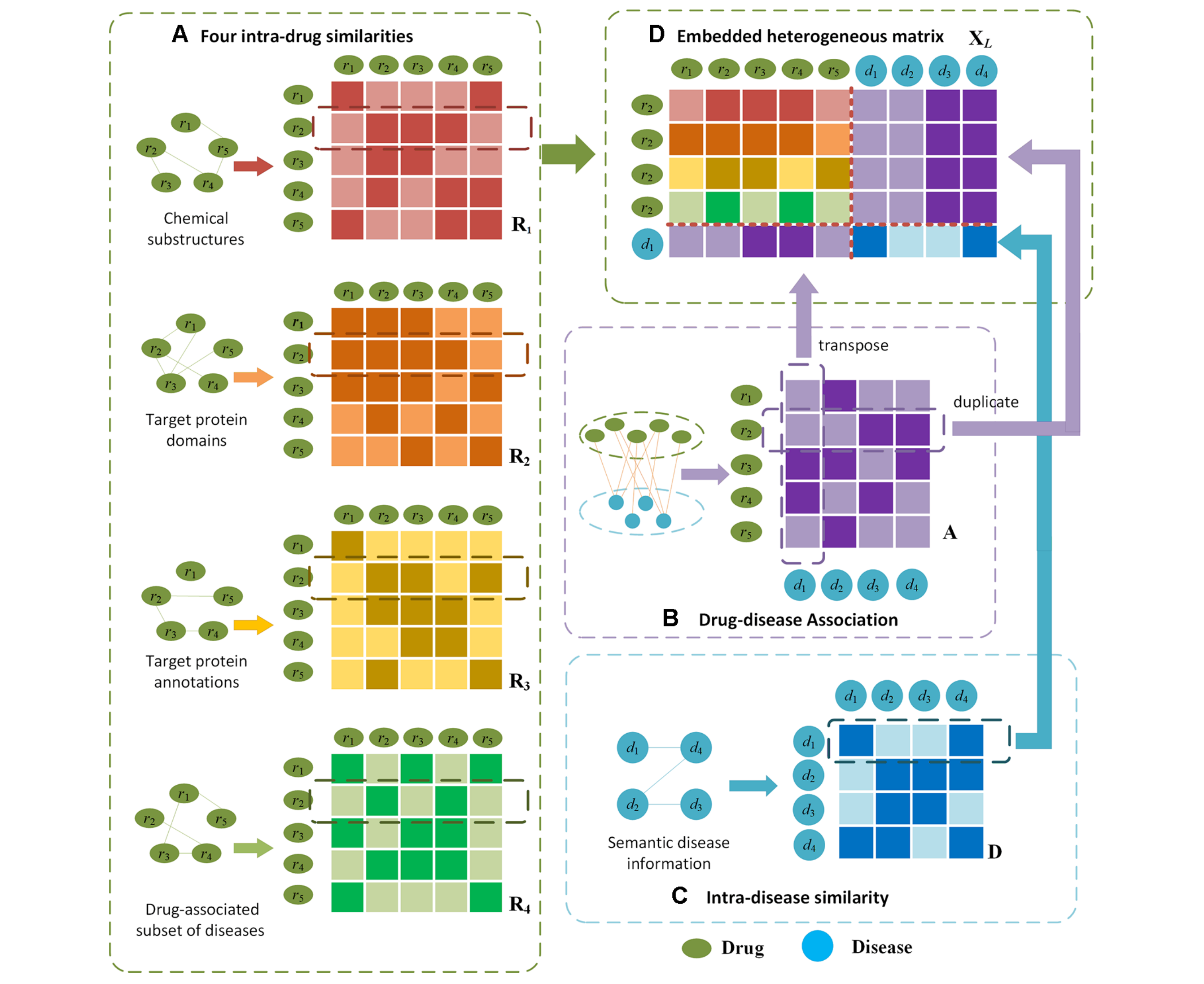

A Four intra-drug similarities
Chemical substructures
R1
Target protein domains
R2
Target protein annotations
R3
Drug-associated subset of diseases
R4
D Embedded heterogeneous matrix XL
transpose
duplicate
A
B Drug-disease Association
Semantic disease information
D
C Intra-disease similarity
Drug
Disease

**FIGURE 2 |** Illustration of the proposed embedding mechanism for heterogenous drug-disease association matrix. Given drug $r_2$ and disease $d_1$ as an example, **(D)** the heterogeneous matrix is obtained by integrating **(A)** four types of intra-drug similarities, **(B)** drug-disease associations and **(C)** intra-disease similarities. In **(A)** and **(C)**, darker colours indicate higher similarities; in **(B)** darker colour represents the drug-disease association is available.

and $r_q$ which are the most similar neighbours of drug $r_i$ in $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, $\boldsymbol{R}_3$ and $\boldsymbol{R}_4$ respectively. We also find $d_l$, the most similar neighbour of $d_k$, in $D$. Similar with $\boldsymbol{X}_{L,11}$, the $m$-th row of $\boldsymbol{R}_1$, $n$th row of $\boldsymbol{R}_2$, $p$-th row of $\boldsymbol{R}_3$, and $q$th row of $\boldsymbol{R}_4$ are combined from top to bottom to form $\boldsymbol{X}_{L-N,11}$. Secondly, the $l$-th column of $\boldsymbol{A}$ indicating the association between the most similar disease $d_l$ and all the drugs is transposed and concatenated under $\boldsymbol{X}_{L-N,11}$ as $\boldsymbol{X}_{L-N,21}$. Thirdly, row vectors $\boldsymbol{A}(m,*)$, $\boldsymbol{A}(n,*)$, $\boldsymbol{A}(p,*)$, $\boldsymbol{A}(q,*)$ are spliced to the right of each row in $\boldsymbol{X}_{L-N,11}$, where $\boldsymbol{A}(m,*)$, $\boldsymbol{A}(n,*)$, $\boldsymbol{A}(p,*)$, $\boldsymbol{A}(q,*)$ indicate the associations between drugs $r_m$,$r_n$,$r_p$ and $r_q$ and all the diseases. Lastly, the $l$-th row of $D$ containing the similarities between disease $d_l$ and all the other diseases is concatenated under $X_{L-N,21}$. In such a way, the final embedding of most similar neighbours of $r_i$ and $d_k$ is formed as $\boldsymbol{X}_{L-N} \in \mathbb{R}^{5\times\left(N^{DR}+N^{DI}\right)}$:

$$\boldsymbol{X}_L = \begin{bmatrix} \boldsymbol{X}_{L-N,11} & \boldsymbol{X}_{L-N,12} \\ \boldsymbol{X}_{L-N,21} & \boldsymbol{X}_{L-N,22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_1(m,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_2(n,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_3(p,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_4(q,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{A}^T(*,l) & \boldsymbol{D}(l,*) \end{bmatrix} \tag{4}$$

In $X_{L-N}$, the most similar neighbours of drugs and diseases serve as the bridge to propagate associations. In **Figure 3**, we use drug $r_2$ and disease $d_1$ whose association is unknown as an example to illustrate the embedding procedure and information

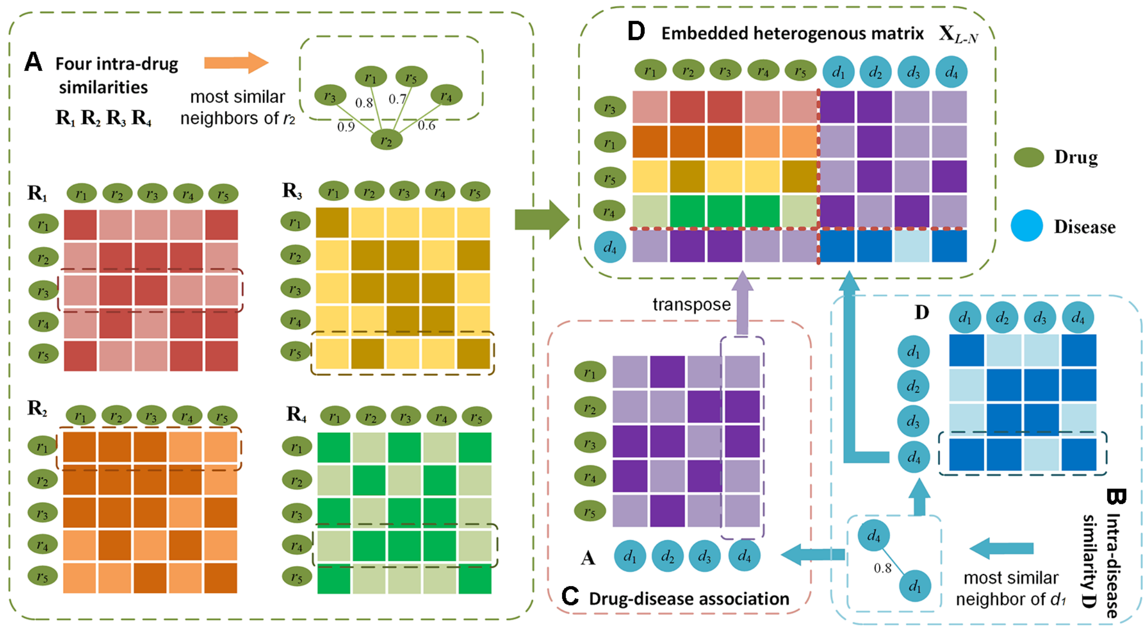

A Four intra-drug similarities
R1 R2 R3 R4
most similar neighbors of r2
r1 r5 r3 r4 r2
0.8 0.7 0.9 0.6
R1
R3
R2
R4
r1 r2 r3 r4 r5
D Embedded heterogenous matrix $X_{L-N}$
d1 d2 d3 d4
Drug
Disease
transpose
D
B Intra-disease similarity D
most similar neighbor of d1
0.8
A
C Drug-disease association

**FIGURE 3 |** Illustration of the embedding procedure for neighbouring heterogeneous matrix. Using drug $r_2$ and disease $d_1$ as an example, **(D)** the final matrix is obtained by finding the most similar neighbours (e.g. $r_3$,$r_1$,$r_5$,$r_4$) of $r_2$ calculated from **(A)** four intra-drug similarities respectively, the most similar neighbour (e.g. $d_4$) of drug $d_1$ by **(B)** intra-drug similarity matrix, and **(C)** drug-disease associations. In **(A)** and **(B)**, darker colours indicate higher similarities; in **(C)** darker colour represents the drug-disease association is available.

propagations. For instance, assume we find that drug $r_2$ likes $r_3$ the most in $\boldsymbol{R}_1$, $r_1$ in $\boldsymbol{R}_2$,$r_5$ in $\boldsymbol{R}_3$, and $r_4$ in $\boldsymbol{R}_4$(**Figure 3A**), and $d_1$ likes $d_4$ the most in $D$ (as shown in **Figure 3B**). In the embedded matrix $\boldsymbol{X}_{L-N}$, the left part indicates that all $r_2^{'}s$ most similar neighbours ($r_3$, $r_1$, $r_5$, $r_4$) are very similar to $r_2$ and $r_3$. Because $d_4$ is associated with bridging drugs $r_2$ and $r_3$ based on $A$ (**Figure 3C**), it can be inferred that there is a high probability that $r_2$ and $d_1$ are associated. The right part shows that the majority of $r_2^{'}s$ most similar neighbours are related with $d_2$. As $d_1^{'}s$ most similar neighbour $d_4$ is closely related to the bridging disease $d_2$ by $\boldsymbol{D}$, it can be derived that $d_1$ is probably related with $r_2$.

## HeteroDualNet for Association Prediction

The architecture of HeterDualNet is given in **Figure 4**. The hetero-layer and hetero-layer-N are obtained by zero padding heterogenous matrices $\boldsymbol{X}_L$ and $\boldsymbol{X}_{L-N}$. One branch in the dual CNN model alternates two convolution and two pooling operations over hetero-layer (**Figure 4A**), the other branch is built where hetero-layer-N is convolved and pooled for neighbouring feature representations (**Figure 4B**). These two branches are connected by a fully connected network to achieve the final association score between $r_i$ and $d_k$ (**Figure 4C**). Same network settings are used in the two branches, thus we introduce the branch with hetero-layer in detail.

**Convolutional module on hetero-layer.** The heterogeneous matrix $\boldsymbol{X}_L$ is firstly padded with zeros to preserve the marginal information of matrices. In the first convolutional layer, we set $N_1$ filters where each filter is with width and length of $n_{w_{c1}}$ and $n_{l_{c1}}$. The hetero-layer is thus denoted as $\boldsymbol{V}_1 \in \mathbb{R}^{(5+2l)\times(N_r+N_d+2p)}$ where $l=\left(n_{w_{c1}}-1\right)/2 \quad p=\left(n_{l_{c1}}-1\right)/2$. The case when $N_1$=3, $n_{w_{c1}}=3$, and $n_{l_{c1}}=5$ is illustrated as an example in **Figure 4A**. The weight parameter matrix of a $n$-th filter in the first convolutional layer is denoted by $\boldsymbol{W}_{1,n} \in \mathbb{R}^{n_{w_{c1}} \times n_{l_{c1}}}$, $n$∈[1,$N_1$]. The step size of a sliding window is set to be $1\times1$. The output of the first convolutional layer is obtained as $\boldsymbol{S}_1 \in \mathbb{R}^{N_1\times5\times(N_r+N_d)}$ where $\boldsymbol{S}_{1,n} \in \mathbb{R}^{5\times(N_r+N_d)}$ is the $n$-th output after $V_1$ is scanned by the $n$-th filter as

$$\boldsymbol{S}_{1,n}=\begin{bmatrix} \boldsymbol{S}_{1,n(1,1)} & \boldsymbol{S}_{1,n(1,2)} & \cdots & \boldsymbol{S}_{1,n(1,N_r+N_d)} \\ \boldsymbol{S}_{1,n(2,1)} & \boldsymbol{S}_{1,n(2,2)} & \cdots & \boldsymbol{S}_{1,n(2,N_r+N_d)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{1,n(5,1)} & \boldsymbol{S}_{1,n(5,2)} & \cdots & \boldsymbol{S}_{1,n(5,N_r+N_d)} \end{bmatrix} \tag{5}$$

where $S_{1,n(i,j)}$ is the element in the $i$-th row and the $j$-th column of $S_{1,n}$ as:

$$\boldsymbol{S}_{1,n(i,j)}=g\left(\boldsymbol{V}_1^{'}\cdot\boldsymbol{W}_{1,n}+b_{1,n}\right) \tag{6}$$

where $b_{1,n}$ is the bias, "g" denotes the dot product, and $g$ is a ReLu function. $V_{1(i,j)}$ is the element in the $i$-th row and the $j$-th column of $V_1$. When the filter slides to the position where $V_{1(i,j)}$ is the center point, $\boldsymbol{V}_{1(i,j)}^{'} \in \mathbb{R}^{n_{w_{c1}} \times n_{l_{c1}}}$ is formed by all the elements in the filter window as follow

$$\boldsymbol{V}_{1(i,j)}^{'}=\begin{bmatrix} \boldsymbol{V}_{1(i-1,j-2)} & \boldsymbol{V}_{1(i-1,j-1)} & \boldsymbol{V}_{1(i-1,j)} & \boldsymbol{V}_{1(i-1,j+1)} & \boldsymbol{V}_{1(i-1,j+2)} \\ \boldsymbol{V}_{1(i,j-2)} & \boldsymbol{V}_{1(i,j-1)} & \boldsymbol{V}_{1(i,j)} & \boldsymbol{V}_{1(i,j+1)} & \boldsymbol{V}_{1(i,j+2)} \\ \boldsymbol{V}_{1(i+1,j-2)} & \boldsymbol{V}_{1(i+1,j-1)} & \boldsymbol{V}_{1(i+1,j)} & \boldsymbol{V}_{1(i+1,j+1)} & \boldsymbol{V}_{1(i+1,j+2)} \end{bmatrix} \tag{7}$$

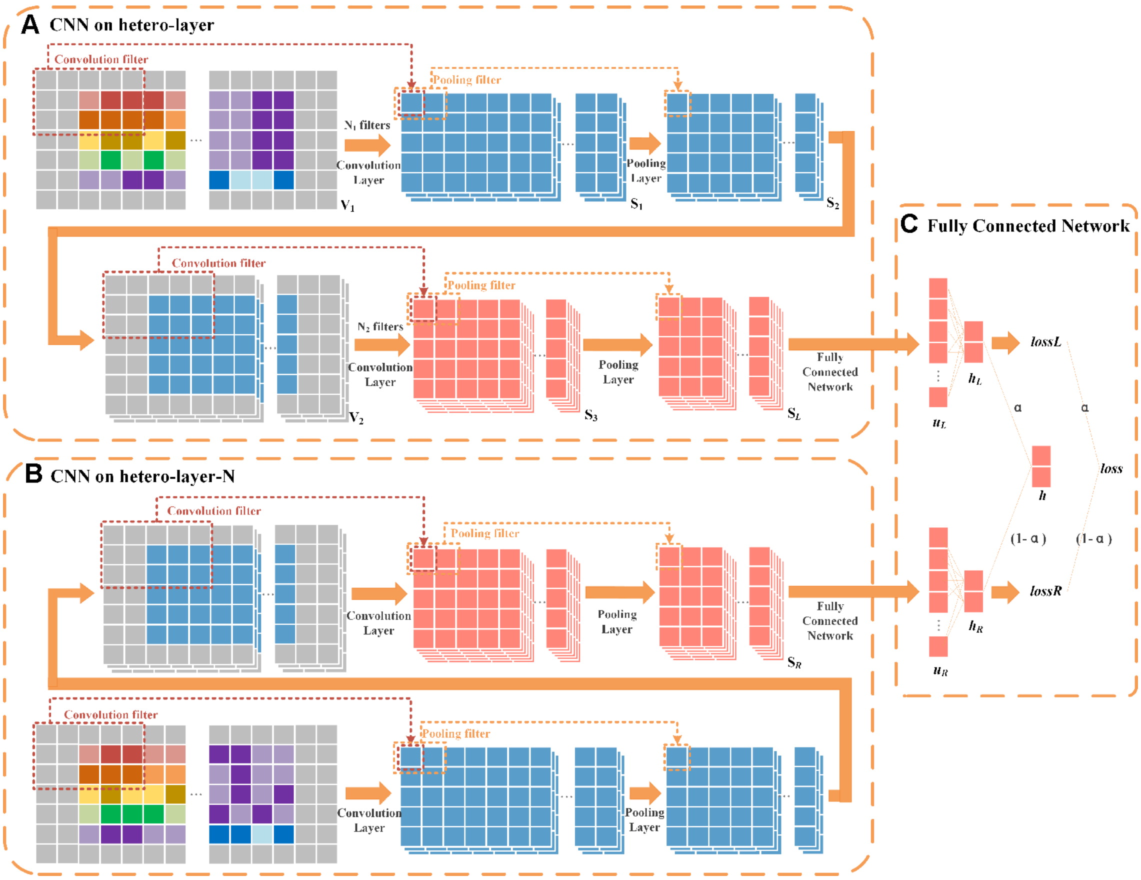


**FIGURE 4 |** Schematic diagram of HeteroDualNet. **(A)** One branch over hetero-layer of drug-disease characteristics and **(B)** one branch over the neighbouring heterogeneous layer (hetero-layer-N) are connected by **(C)** an integration module for final association score prediction. Three 3×5 filters in 1st convolution, six 3×5 filters in 2nd convolution, a sliding window of 1 × 2 in 1st and 2nd pooling are used for illustration.

We set the width and length of the sliding window in the first pooling layer as $n_{w_{p1}}$ and $n_{l_{p1}}$ ($n_{w_{p1}} = 1$ and $n_{l_{p1}} = 2$ as an example in **Figure 4**) and the step size as. The output of the first pooling $\boldsymbol{S}_2 \in \mathbb{R}^{N_1 \times 5 \times (N_r + N_d)/2}$. is obtained by a max-pooling operation where the $n$-th output $\boldsymbol{S}_{2,n} \in \mathbb{R}^{5 \times (N_r + N_d)/2}$ is

$$\boldsymbol{S}_{2,n} = \begin{bmatrix} \boldsymbol{S}_{2,n(1,1)} & \boldsymbol{S}_{2,n(1,2)} & \cdots & \boldsymbol{S}_{2,n(1,(N_r+N_d)/2)} \\ \boldsymbol{S}_{2,n(2,1)} & \boldsymbol{S}_{2,n(2,2)} & \cdots & \boldsymbol{S}_{2,n(2,(N_r+N_d)/2)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{2,n(5,1)} & \boldsymbol{S}_{2,n(5,2)} & \cdots & \boldsymbol{S}_{2,n(5,(N_r+N_d)/2)} \end{bmatrix} \tag{8}$$

where $\boldsymbol{S}_{2,n(i,j)}$ is the maximum value between $\boldsymbol{S}_{1,n(i,2j\text{-}1)}$ and $\boldsymbol{S}_{1,n(i,2j)}$ defined as

$$\boldsymbol{S}_{2,n(i,j)} = max\left(\boldsymbol{S}_{1,n(i,2j-1)}, \boldsymbol{S}_{1,n(i,2j)}\right) \tag{9}$$

By padding $S_{2,n}$ with zeros, $V_2$ is obtained as $\boldsymbol{V}_2 \in \mathbb{R}^{(5+2l) \times (N_r + N_d + 2p)}$ where $l = \left(n_{w_{c2}} - 1\right)/2$ and $p = \left(n_{l_{c2}} - 1\right)/2$. The number of filters is set as $N_2$ in the second convolution. The output of the second convolution is obtained as $\boldsymbol{S}_3 \in \mathbb{R}^{N_2 \times 5 \times (N_r + N_d)/2}$. In the second pooling layer, we set the width and length of the sliding window as $n_{w_{p2}}$ and $n_{l_{p2}}$, and the step size as $n_{w_{p2}} \times n_{l_{p2}}$. For instance, the case when $N_2 = 6, n_{w_{p2}} = 1$ and $n_{l_{p2}} = 2$ is illustrated as an example in **Figure 4**. The output of the second pooling is obtained as $\boldsymbol{S}_4 \in \mathbb{R}^{N_2 \times 5 \times (N_r + N_d)/4}$ which is also the final output. Let $\boldsymbol{S}_L$ represent the final output of this branch, $\boldsymbol{S}_L = \boldsymbol{S}_4$.

**Convolutional module on hetero-layer-N.** The settings of convolution and pooling operations on hetero-layer-N is the same as the above branch. Let $S_R$ denote the final output given $\boldsymbol{X}_{L-N}$ as inputs, $\boldsymbol{S}_R \in \mathbb{R}^{N_2 \times 5 \times (N_r + N_d)/4}$.

**Final integration module**. The integration of two branches is obtained by firstly flattening $\boldsymbol{S}_L$ and $\boldsymbol{S}_R$ as vectors $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathbb{R}^{1 \times (N_2 \times 5 \times (N_r + N_d)/4)}$. $u_L$ and $u_R$ are then fed into a fully connected layer (as shown in **Figure 4C**).

The association score $\boldsymbol{h}_L \in \mathbb{R}^{2 \times 1}$ between drug $r_i$ $r_i$ and disease $d_k$ in one branch is obtained as

$$\boldsymbol{h}_L = softmax\left(\boldsymbol{W}_L \boldsymbol{u}_L^{\mathrm{T}} + \boldsymbol{b}_L\right) \tag{10}$$

where $\boldsymbol{W}_L \in \mathbb{R}^{2 \times (5 \times (N_r + N_d)/4 \times n_2)}$ is the weight parameter matrix, and $b_L$ is a bias vector. $h_L(1)$ contains the probability that $r_i$ is associated with $d_k$ and $h_L(2)$ is the probability that $r_i$ and $d_k$ are not associated. Similarly, the association score $h_R$ of the other branch is calculated by

$$\boldsymbol{h}_R = softmax\left(\boldsymbol{W}_R \boldsymbol{u}_R^{\mathrm{T}} + \boldsymbol{b}_R\right) \tag{11}$$

The final association score $\boldsymbol{h}$ is obtained by a weighted fusion of $\boldsymbol{h}_L$ and $\boldsymbol{h}_R$ as

$$\boldsymbol{h} = \alpha \boldsymbol{h}_L + \left(1-\alpha\right)\boldsymbol{h}_R,\ s.t.\ 0 \le \alpha \le 1 \tag{12}$$

where $\alpha$ is a regulation parameter to balance the contributions of two branches. Let *lossL* and *lossR* denote the losses of two branches as:

$$lossL = \min\left\|\boldsymbol{h}_L - \boldsymbol{y}\right\|_F^2,\ lossR = \min\left\|\boldsymbol{h}_R - \boldsymbol{y}\right\|_F^2 \tag{13}$$

where $\boldsymbol{y} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$ is the probability that drug $r_i$ and disease $d_k$ are associated. If $r_i$ and $d_k$ are associated, $y_0$=0 and $y_1$=1, otherwise $y_0$=1 and $y_1$=0. The final loss *loss* is obtained by

$$loss = min\ \alpha\left\|\boldsymbol{h}_L - \boldsymbol{y}\right\|_F^2 + \left(1-\alpha\right)\left\|\boldsymbol{h}_R - \boldsymbol{y}\right\|_F^2 \tag{14}$$

where the regulation parameter $\alpha$ is the same as that in Equation 12. With the network architecture and loss function, the parameters are randomly initialized and adjusted in the training process until the loss function is minimized. Given three types of drug-disease matrices, the final drug-disease association score can be predicted by the trained HeteroDualNet model.

In order to reduce the impact of overfitting which is caused by the number of parameters in the proposed model based on dual CNN, we adopt the widely used dropout strategy to prevent the overfitting of HeteroDualNet. During each iteration process for training the model, HeteroDualNet randomly ignores some neurons to ensure that the trained model will have a good generalization ability.

## EXPERIMENTAL EVALUATIONS AND DISCUSSIONS

### Experimental Setup

The drug-disease samples with known associations are regarded as one class ($L_1$), while those pairs with unknown associations are considered as the other class ($L_2$). In total, there are 3051 $L_1$ samples, and 763*681-3051 = 516552 $L_2$ samples. Because $L_1$ and $L_2$ samples are largely imbalanced, undersampling strategy is used to address this issue. We divided the data into two subsets. One subset A is composed of 3051 $L_1$ samples and 3051 $L_2$ samples, while the second subset B contains the remaining 516552 – 3051 $L_2$ samples.

Five-fold cross-validation is performed to evaluate the prediction performance of HeteroDualNet and other compared models. The same training and testing data are used for the training and testing of the models. In each round of validation, the samples in subset A are equally divided into five parts where four parts are used as the training dataset, and one part together with subset B are used for testing.

As the calculation of the *4*-th intra-drug similarities matrix $\boldsymbol{R}_4$ involves drug-disease association matrix $\boldsymbol{A}$ and intra-disease matrix $\boldsymbol{D}$ to ensure that there is no testing data information in the training dataset, $\boldsymbol{R}_4$ is recalculated by removing drug-disease samples that appear in training in each round of validation.

### Comparison Methods and Evaluation Metrics

To evaluated the contributions of the proposed HeteroDualNet architecture and heterogenous drug-disease similarity representations, our model is compared with other four prediction methods including TL_HGBI (Wang et al., 2014b), MBiRW (Luo et al., 2016), LRSSL (Liang et al., 2017), and SCMFDD (Zhang et al., 2018). LRSSL is based on three drug features without considering neighbouring information and our proposed fourth intra-drug similarity from drug-related disease correlations. MBiRW used only one type of drug feature. SCMFDD and TL_HGBI used matrix decomposition and heterogeneous networks, but they didn't consider neighbouring information and multiple features.

The prediction performance is comprehensively evaluated by true positive rate (*TPR*), false positive rate (*FPR*), the Receiver Operating Characteristic (ROC) area under curve (ROC AUC), the Precision-Recall area under curve (PR AUC) and recall rate under different top *k* values. TPR and FPR are calculated as

$$TPR = \frac{TP}{TP+FN},\ FPR = \frac{FP}{TN+FP}, \tag{15}$$

where *TP* (*FN)* is the number of positive samples that are correctly identified (misidentified), *TN* (*FP)* is the number of correctly identified (misidentified) negative samples. A sample is regarded as a positive sample when its predicted association score is greater than a threshold $\theta$. If the testing sample's score is smaller than $\theta$, it is identified as a negative sample. The values of *FPR* and *TPR* are calculated by setting different values of $\theta$. The average ROC AUC value of all the evaluated drugs is used as the overall prediction performance of a method.

Since two classes are heavily imbalanced, the evaluation by PR AUC is more appropriate than ROC AUC in our study. Thus, PR AUC is also compared among different methods. *Precision* and *Recall* are defined by

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN} \tag{16}$$

where *Precision* represents the ratio between the number of correctly identified positive samples and all samples which are predicted to be positive samples, and *Recall* represents the ratio of the correctly identified positive samples to all the positive samples. Meanwhile, because the top-ranked results are of greater interest in real practices, which are often considered by biologists for further validation, we also calculate the recall rate in top *k* ranked results. The higher the recall rate for the top *k* disease, the more drug-related diseases can be predicted by the model.

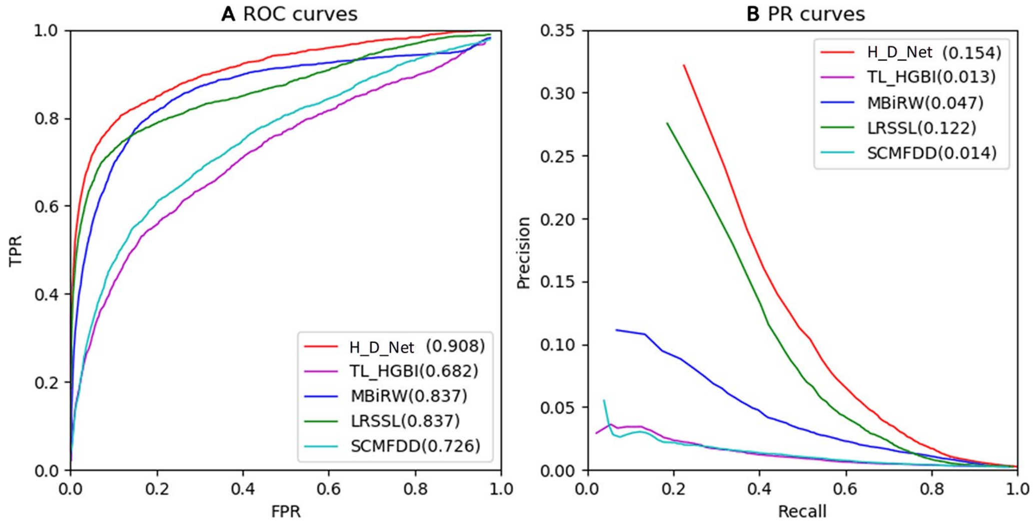


**FIGURE 5 |** Comparison between the proposed HeteroDudalNet model (H_D_Net) against four other methods by Receiver Operating Characteristic (ROC) **(A)** and Precision-Recall (PR) **(B)** curves.

## Experimental Results and Discussion

The ROC and PR of all the methods using all the 763 drugs are shown in **Figure 5**. The AUC results are given in **Table 1**. As shown by **Figure 5A** and **Table 1**, our model achieved the highest AUC of 0.908 among all the methods in comparison, which is 7.1% greater than the second best MBiRW model, 18.2% higher than the SCMFDD method, and 22.6% higher than the TL_HGBI method. As shown by **Figure 5B** and **Table 1**, HeteroDudalNet achieved the best performance where PR AUC reached 0.154, which was 3.2%, 10.7%, 14%, and 14.1% better than the that of LRSSL, MBiRW, SCMFDD and TL_HGBI models respectively.

As shown by the ROC and PR evaluation results, HeteroDudalNet outperformed the second best LRSSL because of the integration of neighbouring information on drugs and diseases and the intra-drug similarity calculated by correlations of drug-related diseases. Compared with LRSSL which considered three types of drug features, the third best model MBiRW considered only one type of drug feature in an adopted a random walk-based model, which resulted in a much lower prediction score. Without considering neighbouring associations and multiple features, SCMFDD and TL_HGBI methods failed to achieve satisfactory prediction performance although they used matrix decomposition and heterogeneous networks.

The average performance over all the 763 drugs in terms of recall rate given different top *k* values is shown in **Figure 6**. The higher the recall rate for the top *k* diseases, the more drug-related diseases can be predicted by a computing model. When increasing the value of *k* from 30 to 240 with a step of 30, the average recall rate of our method is the best among all the models in comparison. When examining the top 30, 60 and 90 diseases, our model achieved recall rates of 69.2%, 77%, and 83.5%, and the second best was obtained by LRSSL with recall rates of 63.4%, 71.3%, and 77.7% respectively. The third-ranked model MBiRW performed slightly worse than LRSSL where the results were 52.9%, 66% and 74.2%. When *k* was increased from 90 to 240, MBiRW started to perform better than LRSSL and achieved its highest recall rate of 88.7% when k was 240, while our model obtained the best rate of 90.9% among all the methods. Overall, the top *k* recall rates of SCMFDD and TL_HGBI were significantly lower than the other techniques in comparison.

As shown by the top *k* recall rate test, our model achieved the best performance, which could be useful for biologists to conduct clinical experiments because the highest ranked list contains more real drug-disease associations. As shown by the results when *k* was smaller than 90, our model and LRSSL outperformed the other methods because of the consideration of multiple drug features. The comprehensive representation of drugs concerning similarities in various perspectives contributes to digging out

**TABLE 1 |** Receiver Operating Characteristic area under curve (ROC AUC) and Precision-Recall area under curve (PR AUC) of all the methods in comparison.

| | Average performance on 763 drugs | | | | |
|---|---|---|---|---|---|
| | **HeteroDualNet** | **TL_HGBI** | **MBiRW** | **LRSSL** | **SCMFDD** |
| ROC AUC | 0.908 | 0.723 | 0.855 | 0.845 | 0.611 |
| PR AUC | 0.154 | 0.031 | 0.045 | 0.089 | 0.006 |

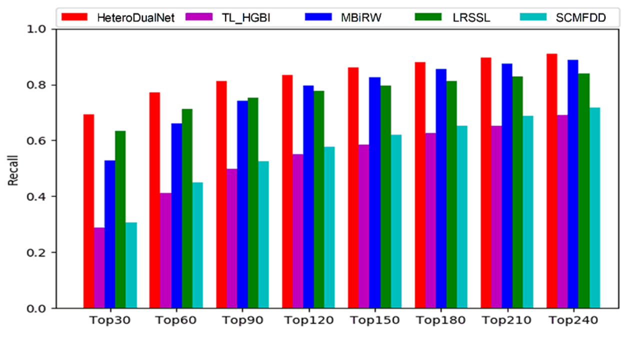


**FIGURE 6 |** The recalls across all the tested drugs at different top *k* cutoffs.

**TABLE 2 |** Top 10 related candidate diseases of ciprofloxacin, ceftriaxone, ofloxacin, ampicillin and cefotaxime.

| Drug name | Rank | Disease name | Description | Rank | Disease | Description |
|---|---|---|---|---|---|---|
| ciprofloxacn | 1 | Pneumonia, Bacterial | CTD | 6 | Gram-Positive Bacterial Infections | CTD |
| | 2 | Salmonella Infections | CTD | 7 | Eye Infections, Bacterial | Literature (Marino et al., 2013) |
| | 3 | Bacterial Infections | CTD | 8 | Soft Tissue Infections | CTD |
| | 4 | Streptococcal Infections | DrugBank | 9 | Enterobacteriaceae Infections | CTD |
| | 5 | Gram-Negative Bacterial Infections | CTD | 10 | Helicobacter Infections | CTD |
| ceftriaxone | 1 | Gram-Negative Bacterial Infections | CTD | 6 | Haemophilus Infections | CTD |
| | 2 | Bacterial Infections | CTD, ClinicalTrials | 7 | Gram-Positive Bacterial Infections | CTD |
| | 3 | Septicemia | DrugBank | 8 | Skin Diseases, Infectious | DrugBank |
| | 4 | Respiratory Tract Infections | CTD | 9 | Wound Infection | ClinicalTrials |
| | 5 | Pseudomonas Infections | DrugBank | 10 | Eye Infections, Bacterial | DrugBank |
| ofloxacin | 1 | Eye Infections, Bacterial | ClinicalTrials, DrugBank | 6 | Pseudomonas Infections | CTD |
| | 2 | Gram-Negative Bacterial Infections | DrugBank | 7 | Bacterial Infections | CTD |
| | 3 | Sinusitis | CTD | 8 | Bacteroides Infections | DrugBank |
| | 4 | Streptococcal Infections | CTD | 9 | Gram-Positive Bacterial Infections | CTD |
| | 5 | Pneumonia, Bacterial | CTD | 10 | Enterobacteriaceae Infections | DrugBank |
| ampicillin | 1 | Pseudomonas Infections | unconfirmed | 6 | Proteus Infections | CTD |
| | 2 | Bacterial Infections | CTD | 7 | Septicemia | DrugBank |
| | 3 | Gram-Positive Bacterial Infections | CTD | 8 | Streptococcal Infections | CTD |
| | 4 | Gram-Negative Bacterial Infections | CTD | 9 | Wound Infection | CTD |
| | 5 | Pneumonia, Bacterial | CTD, ClinicalTrials | 10 | Enterobacteriaceae Infections | DrugBank |
| cefotaxime | 1 | Respiratory Tract Infections | CTD, ClinicalTrials | 6 | Enterobacteriaceae Infections | DrugBank |
| | 2 | Pseudomonas Infections | DrugBank | 7 | Gram-Positive Bacterial Infections | CTD, DrugBank |
| | 3 | Gram-Negative Bacterial Infections | CTD, DrugBank | 8 | Wound Infection | DrugBank |
| | 4 | Septicemia | DrugBank | 9 | Skin Diseases, Infectious | ClinicalTrials |
| | 5 | Bacterial Infections | CTD, ClinicalTrials | 10 | Osteomyelitis | CTD, ClinicalTrials |

*(1) CTD refers to the Comparative Toxicogenomics Database (CTD), which contains a manually managed drug-disease association. (2) DrugBank refers to the drug-disease association held in the DrugBank database, which collects experimental information of the drug. (3) ClinicalTrials means that the association of drugs with the disease is recorded in the online database ClinicalTrials.gov. (4) literature refers to the literature supporting the association of drugs with the disease. (5) unconfirmed means that there is no evidence that the drug is associated with the disease.*

the potential associations between drugs and diseases. When $k > 90$, the number of common features between drug and disease may be decreasing when compared with smaller $k$ values. Thus, considering only multiple features could not always guarantee a good prediction result. MBiRW performed better than LRSSL due to the consideration of global information in a random walk based network. By incorporating three drug characteristics, the calculated correlations between drug-related diseases as intra-drug similarities, and neighbouring information of similar drugs and diseases, our model achieved better results than LRSSL and MBiRW.

## Case Studies of Five Drugs

To further evaluate and demonstrate the effectiveness of the proposed HeteroDualNet in finding reliable disease candidates of drugs, we conducted case studies of five drugs, including ciprofloxacin, ceftriaxone, ofloxacin, ampicillin and cefotaxime. Two public drug disease databases, Comparative Toxicogenomics Database (CTD) and DrugBank, were used to verify and confirm the predicted drug-disease associations by the proposed model. CTD is funded by the National Institute of Environmental Health Sciences which contains information of drugs and drugs' effects on diseases extracted from published literature. DrugBank is supported by the Health Research Institute of Canada, the Alberta Innovation-Health Solutions and Metabolic Innovation Center. Drugs' clinical trial information can be found in DrugBank, which includes drugs and diseases in experiments.

For each of the five drugs, we ranked the predicted diseases according to the relevance scores in descending order. The top 10 ranked diseases are used for verification and listed in **Table 2**. Among all the 50 diseases, 31 disease-drug association information can be found in CTD, and 17 association information can be found in the DrugBank as shown in **Table 2**. The results demonstrated that the predicted candidate diseases are indeed associated with the corresponding drugs. Also, in the CTD database, the association between Ciprofloxacin and Eye Infections, Bacterial can be found in the literature. For the two diseases which cannot be found in CTD and DrugBank, one of them can be verified by ClinicalTrials.gov (https://clinicaltrials.gov/) which records a wealth of clinical research information on various drugs and related diseases by National Institutes of Health (NIH) and the Food and Drug Administration (FDA). Therefore, there is only one disease candidate of drug ampicillin, which is Pseudomonas Infections, cannot be proved by the three databases and is labelled as unconfirmed in **Table 2**. The case studies demonstrated that our model can be used as an effective tool to predict the relations between drugs and diseases. At the same time, it has the capacity to provide computer-aided guidance for biologists in clinical trials.

The future direction for developing userful and powerful computerized prediction methods include establishing web-servers to enable public assessibility (Cheng et al., 2017; Cheng et al., 2018; Xiao et al., 2019; Chou, 2019a; Chou, 2019b). Our future work include providing a web-server for the proposed model to increase the impact of computational model in bioinformatics, medical science and medicinal chemistry.

## CONCLUSION

We present a novel HeteroDualNet model for drug-disease association prediction. Our model incorporates three kinds of drug features, a newly introduced intra-drug similarity based on correlations of drug-related diseases, and neighbouring information of drugs and diseases by constructing embedded drug-disease heterogenous matrices and dual branches in a deep neural network. The evaluation of public dataset and comparison with other four published models demonstrated the improved prediction performance in terms of ROC AUC, PR AUC, and recall rate at top *k*. Case studies of five drugs further proved the effectiveness of our model in finding potential relevant diseases of drugs as validated by database records or literature. Our model can be used as an effective tool to predict the associations between drugs and diseases and provide computer-aided guidance for biologists in clinical trials.

## DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found at https://github.com/LiangXujun/LRSSL.

## AUTHOR CONTRIBUTIONS

PX, HC, and TS conceived the prediction method, and they wrote the paper. NS and TS developed computer programs. PX, TZ, and TS analyzed the results, and PX, HC, and NS revised the paper.

## FUNDING

The work was supported by the Natural Science Foundation of China (61972135), the Natural Science Foundation of Heilongjiang Province (LH2019F049, LH2019A029), the China Postdoctoral Science Foundation (2019M650069), the Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHL-Q18104), the Fundamental Research Foundation of Universities in Heilongjiang Province for Technology Innovation (KJCX201805), and the Fundamental Research Foundation of Universities in Heilongjiang Province for Youth Innovation Team (RCYJTD201805).

## REFERENCES

Adams, C. P., and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff. (Millwood)* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Awais, M., Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., and Chou, K. C. (2019). iPhosH-PseAAC: identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 1–19. doi: 10.1109/TCBB.2019.2919025

Chen, H., Zhang, H., Zhang, Z., Cao, Y., and Tang, W. (2015). Network-based inference methods for drug repositioning. *Comput. Math. Methods Med.* 2015, 130620–130620. doi: 10.1155/2015/130620

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PloS Comput. Biol.* 12, e1004975. doi: 10.1371/journal.pcbi.1004975

Cheng, X., Lin, W. Z., Xiao, X., and Chou, K. C. (2018). pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics* 35 (3), 398–406. doi: 10.1093/bioinformatics/bty628

Cheng, X., Xiao, X., and Chou, K. C. (2017). pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. *Bioinformatics* 34 (9), 1448–1456. doi: 10.1093/bioinformatics/btx711

Chiang, A. P., and Butte, A. J. (2009). Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* 86, 507–510. doi: 10.1038/clpt.2009.103

Chou, K. C., and Shen, H. B. (2009). Recent advances in developing web-servers for predicting protein attributes. *Natural Sci.* 1 (02), 63. doi: 10.4236/ns.2009.12011

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273 (1), 236–247. doi: 10.1016/j.jtbi.2010.12.024

Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* 11 (3), 218–234. doi: 10.2174/1573406411666141229162834

Chou, K. C. (2017). An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Topics Med. Chem.* 17 (21), 2337–2358. doi: 10.2174/1568026617666170414145508

Chou, K. C. (2019a). Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.*, 1–16. doi: 10.1007/s10989-019-09893-5

Chou, K. C. (2019b). Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* 26, 4918–4943. doi: 10.2174/0929867326666190507082559

Dakshanamurthy, S., Issa, N. T., Assefnia, S., Seshasayee, A., Peters, O. J., Madhavan, S., et al. (2012). Predicting new indications for approved drugs using a proteochemometric method. *J. Med. Chem.* 55, 6832–6848. doi: 10.1021/jm300576q

Ehsan, A., Mahmood, M. K., Khan, Y. D., Barukab, O. M., Khan, S. A., and Chou, K. C. (2019). iHyd-PseAAC (EPSV): identifying hydroxylation sites in proteins by extracting enhanced position and sequence variant feature via chou's 5-step rule and general pseudo amino acid composition. *Curr. Genomics* 20 (2), 124–133. doi: 10.2174/1389202920666190325162307

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496–496. doi: 10.1038/msb.2011.26

Grabowski, H. (2004). Are the economics of pharmaceutical research and development changing?: productivity, patents and political pressures. *Pharmacoeconomics* 22, 15–24. doi: 10.2165/00019053-200422002-00003

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nat. Biotechnol.* 32, 40. doi: 10.1038/nbt.2786

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* 93, 335–341. doi: 10.1038/clpt.2013.1

Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., and Chou, K. C. (2019). SPalmitoylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Analytical Biochem.* 568, 14–23. doi: 10.1016/j.ab.2018.12.019

Iwata, H., Sawada, R., Mizutani, S., and Yamanishi, Y. (2015). Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J. Chem. Inf. Model* 55, 446–459. doi: 10.1021/ci500670q

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Brief Bioinform.* 17, 2–12. doi: 10.1093/bib/bbv020

Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., et al. (2017). LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. *Bioinformatics* 33, 1187–1196. doi: 10.1093/bioinformatics/btw770

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 32, 2664–2671. doi: 10.1093/bioinformatics/btw228

Marino, A., Santoro, G., Spataro, F., Lauriano, E. R., Pergolizzi, S., Cimino, F., et al. (2013). Resveratrol role in Staphylococcus aureus -induced corneal inflammation. *Pathog. Dis.* 68, 61–64. doi: 10.1111/2049-632X.12046

Martinez, V., Navarro, C., Cano, C., Fajardo, W., and Blanco, A. (2015). DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.* 63, 41–49. doi: 10.1016/j.artmed.2014.11.003

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221. doi: 10.1093/nar/gku1243

Nosengo, N. (2016). Can you teach old drugs new tricks? *Nature* 534, 314–316. doi: 10.1038/534314a

Padhy, B. M., and Gupta, Y. K. (2011). Drug repositioning: re-investigating existing drugs for new therapeutic indications. *J. Postgrad. Med.* 57, 153–160. doi: 10.4103/0022-3859.81870

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* 9, 203–214. doi: 10.1038/nrd3078

Pritchard, J.-L. E., O'mara, T. A., and Glubb, D. M. (2017). Enhancing the Promise of Drug Repositioning through Genetics. *Front. Pharmacol.* 8, 896–896. doi: 10.3389/fphar.2017.00896

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2018). Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discovery* 18, 41. doi: 10.1038/nrd.2018.168

Shim, J. S., and Liu, J. O. (2014). Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int. J. Biol. Sci.* 10, 654–663. doi: 10.7150/ijbs.9224

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77. doi: 10.1126/scitranslmed.3001318

Tamimi, N. A., and Ellis, P. (2009). Drug development: from concept to marketing! *Nephron Clin. Pract.* 113, c125–c131. doi: 10.1159/000232592

Uniprot, C. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148. doi: 10.1093/nar/gkp846

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, F., Zhang, P., Cao, N., Hu, J., and Sorrentino, R. (2014a). Exploring the associations between drug side-effects and therapeutic indications. *J. BioMed. Inform* 51, 15–23. doi: 10.1016/j.jbi.2014.03.014

Wang, W., Yang, S., Zhang, X., and Li, J. (2014b). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30, 2923–2930. doi: 10.1093/bioinformatics/btu403

Wang, Y., Chen, S., Deng, N., and Wang, Y. (2013). Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PloS One* 8, e78518. doi: 10.1371/journal.pone.0078518

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633. doi: 10.1093/nar/gkp456

Wu, C., Gudivada, R. C., Aronow, B. J., and Jegga, A. G. (2013). Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol.* 7 Suppl 5, S6–S6. doi: 10.1186/1752-0509-7-S5-S6

Xiao, X., Cheng, X., Chen, G., Mao, Q., and Chou, K. C. (2019). pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics* 111 (4), 886–892. doi: 10.1016/j.ygeno.2018.05.017

Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., and Shen, T. (2019). Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*. 35, 4108–4119. doi: 10.1093/bioinformatics/btz182

Yang, L., and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS One* 6, e28025. doi: 10.1371/journal.pone.0028025

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8 Suppl 2, S2. doi: 10.1186/1755-8794-8-S2-S2

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinf.* 19, 233–233. doi: 10.1186/s12859-018-2220-4

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 15 January 2020
doi: 10.3389/fphar.2019.01526

# DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials

*Arash Keshavarzi Arshadi*[1], *Milad Salem*[2], *Jennifer Collins*[1], *Jiann Shiun Yuan*[2*] *and Debopam Chakrabarti*[1*]

[1] Burnett School of Biomedical Sciences, University of Central Florida, Orlando, FL, United States, [2] Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, United States

Antimalarial drugs are becoming less effective due to the emergence of drug resistance. Resistance has been reported for all available malaria drugs, including artemisinin, thus creating a perpetual need for alternative drug candidates. The traditional drug discovery approach of high throughput screening (HTS) of large compound libraries for identification of new drug leads is time-consuming and resource intensive. While virtual *in silico* screening is a solution to this problem, however, the generalization of the models is not ideal. Artificial intelligence (AI), utilizing either structure-based or ligand-based approaches, has demonstrated highly accurate performances in the field of chemical property prediction. Leveraging the existing data, AI would be a suitable alternative to blind-search HTS or fingerprint-based virtual screening. The AI model would learn patterns within the data and help to search for hit compounds efficiently. In this work, we introduce DeepMalaria, a deep-learning based process capable of predicting the anti-*Plasmodium falciparum* inhibitory properties of compounds using their SMILES. A graph-based model is trained on 13,446 publicly available antiplasmodial hit compounds from GlaxoSmithKline (GSK) dataset that are currently being used to find novel drug candidates for malaria. We validated this model by predicting hit compounds from a macrocyclic compound library and already approved drugs that are used for repurposing. We have chosen macrocyclic compounds as these ligand-binding structures are underexplored in malaria drug discovery. The *in silico* pipeline for this process also consists of additional validation of an in-house independent dataset consisting mostly of natural product compounds. Transfer learning from a large dataset was leveraged to improve the performance of the deep learning model. To validate the DeepMalaria generated hits, we used a commonly used SYBR Green I fluorescence assay based phenotypic screening. DeepMalaria was able to detect all the compounds with nanomolar activity and 87.5% of the compounds with greater than 50% inhibition. Further experiments to reveal the compounds' mechanism of action have shown that not only does one of the hit compounds, DC-9237, inhibits all asexual stages of *Plasmodium falciparum*, but is a fast-acting compound which makes it a strong candidate for further optimization.

**Keywords: artificial intelligence, malaria, drug discovery, virtual screening, deep learning, inhibition, toxicity**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Jian Li,*
*Wuxi AppTec, China*
*Timothy John Egan,*
*University of Cape Town,*
*South Africa*

****Correspondence:***
*Jiann Shiun Yuan*
*Jiann-Shiun.Yuan@ucf.edu*
*Debopam Chakrabarti*
*debopam.chakrabarti@ucf.edu*

***Specialty section:***
*This article was submitted to Experimental Pharmacology and Drug Discovery, a section of the journal Frontiers in Pharmacology*

***Received:*** *13 September 2019*
***Accepted:*** *26 November 2019*
***Published:*** *15 January 2020*

***Citation:***
*Keshavarzi Arshadi A, Salem M, Collins J, Yuan JS and Chakrabarti D (2020) DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials. Front. Pharmacol. 10:1526. doi: 10.3389/fphar.2019.01526*

## INTRODUCTION

Malaria is one the deadliest disease afflicting the mankind, with more than 200 million new cases every year, and over 400,000 reported deaths (WHO, 2018). The causative agent of infection, *Plasmodium spp*. parasites have developed resistance to almost all currently marketed drugs including the current treatment choice artemisinin-based combination therapy (ACT) (Fairhurst and Dondorp, 2016). This underscores an urgent need to discover next generation antimalarials (Cowell and Winzeler, 2019). Traditionally, the discovery of new bioactive chemotypes relies on cell or target-based screening (Baniecki et al., 2007) (Swinney, 2013) of natural or synthetic compound libraries. High Throughput Screening (HTS) using either approach entails screening of large library of compounds. This process is often inefficient and not cost effective because of high failure rate at subsequent stages of drug discovery. The real question is, with all the modern technological advancements in drug discovery how can we utilize innovative technologies to find new active compounds more efficiently, thus reducing the cost?

Screening of large diverse compound libraries is likely to yield a higher hit rate. The bioactivity of a compound can also be predicted *in silico* through virtual screening (Shoichet, 2004). In this approach, models are created to predict the activity of a compound based on chemical properties of the compounds. One of the most common descriptors currently used for virtual screening is Extended Connectivity Fingerprint (ECFP) (Rogers and Hahn, 2010). The ECFP uses topological characteristics of a molecule to describe it. The most prevalent use of ECFP in Quantitative Structure-Activity Relationship (QSAR) models involves creating a fingerprint and using a neural network to perform prediction (Ramsundar et al., 2015; Gupta et al., 2016). This approach isolates feature extraction and decision making, thus not allowing the decision-making process to have an effect on the creation of fingerprints.

With the availability of large datasets, such as whole genome sequencing, transcript profiling or HTS, artificial intelligence is expected to have major impacts on various aspects of biomedical research (Jiang et al., 2017; Wainberg et al., 2018; Reddy et al., 2019; Zhavoronkov et al., 2019). Application of AI to various areas of drug discovery would include ligand-based virtual screening (VS) (Mayr et al., 2016; Chen et al., 2018), target prediction (Mayr et al., 2018), structure-based virtual screening (Wallach et al., 2015), de novo molecular design (Kadurin, 2016; Aspuru-Guzik, 2018), or metabolomics approaches (Pirhaji et al., 2016). Deep learning approaches enable end-to-end classification of data via learning feature representation and decision making simultaneously. Deep learning's automatic feature extraction has demonstrated superiority to traditional isolated feature extraction and has resulted in the popularity of these models in many fields such as image recognition, signal classification (Rajpurkar, 2017), and deep processing of natural language (Devlin, 2019).

Recently, Graph Convolutional Neural Networks (GCNN) have shown high accuracy in predicting chemical properties of compounds (Aspuru-Guzik et al., 2015). These models transform the compounds into graphs and learn higher-level abstract representations of the input solely based on the data. Graph convolutional neural networks combine ECFP's concept of creating fingerprints from substructures with deep learning's automatic feature extraction. Compared to ECFP, the GCNN's features are shorter (encoding only the relevant features), contain similarity information for different substructures, and facilitate more accurate predictions (Aspuru-Guzik et al., 2015; Kearnes et al., 2016; Liu et al., 2018).

In this work, we leverage GCNNs to accelerate the process of antimalarial drug discovery. The representative abilities of GCNNs are used to implement a virtual screening pipeline. These models take compounds as input and predict the *P. falciparum* growth inhibition and mammalian HepG2 cell cytotoxicity of the given compounds, aiding in the intelligent selection of scaffolds as input for further analysis. The hyper-parameters of the model are optimized using an external validation on an independent and imbalanced dataset. To overcome the difficulty of low training data, transfer learning is used. The model is initialized with the weights transferred from a model trained on a large unrelated dataset. The compounds are further tested using *in vitro* bioassay for validation of the model.

Another area of drug discovery which increases the probability of detecting high value scaffolds would be the selection of compound libraries. Principal component analysis of about 5 million compounds screened against the malaria parasite *Plasmodium falciparum* in last ten years suggests that not only the libraries used have low diversity but also, they mostly consist of compounds with low molecular weight (Spangenberg et al., 2013). Drug discovery efforts in last few decades using Lipinski "rule of five" compliant synthetic compound libraries are exhibiting diminishing return. Furthermore, biological targets of approved drugs are quite limited. Therefore, for our analysis, we decided to use an unexplored natural product (NP)-inspired class of molecules. NP or NP-inspired compounds have made tremendous impacts in discovery of novel drugs (Butler et al., 2014; Newman and Cragg, 2016). Among the NPs, macrocycles have successful record as efficacious compounds with more than 100 approved drugs in the market (Blanco, 2019). At least 3% of 100,000 NP secondary metabolites are macrocycles (Driggers et al., 2008). Macrocycles are scaffolds with a ring containing at least 12 atoms (Mallinson and Collins, 2012). Macrocycles also contain many desirable properties such as, less rigidity and flexibility, high binding capabilities, having affinity to anions and cations, high bio-availability, and the ability to target protein-protein interactions (PPI) (Choi and Hamilton, 2003; Dougherty et al., 2017; Ermert, 2017; Selwood, 2017).

We present here the application of GCNNs for non-targeted ligand-based virtual screening for antimalarial drug discovery. Our research described in this article creates a practical pipeline for training generalizable virtual screening models, and the use of deep learning techniques such as transfer learning and external validation to improve the model. Results of the model to discover antiplasmodial scaffolds were validated in a prospective manner via comparison to whole cell screening.

# MATERIALS AND METHODS

## Data

### Training Data

GlaxoSmithKline group tested around two million compounds for inhibition of *Plasmodium falciparum* (Pf) Dd2, a chloroquine resistant line, intraerythrocytic life cycle and identified 13,533 bioactive compounds that exhibited greater than 80% inhibition of the *in vitro* growth of the parasite at 2 μM concentrations. This published data are publicly available in the supplementary material 1 of the article (Gamo et al., 2010). DeepMalaria uses this Pf Dd2 inhibition and selectivity data for training. The molecules are classified as one if they possess Dd2 growth inhibitions of 50% and higher and zero if otherwise. The efficacy of these compounds differs in *P. falciparum* strains Dd2 and 3D7 (a *Pf* line sensitive to chloroquine), with most of the molecules in the GSK dataset possessing higher 3D7 inhibition and varying Dd2 inhibition. Therefore, the training data implicitly holds information about the developed resistance, and if the model is trained on Dd2 inhibition data, it would be able to predict compounds that are efficacious in drug resistant strains, a desirable property.

### Validation Data

The validation dataset consists of the results from previously performed HTS in University of Central Florida at the Chakrabarti Laboratory, consisting of natural-products, kinase inhibitors from commercial vendors Asinex (Winston-Salem, NC)and ChemDiv (San Diego, CA) libraries. This dataset contains 4,497 molecules and their inhibition property. Overall, the dataset possesses 112 molecules that have a Pf inhibition greater than 50%. Using this external validation dataset, the realistic capabilities of the model are evaluated in the validation process. The raw data supporting the conclusions of this manuscript will be made available by the authors, without any restriction, to any qualified researcher (**Supplementary Material 4**).

### Compound Library for Test Data

A library of 2,400 macrocyclic compounds was purchased from the commercial vendor Asinex (Winston-Salem, NC) for validation. The compounds selected for purchase were not given any consideration about DeepMalaria prediction to avoid any bias in results (**Supplementary Material 3**).

### Source Data for Transfer Learning

In order to perform transfer learning, a large dataset is chosen as the source to transfer from. One of the largest labeled molecule datasets is publicly available in the PubChem Bioassay (PCBA) repository. Within this dataset, the "PCBA-686979" assay (Wu et al., 2018; Pubchem Database, 2019) contains 303,167 molecules with 20.82% of them being active. The molecules in the mentioned library are not related to *Plasmodium*, and they were screened to find inhibitors of human tyrosyl-DNA phosphodiesterase 1 (TDP1). This enzyme is a target for cancer therapy in spite of not being necessary protein for human cells. This unrelated large and high variance collection is chosen as the source for transfer learning solely based on its size.

## *In Silico*

### Graph Convolutional Neural Network Model

In the research described here, DeepChem's implementation of GCNN was used (Ramsundar et al., 2019). This implementation offers the creation of architectures with graph convolutional layers, graph pooling layers, dropout layers, graph gather layers, and fully connected layers. The molecular graph was sorted via atom index in order to attain the same graph for canonical SMILES. The training data was first cleaned by removing the molecules with missing inhibition data. Two details needed to be considered in the conversion of molecules to graphs; firstly, the nodes represent different atoms and need to contain information of this difference. In order to differentiate between the atom nodes, DeepChem offers 75 different features for describing each atom. In this work, 29 of those features were used containing the type of atom, atom's degree, atom's implicit valence, atom's hybridization, atom's aromatic properties, and total number of Hydrogen connected to the atom. Secondly, in order to convert molecules to graph and not lose special information, chirality was added to the features.

### Data Augmentation and Hyper-Parameter Optimization

The validation dataset for this work, i.e., the "lab dataset" is highly imbalanced. Only 2% of the molecules within the dataset show inhibitory activity. These molecules are also the most important part of the dataset, since the goal of the model is to find active molecules. In order to have a fair validation on this dataset, the data needs to be balanced first. The data augmentation process created more copies of the active molecules after shuffling the atom orders. This balancing process is done via SMILES Enumeration (Bjerrum, 2017), creating on average 38 copies of each active molecule.

The augmented validation dataset can be used for finding the optimum topology, hyper-parameters, and epochs for training. Starting with the topology, the hyper-parameters that can be defined are the number of convolution layers, the size of each convolution layer, number of neurons in the dense layer, and the dropout of each layer. The remaining hyper-parameters that can be defined are the learning rate and batch size. To perform hyper-parameter optimization and find a fitting architecture, grid search was performed. Different values were chosen for each hyper-parameter, the model was trained on the training dataset and tested on the validation dataset. The set of hyper-parameters that has the best performance was chosen, and the architecture and variables of the model were finalized.

### Transfer Learning

Training a deep learning model often requires a large amount of data as the algorithms contain numerous variables that are optimized during training. DeepMalaria's training set is in the order of a few thousands, when compared to the image domain datasets it is considered to be rather small amount of data. This amount of data makes the training of the model to be sensitive to its initial weights. In order to overcome this challenge, transfer learning was used from a larger source dataset. It has been shown that the source dataset does not necessarily need to have correlation with the target dataset. The patterns within the molecules of the transfer dataset (PCBA) can help initialize the GCNN and make the training on the target dataset (GSK) to be more efficient. After the optimized architecture for model is found, the model is trained on the source dataset for 50 epochs, then the weights are saved and restored in the beginning of training on the training dataset.

## Evaluation of the Model

In order to assess the performance of the model, evaluation metrics are needed. One evaluation metric that is commonly used for classification task is accuracy. If the model can correctly classify active compounds as active (true positive or "TP") and inactive compounds as inactive (true negative or "TN"), it would have a high accuracy. If the model is missing the active molecules and is incorrectly classifying them as inactive (false negative or "FN"), or if the model is predicting inactive molecules to be active (false positive or "FP"), the accuracy would be decreased. **Table 1** shows these categories for the results of classification.

With these definitions in mind, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

In the field of drug discovery, having a high TP and a low FN is highly important, since the purpose of the model is to predict the active molecules that are few in number. One metric that can represent the ability of the model to capture active molecules is recall, as defined below:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Since the test dataset is imbalanced, accuracy would be a misleading metric. An untrained model can classify every input as inactive and still have an accuracy of nearly 97%. Furthermore, recall alone would not be enough to evaluate models in imbalanced setting because it does not contain any information of the performance of the model on the inactive molecules. To fully display model's behavior, normalized confusion matrix is used to show the percentage of data classified as each classification category. Additionally, the Area Under the Receiver operation Characteristic Curve (ROC-AUC or AUC) is used as a fair score metric.

## Cytotoxicity Prediction

In order to predict the cytotoxicity of the compounds, the model was trained with the same parameters used for inhibition. However, the dependent variable in the GSK dataset was changed to contain cytotoxicity information. Inhibition percentages over 50 are considered active against the human cell line (given the label 1) and otherwise considered non-active (given the label 0). Since only the active compounds with nanomolar potency against *Plasmodium falciparum* would go to next step of evaluation, the model was tested on the nanomolar active hits to predict their cytotoxicity and is prospectively evaluated.

**TABLE 1** | Classification categories.

| | Truly active | Truly inactive |
|---|---|---|
| Predicted as Active | TP | FP |
| Predicted as Inactive | FN | TN |

# *In Vitro*

## *Plasmodium* Growth Inhibition Assays

*P. falciparum* cultures were maintained under standard culture conditions in RPMI 1640 medium supplemented with 25 mM HEPES, pH 7.4, 26 mM $NaHCO_3$, 2% dextrose, 15 mg/L hypoxanthine, 25 mg/L gentamycin, and 0.5% Albumax II maintained at 37°C in 5% $CO_2$ and 95% air. Initially, a fixed concentration phenotypic screening was performed against multidrug resistant *P. falciparum* strain Dd2 (resistant to chloroquine, pyrimethamine, and mefloquine) using a SYBR Green I assay (Johnson et al., 2007; Vossen et al., 2010). An EVO-150 robotic liquid handler (Tecan, Morrisville, NC) was used to aliquot compounds at a final concentration of 1 µM followed by addition of culture to 96-well plates (Greiner, Monroe, NC) at 1% parasitemia, 2% hematocrit. Plates were incubated for 72 h at 37°C in a humidified atmosphere 5% $CO_2$/95% air prior to freezing. Plates were subsequently thawed and 1X SYBR Green I was added with lysis buffer (20 mM Tris-HCl, 0.08% saponin, 5 mM EDTA, 0.8% Triton X-100). After incubation at room temperature in dark for 1 h, the fluorescent emission was measured at excition and emission wavelengths of 485 nm and 530 nm, respectively, using a BioTek Synergy neo2 (Winooski, VT) plate reader. Preliminary hits exhibiting greater than 50% inhibition were then screened to determine $EC_{50}$ values. For $EC_{50}$, determination compounds were serially diluted in growth medium starting at 5 µM Untreated cultures and the ones treated with chloroquine at 1 µM rved as controls. Curve fitting was performed using GraphPad Prism and $EC_{50}$ value was determined for each compound.

## Cytotoxicity Determination

Selectivity was determined by counter-screening against human hepatoma cell line HepG2 in a MTS (3-(4,5-Dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) based cytotoxicity assay (CellTiter 96® Aqueous One, Promega, Madison, WI) (Riss et al., 2004). Briefly, microtiter plates were seeded with 1,500 cells per well in a 384 well plate and incubated for 24 h at 37°C in 5% $CO_2$/95% air atmosphere. The next day, compounds were added at seven different serially diluted concentrations starting at 25 µM d incubated for additional 48 h at stated conditions. MTS solution was next added to each well, incubated for additional 3 h at 37°C, and absorbance was read at 490 nm using BioTek Synergy neo 2 plate reader. Untreated cells served as control. Curve fitting was performed using GraphPad Prism and $EC_{50}$ value was determined.

## Stage-Specific Activity Assay

*P. falciparum* Dd2 cell line was synchronized using a combination of 5% sorbitol and magnetic column separation as described (Roberts et al., 2016). Cultures at 2% parasitemia and 2% hematocrit were added to microtiter plate wells and measurements began 6 h post-invasion (6 hpi). Hit compounds were added at 3X $EC_{50}$ concentration at specified time points. Controls are DHA (3X $EC_{50}$) and untreated cultures. Giemsa-stained thin smears were made for each time point and an aliquot of the culture was fixed for flow cytometric analysis using 0.04% glutaraldehyde in PBS. After fixing and aspirating, cells were permeabilized with 0.25% Triton-100 followed by treatment with RNase (0.05 mg/ml) for 3 h. Next, YOYO 1 (10.24 μM) fluorescent dye was added and samples were analyzed (Bouillon et al., 2013) using Beckman Coulter (Indianapolis, IN) CytoFlex S flowcytometer.

### Rate of Killing Determination

*Pf* Dd2 culture was synchronized as described (Roberts et al., 2016), plated into 24 well plates at 2% parasitemia and 2% hematocrit at 6 hpi. Compounds were added at 6, 18, or 30 hpi with a final concentration of 5X $EC_{50}$. Each well was exposed to the inhibitor for either 6 or 12 h using dihydroartemisinin (DHA) (25 μM) and untreated culture as controls. After washing the compounds off, the media was changed twice a day. The parasitemia was tracked for 6 days after addition of compounds. Thin smears were stained with Giemsa, and parasitemia was counted using microscope.

# RESULTS

## Overview

This work consists of two main sections: *in silico* and *in vitro*. In the *in silico* approach, DeepMalaria enables virtual screening of molecules on *Plasmodium falciparum* using a deep learning model. At the core, the GCNN model acts as a classifier, predicting the inhibition of input molecules and classifying them as "active" or "inactive." In order to optimize the hyper-parameters of the deep learning model, the model is validated externally on an independent and augmented validation dataset. The optimized model is trained on the large transfer dataset to extract useful initialization weights from it. Then, the pre-trained GCNN model is trained on the training dataset. The overview of our method is shown in **Figure 1**. This architecture enables the use of transfer learning for in silico screening, thus we coin the term "Transilico" for it. The code for this work can be accessed through www.transilico.com.

## *In Silico* Training

The results of the grid search for hyper-parameter optimization are shown in **Figure 2**. Overall, 144 different combinations of hyper-parameters were chosen for training and the trained model was tested on the validation dataset.

Trial 121 is among the hyper-parameters that yielded high average ROC-AUC scores, and it achieves the highest score between all trials. These hyper-parameters were chosen as the optimum variables and are shown in **Table 2**.

Having defined the architecture of the GCNN model, the model was trained on the transfer dataset. The weights were then saved and loaded for the main training process to start. The pre-trained model was fine-tuned on the training dataset with the batch size of 32. At each epoch the AUC score on the training set and the validation set were calculated and recorded. The results are shown in **Figure 3**.

As evident from **Figure 3**, the model starts to perform differently on the validation set from the training set after the 2nd epoch. While the score on the training set rises and model learns the training set more, the performance on the validation dataset drops. These results demonstrate over fitting happening after the 2nd epoch. Therefore, the model from this epoch is loaded as the trained model and the optimum duration of training is found.

## Phenotypic Screening Identifies Selective Compounds

Evaluation of the model was performed by phenotypic testing of compounds from a commercial macrocyclic compound library. This NP-inspired library is considered a bridge between small compounds and biomolecules (Driggers et al., 2008), thus increasing the possibility of targeting unknown biomolecules in *Plasmodium*. To rule out any validation bias in this experiment, we did not consider the *in-silico* results when buying the library, and all compounds were purchased based on their druggability as identified using traditional cluster analysis (data not shown). To compare the predictions from the DeepMalaria with the outcome of a traditional *Plasmodium* phenotypic cell-based screening, all 2,400 compounds were tested for antiplasmodial activities using SYBR green I fluorescence assay. Multi-drug resistant *Pf* Dd2 strain was used to provide clinically relevant results. Of the 2,400 compounds, 49 compounds exhibited growth inhibition of greater than 50% at 1 μM. This is a comparatively high hit rate (~2%) and provides evidence for the potential of macrocycles as a new class of antimalarial compounds. The 49 hits underwent $EC_{50}$ determination and five compounds showed activity under 1 μM (**Table 4**). DC-9235, DC-9239, and DC-9236 are analogs and considered amino acid macrocyclic scaffolds which are novel antimalarial candidates, and further hit to lead development would increase the potency of core structure. Other compounds are not analogs of each other, suggesting discovery of four unique antimalarial macrocyclic scaffolds. The selectivity of the compounds for malaria parasite was determined by counterscreening against human hepatoma HepG2 cell line using MTS proliferation assay. All 6 hits exhibited greater than 15-fold selectivity index suggesting none is consider to be cytotoxic for HepG2 cell line (**Table 4**).

## *In Silico* and *In Vitro* Results Are Comparable

After *in vitro* phenotypic screening, the ground truth labels for the test dataset are found. The model can now be evaluated both retrospectively and prospectively, via its prediction on the validation set and the test set. The results of this evaluation are shown in **Table 3**.

The model yields a high recall in both the validation and the test set, showing the ability of the model in finding active compounds. To fully display the performance of the model, the confusion matrices of the validation and test set are shown in **Figure 4**.

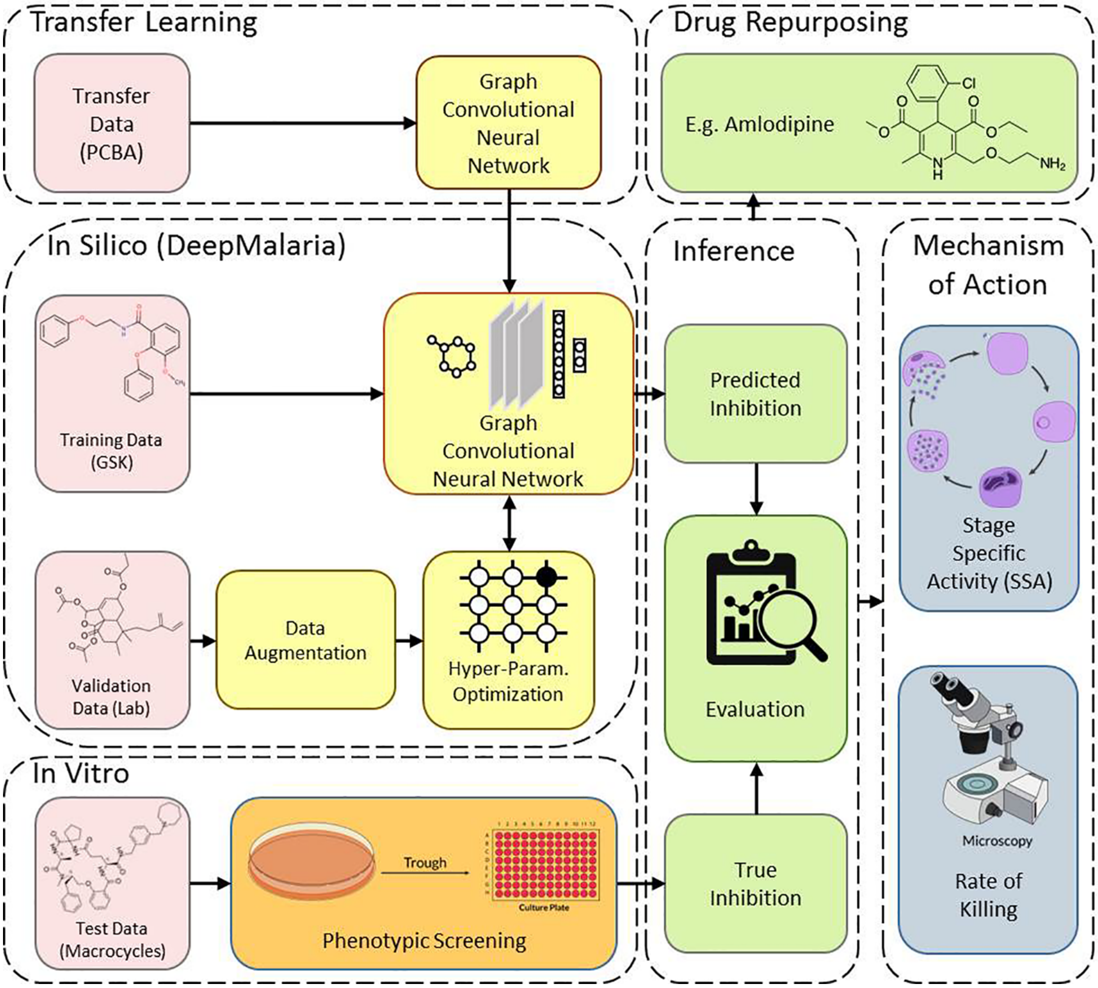


**FIGURE 1 |** Overview of Transilico architecture used to train DeepMalaria. *In silico*, the validation dataset is augmented and is used to determine the hyper-parameters of the model. The model is pre-trained on the transfer dataset and fine-tuned on the training dataset. *In vitro*, the test compounds are tested on *Pf*. The results are compared to predictions. The trained model can be applied for drug repurposing. The mechanism of action for the hits are determined.

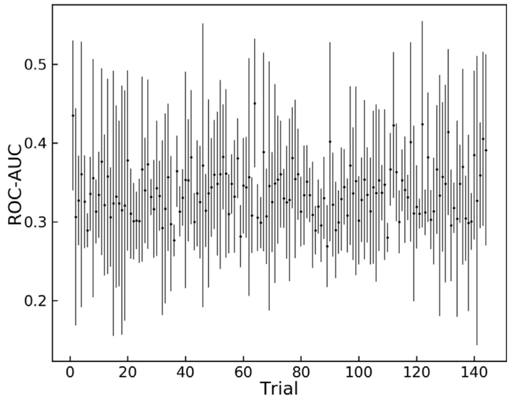


**FIGURE 2 |** Grid search results for different sets of hyper-parameters. 144 different sets of hyper parameters for the model are tested on the augmented validation dataset.

**Figure 4** shows similar behavior of the model on active molecules in the validation dataset and the test set, achieving the goal of the external validation process in DeepMalaria. Moreover, the model is inclined to predict the input as active, yielding a higher false positive rate than false negative rate. This behavior is essential in a drug discovery model, since finding the active molecules are of priority, and falsely predicting them as inactive will likely be counterproductive.

As shown in **Table 4**, the model was able to correctly predict all of the 6 compounds with nanomolar activity. Based on the results in **Tables 3** and **4**, it is evident that DeepMalaria is capable of virtually identifying potent compounds with high accuracy. From 44.13% accuracy in the whole library, to 87.75% accuracy for hits with at least 50% growth inhibition, and finally 100% accuracy for all nanomolar active compounds, DeepMalaria is prone to have less false negative when screening more potent set of molecules. Thus, it is unlikely that the model might miss highly active compounds. (**Figure 5**). Since the identification of

**TABLE 2 |** Finalized hyper-parameters from grid search.

| Hyper-Parameter | Optimum Value | Hyper-Parameter | Optimum Value |
|---|---|---|---|
| # of Conv. Layers | 3 | Dropout | 0 |
| Conv. Layer Sizes | 64, 64, 64 | Learning Rate | 0.0001 |
| # of Neurons | 256 | Batch Size | 128 |

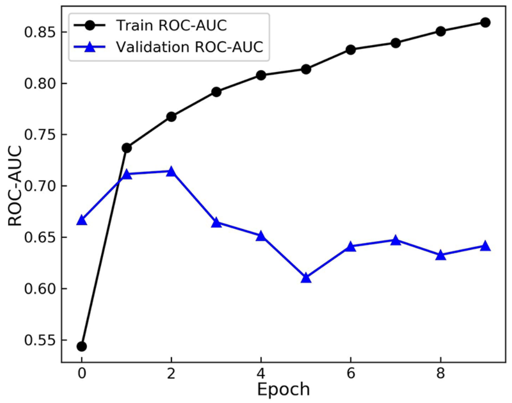


**FIGURE 3 |** Area Under the Receiver operation Characteristic Curve (AUC) scores of the model during training. The model is evaluated on the training and augmented validation dataset at the end of each epoch. The model starts to over-fit after 2 epochs.

**TABLE 3 |** Results of the trained model.

| | # of Active | TP | TN | FP | FN | Accuracy | Recall |
|---|---|---|---|---|---|---|---|
| Validation Dataset | 112 | 81 | 2620 | 1756 | 31 | 60.06 | 72.32 |
| Test Dataset | 49 | 43 | 1016 | 1335 | 6 | 44.13 | 87.75 |

highly potent hit compounds is a goal of all drug discovery programs, predicting 100% of the nanomolar active hits proves the utility of AI as a rapid and low-cost alternative to traditional methods of bioactive hits discovery.

## Comparison to Other Methods

The external validation process can also be used for traditional approaches of virtual screening. As in traditional approaches, a Random Forest (RF) and a Fully Connected Neural Network (NN) models are trained on the ECFP4 of the molecules (with size 1024) after optimized hyper-parameters were found. RF is chosen since it offers a fair amount of control over over-fitting. Both models pass through the process of external validation to be given the optimum hyper-parameters. Furthermore, in order to evaluate the impact of transfer learning, a model without transferred weights is trained. The results are shown in **Table 5**.

The RF and NN models predict most of the input molecules as active, resulting in an impractical model. The GCNN in DeepMalaria can outperform RF without transfer learning, showing the superiority of learnt features during training (in GCNN) to isolated feature extraction (ECFP). When transfer learning was used the model gains a noticeable boost in performance, correctly predicting more active and inactive molecules. This shows the effectiveness of DeepMalaria's process in early hit prediction.

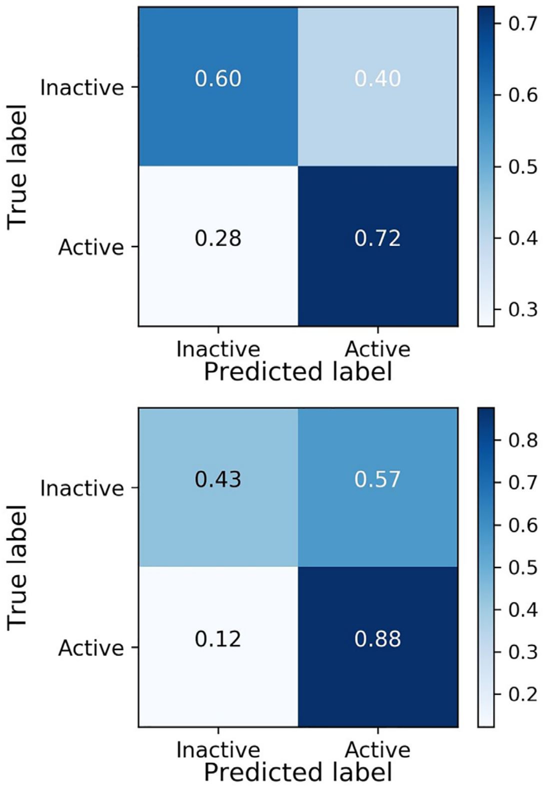


**FIGURE 4 |** Confusion matrices of validation dataset **(A)** and test dataset **(B)**.

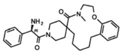
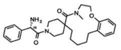
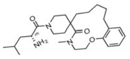
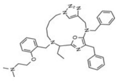
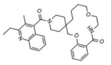
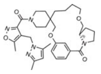
## Stage Specific Activity Determination of Active Scaffold

A disadvantage of phenotypic screening compared to target-based screening is that the biological target of the active compound is unknown. However, analysis of the development stages affected by these compounds and the rate of killing may provide insight into their mechanism of action. We hypothesized that macrocyclic compounds are likely target new macromolecules in the plasmodial life cycle because of their unusual standing as a bridge between small molecule inhibitors and larger biomolecules (Driggers et al., 2008). We explore this by assessing the stage-specific activity and the rate of killing of the four novel antiplasmodial scaffolds.

### Macrocyclic Hits Inhibit Multiple *Plasmodium* Developmental Stages

Only few of the marketed antimalarials are able to target multiple stages of the intraerythrocytic life cycle including the early ring stage (Roberts et al., 2017). Additionally, 4 out of 12 current antimalarials, including artemisinin, inhibit growth of the early ring stage (Wilson et al., 2013). To determine the stage specific activity of the hit compounds identified in this work, synchronous culture was exposed to compounds at different time points of the life cycle and the maturation of the parasite was assessed by flowcytometric and microscopic analysis. As seen in **Figure 6**, flowcytometric and Giemsa-stained microscopic data suggest that the control culture

**TABLE 4 |** *In vitro* and in silico results of the compound with Nano-molar activity.

| ID | Inhibition One-point (%) | Toxicity SI | $EC_{50}$ *μM* HepG2 | $EC_{50}$ *μM* Pf | DeepMalaria antimalarial Prediction | DeepMalaria Softmax Output | DeepMalaria Toxicity Prediction |
|---|---|---|---|---|---|---|---|
| *DC-9239*<br> | 79 | 25 | 24.2 | 1.09± 60nM | Active | 0.70 | Negative |
| *DC-9235*<br> | 92 | >40 | >25 | 0.79± 61nM | Active | 0.71 | Negative |
| *DC-9236*<br> | 98 | 15.9 | 6.5 | 0.41± 30nM | Active | 0.85 | Negative |
| *DC-9237*<br> | 95 | 35.3 | 17 | 0.49± 44nM | Active | 0.68 | Negative |
| *DC-5931*<br> | 80 | >40 | >25 | 0.52± 25nM | Active | 0.96 | Negative |
| *DC-5921*<br> | 90 | >40 | >25 | 0.9± 10nM | Active | 0.91 | Negative |

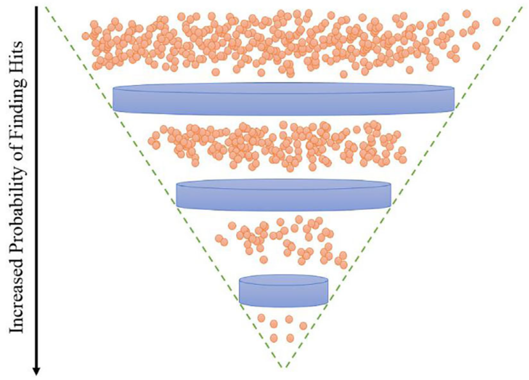


**FIGURE 5 |** DeepMalaria finds potent hits with higher recall. 87.85% for hits with inhibition of 50% or more, 100% for nanomolar active hits.

matured as expected progressing from ring (6 h) to early trophozoite (18 h) to trophozoite (30 h) to multinucleated schizont (42 h). At 54 h parasites are at the ring stage upon reinvasion with a concomitant increase in parasitemia as evident from an increase in the flow cytometric peak. In contrast, DC-9237 inhibits all asexual blood stages of Dd2 including the early ring stage. DC-9236 inhibits primarily the mature schizonts and both DC-5921 and DC-5931 inhibit the early stages (**Supplementary Material 1**). This multistage active antiplasmodial chemotype identification from a single library is further evidence of the utility of macrocyclic compounds as candidate antimalarial scaffolds since not many compounds would target all or early asexual stages of *P. falciparum*.

**TABLE 5 |** Comparison of different models on test dataset.

| | Featurization | Accuracy | Recall | ROC-AUC |
|---|---|---|---|---|
| Random Forest | ECFP4 | 14.08 | 89.79 | 0.51 |
| DeepMalaria without Transfer Learning | GraphConv | 33.46 | 77.55 | 0.55 |
| DeepMalaria | GraphConv | 44.13 | 87.75 | 0.69 |

## DC-9237 Is a Fast-Acting Compound

An attractive property of a successful antimalarial compound is rapid clearance of parasites, reducing the need for additional doses. Using synchronized Dd2 culture, we measured the rate of parasite killing at different stages of intraerythrocytic development. As evident from **Figure 7**, compared to the control, DC-9237 inhibited growth at all asexual stages after 6 h of exposure and parasite population did not resume growth even after 6 days. In contrast, the remaining compounds did not completely inhibit growth even at 12 h of exposure, suggesting a low rate of elimination. DC-9236 is showing higher elimination in the second life cycle. The types of compounds which are not effective at the first cell cycle, thus causing a delayed parasite death, are known to target apicoplast which is a vestigial chloroplast-like organelle (Kennedy et al., 2019b).

## Deepmalaria Identifies Drug Repurposing Candidates

Drug repurposing is a very important aspect of modern drug discovery that reduces the cost significantly. Compounds that have been already approved would make suitable drugs with new medical indication since they need lower developmental costs and

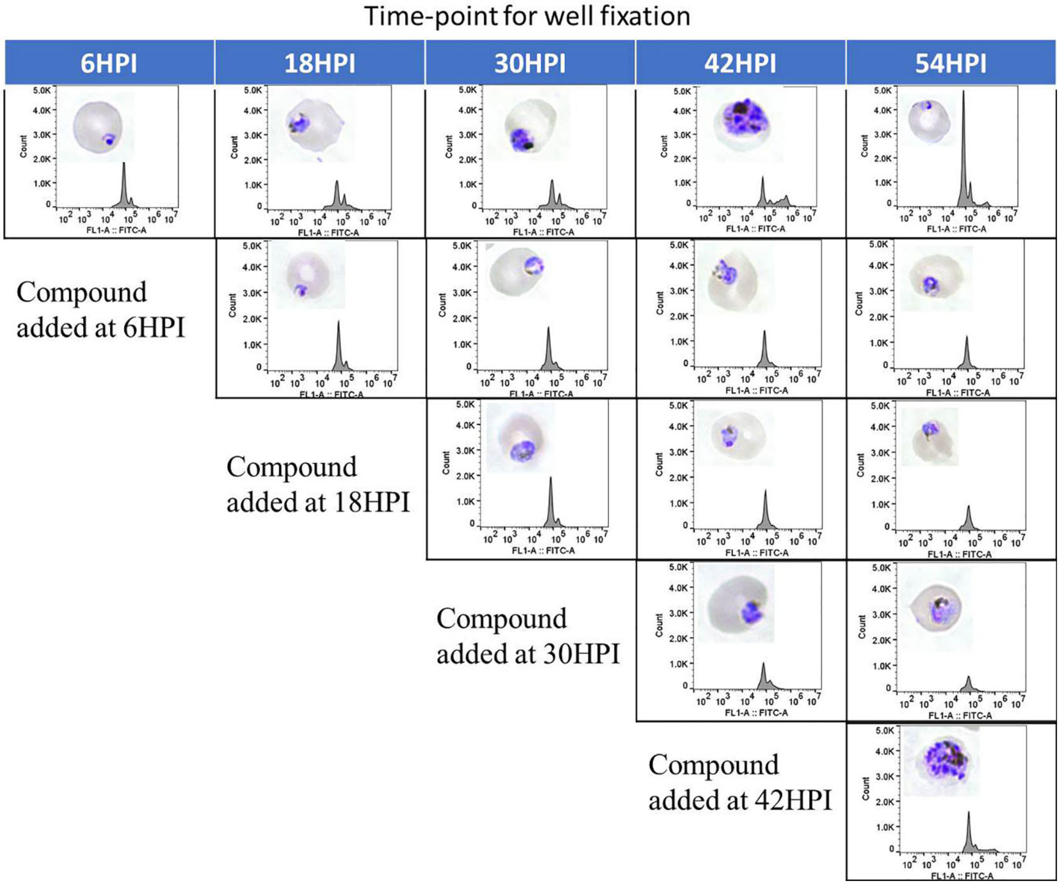


**FIGURE 6 |** Stage Specific Activity of DC-9237. At different time-points of the malaria parasite intraerythrocytic developmental cycle, DC-9237 was added at a concentration of 3x$EC_{50}$. Additions were at 6, 18, 30, and 42 h post-invasion (hpi). Samples were processed 12 h later and analyzed by Giemsa staining and flow cytometry of YOYO-1 stained samples.

faster approval processes (Pushpakom et al., 2019). It has been reported that 226 FDA approved drugs are active against different stages and cell lines of *Plasmodium falciparum* (Chong et al., 2006; Derbyshire et al., 2012). After removing all inorganic molecules and also inactive ones against blood stages of Dd2 cell line, 211 drugs were screened virtually using DeepMalaria. The model showed 74% accuracy in predicting those 211 compounds as repositioning candidates (**Supplementary Material 2**). Pazhayam et al. proposed eight of those compounds as stronger candidates because of sharing common targets between the host and parasite (Pazhayam et al., 2019). These drugs include Azithromycin, Cyclosporin A, Esomeprazole, Pentamidine, Omeprazole, Auranofin, Loperamide and Amlodipine. As expected, the model predicts all of the eight candidates as potential antimalarials. This further validates the promise of DeepMalaria as a powerful tool for drug repurposing (**Figure 8**).

## DISCUSSION

Options for malaria therapy are increasingly becoming limited because of widespread drug resistance. Even artemisinin-based combination therapies (ACTs), the front-line therapeutic choices for uncomplicated *P. falciparum* malaria, are gradually becoming ineffective in many countries of Southeast Asia (Cui, 2011; Ashley et al., 2014). Reports of failure of dihydroartemisinin-piperaquine drug combination therapy in Cambodia leaves us with very few therapeutic choices (Saunders et al., 2014). This bleak situation emphasizes the urgent need to develop new antimalarials that act on novel targets. Although recent increase in novel antimalarial discovery efforts has led to quite a few lead compounds in preclinical development (Ashley and Phyo, 2018), the need for new antimalarial drugs will continue to exist because of expected loss of new drugs due to future emergence of resistance.

In this work, a deep learning model was trained on publicly available data to predict *Plasmodium falciparum* inhibition of compounds. A validation dataset was created from previous experiments and was augmented to assist in hyper-parameter optimization. Transfer learning from a large corpus of unrelated data was employed to facilitate the training of the deep learning model. The model was tested on an independent macrocyclic test dataset in order to find new drug candidates. DeepMalaria

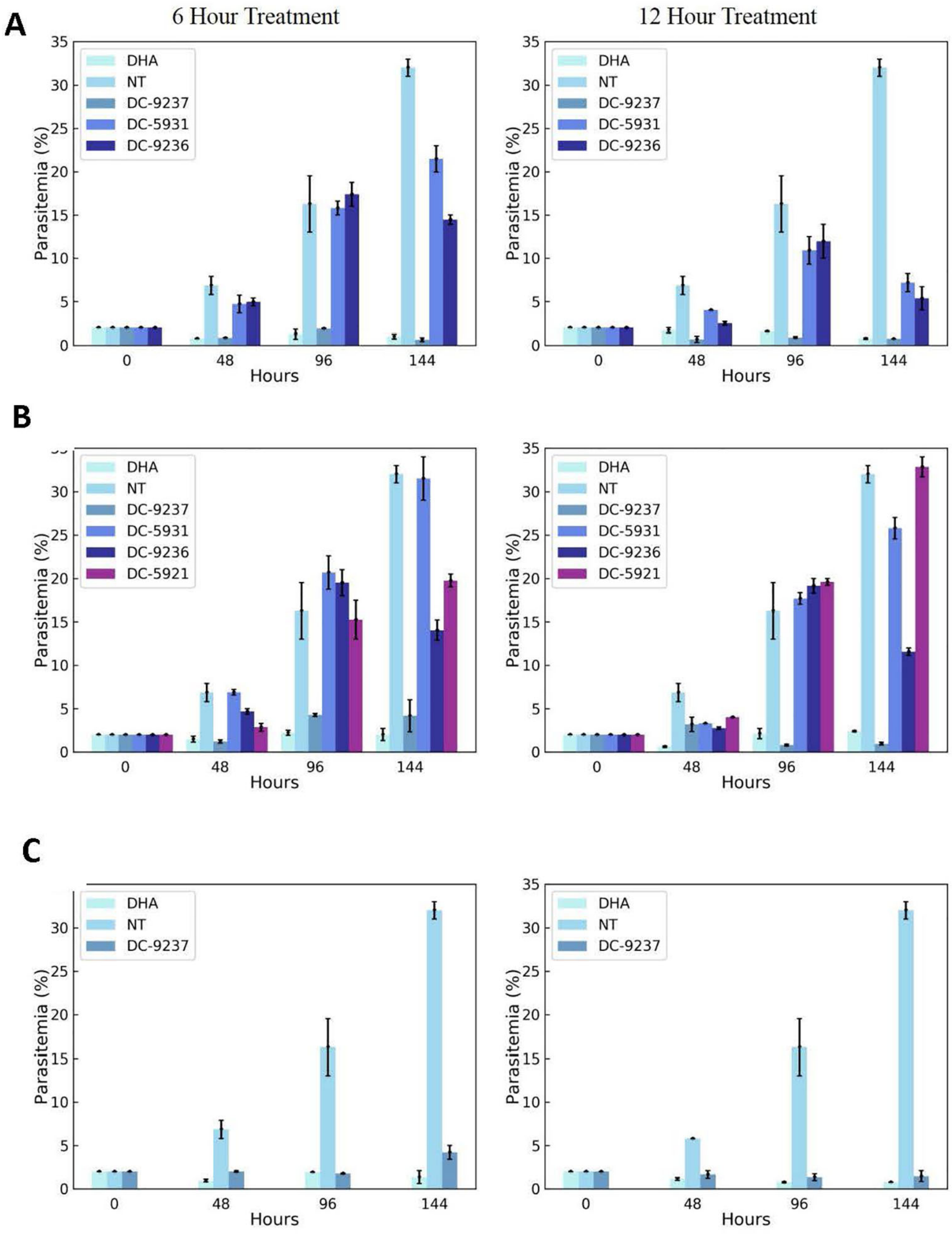


**FIGURE 7 |** Assessment of rate of killing. Synchronous cultures were subjected to 3 x $EC_{50}$ concentrations of macrocyclic compounds or dihydroartemisinin (DHA) for 6 or 12 h, followed by washing to remove the inhibitor and incubating in the growth medium in the absence of compounds to monitor recovery. **(A)** Compounds added at the schizont stage, **(B)** treatment at the trophozoite stage, and **(C)** ring stage culture exposed to the compound. Compounds were added at a stage where they exhibit block in cell cycle progression.

was able to find 72.32% of active molecules from the validation dataset and 87.75% of that of the test dataset, while maintaining an acceptable accuracy in an imbalanced setting. The results show that deep learning automatic feature extraction can learn patterns within the molecules that are generalizable to new and unseen datasets, outperforming the traditional approach of classifying fingerprints. DeepMalaria has shown increasing accuracy when predicting more potent compounds, a very important characteristic which did not let any nanomolar active/non-cytotoxic compound to be missed. Furthermore, the hit compounds were narrowed down to one fast-acting compounds working at all stages of *P. falciparum* growth. Also, DC-9236 showed inhibition in the second developmental cycle of Pf causing delayed death most likely because of its action on

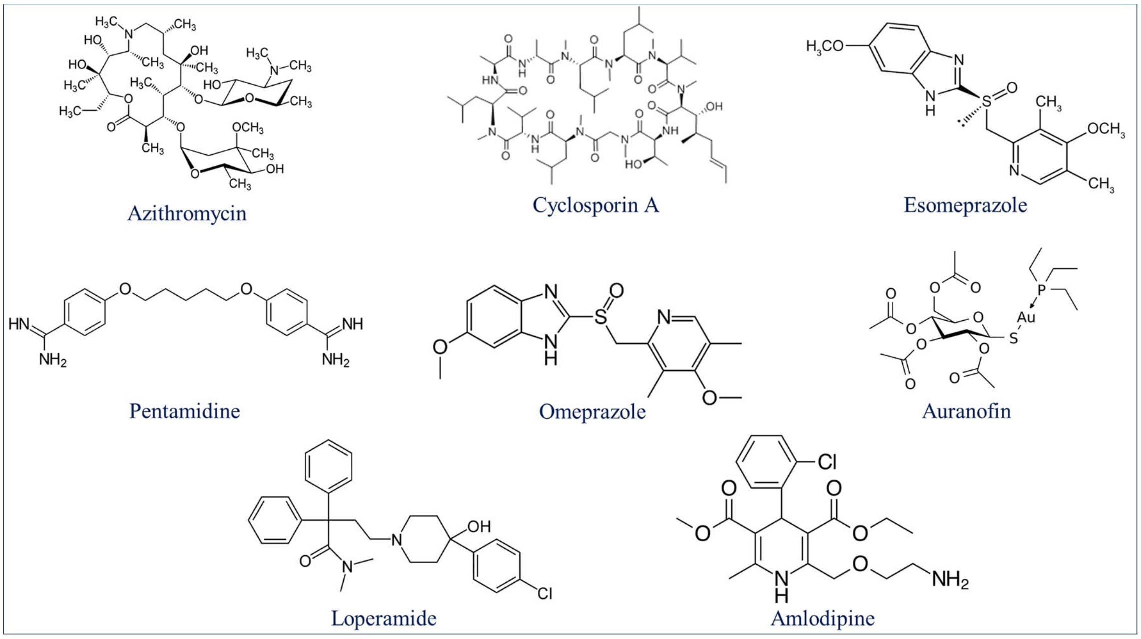


**FIGURE 8 |** Eight drugs suggested by Pazhayam et al. (2019) with probable antimalarial activity. DeepMalaria was able to predict all scaffolds as hits for *Plasmodium falciparum*.

the apicoplast. Compounds with delayed death characteristics would be a very good candidate for combination therapy (Kennedy et al., 2019a).

We demonstrated the potential of deep learning and the Transilico architecture to accelerate the process of active compound identification in early drug discovery (www.transilico.com). Since last decade AI and especially deep learning is generating new hope in small molecule early drug discovery (Leelananda and Lindert, 2016). There is an increased interest to use machine learning and related technologies to rapidly discover novel pharmacophores thus avoiding the expense of HTS (Fleming, 2018). Given the pressing need for novel antimalarials, accelerated hit identification through the use of AI as has been presented in this work would be of great interest. Artificial intelligence has revolutionized many fields of medicine including drug discovery (Wang and Shen, 2017; Doan and Carpenter, 2019; Topol, 2019). Expensive HTS, low hit rate of synthetic libraries, incompatibility of natural products with HTS, non-diverse libraries etc., are some of the reasons for limited success of many of today's drug discovery efforts (Koehn and Carter, 2005; Li and Vederas, 2009; Schneider, 2017). However, many AI-based approaches of early drug discovery such as structural based VS and de novo design of molecules are still relatively unexplored in malaria therapeutics development. After the success of abstract and superior performance of deep learning feature extraction, Generative Adversarial Networks and Variational Auto-Encoders would act as good candidates for leveraging this abstract representation for bioactive molecule identification.

To establish AI for accelerated malaria drug lead discovery, we used a commercial macrocyclic compound library for validation. Peptidic macrocycles compounds have characteristics of both small molecules and polypeptides, and have not been investigated for antimalarial therapeutics discovery. High hit rate of macrocyclic compound screening suggests their utility as antimalarials. Other classes of macrocyclic compounds such as cyclic peptides would also be a good candidate for further study (Whitty et al., 2017). It is of note that, although the model was trained on a synthetic library, DeepMalaria was highly accurate in discovering natural products hits. This article is the first report regarding the use of transfer learning for malaria drug discovery and will be a model for future projects of AI-based drug discovery. Additionally, our model would aid in drug repurposing as it showed strength in predicting potential antimalarial activities of already approved drugs.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**. The code for this work can be accessed through www.transilico.com.

## AUTHOR CONTRIBUTIONS

AA conducted all biological experiments, acquired data for Deep Malaria, optimized the *in-silico* results, and wrote the relevant parts of the manuscript. MS trained and optimized the *in silico* model and wrote the relevant parts of the manuscript. JC screened most of evaluation data. DC provided guidance for biological experiments for malaria and DeepMalaria data acquisition. JY provided guidance in the opportunities of deep learning in a multidiscipline collaboration. All authors have read the submitted manuscript.

## FUNDING

Funding for this research is from intramural funding (University of Central Florida, College of Medicine).

## ACKNOWLEDGMENTS

We would like to thank Farnam Kavehei for providing graphical figures in **Figure 1**. Also, we appreciate Dr. Taj Azarian for his critical comments on the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.01526/full#supplementary-material

## REFERENCES

Ashley, E. A., and Phyo, A. P. (2018). Drugs in Development for Malaria. *Drugs* 78, 861–879. doi: 10.1007/s40265-018-0911-9

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., et al. (2014). Spread of artemisinin resistance in Plasmodium falciparum malaria. *N. Engl. J. Med.* 371, 411–423. doi: 10.1056/NEJMoa1314981

Aspuru-Guzik, A., Duvenaud, D., Maclaurin, D., Aguilera-Iparraguire, J., Gomez-Bombarelli, R., Hirzel, T. D., et al. (2015). Convolutional networks on graphs for learning molecular fingerprints. in *Advances in Neural Information Processing Systems*. (Curran Associates, Inc) 28, 2224–2232.

Aspuru-Guzik, A. (2018). Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. arXiv:1705.10843

Baniecki, M. L., Wirth, D. F., and Clardy, J. (2007). High-throughput Plasmodium falciparum growth assay for malaria drug discovery. *Antimicrob. Agents Chemother.* 51, 716–723. doi: 10.1128/AAC.01144-06

Bjerrum, E. J. (2017). SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. arXiv:1703.07076

Blanco, M. J. (2019). Building upon nature's framework: overview of key strategies toward increasing drug-like properties of natural product cyclopeptides and macrocycles. *Methods Mol. Biol.* 2001, 203–233. doi: 10.1007/978-1-4939-9504-2_10

Bouillon, A., Gorgette, O., Mercereau-Puijalon, O., and Barale, J. C. (2013). Screening and evaluation of inhibitors of Plasmodium falciparum merozoite egress and invasion using cytometry. *Methods Mol. Biol.* 923, 523–534. doi: 10.1007/978-1-62703-026-7_36

Butler, M. S., Robertson, A. A., and Cooper, M. A. (2014). Natural product and natural product derived drugs in clinical trials. *Nat. Prod. Rep.* 31, 1612–1661. doi: 10.1039/c4np00064a

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Choi, K., and Hamilton, A. D. (2003). Macrocyclic anion receptors based on directed hydrogen bonding interactions. *Coord. Chem. Rev.* 240, 101–110.

Chong, C. R., Chen, X., Shi, L., Liu, J. O., and Sullivan, D. J.Jr. (2006). A clinical drug library screen identifies astemizole as an antimalarial agent. *Nat. Chem. Biol.* 2, 415–416.

Cowell, A. N., and Winzeler, E. A. (2019). The genomic architecture of antimalarial drug resistance. *Briefings Funct. Genomics* 18 (5), 314–328. doi: 10.1093/bfgp/elz008

Cui, W. (2011). WHO urges the phasing out of artemisinin based monotherapy for malaria to reduce resistance. *BMJ* 342, d2793. doi: 10.1136/bmj.d2793

Derbyshire, E. R., Prudencio, M., Mota, M. M., and Clardy, J. (2012). Liver-stage malaria parasites vulnerable to diverse chemical scaffolds. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8511–8516. doi: 10.1073/pnas.1118370109

Devlin, J. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1703.07076

Doan, M., and Carpenter, A. E. (2019). Leveraging machine vision in cell-based diagnostics to do more with less. *Nat. Mater.* 18, 414–418. doi: 10.1038/s41563-019-0339-y

Dougherty, P. G., Qian, Z., and Pei, D. (2017). Macrocycles as protein–protein interaction inhibitors. *Biochem. J.* 474, 1109–1125. doi: 10.1042/BCJ20160619

Driggers, E. M., Hale, S. P., Lee, J., and Terrett, N. K. (2008). The exploration of macrocycles for drug discovery — an underexploited structural class. *Nat. Rev. Drug Discovery* 7, 608–624. doi: 10.1038/nrd2590

Ermert, P. (2017). Design, properties and recent application of macrocycles in medicinal chemistry. *CHIMIA Int. J. Chem.* 71, 678–702. doi: 10.2533/chimia.2017.678

Fairhurst, R. M., and Dondorp, A. M. (2016). Artemisinin-Resistant Plasmodium falciparum Malaria. *Microbiol. Spectr.* 4 (3), 1–16. doi: 10.1128/microbiolspec.EI10-0013-2016

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature* 557, S55–S57. doi: 10.1038/d41586-018-05267-x

Gamo, F.-J., Sanz, L. M., Vidal, J., De Cozar, C., Alvarez, E., Lavandera, J.-L., et al. (2010). Thousands of chemical starting points for antimalarial lead identification. *Nature* 465, 305–310. doi: 10.1038/nature09107

Gupta, M. K., Gupta, S., and Rawal, R. K. (2016). Impact of artificial neural networks in QSAR and computational modeling. *Artificial Neural Network for Drug Design, Delivery and Disposition*. (Academic Press) 8, 153–179. doi: 10.1016/B978-0-12-801559-9.00008-9

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* 2, 230–243. doi: 10.1136/svn-2017-000101

Johnson, J. D., Dennull, R. A., Gerena, L., Lopez-Sanchez, M., Roncal, N. E., and Waters, N. C. (2007). Assessment and continued validation of the malaria SYBR green I-based fluorescence assay for use in malaria drug screening. *Antimicrob. Agents Chemother.* 51, 1926–1933. doi: 10.1128/AAC.01607-06

Kadurin (2016). The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8 (7), 10883–10890. doi: 10.18632/oncotarget.14073

Kearnes, S., Mccloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J. Computer-Aided Mol. Des.* 30, 595–608. doi: 10.1007/s10822-016-9938-8

Kennedy, K., Cobbold, S. A., Hanssen, E., Birnbaum, J., Spillman, N. J., Mchugh, E., et al. (2019a). Delayed death in the malaria parasite Plasmodium falciparum is caused by disruption of prenylation-dependent intracellular trafficking. *PloS Biol.* 17, e3000376. doi: 10.1371/journal.pbio.3000376

Kennedy, K., Crisafulli, E. M., and Ralph, S. A. (2019b). Delayed Death by Plastid Inhibition in Apicomplexan Parasites. *Trends Parasitol.* 35, 747–759. doi: 10.1016/j.pt.2019.07.010

Koehn, F. E., and Carter, G. T. (2005). The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discovery* 4, 206–220. doi: 10.1038/nrd1657

Leelananda, S. P., and Lindert, S. (2016). Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12, 2694–2718. doi: 10.3762/bjoc.12.267

Li, J. W., and Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science* 325, 161–165. doi: 10.1126/science.1168243

Liu, K., Sun, X., Jia, L., Ma, J., Xing, H., Wu, J., et al. (2018). Chemi-net: a graph convolutional network for accurate drug property prediction. Arxiv. 1803.06236

Mallinson, J., and Collins, I. (2012). Macrocycles in new drug discovery. *Future Med. Chem.* 4, 1409–1438. doi: 10.4155/fmc.12.93

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Front. In Environ. Sci.* 3, 80. doi: 10.3389/fenvs.2015.00080

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9, 5441–5451. doi: 10.1039/C8SC00148K

Newman, D. J., and Cragg, G. M. (2016). Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* 79, 629–661. doi: 10.1021/acs.jnatprod.5b01055

Pazhayam, N. M., Chhibber-Goel, J., and Sharma, A. (2019). New leads for drug repurposing against malaria. *Drug Discovery Today* 24, 263–271. doi: 10.1016/j.drudis.2018.08.006

Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C. B., et al. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nat. Methods* 13, 770–776. doi: 10.1038/nmeth.3940

Pubchem Database (2019). National Center for Biotechnology Information. *Pubchem Database*. https://pubchem.ncbi.nlm.nih.gov/bioassay/686979 (accessed on Dec. 18, 2019).

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discovery* 18, 41–58. doi: 10.1038/nrd.2018.168

Rajpurkar, P. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. Arxiv. 1707.01836

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively Multitask Networks for Drug Discovery. Arxiv. 1502.02072

Ramsundar, B., Eastman, P., Walters, P., and Pande, V. (2019). *Deep Learning for the Life Sciences : Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More* (Sebastopol, CA: O'Reilly Media). (http://oreilly.com).

Reddy, S., Fox, J., and Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* 112, 22–28. doi: 10.1177/0141076818815510

Riss, T. L., Moravec, R. A., Niles, A. L., Duellman, S., Benink, H. A., Worzella, T. J., et al. (2004). "Cell Viability Assays," in *Assay Guidance Manual*. Eds. G. S. Sittampalan, A. Grossman and K. Bricombet (Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences).

Roberts, B. F., Iyamu, I. D., Lee, S., Lee, E., Ayong, L., Kyle, D. E., et al. (2016). Spirocyclic chromanes exhibit antiplasmodial activities and inhibit all intraerythrocytic life cycle stages. *Int. J. Parasitol. Drugs Drug Resist.* 6, 85–92. doi: 10.1016/j.ijpddr.2016.02.004

Roberts, B. F., Zheng, Y., Cleaveleand, J., Lee, S., Lee, E., Ayong, L., et al. (2017). 4-Nitro styrylquinoline is an antimalarial inhibiting multiple stages of Plasmodium falciparum asexual life cycle. *Int. J. Parasitol. Drugs Drug Resist.* 7, 120–129. doi: 10.1016/j.ijpddr.2017.02.002

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Saunders, D. L., Vanachayangkul, P., Lon, C.Program, U.S.a.M.M.R. and National Center for Parasitology, E, , Malaria, C., et al. (2014). Dihydroartemisinin-piperaquine failure in Cambodia. *N. Engl. J. Med.* 371, 484–485. doi: 10.1056/NEJMc1403007

Schneider, G. (2017). Automating drug discovery. *Nat. Rev. Drug Discovery* 17, 97–113. doi: 10.1038/nrd.2017.232

Selwood, D. L. (2017). Macrocycles, the edge of drug-likeness chemical space or Goldilocks zone? *Chem. Biol. Drug Des.* 89, 164–168. doi: 10.1111/cbdd.12922

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature* 432, 862–865. doi: 10.1038/nature03197

Spangenberg, T., Burrows, J. N., Kowalczyk, P., Mcdonald, S., Wells, T. N. C., and Willis, P. (2013). The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases. *PloS One* 8, e62906. doi: 10.1371/journal.pone.0062906

Swinney, D. C. (2013). Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines. *Clin. Pharmacol. Ther.* 93, 299–301. doi: 10.1038/clpt.2012.236

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* Basic Books.

Vossen, M. G., Pferschy, S., Chiba, P., and Noedl, H. (2010). The SYBR Green I malaria drug sensitivity assay: performance in low parasitemia samples. *Am. J. Trop. Med. Hyg.* 82, 398–401. doi: 10.4269/ajtmh.2010.09-0417

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Arxiv. doi: arXiv:1510.02855

Wang, Q., and Shen, D. (2017). Computational medicine: a cybernetic eye for rare disease. *Nat. Biomed. Eng.* 1, 0032. doi: 10.1038/s41551-017-0032

Whitty, A., Viarengo, L. A., and Zhong, M. (2017). Progress towards the broad use of non-peptide synthetic macrocycles in drug discovery. *Org. Biomol. Chem.* 15, 7729–7735. doi: 10.1039/C7OB00056A

WHO. (2018). World malaria report, World Health Organization. https://www.who.int/malaria/publications/world-malaria-report-2018/en/.

Wilson, D. W., Langer, C., Goodman, C. D., Mcfadden, G. I., and Beeson, J. G. (2013). Defining the timing of action of antimalarial drugs against plasmodium falciparum. *Antimicrob. Agents Chemother.* 57, 1455. doi: 10.1128/AAC.01881-12

Wu, Z., Ramsundar, B., Feinberg, , Evan n., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi: 10.1039/C7SC02664A

Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., and Aliper, A. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Res. Rev.* 49, 49–66. doi: 10.1016/j.arr.2018.11.003

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 28 January 2020
doi: 10.3389/fphar.2019.01631

# Capsule Networks Showed Excellent Performance in the Classification of hERG Blockers/Nonblockers

*Yiwei Wang[1,2†], Lei Huang[3,4†], Siwen Jiang[3], Yifei Wang[1], Jun Zou[1], Hongguang Fu[3] and Shengyong Yang[1*]*

*[1] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [2] College of Preclinical Medicine, Southwest Medical University, Luzhou, China, [3] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, [4] Basic Teaching Department, Sichuan College of Architectural Technology, Deyang, China*

Capsule networks (CapsNets), a new class of deep neural network architectures proposed recently by Hinton et al., have shown a great performance in many fields, particularly in image recognition and natural language processing. However, CapsNets have not yet been applied to drug discovery-related studies. As the first attempt, we in this investigation adopted CapsNets to develop classification models of hERG blockers/nonblockers; drugs with hERG blockade activity are thought to have a potential risk of cardiotoxicity. Two capsule network architectures were established: convolution-capsule network (Conv-CapsNet) and restricted Boltzmann machine-capsule networks (RBM-CapsNet), in which convolution and a restricted Boltzmann machine (RBM) were used as feature extractors, respectively. Two prediction models of hERG blockers/nonblockers were then developed by Conv-CapsNet and RBM-CapsNet with the Doddareddy's training set composed of 2,389 compounds. The established models showed excellent performance in an independent test set comprising 255 compounds, with prediction accuracies of 91.8 and 92.2% for Conv-CapsNet and RBM-CapsNet models, respectively. Various comparisons were also made between our models and those developed by other machine learning methods including deep belief network (DBN), convolutional neural network (CNN), multilayer perceptron (MLP), support vector machine (SVM), k-nearest neighbors (kNN), logistic regression (LR), and LightGBM, and with different training sets. All the results showed that the models by Conv-CapsNet and RBM-CapsNet are among the best classification models. Overall, the excellent performance of capsule networks achieved in this investigation highlights their potential in drug discovery-related studies.

**Keywords: deep learning, hERG, classification model, Capsule network, convolution-capsule network, restricted Boltzmann machine-capsule networks**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Mingyue Zheng,*
*Chinese Academy of Sciences, China*
*Youjun Xu,*
*Peking University, China*

****Correspondence:***
*Shengyong Yang*
*yangsy@scu.edu.cn*

*†These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *27 September 2019*
***Accepted:*** *13 December 2019*
***Published:*** *28 January 2020*

***Citation:***
*Wang Y, Huang L, Jiang S, Wang Y, Zou J, Fu H and Yang S (2020) Capsule Networks Showed Excellent Performance in the Classification of hERG Blockers/Nonblockers. Front. Pharmacol. 10:1631. doi: 10.3389/fphar.2019.01631*

# INTRODUCTION

The human ether-a-go-go-related gene (hERG) encodes a potassium channel protein, which is important for cardiac electrical activity and the coordination of heartbeat. Blockade of the hERG potassium channel can result in a potentially fatal disorder called long QT syndrome, as well as serious cardiotoxicity, which has led to the withdrawal of several marketed drugs and the failure of many drug research and development projects (Fermini and Fossa, 2003; Recanatini et al., 2005; Sanguinetti and Tristani-Firouzi, 2006; Bowes et al., 2012; Nachimuthu et al., 2012; Zhang et al., 2012; Shah, 2013; Kalyaanamoorthy and Barakat, 2018; Mladenka et al., 2018). Therefore, drug candidates that can bind with hERG should be eliminated as early as possible in drug discovery studies. At present, various *in vitro* experimental assays, such as fluorescent measurements (Dorn et al., 2005), radioligand binding assay (Yu et al., 2014), and patch-clamp electrophysiology (Stoelzle et al., 2011; Gillie et al., 2013; Danker and Moller, 2014), have been developed to measure the hERG binding affinity of chemicals. Nevertheless, these assays are often expensive and time-consuming, implying that they are not suitable for the evaluation of hERG binding affinity for a large number of chemicals in the early stage of drug discovery. Furthermore, the preconditions for the use of these analytical techniques are that the chemical compounds have been synthesized and are available in hand, which are usually not applicable in the era of virtual high-throughput screening. An alternative strategy is to use *in silico* methods; compared with experimental assays, *in silico* methods are cheaper and faster, and also do not involve any of the aforementioned preconditions.

To date, various *in silico* prediction models have been developed for hERG channel blockade. These models can be classified into structure-based and ligand-based models. Structure-based models utilize molecular docking to predict the binding mode and binding affinity of compounds to hERG. However, the structure-based methods often have some limitations such as protein flexibility, inaccurate scoring function, and solvent effect (Jia et al., 2008; Li et al., 2013). Ligand-based models can further be classified into several categories based on structural and functional features (Zolotoy et al., 2003; Aronov, 2005), quantitative structure-activity relationship (QSAR) models (Perry et al., 2006; Yoshida and Niwa, 2006; Tan et al., 2012), pharmacophore models (Cavalli et al., 2002; Aronov, 2006; Durdagi et al., 2011; Yamakawa et al., 2012; Kratz et al., 2014; Wang et al., 2016), and machine learning models (Wang et al., 2008; Klon, 2010; Wacker and Noskov, 2018). Compared with other models, machine learning models have attracted more attention in recent years due to the remarkable performance of machine learning methods in the handling of classification issues. For example, Wang et al. (2012) established binary classification models using Naïve Bayes (NB) classification and recursive partitioning (RP) methods, which achieved prediction accuracies of 85–89% in their test sets. Zhang and coworkers (Zhang et al., 2016) used five machine learning methods to develop models that can discriminate hERG blockers from nonblockers, and they found that k-nearest neighbors (kNN) and support vector machine (SVM) methods showed a better performance than others. Broccatelli et al. (2012) derived several classification models of hERG blocker/nonblocker by using random forests (RF), SVM, and kNN algorithms with descriptor selections *via* genetic algorithm (GA) methods, and their prediction accuracies ranged from 83 to 86%. Didziapetris and Lanevskij (2016) employed a gradient-boosting machine (GBM) statistical technique to classify hERG blockers/nonblockers, and this offered overall prediction accuracies of 72–78% against different test sets. Very recently, Siramshetty et al. (2018) employed three methods (kNN, RF, and SVM) with different molecular descriptors, activity thresholds, and training set compositions to develop predictive models of hERG blockers/nonblockers, and their models showed better performance than previously reported ones.

There have been remarkable advances in deep learning methods since a fast learning algorithm for deep belief nets was proposed by Hinton in 2006 (Hinton et al., 2006a). They have widely been applied to fields particularly computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, and various games (Collobert and Weston, 2008; Bengio, 2009; Dahl et al., 2012; Hinton et al., 2012; LeCun et al., 2015; Defferrard et al., 2016; Mamoshina et al., 2016), where they have produced results comparable to or in some cases superior to human experts. In recent years, deep learning has also been applied to drug discovery, and it has demonstrated its potentials (Lusci et al., 2013; Ma et al., 2015; Xu et al., 2015; Aliper et al., 2016; Mayr et al., 2016; Pereira et al., 2016; Subramanian et al., 2016; Kadurin et al., 2017; Ragoza et al., 2017; Ramsundar et al., 2017; Xu et al., 2017; Ghasemi et al., 2018; Harel and Radinsky, 2018; Hu et al., 2018; Popova et al., 2018; Preuer et al., 2018; Russo et al., 2018; Segler et al., 2018; Shin et al., 2018; Cai et al., 2019; Wang et al., 2019a; Yang et al., 2019). However, there are still some issues that limit the application of deep learning in drug discovery. For example, deep learning usually requires a large number of samples for model training. Unfortunately, there are often a very limited number of agents (usually hundreds or thousands) in drug discovery-related studies due to high cost and the lengthy process involved in obtaining samples and their associated properties. In addition, commonly used deep learning algorithms or networks, such as convolutional neural network (CNN), are primarily designed for two-dimensional (2D) image recognition. In these networks, some special algorithms, such as the pooling algorithm in CNN, are adopted to reduce the dimensionality of the representation, which might lead to a loss of information.

To overcome the shortcomings of traditional deep learning networks, Hinton group (Sabour et al., 2017) proposed new deep learning architectures known as capsule networks (CapsNets), which introduced a novel building block that is used in deep learning to improve the model hierarchical relationships inside the internal knowledge representation of a neural network. CapsNets have shown great potential in some fields (Xi et al., 2017; Afshar et al., 2018; Lalonde and Bagci, 2018; Qiao et al., 2018; Vesperini et al., 2018; Zhao et al., 2018; Wang et al., 2019b; Peng et al., 2019). However, CapsNets have not yet been applied to drug discovery-related studies. As the first attempt, in this study, we established two classification models of hERG blockers/nonblockers by using modified capsule network architectures. The models were evaluated using a test set and an external validation set, which are independent of the training set. Furthermore, our models were also compared with others.

The rest of this paper is organized as follows. The *Materials and Methods* section describes the implementation of the two capsule networks [convolution-capsule network (Conv-CapsNet) and RBM-CapsNet] developed in this study, as well as the data sets used and computational modeling details. The modeling, evaluation, and comparison with other models are presented in the *Results* section. The strengths of the capsule networks are analyzed in the *Discussion* section which is followed by a final summary.

# MATERIALS AND METHODS

## Convolution-Capsule Network

### Architecture

The architecture of Conv-CapsNet is schematically shown in **Figure 1**, which is similar in nature to that of the Hinton's original Capsule Network, except for one additional hidden feature layer. Apparently, Conv-CapsNet contains four layers: a convolutional layer, a hidden feature layer, a PrimaryCaps layer, and a DigitCaps layer. It is composed of 179 nodes for input, which are based on the feature vector size of the molecules. With mapping from the input vector, the hidden feature layer with 128 dimensional nodes was generated by one convolutional operation and one fully connected operation. The PrimaryCaps layer comprises eight capsules ($u_i$), and each capsule in this layer includes eight-dimensional features. Furthermore, we computed the contribution ($\hat{u}_{j|i}$) of each capsule ($u_i$) in PrimaryCaps to that ($v_j$) in DigitCaps by using Eq. 1.

$$\hat{u}_{j|i} = W_{ij} \cdot u_i \tag{1}$$

The final layer (DigitCaps) has a two-dimensional capsule ($v_j$) per digit class (two classes in this investigation). Each of these capsules received input from all the capsules in the PrimaryCaps layer through Eq. 2-1, Eq. 2-2, and Eq. 2-3.

$$c_{ij} = \frac{exp(b_{ij})}{\Sigma_k\, exp(b_{ik})} \tag{2 – 1}$$

$$s_j = \Sigma_k c_{ij} \hat{u}_{j|i} \tag{2 – 2}$$

$$v_j = \frac{\| s_j \|^2}{1 + \| s_j \|^2} \frac{s_j}{\| s_j \|} \tag{2 – 3}$$

Finally, we computed the length of each digit capsule to predict the class of chemical molecules from Eq 3.

$$L_k = T_k\, max(0, m^+ - \| v_k \|)^2 + \lambda(1 - T_k)$$
$$\times\ max(0, \| v_k \| - m^-)^2 \tag{3}$$

In view of the small size of the dataset in this account, we added the L2 regularization behind the convolutional operation to prevent the network from overfitting (Ng, 2004).

### Hyperparameter Optimization

For the hyperparameter optimization of the Conv-CapsNet architecture, the different numbers of filters in the convolutional layer, nodes in the hidden feature layer, and dimensions in PrimaryCaps were explored. Additionally, the dynamic routing iterations between two capsule layers were tested from 1 to 3 with an increment of 1. For each group of the parameter settings, the performance of the model was evaluated by five-fold cross-validation based on the training

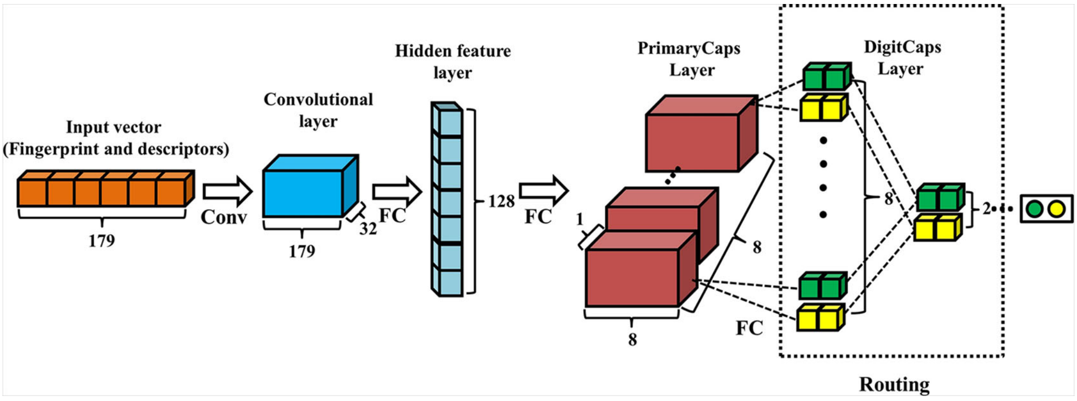


**FIGURE 1 |** Architecture of convolution-capsule networks (Conv-CapsNet). The input is one-dimensional vector containing 179 components. The convolution layer has 32 filters of size 1×3. The hidden feature layer and PrimaryCaps layer consist of 128 and 64 nodes, respectively. The weight matrix between PrimaryCaps layer and DigitCaps layer is 8×8×2×2, and two dynamic routing iterations were adopted.

set. Once the highest accuracy was achieved with all the candidate settings, the best setting was subsequently applied to the test set and external validation set. We employed early stopping to reduce the overfitting problem, which is a technique commonly used for the reduction of overfitting (Caruana et al., 2001). With the early stopping, original training set was randomly divided into a new training set and a validation set (4:1). When the error in the validation set was less than that from the previous iteration, the training was immediately stopped. The final optimal hyperparameters for Conv-CapsNet are listed in **Table 1**.

### Model Training of Conv-CapsNet

The Conv-CapsNet weights were randomly initialized using a truncated normal distribution with the standard deviation being set as 0.01 during training. Both the convolutional and hidden feature layers adopted the rectified linear unit (Relu) as the activation function. To reduce the internal-covariate-shift, we used batch normalization to normalize the input distribution of each layer to a standard Gaussian distribution (Hinton et al., 2011; Ioffe and Szegedy, 2015). The adaptive moment estimation (Adam) method was employed for optimization (Kingma and Ba, 2014).

**Table 2** summarizes the algorithm and training procedure for Conv-CapsNet. *CW*, *W*1, and *W*2 represent the parameters in the convolutional, hidden feature, and PrimaryCaps layers, respectively. The convolutional and the first two fully connected operations are represented by *conv*, *fc*1, and *fc*2, respectively; *conv_layer, hf_layer*, and *pc_layer* denote the output from the convolutional, hidden feature, and PrimaryCaps layers, respectively. Through a feature vector extraction process in the convolutional layer, the hidden feature layer, and the PrimaryCaps layer (lines 1–4), *pc_layer* was packed as capsules *u* (line 5). Here, $\hat{u}$ denotes the contribution of one layer to the next layer. Next, the routing algorithm was used to generate the digit capsules (lines 6–13). Len is the length of the output of DigitCaps layer (lines 14). Lines 15–20 are for the network parameter update using a gradient step.

**TABLE 1** | Hyperparameter settings of convolution-capsule networks (Conv-CapsNet).

| Hyperparameter | Setting |
|---|---|
| L2 normalization term | 0.001 |
| Activation | Relu |
| Batch size | 148 |
| Iteration epoch | 300 |
| Learning rate of network | 0.001 |
| Optimizer | Adam |
| Filter | 32 |
| Kernel_size | 3 |
| Number of nodes in the hidden feature layer | 128 |
| Number of nodes in the PrimaryCaps layer | 64 |
| Routing time | 2 |
| Dimension of each capsule | 8 |
| Length of PrimaryCaps | 2 |
| Length of DigitCaps | 2 |

**TABLE 2** | Algorithm and training procedure of convolution-capsule networks (Conv-CapsNet).

```
Algorithm: Conv-CapsNet training algorithm, using a mini-batch stochastic gradient descent (SGD) for simplicity.
Input: mini batch feature vector (x);
       Number of Conv-CapsNet training epoch (S);
       Number of dynamic routing iterations (iter).
Output: Length of each capsules (Len).
1:  For n=1 to S do
2:    $conv\_layer \leftarrow conv(x, CW)$
3:    $hf\_layer \leftarrow fc1(conv\_layer, W1)$
4:    $pc\_layer \leftarrow fc2(hf\_layer, W2)$
5:    $u \leftarrow \mathrm{Encapule}(pc\_layer)$
6:    For all capsule $i$ in PrimaryCaps layer:$\hat{u}_{j|i} \leftarrow W_{ij}u_i$………… {contribution computes Eq. 1}
7:    For all capsule $i$ in PrimaryCaps layer and capsule $j$ in DigitCaps layer:$b_{ij} \leftarrow 0$
8:    For $m=1$ to iter do
9:    For all capsule $i$ in PrimaryCaps layer: $c_i \leftarrow softmax(b_i)$ ……{softmax computes Eq. 2-1}
10:   For all capsule $j$ in DigitCaps layer:$s_j \leftarrow \sum_i c_{ij}\hat{u}_{j|i}$ …{dynamic computes Eq. 2-2}
11:   For all capsule $j$ in DigitCaps layer: $v_j \leftarrow squash\ (s_j)$ ………{squash computes Eq. 2-3}
12:   For all capsule $i$ in PrimaryCaps layer and capsule $j$ in DigitCaps layer: $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} \cdot v_j$
13:     End for
14:     $Len \leftarrow Length\ \ of\ \ v$
15:     $L \leftarrow loss\ \ of\ \ v$…………………………………………{loss computes Eq. 3}
16:     $W \leftarrow W - \partial L/\partial W$
17:     $CW \leftarrow CW - \partial L/\partial CW$
18:     $W1 \leftarrow W1 - \partial L/\partial W1$
19:     $W2 \leftarrow W2 - \partial L/\partial W2$
20: End for
```

## Restricted Boltzmann Machine-Capsule Network

### Architecture

**Figure 2** displays the architecture of RBM-CapsNet, which consists of three layers: a hidden feature layer, a PrimaryCaps layer, and a DigitCaps layer. In RBM-CapsNet, two restricted Boltzmann machines (RBMs) replaced the convolutional and fully connected operations in Conv-CapsNet. The first RBM encodes the original vector (179-dimension) for the feature space (the hidden feature layer), which is subsequently used as the input for the next RBM. The RBMs used energy function (Eq. 4) as the loss function (Hinton and Salakhutdinov, 2006b).

$$E(v,h) = -\left(a^T \cdot v + b^T \cdot h + v^T \cdot \omega \cdot h\right) \tag{4}$$

The capsule networks still consist of PrimaryCaps and DigitCaps, which are the same as in Conv-CapsNet. The detailed definitions of all the parameters in Eq. 1, 2, 3, and 4 are listed in the **Supplementary Material**.

### Hyperparameter Optimization

To optimize the hyperparameters in the RBM-CapsNet architecture, all the combinations of one to five RBM operations and 32, 64, 128, 256, and 512 nodes in each RBM were tested. The basic optimization procedure for the hyperparameters related to the capsules is very similar with that for Conv-CapsNet. The performance of each RBM-

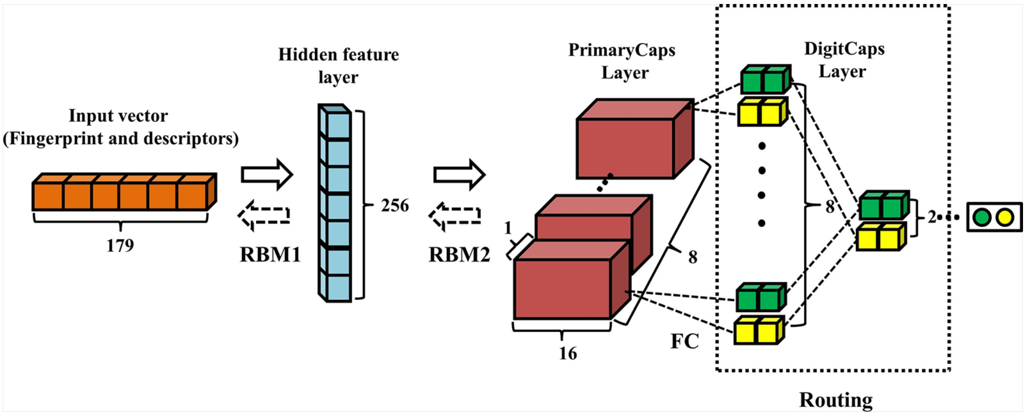


**FIGURE 2 |** Architecture of restricted Boltzmann machine-capsule networks (RBM-CapsNet). The input is one-dimensional vector containing 179 components. The hidden feature layer and PrimaryCaps layer consist of 256 and 128 nodes, respectively. The weight matrix between PrimaryCaps layer and DigitCaps layer is 8×8×2×2, and two dynamic routing iterations were adopted.

CapsNet architecture was examined by five-fold cross-validation. The candidate RBM-CapsNet architecture that provided the highest accuracy was validated using the test set and external validation set. The detailed information on the optimized hyperparameters in RBM-CapsNet is summarized in **Table 3**.

## Model Training of Restricted Boltzmann Machine-Capsule Network

The training process was divided into two stages. First, two RBMs were pre-trained one by one with the loss function shown in Eq. 4. Second, the parameters of RBMs from pre-training were taken as initial values and the whole network was fine-tuned by back-propagation algorithm with end-to-end (Rumelhart et al., 1986).

**Table 4** summarizes the algorithm and training procedure for RBM-CapsNet. $\theta1$ and $\theta2$ represent the parameters of the hidden feature and PrimaryCaps layers, respectively. $\phi1$ and $\phi2$ represent the operations in RBM1 and RBM2, respectively. The *hf_layer* and *pc_layer* denote the output from the hidden feature and PrimaryCaps layers, respectively. After training RBM1 and RBM2 individually (lines 1–6), the *pc_layer* was packed as capsules $u$ (line 10). The routing algorithm was then used to generate the digit capsules (lines 11–18). Len is the length of the output of DigitCaps layer (lines 19). Lines 20 to 24 are for a network parameter update using a gradient step ($\partial L/\partial W$ represents the gradient of the contribution matrix, and $\partial L/\partial \theta1$ and $\partial L/\partial \theta2$ represent the gradients of the parameters for the hidden feature and PrimaryCaps layers, respectively).

**TABLE 3 |** Hyperparameter settings of restricted Boltzmann machine-capsule networks (RBM-CapsNet).

| Hyperparameter | Setting |
|---|---|
| Numbers of RBM | 2 |
| Number of nodes in the hidden feature layer | 256 |
| Number of nodes in the PrimaryCaps layer | 128 |
| Iteration of RBM | 100 |
| Iteration of network | 200 |
| Learning rate of RBM | 0.001 |
| Learning rate of network | 0.005 |
| Activation | Relu |
| Batch size | 148 |
| Optimizer | Adam |
| Routing time | 2 |
| Dimension of each capsule | 8 |
| Length of PrimaryCaps | 2 |
| Length of DigitCaps | 2 |

# Data Sets

In this investigation, the Doddareddy's hERG blockade data set was used to establish our models (Doddareddy et al., 2010), which includes literature compounds tested on the hERG channel and Food and Drug Administration (FDA)-approved drugs. This data set contains a total of 2,644 compounds, including 1,112 positives (hERG blocker, $IC_{50}$ < 10 μM) and 1,532 negatives (hERG nonblocker, $IC_{50}$ > 30 μM). Doddareddy et al. partitioned this data set into a training set and a test set (Doddareddy et al., 2010). For comparison, the same partition scheme for the training and test sets as that by Doddareddy et al. was adopted in this investigation. Furthermore, we used Doddareddy's experimentally validated dataset (a total of 60 compounds: 50 agents from the Chembridge database and 10 from an in-house compound library) as an external validation set to assess the generalization ability of our models. In order to compare the performance of our models with others reported in the literature, we also used the same data sets as those in the literature, including Hou's (Wang et al., 2012; Wang et al., 2016), Zhang's (Zhang et al., 2016), Sun's (Sun et al., 2017), Siramshetty's (Siramshetty et al., 2018), and Cai's (Cai et al., 2019) data sets. Here, it is necessary to mention that an integrated data set of hERG blockade, which is the largest database to date, has been collected by Sato et al. (2018). However, we did not use this data set because it was not accessible. Another reason was that this data set has not been used to develop prediction models so far, and hence, a comparison study involving the data set was not feasible.

**TABLE 4** | Algorithm and training procedure of restricted Boltzmann machine-capsule networks (RBM-CapsNet).

```
Algorithm: RBM-CapsNet training algorithm, using a mini-batch stochastic gradient descent (SGD) for simplicity.
Input: mini batch feature vector (x);
       Number of RBM training epoch (S1);
       Number of Capsule training epoch (S2);
       Number of dynamic routing iterations (iter).
Output: Length of each capsules (Len).
For n=1 to S1 do
    $hf\ layer \leftarrow \phi 1(x, \theta 1)$............................................{RBM1 training}
End for
For n=1 to S1 do
    $pc\ layer \leftarrow \phi 2(hf\ layer, \theta 2)$......................................{RBM2 training}
End for
For n=1 to S2 do
    $hf\ layer \leftarrow \phi 1(x, \theta 1)$
    $pc\ layer \leftarrow \phi 2(h1\ layer, \theta 2)$
    $u \leftarrow$ Encapule ($pc\ layer$ )
    For all capsule $i$ in PrimaryCaps layer:$\hat{u}_{j|i} \leftarrow W_{ij}u_i$...........{contribution computes Eq. 1}
    For all capsule $i$ in PrimaryCaps layer and capsule $j$ in DigitCaps layer:$b_{ij} \leftarrow 0$
    For $m=1$ to iter do
        For all capsule $i$ in PrimaryCaps layer: ......{softmax computes Eq. 2-1}
        For all capsule $j$ in DigitCaps layer: $s_j \leftarrow \sum_i c_{ij}\hat{u}_{j|i}$.........{dynamic computes Eq. 2-2}
        For all capsule $j$ in DigitCaps layer:$v_j \leftarrow squash(s_j)$...........{squash computes Eq. 2-3}
        For all capsule $i$ in PrimaryCaps layer and capsule $j$ in DigitCaps layer: $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} \cdot v_j$
    End for
    $Len \leftarrow Length\ of\ v$
    $L \leftarrow loss\ of\ v$...........................................{loss computes Eq. 3}
    $W \leftarrow W - \partial L / \partial W$
    $\theta 1 \leftarrow \theta 1 - \partial L / \partial \theta 1$
    $\theta 2 \leftarrow \theta 2 - \partial L / \partial \theta 2$
End for
```

## Molecular Characterization

In this investigation, a combination of MACCS (MDL Molecular Access) molecular fingerprints (166 bits) and 13 molecular descriptors was utilized to characterize the chemical compounds, which has been used by Zhang et al. and showed a good predictive performance in hERG blockade classification modeling (Zhang et al., 2016). Another reason why we adopted this characterization method (MACCS+13 descriptors, a total of 179 features) is because of their short length which is important for the reduction of the number of parameters in the modeling and the training time. By the way, the 13 molecular descriptors were selected because they are thought to be very related to ADMET properties and have been widely used in the modeling of various ADMET properties (Hou and Wang, 2008; Hou et al., 2009; Wang et al., 2012; Zhang et al., 2016). A detailed list of these descriptors are given as follows: the octanol-water partitioning coefficient, apparent partition coefficient at pH = 7.4, molecular solubility, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of rings, number of aromatic rings, sum of the oxygen and nitrogen atoms, polar surface area, molecular fractional polar surface area, and molecular surface area.

All the molecular fingerprints and molecular descriptors were computed with RDKit (Landrum, 2018) and PaDEL-Descriptor (Yap, 2011), respectively. Because the values of the different descriptors might span significantly different numerical ranges, their values were scaled to the same range (0, 1) by using the following formula:

$$x^* = \frac{x - min}{max - min} \tag{5}$$

where $x$ is the original value, $x^*$ is the scaled value, and max and min are the maximum and minimum values of a descriptor, respectively.

## Model Assessment

All the models were assessed based on their accuracy (Q), sensitivity (SE), and specificity (SP). Q reflects the total prediction effect of a classifier. SE and SP represent the predictive power for positives and negatives, respectively. The definitions are given as follows (TP, true positive/blocker; TN, true negative/nonblocker; FP, false positive/blocker; and FN, false negative/nonblocker):

$$Q = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$SE = \frac{TP}{TP + FN} \tag{7}$$

$$SP = \frac{TN}{FP + TN} \tag{8}$$

The classification capability of models was measured by area under the receive operating characteristic curve (AUC), which is an important indicator to illustrate the classification performance by changing its discrimination threshold.

Another measurement of the quality of binary (two-class) classifications is the Matthew's correlation coefficient (MCC). The MCC considers the balance ratios of the four confusion matrix categories (TP, TN, FP, and FN), and reflects the predictive power of models objectively without the influence of the disproportionate ratio of positives to negatives in the dataset. The MCC was calculated by using the following equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(FP + TN)(FP + TP)(FN + TN)(FN + TP)}} \tag{9}$$

## Computations

All the calculations were carried out with a single dual-processor, 16-core 2.1 GHz Intel® Xeon® E5-2683 v4 CPU with 126 GB memory and two NVIDIA Tesla P100 GPU accelerators. The software modules that were used to implement this project included Scikit-learn 0.18.1, Python 3.6.4, Anaconda 5.1.0 (64-bit), and TensorFlow 1.4.0.

# RESULTS

## Selection of the Optimal Capsule Network Architectures and Model Development

Hinton et al raised the concept of capsule network and proposed the first capsule network architecture prototype (Sabour et al., 2017). To find the optimal capsule network architectures for the modeling of hERG blockade, we tried to construct a number of capsule networks with different architectures following Hinton's principle. Here, the Doddareddy's training set (positives: 1,004; negatives: 1,385) was adopted to train all the models, and the five-fold cross-validation method was used to monitor the training processes. In the five-fold cross-validation, the training set was randomly divided into five subsets. Of the five subsets, four subsets were used as the training data, and the remaining subset was used as the validation data for testing the model. The cross-validation process was repeated five times, with each of the five subsets used exactly once as the validation data. The average of the results from the five runs was calculated to produce a single estimation. The five-fold cross-validation results for the training set are given in **Table 5**. According to these results, Conv-CapsNet and RBM-CapsNet showed the best performance. For the Conv-CapsNet model, the prediction accuracies for the hERG blockers (SE) and the hERG nonblockers (SP) were 88.6 and 89.1%, respectively, and the overall prediction accuracy (Q) was 88.9%. For the RBM-CapsNet model, the prediction accuracies for hERG blockers and nonblockers were 84.3 and 89%, respectively and the overall prediction accuracy was 87.0%. Importantly, the MCC values of Conv-CapsNet and RBM-CapsNet were 0.774 and 0.734, respectively, which were also the highest among all the MCC values (**Table 5**); a higher MCC value often indicates a better prediction power of model. Therefore, the architectures of Conv-CapsNet and RBM-CapsNet were chosen as our capsule network architectures, and a detailed description for these architectures was given in the *Materials and Methods* section.

## Validation of Our Models' Prediction Ability Against hERG Blockers/Nonblockers by Doddareddy's Test Set and External Validation Set

In the above subsection, we obtained the optimal architectures of capsule networks. With these capsule network architectures, two classification models of hERG blockers/nonblockers, Conv-CapsNet and RBM-CapsNet models have been developed. To verify the predictive ability of these models, two test sets that are independent of the training set were used: Doddareddy's test set (positives: 108; negatives: 147) and external validation set (positives: 18; negatives: 42).

**Table 6** summarizes the prediction results of the Conv-CapsNet and RBM-CapsNet models. From **Table 6**, we can see that both models show excellent prediction ability to the Doddareddy's test set and external validation set. With the Conv-CapsNet model, of the 108 blockers in the test set, 102 were correctly predicted, indicating a prediction accuracy of 94.4% for the blockers (SE). For the 147 nonblockers, 132 (TN) were properly predicted. The accuracy for the prediction of nonblockers (SP) was 89.8%. Of all the 255 agents (blockers and nonblockers), 234 were correctly predicted and 20 were wrongly predicted (see **Table 6**). The overall prediction accuracy (Q) and AUC measure were 91.8% and 0.940 (see **Figure 3**), respectively. For the external validation set, of the 18 blockers, 16 (TP) were correctly discriminated from nonblockers. The prediction accuracy for the blockers (SE) was 88.9%. Of the 42 nonblockers, 30 (TN) were correctly predicted, indicating a prediction accuracy of 71.4% for the nonblockers (SP). Totally, 46 out of 60 compounds were correctly predicted. The overall prediction accuracy (Q) and AUC measure were 76.7% and 0.806, respectively. With the RBM-CapsNet model, in the test set, 99 (TP) out of 108 blockers were correctly predicted, indicating a prediction accuracy of 91.7%. Out of 147 nonblockers, 136 (TN) were correctly predicted, indicating a prediction accuracy of 92.5% for nonblockers. This model achieved an overall prediction accuracy of 92.2%. For the external validation set, the prediction accuracies for blockers (SE) and nonblockers (SP) were 94.4 and 71.4%, respectively.

**TABLE 5** | Prediction results of hERG blockers/nonblockers classification models developed by capsule networks with different architectures.

| Capsule network architecture | SE | SP | MCC | SD | Q (%) |
|---|---|---|---|---|---|
| Original CapsNet | 80.4% | 86.7% | 0.673 | 0.0141 | 84.1% |
| FC+FC | 82.6% | 86.7% | 0.694 | 0.0195 | 85.0% |
| Conv+FC | 82.2% | 86.4% | 0.687 | 0.0166 | 84.6% |
| Conv+FC+FC **(Conv-CapsNet)** | **88.6%** | **89.1%** | **0.774** | **0.0109** | **88.9%** |
| Conv+Conv+FC+FC | 84.5% | 85.3% | 0.693 | 0.0142 | 84.9% |
| Conv+Conv+Conv+FC+FC | 81.9% | 86.9% | 0.685 | 0.0173 | 84.9% |
| One RBM | 83.1% | 86.5% | 0.694 | 0.0182 | 84.9% |
| Two RBMs **(RBM-CapsNet)** | **84.3%** | **89.0%** | **0.734** | **0.0160** | **87.0%** |
| Three RBMs | 84.5% | 85.5% | 0.696 | 0.0160 | 85.0% |
| Four RBMs | 81.2% | 86.0% | 0.673 | 0.0108 | 83.9% |
| Five RBMs | 84.1% | 86.4% | 0.701 | 0.0156 | 85.4% |

**Conv, convolutional operation; FC, fully connected operation; RBM, restricted Boltzmann machine; Conv-CapsNet, convolution-capsule network; RBM-CapsNet, restricted Boltzmann machine-capsule network (The training set used was the Doddareddy's training set, and five-fold cross-validation was used to monitor the training performance. SE (%), sensitivity; SP (%), specificity; MCC, Matthew's correlation coefficient; SD, standard deviation; Q (%), overall accuracy). Conv-CapsNet and Conv-CapsNet showed the best performance.*

**TABLE 6** | Prediction accuracies of hERG blockade classification models developed by different methods with the same Doddareddy's training set.

| Model | SE | SP | MCC | Q (%) | AUC |
|---|---|---|---|---|---|
| | | Doddareddy's test set (255/P:108, N:147) | | | |
| Conv-CapsNet | 94.4% | 89.8% | 0.835 | 91.8% | 0.940 |
| RBM-CapsNet | 91.7% | 92.5% | 0.840 | 92.2% | 0.944 |
| CNN | 87.0% | 85.0% | 0.715 | 85.9% | 0.933 |
| MLP | 82.4% | 86.4% | 0.687 | 84.7% | 0.920 |
| DBN | 72.2% | 80.8% | 0.533 | 80.8% | 0.903 |
| SVM | 90.7% | 84.4% | 0.743 | 87.1% | 0.933 |
| kNN | 69.4% | 96.6% | 0.703 | 85.1% | 0.928 |
| Logistic regression | 88.8% | 83.7% | 0.710 | 85.5% | 0.858 |
| LightGBM | 79.6% | 82.3% | 0.617 | 81.2% | 0.810 |
| | | Doddareddy's external validation (60/P:18, N:42) | | | |
| Conv-CapsNet | 88.9% | 71.4% | 0.554 | 76.7% | 0.806 |
| RBM-CapsNet | 94.4% | 71.4% | 0.604 | 78.7% | 0.811 |
| CNN | 94.4% | 52.4% | 0.441 | 65.0% | 0.725 |
| MLP | 88.9% | 57.1% | 0.426 | 66.7% | 0.707 |
| DBN | 88.9% | 52.4% | 0.386 | 63.3% | 0.683 |
| SVM | 88.9% | 52.4% | 0.386 | 63.3% | 0.660 |
| kNN | 77.8% | 52.4% | 0.279 | 60.0% | 0.624 |
| Logistic regression | 83.3% | 52.4% | 0.332 | 61.7% | 0.623 |
| LightGBM | 61.1% | 59.5% | 0.190 | 60.0% | 0.609 |

*(TP, true positive; TN, true negative; FP, false positive; FN, false negative; SE (%), sensitivity, SE = TP/(TP + FN); SP (%), specificity, SP = TN/(TN + FP); Q (%), overall accuracy, Q = [TP + TN)/(TP + TN + FP + FN)].*

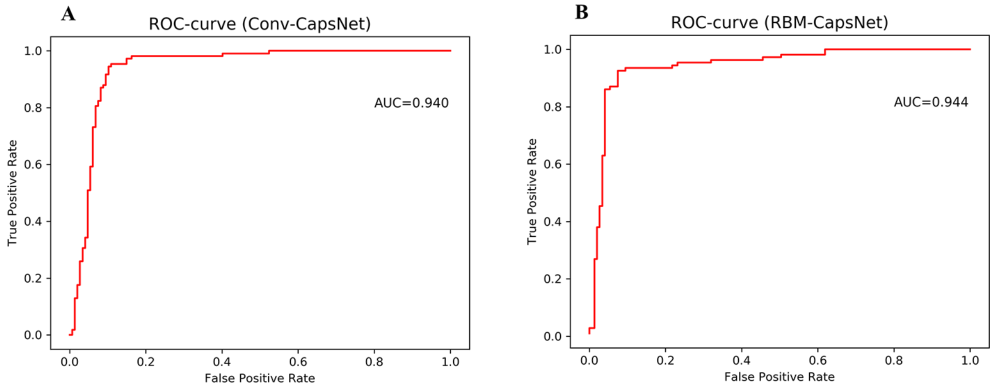


**FIGURE 3 |** Receiver operating characteristic (ROC) curves for Doddareddy's test set by **(A)** convolution-capsule networks (Conv-CapsNet) and **(B)** restricted Boltzmann machine-capsule networks (RBM-CapsNet), respectively.

The overall prediction accuracy (Q) and the MCC values were 78.7% and 0.604, respectively. AUC for the test and external validation sets were 0.944 (see **Figure 3**) and 0.811, respectively. All of these results clearly demonstrate that the established Conv-CapsNet and RBM-CapsNet models can not only correctly classify the training set compounds but also have an outstanding predictability for external agents outside of the training set.

## Comparison of Our Models With Other Models Developed With the Same Doddareddy's Training Set

To compare the performance of our models with that of others, we adopted commonly used machine learning methods to develop prediction models of hERG blockers/nonblockers with the same Doddareddy's training set. These machine learning methods include deep belief network (DBN), CNN, multilayer perceptron (MLP), SVM, kNN, logistic regression (LR), and LightGBM. Hyperparameters for these methods were optimized by five-fold cross-validation in advance, and the optimal hyperparameter values are listed in **Tables S1–S4**, respectively. The prediction results to the Doddareddy's test set and external validation set are also given in **Table 6**. From **Table 6**, we can see that the prediction accuracies of the seven models are obviously lower than those of our Conv-CapsNet and RBM-CapsNet models.

## Comparison of Our Models With Other Models Developed With Training Sets Different From Doddareddy's Training Set

It has been well known that the performance of a prediction model is often strongly dependent on the training set used. Therefore, to make a more objective comparison, we collected various hERG blockade classification models developed with training sets different from Doddareddy's training set. With these training sets, we established a series of new prediction models by the Conv-CapsNet and RBM-CapsNet methods. To avoid a possible influence of molecular features, the same molecular features used in the literature were used. **Table 7** summarizes the prediction accuracies of various models reported in the literature together with those of models by Conv-CapsNet and RBM-CapsNet.

Entry 1–3 of **Table 7** list models developed with Hou's training set 1 (positives: 283; negatives: 109), training set 2 (positives: 272; negatives: 120), and training set 3 (positive: 314; negative: 306), respectively. In Hou's training sets 1 and 2, a threshold of 40 μM was used to distinguish hERG blockers and nonblockers (blockers: $IC_{50}$ < 40 μM; nonblockers: $IC_{50} \geq 40$ μM). With training sets 1 and 2, Hou et al. established three models by RP, NB, and SVM methods, and the SVM models showed the best performance on their test sets. In Hou's training set 3, a threshold of 30 μM was used to define hERG blockers and nonblockers. A Bayesian classification model developed by Hou et al. with Hou's training set 3 gave a prediction accuracy of 85% on their test set. With Hou's training sets 1–3, we also separately established models by Conv-CapsNet and RBM-CapsNet methods. As shown in **Table 7**, our models showed comparable or superior performance compared with Hou's models. Entry 4 in **Table 7** shows models established by Zhang's training set (positives: 717; negatives: 210), in which a threshold of 30 μM was used to define hERG blockers and nonblockers. With the training set, Zhang et al. built two models by using SVM and kNN methods, which gave prediction accuracies of 83.5 and 82.2%, respectively, on their test set. Our models, developed by Conv-CapsNet and RBM-CapsNet, exhibited a better performance on the same test set (prediction accuracies: 84.5 and 85.2%, respectively). Entry 5 in **Table 7** displays models developed with Sun's training set, which is a big data set consisting of 3,024 agents (positives: 483; negatives: 2,541) with a threshold of 30 μM for defining hERG blockers and nonblockers. With the training set, Siramshetty et al. established two models by using LibSVM and RF methods, and their prediction accuracies on the test set were 71.0 and 74.0%, respectively. Our models offered much higher prediction accuracies (Conv-CapsNet: 83.3%; RBM-CapsNet: 86.3%). Entry 6 in **Table 7** shows models built with Siramshetty's

**TABLE 7 |** Prediction results of various hERG blockade classification models developed with training sets different from Doddareddy's training set.

| Entry | Model | Training set | Test set | SE | SP | Q |
|---|---|---|---|---|---|---|
| 1 | RP (Wang et al., 2016) | Hou's training set 1 | Hou's test set 1 | 79.8% | 75.8% | 78.5% |
| | NB (Wang et al., 2016) | (P: 283; N: 109) | (P: 129; N: 66) | 82.2% | 75.8% | 80.0% |
| | SVM (Wang et al., 2016) | | | 90.7% | 65.2% | 82.1% |
| | Conv-CapsNet | | | 85.7% | 78.8% | 82.0% |
| | RBM-CapsNet | | | 84.1% | 80.3% | 82.0% |
| 2 | RP (Wang et al., 2016) | Hou's training set 2 | Hou's test set 2 | 80.0% | 74.5% | 78.5% |
| | NB (Wang et al., 2016) | (P: 272; N: 120) | (P: 140; N: 55) | 81.4% | 80.0% | 81.0% |
| | SVM (Wang et al., 2016) | | | 85.0% | 74.5% | 82.1% |
| | Conv-CapsNet | | | 82.1% | 81.8% | 82.0% |
| | RBM-CapsNet | | | 81.4% | 83.6% | 82.0% |
| 3 | Bayesian (Wang et al., 2012) | Hou's training set 3 | Hou's test set 3 | 86.9% | 83.1% | 85.0% |
| | Conv-CapsNet | (P: 314; N: 306) | (P: 63; N: 57) | 87.3% | 86.0% | 86.8% |
| | RBM-CapsNet | | | 88.9% | 84.2% | 86.8% |
| 4 | SVM (Zhang et al., 2016) | Zhang's training set | Zhang's test set | 95.8% | 34.0% | 83.5% |
| | kNN (Zhang et al., 2016) | (P: 717; N: 210) | (P: 188; N: 48) | 92.6% | 40.4% | 82.2% |
| | Conv-CapsNet | | | 88.8% | 66.7% | 84.5% |
| | RBM-CapsNet | | | 90.4% | 64.6% | 85.2% |
| 5 | LibSVM (Siramshetty et al., 2018) | Sun's training set | Sun's test set | 68.0% | 85.0% | 71.0% |
| | RF (Siramshetty et al., 2018) | (P: 483; N: 2541) | (P: 53; N: 13) | 72.0% | 85.0% | 74.0% |
| | Conv-CapsNet | | | 83.0% | 84.6% | 83.3% |
| | RBM-CapsNet | | | 86.8% | 84.6% | 86.3% |
| 6 | LibSVM (Siramshetty et al., 2018) | Siramshetty's training set | Doddareddy's test set | 64.0% | 89.0% | 78.0% |
| | RF (Siramshetty et al., 2018) | T3 (P: 1406; N: 1708) | (P: 108; N: 147) | 68.0% | 91.0% | 81.0% |
| | Conv-CapsNet | | | 85.2% | 88.4% | 87.1% |
| | RBM-CapsNet | | | 83.3% | 91.2% | 87.8% |

training set T3 which were extracted from the ChEMBL database. In this training set, agents with a binding affinity of less than 1 μM were defined as hERG blockers, and those with a binding affinity of greater than 10 μM were defined as hERG nonblockers. With the training set, Siramshetty et al. established two models by using LibSVM and RF methods, and their prediction accuracies on their test set were 78.0 and 81.0%, respectively. Our Conv-CapsNet and RBM-CapsNet models gave prediction accuracies of 87.1 and 87.8%, respectively, which are obviously higher than those of LibSVM and RF models. Very recently, Cai et al. developed a deep learning model, termed deephERG, to predict hERG blockers with a large dataset containing 7,889 compounds (Cai et al., 2019). To make a comparison, we also used the same datasets to train and test hERG blocker prediction models. With the same validation set and evaluation method as those in Cai's work, our Conv-CapsNet (AUC = 0.974) and RBM-CapsNet (AUC = 0.978) showed a better performance than their deephERG (AUC = 0.967) (see **Table S5**). Collectively, for different training sets given here, the models developed by Conv-CapsNet and RBM-CapsNet were among the best models established by various machine learning methods.

## DISCUSSION

Since the first capsule networks were proposed by Hinton's group in 2017 (Sabour et al., 2017), they have attracted considerable attention because of their performance. For example, despite the simple three-layer architecture of the original capsule networks, they have achieved state-of-the-art results with 0.25% test error on Mixed National Institute of Standards and Technology database (MNIST) without data augmentation, which is better than the previous baseline of 0.39% (Sabour et al., 2017). The excellent performance of capsule networks is mainly due to the introduction of the capsules and dynamic routing algorithms. A capsule is a set of neurons that forms a vector. These vectors contain information including the magnitude/prevalence, spatial orientation, and other attributes of the extracted feature. In the capsule networks, capsules are "routed" to any capsule in the next layer *via* a dynamic routing algorithm, which takes into account the agreement between these capsule vectors, thus forming meaningful part-to-whole relationships not found in standard CNNs. In other words, capsule networks are capable of catching and holding more fine information than traditional deep neuron networks, one benefit of which is that the amount of input data can be significantly reduced.

Although CapsNets were just proposed very recently, they have already been successfully applied in many fields (Afshar et al., 2018; Kumar, 2018; Lalonde and Bagci, 2018; Li et al., 2018; Liu et al., 2018; Mobiny and Van Nguyen, 2018; Qiao et al., 2018; Zhao et al., 2018; Peng et al., 2019). Among these applications, majorities are related to image recognition. For example, Afshar et al. (2018) established a CapsNet for brain tumor classification by recognizing brain magnetic resonance imaging (MRI) images and proved that it could successfully overcome the defects of CNNs. Kumar (2018) proposed a novel method for traffic sign detection using a CapsNet that achieved outstanding performance, the input of which was traffic sign images. Li et al. (2018) built a CapsNet to recognize rice composites from unmanned aerial vehicle (UAV) images. This is understandable because CapsNets were originally developed to overcome the defects associated with image recognition in the traditional deep learning networks.

In image recognition, the input data is a two-dimensional array. In this two-dimensional array data, adjacent data points are often highly correlated. Small changes in any points generally do not affect image recognition in traditional deep learning methods. However, in issues related to drug discovery, such as the evaluation of ADMET properties (like the prediction of hERG blockers), one-dimensional vectors that describe small molecular structures and properties are usually used as the network input, for example, molecular fingerprints and descriptors. Generally, there is no direct logical relationship between the components in each vector for this kind of input. Importantly small changes in vector components might have a significant impact on the entire molecular structure and its associated properties. Nevertheless, these small changes in vector components are often overlooked in traditional deep learning methods. In addition, the relative positions of vector components are often critical though there is no direct logical relationship between them because a vector component represents a substructure or property. In this situation, capsule networks, which adopt vector neurons, are expected to have a better performance in handling this kind of issue (like the hERG blocker modeling) than other deep scalar neuron networks.

As expected, the two established capsule networks, Conv-CapsNet and RBM-CapsNet, showed excellent performance in the classification of hERG blockade. Although this is the first application of capsule networks in the classification of hERG blockers/nonblockers, the established models are still among the best classification models for hERG blockers/nonblockers. There can be no doubt that the use of capsules or vector neurons is one of the main reasons that contribute to the excellent performance of our models. Here each capsule represents a combination of substructures and/or properties. Analogy to the case in image recognition, the length of each capsule is the probability that the combination of substructures or properties exists in a molecule, and the orientation may represent the relative position of the combination of substructures in a compound. Obviously, our capsule networks can learn some combinations of substructures and/or properties that are important for the hERG blockers or nonblockers. Even so, we have to acknowledge that the prediction models of hERG blockers/nonblockers developed by the new capsule networks are still like a black box. Some questions regarding the models are difficult to answer. For example, we can't exactly know what the combination of substructures and/or properties is, and which features are important to the model and which samples are hard to classify. Overall, the application of capsule networks in drug discovery is still in its infancy. Further improvement of capsule networks and applications in drug discovery are necessary in future studies.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

SY designed the study. LH designed the algorithms. YwW and SJ executed the experiment and performed the data analysis. YwW mainly wrote the manuscript. SY, JZ, YfW and HF revised the manuscript. All authors discussed and commented on the manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (61876034, 81573349, 81773633, 21772130, and 81930125), National Science and Technology Major Project (2018ZX09711003-003-006; 2018ZX09711002-014-002, 2018ZX09201018, 2019ZX09301-135 and 2018ZX09711002-011-019), and 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.01631/full#supplementary-material

## REFERENCES

Afshar, P., Mohammadi, A., and Plataniotis, K. N. (2018). "Brain tumor type classification via capsule networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens: IEEE), 3129–3133. doi: 10.1109/icip.2018.8451379

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Aronov, A. (2005). Predictive in silico modeling for hERG channel blockers. *Drug Discov. Today* 10, 149–155. doi: 10.1016/s1359-6446(04)03278-7

Aronov, A. M. (2006). Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J. Chem. Inf. Model.* 49, 6917–6921. doi: 10.1021/jm060500o

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., et al. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discov.* 11, 909–922. doi: 10.1038/nrd3845

Broccatelli, F., Mannhold, R., Moriconi, A., Giuli, S., and Carosati, E. (2012). QSAR modeling and data mining link Torsades de Pointes risk to the interplay of extent of metabolism, active transport, and HERG liability. *Mol. Pharm.* 9, 2290–2301. doi: 10.1021/mp300156r

Cai, C., Guo, P., Zhou, Y., Zhou, J., Wang, Q., Zhang, F., et al. (2019). Deep learning-based prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model.* 59, 1073–1084. doi: 10.1021/acs.jcim.8b00769

Caruana, R., Lawrence, S., and Giles, C. L. (2001). "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping". In Proceedings of the 13th Conference on Neural Information Processing Systems (NIPS 2000), Denver, CO, USA. MIT Press, 402–408. doi: 10.5555/3008751.3008807

Cavalli, A., Poluzzi, E., De Ponti, F., and Recanatini, M. (2002). Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a

CoMFA study of HERG K+ channel blockers. *J. Med. Chem.* 45, 3844–3853. doi: 10.1021/jm0208875

Collobert, R., and Weston, J. (2008). "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning* (Helsinki: ACM), 160–167. doi: 10.1145/1390156.1390177

Dahl, G. E., Dong, Y., Li, D., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* 20, 30–42. doi: 10.1109/tasl.2011.2134090

Danker, T., and Moller, C. (2014). Early identification of hERG liability in drug discovery programs by automated patch clamp. *Front. Pharmacol.* 5, 203. doi: 10.3389/fphar.2014.00203

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral filtering,". In Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. 3844–3852. doi: 10.5555/3157382.3157527

Didziapetris, R., and Lanevskij, K. (2016). Compilation and physicochemical classification analysis of a diverse hERG inhibition database. *J. Comput. Aided Mol. Des.* 30, 1175–1188. doi: 10.1007/s10822-016-9986-0

Doddareddy, M. R., Klaasse, E. C., Shagufta,, Ijzerman, A. P., and Bender, A. (2010). Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *Chem. Med. Chem.* 5, 716–729. doi: 10.1002/cmdc.201000024

Dorn, A., Hermann, F., Ebneth, A., Bothmann, H., Trube, G., Christensen, K., et al. (2005). Evaluation of a high-throughput fluorescence assay method for HERG potassium channel inhibition. *J. Biomol. Screen* 10, 339–347. doi: 10.1177/1087057104272045

Durdagi, S., Duff, H. J., and Noskov, S. Y. (2011). Combined receptor and ligand-based approach to the universal pharmacophore model development for studies of drug blockade to the hERG1 pore domain. *J. Chem. Inf. Model.* 51, 463–474. doi: 10.1021/ci100409y

Fermini, B., and Fossa, A. A. (2003). The impact of drug-induced QT interval prolongation on drug discovery and development. *Nat. Rev. Drug. Discov.* 2, 439–447. doi: 10.1038/nrd1108

Ghasemi, F., Mehridehnavi, A., Perez-Garrido, A., and Perez-Sanchez, H. (2018). Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discov. Today* 23, 1784–1790. doi: 10.1016/j.drudis.2018.06.016

Gillie, D. J., Novick, S. J., Donovan, B. T., Payne, L. A., and Townsend, C. (2013). Development of a high-throughput electrophysiological assay for the human ether-a-go-go related potassium channel hERG. *J. Pharmacol. Toxicol. Methods* 67, 33–44. doi: 10.1016/j.vascn.2012.10.002

Harel, S., and Radinsky, K. (2018). Prototype-based compound discovery using deep generative models. *Mol. Pharm.* 15, 4406–4416. doi: 10.1021/acs.molpharmaceut.8b00474

Hinton, G. E., and Salakhutdinov, R. R. (2006b). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006a). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011). "Transforming auto-encoders," in *International Conference on Artificial Neural Networks* (Berlin, Heidelberg: Springer), 44–51.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Hou, T., and Wang, J. (2008). Structure-ADME relationship: still a long way to go? *Expert Opin. Drug Metab. Toxicol.* 4, 759–770. doi: 10.1517/17425255.4.6.759

Hou, T., Li, Y., Zhang, W., and Wang, J. (2009). Recent developments of in silico predictions of intestinal absorption and oral bioavailability. *Comb. Chem. High T. Scr.* 12, 497–506. doi: 10.2174/138620709788489082

Hu, Q., Feng, M., Lai, L., and Pei, J. (2018). Prediction of drug-likeness using deep autoencoder neural networks. *Front. Genet.* 9, 585. doi: 10.3389/fgene.2018.00585

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167.* doi: arXiv:1502.03167v3

Jia, R., Yang, L. J., and Yang, S. Y. (2008). Binding energy contributions of the conserved bridging water molecules in CDK2-inhibitor complexes: a combined QM/MM study. *Chem. Phys. Lett.* 460, 300–305. doi: 10.1016/j.cplett.2008.06.002

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* 14, 3098–3104. doi: 10.1021/acs.molpharmaceut.7b00346

Kalyaanamoorthy, S., and Barakat, K. H. (2018). Development of safe drugs: the hERG challenge. *Med. Res. Rev.* 38, 525–555. doi: 10.1002/med.21445

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: arXiv:1412.6980v9

Klon, A. E. (2010). Machine learning algorithms for the prediction of hERG and CYP450 binding in drug development. *Expert Opin. Drug Metab. Toxicol.* 6, 821–833. doi: 10.1517/17425255.2010.489550

Kratz, J. M., Schuster, D., Edtbauer, M., Saxena, P., Mair, C. E., Kirchebner, J., et al. (2014). Experimentally validated HERG pharmacophore models as cardiotoxicity prediction tools. *J. Chem. Inf. Model.* 54, 2887–2901. doi: 10.1021/ci5001955

Kumar, A. D. (2018). Novel deep learning model for traffic sign detection using capsule networks. *arXiv preprint arXiv:1805.04424.*. doi: arXiv:1805.04424v1

Lalonde, R., and Bagci, U. (2018). Capsules for object segmentation.*arXiv preprint arXiv:1804.04241*. doi: arXiv:1804.04241v1

Landrum, G. (2018). Provided by GitHub and SourceForge. *RDKit: Open-Source Cheminformatics Software [Online]*. Available: http://www.rdkit.org.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, G. B., Yang, L. L., Wang, W. J., Li, L. L., and Yang, S. Y. (2013). ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J. Chem. Inf. Model.* 53, 592–600. doi: 10.1021/ci300493w

Li, Y., Qian, M., Liu, P., Cai, Q., Li, X., Guo, J., et al. (2018). The recognition of rice images by UAV based on capsule network. *Cluster Comput.* 6, 1–10. doi: 10.1007/s10586-018-2482-7

Liu, Y., Tang, J., Song, Y., and Dai, L. (2018). A capsule based approach for polyphonic sound event detection. *arXiv preprint arXiv:1807.07436.* doi: arXiv:1807.07436v2

Lusci, A., Pollastri, G., and Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* 53, 1563–1575. doi: 10.1021/ci400187y

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* 55, 263–274. doi: 10.1021/ci500747n

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Mol. Pharm.* 13, 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep Learning. *Front. Environ. Sci.* 3, 8. doi: 10.3389/fenvs.2015.00080

Mladenka, P., Applova, L., Patocka, J., Costa, V. M., Remiao, F., Pourova, J., et al. (2018). Comprehensive review of cardiovascular toxicity of drugs and related agents. *Med. Res. Rev.* 38, 1332–1403. doi: 10.1002/med.21476

Mobiny, A., and Van Nguyen, H. (2018). "Fast capsNet for lung cancer screening," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 741–749.

Nachimuthu, S., Assar, M. D., and Schussler, J. M. (2012). Drug-induced QT interval prolongation: mechanisms and clinical management. *Ther. Adv. Drug Saf.* 3, 241–253. doi: 10.1177/2042098612454283

Ng, A. Y. (2004). "Feature selection, L 1 vs. L 2 regularization, and rotational invariance." In Proceedings of the 21th international conference on Machine learning(ACM), Banff. 78. doi: 10.1145/1015330.1015435

Peng, C., Zheng, Y., and Huang, D. S. (2019). Capsule network-based modeling of multi-omics data for discovery of breast cancer-related genes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2909905

Pereira, J. C., Caffarena, E. R., and Dos Santos, C. N. (2016). Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506. doi: 10.1021/acs.jcim.6b00355

Perry, M., Stansfeld, P. J., Leaney, J., Wood, C., De Groot, M. J., Leishman, D., et al. (2006). Drug binding interactions in the inner cavity of HERG channels:

molecular insights from structure-activity relationships of clofilium and ibutilide analogs. *Mol. Pharm.* 69, 509–519. doi: 10.1124/mol.105.016741

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4, eaap7885. doi: 10.1126/sciadv.aap7885

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 34, 1538–1546. doi: 10.1093/bioinformatics/btx806

Qiao, K., Zhang, C., Wang, L., Yan, B., Chen, J., et al. (2018). Accurate reconstruction of image stimuli from human fMRI based on the decoding model with capsule network architecture. *arXiv preprint arXiv:1801.00602.* doi: arXiv:1801.00602v1

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* 57, 942–957. doi: 10.1021/acs.jcim.6b00740

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., et al. (2017). Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* 57, 2068–2076. doi: 10.1021/acs.jcim.7b00146

Recanatini, M., Poluzzi, E., Masetti, M., Cavalli, A., and De Ponti, F. (2005). QT prolongation through hERG K(+) channel blockade: current knowledge and strategies for the early prediction during drug development. *Med. Res. Rev.* 25, 133–166. doi: 10.1002/med.20019

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H., and Ekins, S. (2018). Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Mol. Pharm.* 15, 4361–4370. doi: 10.1021/acs.molpharmaceut.8b00546

Sabour, S., Frosst, N., and Hinton, G. E. (2017). "Dynamic Routing Between Capsules,". In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 3859–3869. doi: 10.5555/3294996.3295142.

Sanguinetti, M. C., and Tristani-Firouzi, M. (2006). hERG potassium channels and cardiac arrhythmia. *Nature* 440, 463–469. doi: 10.1038/nature04710

Sato, T., Yuki, H., Ogura, K., and Honma, T. (2018). Construction of an integrated database for hERG blocking small molecules. *PLoS One* 13, e0199348. doi: 10.1371/journal.pone.0199348

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi: 10.1021/acscentsci.7b00512

Shah, R. R. (2013). Drug-induced QT interval prolongation: does ethnicity of the thorough QT study population matter? *Br. J. Clin. Pharmacol.* 75, 347–358. doi: 10.1111/j.1365-2125.2012.04415.x

Shin, M., Jang, D., Nam, H., Lee, K. H., and Lee, D. (2018). Predicting the absorption potential of chemical compounds through a deep learning approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 432–440. doi: 10.1109/TCBB.2016.2535233

Siramshetty, V. B., Chen, Q., Devarakonda, P., and Preissner, R. (2018). The Catch-22 of predicting hERG blockade using publicly accessible bioactivity data. *J. Chem. Inf. Model.* 58, 1224–1233. doi: 10.1021/acs.jcim.8b00150

Stoelzle, S., Obergrussberger, A., Bruggemann, A., Haarmann, C., George, M., Kettenhofen, R., et al. (2011). State-of-the-art automated patch clamp devices: heat activation, action potentials, and high Throughput in ion channel screening. *Front. Pharmacol.* 2, 76. doi: 10.3389/fphar.2011.00076

Subramanian, G., Ramsundar, B., Pande, V., and Denny, R. A. (2016). Computational modeling of beta-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* 56, 1936–1949. doi: 10.1021/acs.jcim.6b00290

Sun, H., Huang, R., Xia, M., Shahane, S., Southall, N., Wang, Y., et al. (2017). Prediction of hERG liability - Using SVM classification, bootstrapping and jackknifing. *Mol. Inform.* 36, 1600126. doi: 10.1002/minf.201600126

Tan, Y., Chen, Y., You, Q., Sun, H., and Li, M. (2012). Predicting the potency of hERG K (+) channel inhibition by combining 3D-QSAR pharmacophore and 2D-QSAR models. *J. Mol. Model.* 18, 1023–1036. doi: 10.1007/s00894-011-1136-y

Vesperini, F., Gabrielli, L., Principi, E., and Squartini, S. (2018). Polyphonic sound event detection by using capsule neural network. *arXiv preprint arXiv:1810.06325.* doi: arXiv:1810.06325

Wacker, S., and Noskov, S. Y. (2018). Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel. *Comput. Toxicol.* 6, 55–63. doi: 10.1016/j.comtox.2017.05.001

Wang, M., Yang, X.-G., and Xue, Y. (2008). Identifying hERG potassium channel inhibitors by machine learning methods. *QSAR Comb. Sci.* 27, 1028–1035. doi: 10.1002/qsar.200810015

Wang, S., Li, Y., Wang, J., Chen, L., Zhang, L., Yu, H., et al. (2012). ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharm.* 9, 996–1010. doi: 10.1021/mp300023x

Wang, S., Sun, H., Liu, H., Li, D., Li, Y., and Hou, T. (2016). ADMET evaluation in drug discovery. 16. predicting hERG blockers by combining multiple pharmacophores and machine learning approaches. *Mol. Pharm.* 13, 2855–2866. doi: 10.1021/acs.molpharmaceut.6b00471

Wang, D., Cui, C., Ding, X., Xiong, Z., Zheng, M., Luo, X., et al. (2019). Improving the virtual screening ability of target-specific scoring functions using deep learning methods. *Front. Pharmacol.* 10, 1–11. doi: 10.3389/fphar.2019.00924

Wang, D., Liang, Y., and Xu, D. (2019). Capsule network for protein post-translational modification site prediction. *Bioinformatics* 35, 2386–2394. doi: 10.1093/bioinformatics/bty977

Xi, E., Bing, S., and Jin, Y. (2017). Capsule network performance on complex data. *arXiv preprint arXiv:1712.03480.* doi: arXiv:1712.03480

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., and Lai, L. (2015). Deep learning for drug-induced liver injury. *J. Chem. Inf. Model.* 55, 2085–2093. doi: 10.1021/acs.jcim.5b00238

Xu, Y., Pei, J., and Lai, L. (2017). Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *J. Chem. Inf. Model.* 57, 2672–2685. doi: 10.1021/acs.jcim.7b00244

Yamakawa, Y., Furutani, K., Inanobe, A., Ohno, Y., and Kurachi, Y. (2012). Pharmacophore modeling for hERG channel facilitation. *Biochem. Biophys. Res. Commun.* 418, 161–166. doi: 10.1016/j.bbrc.2011.12.153

Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* 119, 10520–10594. doi: 10.1021/acs.chemrev.8b00728

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466–1474. doi: 10.1002/jcc.21707

Yoshida, K., and Niwa, T. (2006). Quantitative structure– activity relationship studies on inhibition of hERG potassium channels. *J. Chem. Inf. Model.* 46, 1371–1378. doi: 10.1021/ci050450g

Yu, Z., Klaasse, E., Heitman, L. H., and Ijzerman, A. P. (2014). Allosteric modulators of the hERG K (+) channel: radioligand binding assays reveal allosteric characteristics of dofetilide analogs. *Toxicol. Appl. Pharmacol.* 274, 78–86. doi: 10.1016/j.taap.2013.10.024

Zhang, W., Roederer, M. W., Chen, W.-Q., Fan, L., and Zhou, H.-H. (2012). Pharmacogenetics of drugs withdrawn from the market. *Pharmacogenomics* 13, 223–231. doi: 10.2217/pgs.11.137

Zhang, C., Zhou, Y., Gu, S., Wu, Z., Wu, W., Liu, C., et al. (2016). In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicol. Res. (Camb)* 5, 570–582. doi: 10.1039/c5tx00294j

Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., et al. (2018). Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538.* doi: arXiv:1804.00538v4

Zolotoy, A. B., Plouvier, B. P., Beatch, G. B., Hayes, E. S., Wall, R. A., Walker, M. J. A., et al. (2003). Physicochemical determinants for drug induced blockade of HERG potassium channels: effect of charge and charge shielding. *Curr. Med. Chem. Cardiovasc. Hematol. Agents* 1, 225–241. doi: 10.2174/1568016033477432

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 28 January 2020
doi: 10.3389/fphar.2019.01592

# A Novel Approach for Drug-Target Interactions Prediction Based on Multimodal Deep Autoencoder

*Huiqing Wang*[1*], *Jingjing Wang*[1], *Chunlin Dong*[2], *Yuanyuan Lian*[1], *Dan Liu*[1] *and Zhiliang Yan*[1]

*[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, China, [2] Dryland Agriculture Research Center, Shanxi Academy of Agricultural Sciences, Taiyuan, China*

Drug targets are biomacromolecules or biomolecular structures that bind to specific drugs and produce therapeutic effects. Therefore, the prediction of drug-target interactions (DTIs) is important for disease therapy. Incorporating multiple similarity measures for drugs and targets is of essence for improving the accuracy of prediction of DTIs. However, existing studies with multiple similarity measures ignored the global structure information of similarity measures, and required manual extraction features of drug-target pairs, ignoring the non-linear relationship among features. In this paper, we proposed a novel approach MDADTI for DTIs prediction based on MDA. MDADTI applied random walk with restart method and positive pointwise mutual information to calculate the topological similarity matrices of drugs and targets, capturing the global structure information of similarity measures. Then, MDADTI applied multimodal deep autoencoder to fuse multiple topological similarity matrices of drugs and targets, automatically learned the low-dimensional features of drugs and targets, and applied deep neural network to predict DTIs. The results of 5-repeats of 10-fold cross-validation under three different cross-validation settings indicated that MDADTI is superior to the other four baseline methods. In addition, we validated the predictions of the MDADTI in six drug-target interactions reference databases, and the results showed that MDADTI can effectively identify unknown DTIs.

**Keywords: drug-target interactions, multiple similarity measures, random walk with restart, positive pointwise mutual information, multimodal deep autoencoder**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Vladimir G. Gligorijevic,*
*Flatiron Institute, United States*
*Youjun Xu,*
*Peking University, China*

****Correspondence:***
*Huiqing Wang*
*1013208257@qq.com*

***Specialty section:***
*This article was submitted to Experimental Pharmacology and Drug Discovery, a section of the journal Frontiers in Pharmacology*

***Received:*** *16 June 2019*
***Accepted:*** *09 December 2019*
***Published:*** *28 January 2020*

***Citation:***
*Wang H, Wang J, Dong C, Lian Y, Liu D and Yan Z (2020) A Novel Approach for Drug-Target Interactions Prediction Based on Multimodal Deep Autoencoder. Front. Pharmacol. 10:1592. doi: 10.3389/fphar.2019.01592*

## INTRODUCTION

Drug targets are a kind of biological macromolecule in the body that have a pharmacodynamics function by interacting with drugs, such as certain proteins and nucleic acids. Drugs achieve disease treatment by binding specific targets and changing gene function of their targets. The prediction of drug-target interactions (DTIs) is a crucial process in drug discovery and it can facilitate the understanding of drug action mechanism, disease pathology, and drug side effect (Keiser et al., 2009; Lounkine et al., 2012; Núñez et al., 2012). Drug targets are the main carriers of drug action in drug therapy; thus, the prediction of DTIs is of great significance for disease therapy.

Drug-target interactions prediction can be viewed as a binary classification problem, where the goal is to learn a classifier that can distinguish true and false DTIs. For this problem, drug-drug similarities and target-target similarities are helpful, assuming that similar drugs tend to share similar targets and vice versa (Klabunde, 2007). Many studies applied a single similarity measure of drugs and that of targets, i.e., chemical structural similarity of drugs and amino acid sequence similarity of targets, to predict DTIs (Jacob and Vert, 2008; Yamanishi et al., 2008; Bleakley and Yamanishi, 2009; Xia et al., 2010; van Laarhoven et al., 2011; Gönen, 2012). However, both drugs and targets have different types of similarity measures and they utilize different attributes of drugs and targets, such as gene expression similarities of drugs and target proteins, drug side-effect-based similarity, proximity in protein-protein interactions and so on. It is demonstrated that drugs with similar expression patterns are likely to share common target proteins (Hizukuri et al., 2015; Vilar and Hripcsak, 2016) and drugs with similar target protein binding profiles tend to cause similar side effects, implying a direct correlation between target protein binding and side-effect similarity (Campillos et al., 2008; Hizukuri et al., 2015). Thus, only utilizing chemical structural similarity of drugs and amino acid sequence similarity of targets may miss information that is relevant to predicting new interactions.

With the development of high-throughput sequencing technology, massive multi-omics data have been generated, which provide abundant resources for predicting DTIs, including drug-side-effect association data from SIDER2 (Kuhn et al., 2015), drug-disease association data and target protein-disease association data from KEGG Disease (Kanehisa et al., 2016), protein-protein interaction data from HIPPIE (Alanis-Lobato et al., 2016), etc. Based on these data, a variety of similarity measures for drugs and targets can be calculated, which describe characteristics of drugs and targets from various aspects, and there is information complementarity among them. Thus, methods for predicting DTIs using multiple similarity measures of drugs and multiple similarity measures of targets are generated.

Perlman et al. used forward selection and backward elimination for feature selection. They selected 10 features from 15 features consisting of 5 similarity measures of drugs and 3 similarity measures of targets, and they applied logistic regression classifier to predict DTIs (Perlman et al., 2011). Olayan et al. used multiple similarity networks of drugs and multiple similarity networks of targets to construct a heterogeneous network with the known drug-target interaction network, and then they manually extracted 12 different path-category-based features from it; finally, they applied random forest to predict DTIs (Olayan et al., 2017). Nascimento et al. linearly weighted 10 drug similarity measures and 10 target similarity measures to obtain the feature of drugs and targets, respectively, and then they computed the Kronecker product of them as the feature of drug-target pairs that were fed into Kronecker regularized least squares (KronRLS) to predict DTIs (Nascimento et al., 2016). Hao et al. used Similarity Network Fusion (SNF) method to fuse two similarity measures of drugs and two similarity measures of targets into one drug similarity measure and one target similarity measure, respectively, forming features of drugs and targets, and then input them into dual network integrated logistic matrix factorization (DNILMF) to predict DTIs (Hao et al., 2017). Zheng et al. linearly weighted two similarity measures of drugs and three similarity measures of targets as the feature of drugs and targets, respectively, and then they applied Multiple Similarities Collaborative Matrix Factorization (MSCMF) to predict DTIs (Zheng et al., 2013). Compared with methods using a single similarity measure of drugs and targets, these methods achieved more accurate predictions because of fusing multiple similarity measures.

The similarity measure of drugs (targets) can be regarded as a similarity network with drugs as nodes and drug-drug similarity values as the weights of edges. These methods mentioned above directly applied multiple similarity measures to predict DTIs that only calculated the similarity between two nodes in isolation and did not consider the global topological connectivity patterns within network, ignoring the global structure information of the similarity network. Researches demonstrated that considering the global structure of network can improve the performance (Köhler et al., 2008; Fang and Gough, 2013; Peng et al., 2018). In addition, these methods relied on manual extraction features of drug-target pairs, ignoring the non-linear relationship among features, and failed to provide satisfactory prediction results.

Deep learning is a deep neural network structure with multiple hidden layers. It combines low-level features to form more abstract high-level features, discovering effective feature representations of data. Compared with traditional machine learning methods, the greatest advantage of deep learning methods is that they can extract features automatically, which do not need to perform data processing, such as feature selection, dimension reduction, format conversion, and so on. A number of studies applied deep learning to learn high-level features from the training data automatically and predict bioinformatics tasks (Pan et al., 2016; Deng et al., 2017; Fu and Peng, 2017; Gligorijević et al., 2018). Fu et al. used stacked autoencoder to learn high-level features from miRNA and disease similarity automatically, and then these features were passed to Deep Neural Network (DNN) to predict miRNA-disease associations (Fu and Peng, 2017). Pan et al. extracted raw sequence composition features from RNA and protein sequences, then applied stacked autoencoder to learn hidden high-level features, which are fed into random forest to predict RNA-protein interactions (Pan et al., 2016). These studies demonstrated that deep learning has powerful ability to learn high-level features from original data automatically, which greatly enhanced the performance of the methods and made them show satisfactory results. Gligorijević et al. proposed a new deep learning model-Multimodal Deep Autoencoder (MDA). They applied MDA to learn low-dimensional features of proteins from multiple networks and realized the fusion of multiple networks. Finally, they trained SVM with low-dimensional features of proteins to predict protein functions and achieved great performance (Gligorijević et al., 2018).

Therefore, to automatically learn features from multiple similarity measures to predict DTIs, we introduced MDA and proposed MDADTI, a novel approach for drug-target interactions prediction based on MDA. MDADTI applied Random Walk with Restart (RWR) method and Positive Pointwise Mutual Information (PPMI) to calculate topological similarity matrices of drugs (targets), capturing the global structure information of similarity measures. Then it fused multiple topological similarity matrices of drugs and targets with MDA to automatically learn the low-dimensional features of drugs and targets. Finally, it sent them to Deep Neural Network (DNN) for predicting DTIs. Furthermore, we validated the predictions of the MDADTI in drug-target interactions reference databases.

## MATERIALS AND METHODS

Multiple similarity measures of drugs (targets) describe the drug-drug similarity from various aspects, such as drug side-effects and chemical structure. Multiple similarity measures can provide complementary information for drugs or targets. Combining multiple similarity measures can improve prediction accuracy. Existing methods for predicting DTIs with multiple similarity measures directly took multiple similarity measures as inputs, ignoring their global structure information. Moreover, they required manual extraction features of drug-target pairs, limiting the size of the dataset used to train the model, ignoring the non-linear relationship among features, resulting in the lower predictive performance. Multimodal Deep Autoencoder (MDA) can fuse multiple similarities and learn high-level features automatically. This paper proposed a novel approach MDADTI based on MDA to predict drug-target interactions. MDADTI first applied Random Walk with Restart (RWR) method and Positive Pointwise Mutual Information (PPMI) to calculate topological similarity matrices of drugs (targets), capturing global structural information of each similarity measure; then it fused multiple topological similarity matrices of drugs (targets) with MDA, and realized the automatic learning and dimension reduction of drug features (target features); finally, the extracted low-dimensional features were sent into Deep Neural Network (DNN) to predict DTIs. **Figure 1** shows the overall framework of the MDADTI method.

### Materials

We evaluated the performance of our method with five datasets, including enzyme (E), ion channels (IC), G-protein-coupled receptors (GPCR), nuclear receptors (NR), and DrugBank_FDA. Each dataset contains 3 types of data: (1) DTIs data; (2) multiple similarity measures for drugs; (3) multiple similarity measures for targets.

These five datasets (E, NR, IC, GPCR, and DrugBank_FDA) were provided by Olayan et al., 2017. The DTIs data of E, NR, IC, and GPCR were originally collected by Yamanishi et al., 2008 and have been applied to many drug-target interactions prediction studies (Mei et al., 2012; Ba-Alawi et al., 2016; Lim

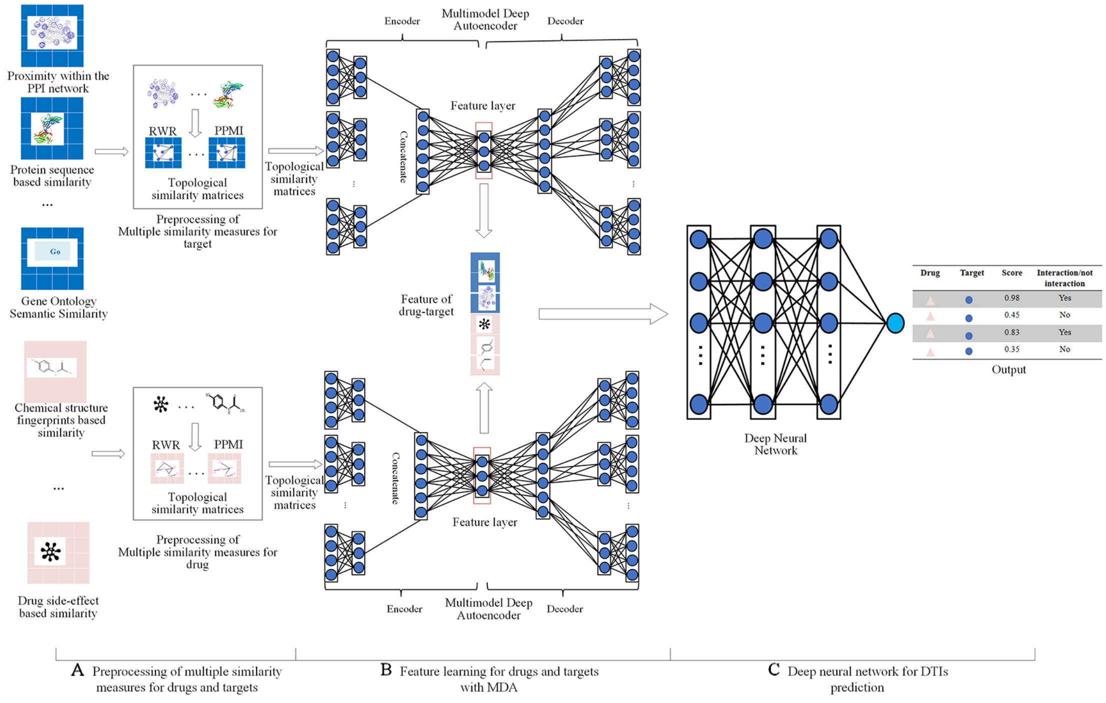

Proximity within the PPI network
Protein sequence based similarity
...
Gene Ontology Semantic Similarity
Go
RWR
PPMI
Topological similarity matrices
Preprocessing of Multiple similarity measures for target
Topological similarity matrices
Encoder
Multimodel Deep Autoencoder
Decoder
Feature layer
Concatenate
Feature of drug-target
Chemical structure fingerprints based similarity
...
Drug side-effect based similarity
RWR
PPMI
Topological similarity matrices
Preprocessing of Multiple similarity measures for drug
Topological similarity matrices
Concatenate
Feature layer
Encoder
Multimodel Deep Autoencoder
Decoder
Deep Neural Network
Drug
Target
Score
Interaction/not interaction
0.98
Yes
0.45
No
0.83
Yes
0.35
No
Output
A Preprocessing of multiple similarity measures for drugs and targets
B Feature learning for drugs and targets with MDA
C Deep neural network for DTIs prediction

**FIGURE 1 |** The overall framework of MDADTI method. **(A)** MDADTI applied RWR method and PPMI to calculate topological similarity matrices of drugs (targets); **(B)** MDA was applied to fuse multiple topological similarity matrices of drugs (targets) and automatically learned the low-dimensional features of drugs (targets); **(C)** DNN was applied to predict DTIs.

et al., 2016; Lu et al., 2017). The multiple similarity measures for drugs and targets we used in this paper in the four datasets were computed by Nascimento et al., 2016 in the first place. DrugBank_FDA dataset was extracted from 5.0.3 version of DrugBank database (Wishart et al., 2007). It only included DTIs information of drugs approved by the FDA and single human target proteins; these proteins are not part of protein complexes. Multiple similarity measures of DrugBank_FDA for drugs and targets were computed by Olayan et al., 2017.

**Table 1** is the summary of drug-target interactions data in five datasets. As can be seen from **Table 1**, the number of negative interactions is larger than that of positive interactions in these five datasets called imbalanced data, which can reduce the predictive performance. Thus, we applied Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to manage the imbalanced datasets. SMOTE can syntactically generate positive samples of these datasets to balance the minority class and enhance the prediction efficiency of the classifier (Waris et al., 2016; Khan et al., 2017).

**Table 2** shows a variety of similarity measures for drugs and targets in five datasets used in this paper. In NR, GPCR, IC, and E datasets, for drugs, the similarities of drugs were calculated based on distinct chemical structure fingerprints, side-effects profiles; nine various similarity measures of drugs were obtained. For targets, various amino acid sequence profiles of proteins, different parameterizations of the Mismatch (MIS) and the Spectrum (SPEC) kernels, and target proteins functional annotation based on Gene Ontology (GO) terms, proximity in the protein-protein interaction (PPI) network were considered as target information source to measure and calculate the similarities of targets; nine various similarity measures of targets were obtained.

In DrugBank_FDA dataset, different similarity measures between drugs were computed based on the following: different types of molecular fingerprints, drug interaction profile, drug side-effects profile, drug profile of the anatomical therapeutic class (ATC) coding system, drug-induced gene expression profile, drug-disease profiles, and drug pathways profiles; 25 various similarity measures of drugs were obtained. Furthermore, different similarity measures of target proteins were calculated based on the following protein amino acid sequence, their GO annotations, proximity in the PPI network, protein domain profiles and gene expression similarity profiles of protein encoding genes; 17 various similarity measures of targets were obtained. Chemical structures of drugs were extracted from DrugBank (Wishart et al., 2007), while the target protein sequences were extracted from UniProt (Boutet et al., 2016).

# Methods

## Problem Description

We defined a set of DTIs and it is composed of a set of drugs D = { $d_i$,$i$ = 1,......, $n_d$ } and a set of targets T = { $t_j$,$j$ = 1,......,$n_t$ }, where $n_d$ represents the number of drugs and $n_t$ represents the number of targets. We also defined the interactions between D and T as a binary matrix Y whose element values are 0 or 1, where $y_{ij} = 1$ represents the drug $d_i$ interacts with the target $t_j$. We defined the set of similarity matrices between drugs in D as $\hat{D}_S$, whose dimensions are $n_d{*}n_d$; Similarly, we also defined the set of similarity matrices between targets in T as $\hat{T}_S$, whose dimensions are $n_t{*}n_t$. Element values in different similarity matrices represent how much drugs or targets are similar to each other based on different measures. The values of all elements in each matrix are in the range of [0, 1]. Our goal is to predict novel (i.e., unknown) interactions in Y based on the matrix Y, similarity matrices of drugs in $\hat{D}_S$ and similarity matrices of targets in $\hat{T}_S$.

## Preprocessing of Multiple Similarity Measures

A similarity matrix of drugs can be regarded as a similarity network with drugs as nodes and drug-drug similarity values as the weights of edges. The similarity network of drugs only calculates the similarity between two drug nodes in isolation and does not consider the relation among more drugs, thus cannot directly include the global structure information of the network. The topological similarity of drugs can describe the topological similarity between all pair of drug nodes in the similarity network. If the topological similarity value between two drug nodes is much larger, it indicates that they have similar positions in the similarity network and have similar functions. The topological similarity of drugs includes both the original information of the similarity network and its global structure information. Therefore, the topological similarity of drugs can solve the problem of losing information caused by original similarity network, which only considers the similarity between two drugs nodes and ignores the global structure of the similarity network. In this paper, we applied Random Walk with Restart (RWR) method and Positive Pointwise Mutual Information (PPMI) (Cao et al., 2016; Fan et al., 2019) to calculate the topological similarity of drugs in each similarity network and capture the global structure information of the similarity network. The detailed process is as follows:

(1) Given a similarity network $\widehat{D_S} = \{S^{(1)},\ldots\ldots,S^{(n)}\}$, we performed RWR on each similarity network $S^{(j)}$ in $\hat{D}_S$ to

**TABLE 1** | Summary of drug-target interaction data.

| Datasets | Number of drugs | Number of targets | Number of positive interactions | Number of negative interactions | Total number of interactions |
|---|---|---|---|---|---|
| NR | 54 | 26 | 90 | 1314 | 1404 |
| GPRC | 223 | 95 | 635 | 20550 | 21185 |
| IC | 210 | 204 | 1476 | 41364 | 42840 |
| E | 445 | 664 | 2926 | 292554 | 295480 |
| DrugBank_FDA | 1482 | 1408 | 9881 | 2076775 | 2086656 |

**TABLE 2 |** Summary of multiple similarity measures of drugs and targets.

| Dataset | Entity | Information source | Similarity measures |
|---|---|---|---|
| | Drug | Chemical structure fingerprints | TAN-Tanimoto Kernel<br>LAMBDA-Lambda-k Kernel<br>MARG-Marginalized Kernel<br>MINMAX-MinMax Kernel<br>SIMCOMP-Graph kernel<br>SPEC-Spectrum Kernel |
| | | Side-effects | AERS-bit-AERS bit<br>AERS-freq-AERS freq<br>SIDER-Side-effects Similarity |
| | Target | Functional annotation | GO - Gene Ontology Semantic Similarity |
| | | Sequences | MIS-k3m1-Mismatch kernel (k = 3, m = 1)<br>MIS-k4m1-Mismatch kernel (k = 4, m = 1)<br>MIS-k3m2-Mismatch kernel (k = 3, m = 2)<br>MIS-k3m2-Mismatch kernel (k = 4, m = 2)<br>SPEC-k3-Spectrum kernel (k = 3)<br>SPEC-k4-Spectrum kernel (k = 4)<br>SW-Smith-Waterman alignment score |
| | | Protein-protein Interactions | PPI-Proximity in protein-protein network |
| DrugBank_FDA | Drug | Molecular fingerprints | CDK_Standard, CDK_Graph, CDK_Extended, CDK_Hybridization, KR, MACCS, PubChem, SIMCOMP, EC4, FC4, EC6, FC6, Lambda, Marginalized, MinMaxTanimoto, Tanimoto, Spectrum |
| | | ATC code | _FDA_FirstLevel,<br>FDA |
| | | Drug interaction profile | D_interactions_FDA |
| | | side-effects | SIDER-Side-effects Similarity |
| | | Drug- induced gene expression | Cmap_v2_MCF7 |
| | | Drug pathways profiles | KEGG_Drug_2_Pathway |
| | | Drug disease profiles | KEGG_Drug_Compound_DGroup_2_Disease |
| | Target | Amino acid sequence | mismatch_kernel_3_1,<br>mismatch_kernel_3_2,<br>mismatch_kernel_4_1,<br>mismatch_kernel_4_2,<br>spectrum_kernel_3,<br>spectrum_kernel_4<br>Merged_SWAlign_Edited |
| | | GO annotations | CC_WANG_BMA_GO_similarity,<br>BP_Wang_BMA_combined,<br>MF_Wang_BMA_combined |
| | | Proximity in the PPI network | shortest_path_networkX_distance_UP_ID_Sim_Perlman,<br>shortest_path_networkX_distance_UP_ID_Sim_Dnorm |
| | | Protein domain profiles | protein2ipr_binaryMatrix_cosSim,<br>protein2ipr_binaryMatrix_jaccardSim |
| | | Gene expression similarity profiles | Cmap_v2_MCF7 |
| | | Target disease profiles | KEGG_Gene_2_Disease |
| | | Target pathway profiles | KEGG_Gene_2_Pathway |

obtain the topology structure feature of drug nodes. The RWR approach can be formulated as the following recurrence relation:

$$p_i^{(t)} = \alpha p_i^{(t-1)} \widehat{S^{(j)}} + (1-\alpha) p_i^{(0)} \tag{1}$$

where $p_i^{(t)}$ is a row vector of drug $i$ and its $e$th element indicates the probability of reaching the $e$th drug node after $t$ steps starting from drug $i$, $p_i^{(0)}$ is the initial one-hot vector, $\alpha$ is the probability of restart, and $\widehat{S^{(j)}}$ is the one-step probability transition matrix obtained by applying row-wise normalization of the similarity matrix $S^{(j)}$. The calculation formula of the topological structure feature of drug node $i$ is as follows:

$$p_i = \sum_{t=1}^{T} p_i^{(t)} \tag{2}$$

where $T$ is the total number of random walk steps. Repeat this process for each node $i$ in the similarity network $S^{(j)}$ to obtain topology feature matrix $P^{(S^{(j)})} \in R^{n_d \times n_d}$.

(2) Based on the topological structure feature matrix $P^{(S^{(j)})}$, we applied PPMI (Chen et al., 2016) to calculate the topological similarity between all pair of nodes, and obtained the topological similarity matrix $X^{(S^{(j)})} \in R^{n_d \times n_d}$ of the similarity network $S^{(j)}$, capturing the global structure information. The topological similarity between node $i$ and node $k$ is defined as:

$$X_{ik}^{(S^{(j)})} = \max\left(0, \log_2 \frac{P_{ik}^{(S^{(j)})} \sum_i \sum_k P_{ik}^{(S^{(j)})}}{\sum_i P_{ik}^{(S^{(j)})} \sum_k P_{ik}^{(S^{(j)})}}\right) \tag{3}$$

where $P_{ik}^{(S^{(j)})}$represents the elements of the $i$th row and the $k$th column of the topological structure feature matrix $\mathrm{P}^{(S^{(j)})}$.

The preprocessing procedure for multiple similarity measures of targets is the same as that of drugs.

## Feature Learning for Drugs and Targets With MDA

Fusing multiple drug-drug similarity measures and multiple target-target similarity measures contributes to obtaining abundant information about drugs and targets. Capturing non-linear relationships among features can improve the accuracy of DTIs prediction. Therefore, we applied MDA to fuse multiple similarity measures of drugs and targets and automatically learn low-dimensional feature matrices of drugs and targets, respectively, capturing the non-linear relationship among features. After the pretreatment, we obtained multiple topological similarity matrices of drugs $X^{(S^{(j)})} \in R^{n_d \times n_d}, j \in [1, \ldots \ldots n]$ that contain both original information of similarity measures and their global structure information. In this paper, we applied MDA to fuse multiple topological similarity matrices of drugs and automatically learn the low-dimensional feature matrix of drugs $H_c^{(d)} \in R^{n_d \times d}$. As an unsupervised neural network model, MDA uses backpropagation algorithm to train and adjust the model parameters, so that the input data can still be restored to the original features by encoding and decoding process. The structure of MDA is shown in **Figure 2**.

Encoding is the process that MDA learns the hidden features of input data with multi-layer non-linear functions. We first calculated the non-linear embedding $H^{(S^{(j)})}$ of each topological similarity matrix $X^{(S^{(j)})}$ in the first hidden layer of MDA:

$$H^{(S^{(j)})} = \sigma\left(W_1^{(S^{(j)})} X^{(S^{(j)})} + B_1^{(S^{(j)})}\right) \tag{4}$$

where $W_1^{(S^{(j)})}$ and $B_1^{(S^{(j)})}$ are weight matrix and bias matrix, j∈[1,…n], $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function.

Then, we computed the low-dimensional feature matrix of drugs $H_c^{(d)} \in R^{n_d \times d}$ by applying multiple non-linear functions (i.e., multiple hidden layers) on the feature representation obtained by concatenating features from all topological similarity matrices obtained in the previous layer:

$$H_c^{(d)} = \sigma\left(W_1\left[H^{(S^{(1)})}, \ldots\ldots, H^{(S^{(n)})}\right] + B_1\right) \tag{5}$$

where $[H^{(s^{(1)})}, \ldots\ldots, H^{(s^{(n)})}]$is the concatenated matrix of N embedding $H^{(S^{(j)})}$ obtained in the previous layer; $W_1$ and $B_1$ are weight matrix and bias matrix, and $\sigma(x)$ is the sigmoid activation function.

Decoding is the process that MDA reconstructs input data from hidden features with multi-layer non-linear functions. Hidden features are obtained through encoding process. We reconstructed multiple topological similarity matrices $\widehat{X^{(S^{(j)})}}$ from the feature matrix $H_c^{(d)}$ of drugs with a multi-layer non-linear function:

$$\widehat{X^{(S^{(j)})}} = \sigma\left(W_2^{(S^{(j)})} H_c^{(d)} + B_2^{(S^{(j)})}\right) \tag{6}$$

where $W_2^{(S^{(j)})}$ and $B_2^{(S^{(j)})}$ are weight matrix and bias matrix, j∈ [1, ……, n], $\sigma(x)$ is the sigmoid activation function.

To get the feature matrix of drugs $H_c^{(d)}$, MDA obtained the unknown parameters θ in the encoding and decoding process by minimizing the reconstruction error of $X^{(S^{(j)})}$ and $\widehat{X^{(S^{(j)})}}$:

$$\hat{\theta} = argminL(\theta) = {}^{argmin}_{\quad\theta} \sum_{j=1}^{n} loss\left(X^{(S^{(j)})}, \widehat{X^{(S^{(j)})}}\right) \tag{7}$$

where $\theta = \{W_1, B_1, W_1^{(S^{(j)})}, B_1^{(S^{(j)})}, W_2^{(S^{(j)})}, B_2^{(S^{(j)})}\}$ is the set of unknown parameters in the encoding and decoding process, and $n$ represents the number of drug topological similarity matrices, and $loss(*)$ is cross-entropy function.

The learning process of the feature matrix $H_c^{(t)}$of targets is the same as that of feature matrix $H_c^{(d)}$of drugs. The hyperparameters of training MDA include epoch, batch size, and learning rate with values of 100, 32, and 0.001, respectively.

## Deep Neural Network for DTIs Prediction

We formulated the problem of DTIs prediction as a binary classification problem. We introduced Deep Neural network

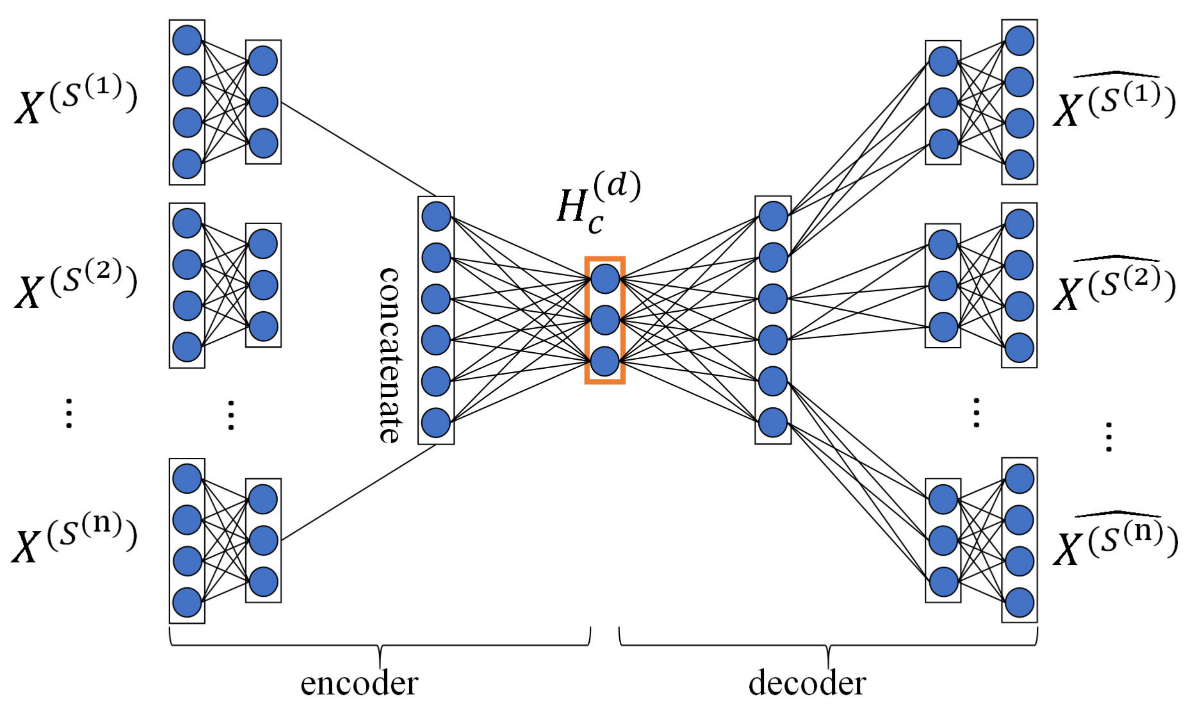


**FIGURE 2 |** Structure diagram of MDA. The MDA consists of two parts: encoder and decoder, the inputs of encoder are multiple topological similarity matrices $X^{(S^{(i)})}$, the hidden layer in the red box is feature layer whose output is the low-dimensional feature matrix of drugs $H_c^{(d)}$, the output of decoder are multiple reconstructed topological similarity matrices $\widehat{X^{(S^{(i)})}}$.

(DNN) to predict DTIs. The DNN of our method consists of 5 fully-connected layers, including 1 input layer, 3 hidden layers, and 1 output layer. The choice of the number of hidden layers depends on experiments. After a lot of experiments, MDADTI achieved best predicted results when DNN consists of 3 hidden layers and the number of each layer is 300, 200, and 100. All the neuron units in the layer *i* are connected to the previous layer (*i-1*) and then generated outputs with non-linear transformation function *f*:

$$o_j = f\left(\sum_{i=1}^{H} w_i o_i + b_i\right) \qquad (8)$$

where *H* is the number of neurons in hidden layer; $\{w_{i,} b_i\}_{i=1}^{H}$ are the weights and bias of neuron j which sums up all the hidden units; *f*(*) is Relu activation function, which is a non-linear function that can capture hidden patterns in the input data (Chen et al., 2016) and can reduce gradient vanishing at the same time.

In order to predict DTIs, we concatenated the feature matrix of drugs $H_c^{(d)}$and the feature matrix of targets $H_c^{(t)}$to get the feature matrix of drug-target pairs $H_c$. Then we used $H_c$ to train DNN, and the final output layer utilized *sigmoid* $= \frac{1}{1+e^{-x}}$ function to predict the interaction possibility of the drug-target pair. If the probability exceeds 0.5, we determine that there is potential interaction between the drug and the target.

## Model Training

MDADTI was trained using the Keras 1.0.1 library with Tensorflow as the backend. The model utilized a backpropagation algorithm to calculate the loss function value between the output and the label, then it calculated its gradient relative to each neuron, and updated the weight according to the gradient direction. We chose cross-entropy function as the loss function:

$$C = -\frac{1}{n}\sum_{x}\sum_{t}\left[y \ln \alpha + (1-y)\ln(1-\alpha)\right] \qquad (9)$$

where *C* is the output of cross-entropy cost function, *x* represents the index of the training samples (i.e., drug-target pairs), *t* represents the index of different labels, *y* represents the true label for sample *x* whose value is 0 or 1, and *a* represents the predicted output for sample *x*. Since the closer the predicted output is to the true label, the smaller C value we can get, our goal is to minimize the cross-entropy function to get the best prediction of DTIs.

In the process of training the model, choosing a good optimizer not only accelerate the training of the model but also contribute to obtaining relatively good experimental results. It is observed that momentum-based stochastic gradient descent (SGD) can effectively train deep learning models (Sutskever et al., 2013). Thus, we chose SGD with momentum as optimizer to minimize the objective function.

Overfitting is a common problem in deep learning. It means that the model works well on the training set, and its predictive effect on the test set is poor, which results in weak generalization ability of the model. We used Dropout and EarlyStopping to prevent overfitting. Dropout (Srivastava et al., 2014) is a common regularization technique in neural networks, referring to randomly 'dropping' (i.e., setting to zero) the output of a neuron with some fixed probability $p$. It means that the start-up effects on the downstream of these neurons are neglected in the forward propagation, and the weights are not updated in the backpropagation. The effect of dropout is that the network is less sensitive to the change of the weight of a certain neuron; it also leads to increased generalization ability and reduced overfitting. We used dropout in each fully connected layer of DNN and set dropout rate of $p = 0.5$, which seems to be close to optimal for a wide range of networks and tasks (Srivastava et al., 2014). EarlyStopping refers to stopping training model when the performance of the model on the validation set begins to decline. Thus, the overfitting problem caused by overtraining can be avoided. We implemented EarlyStopping by training our model with the training set and computing the accuracy on the validation set. We monitored the accuracy of MDADTI on validation set at the end of every epoch and stop the training when accuracy does not rise for 10 consecutive epochs.

# RESULTS

## Experimental Setup and Model Evaluation

In this paper, we applied the area under the ROC (receiver-operating characteristics) curve (AUC) and the area under the precision-recall curve (AUPR) to evaluate the performance of MDADTI model. An AUC value of 1 indicates that the performance is perfect, and an AUC value of 0.5 indicates random predictive performance. Similar to the AUC score, AUPR values closer to 1 indicates that the performance is better. The calculation formulas for True Positive Rate (TPR), False Positive Rate (FPR), and precision and recall related to AUC and AUPR are as follows:

$$\mathrm{TPR} = \mathrm{recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{10}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN}) \tag{11}$$

$$\mathrm{precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{12}$$

where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative; these formulas are based on the confusion matrix.

The performance of DTIs prediction methods was evaluated under 5-repeats of 10-fold cross-validation (CV), and both AUC and AUPR were used as the evaluation metrics. We calculated an AUC score in each repetition of CV and reported a final AUC score that was the average over the five repetitions. The AUPR score was calculated in the same manner. The drug-target interaction matrix Y has $n_d$ rows for drugs and $n_t$ columns for targets. We conducted CV under three different settings as follows:

- CVS1: CV on drug-target pairs—random entries in Y (i.e., drug-target pairs) were selected for testing.
- CVS2: CV on drugs—random rows in Y (i.e., drugs) were blinded for testing.
- CVS3: CV on targets—random columns in Y (i.e., targets) were blinded for testing.

Under CVS1, we applied 5-repeats of stratified 10-fold cross-validation to evaluate the performance of MDADTI model. In each round, we used 90% of elements in Y as training data and the remaining 10% of elements as test data. Under CVS2, in each round, we used 90% of rows in Y as training data and the remaining 10% of rows as test data. Under CVS3, in each round, we used 90% of columns in Y as training data and the remaining 10% of columns as test data. These three settings CVS1, CVS2 and CVS3 refer to the DTIs prediction for 1) new (unknown) pairs, 2) new drugs, and 3) new targets, respectively.

For datasets GPCR, IC, E, and DrugBank_FDA, in order to determine the layer configurations of MDA (the number of layers and the number of neurons in each layer) in MDADTI model, we applied 5-repeats of stratified 10-fold cross-validation under CVS1 to evaluate the performance of MDADTI models with different layer configurations of MDA. Stratified 10-fold cross-validation can make the category ratio in each fold be consistent with that in the whole dataset.

For the small dataset NR, considering the overfitting problem on the small dataset of MDADTI model, for each CV setting, we applied transfer learning strategy (Pan and Yang, 2009) to predict DTIs. We first pretrained MDADTI model under CVS1 setting with the drug-target interactions in the E dataset. Then we froze all layers of the pretrained models except the output layer, i.e., only set weights of the output layer to be trainable. Finally, we finetuned the pretrained model with drug-target interactions data in NR dataset and predicted DTIs under CVS1. The transfer learning process under CVS2 and CVS3 settings are the same as that under CVS1 setting.

In order to focus on the differences between MDADTI and other methods on NR, GPCR, IC, E, and DrugBank_FDA datasets, we applied 5-repeats of 10-fold cross-validation under three different settings to compare the performance of MDADTI with DDR (Olayan et al., 2017), KronRLS-MKL(Nascimento et al., 2016), NRLMF(Liu et al., 2016), and BLM-NII (Mei et al., 2012).

## The Results of MDADTI With Different Layer Configurations of MDA

For GPCR, IC, E, and DrugBank_FDA datasets, in order to determine the layer configurations of two MDAs for extracting drug and target features in the MDADTI model, we applied 5-

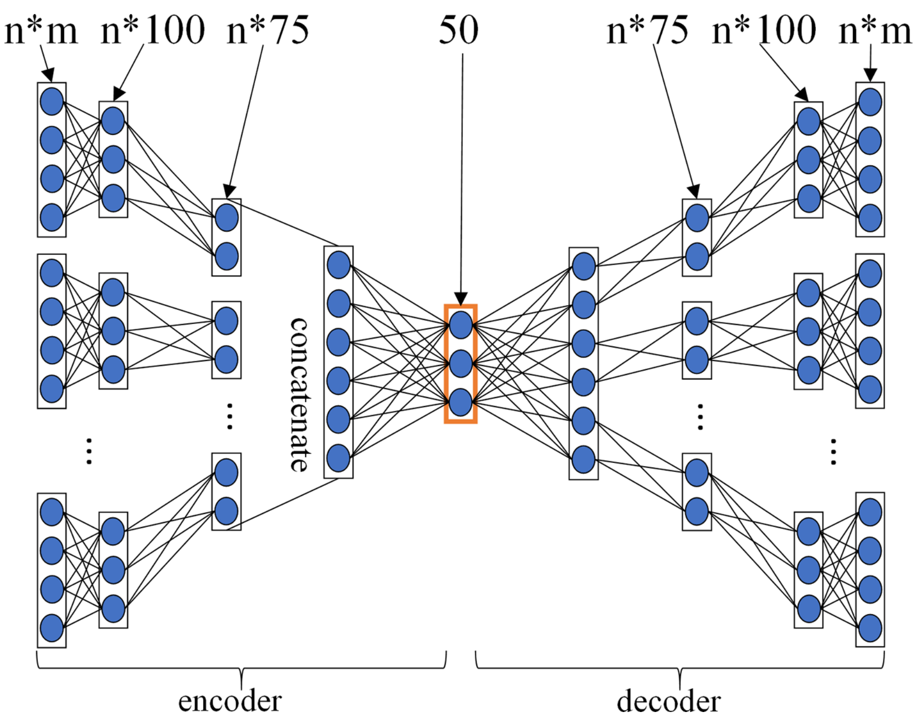


**FIGURE 3 |** The layer configurations diagram of MDA. The layer configurations are [n∗m, n∗100, n∗75, 50, n∗75, n∗100, n∗m]. It consists of 7 layers of neurons, including 1 input layer n∗m, where n is the number of input similarity measures, and m is the number of columns of each similarity matrix, i.e. the number of drugs (targets), 1 output layer n∗m, 2 encoding layers n∗100 and n∗75, 2 decoding layers n∗75 and n∗100, 1 feature layer with 50 neurons.

repeats of 10-fold cross-validation under CVS1 to evaluate the performance of MDADTI models with different layer configurations of MDAs. **Figure 3** is a layer configurations diagram of MDA whose layer configurations is [n∗m, n∗100, n∗75, 50, n∗75, n∗100, n∗m]. It consists of 7 layers of neurons, including 1 input layer n∗m, where n is the number of input similarity measures and m is the number of columns of each similarity matrix, i.e., the number of drugs (targets); 1 output layer n∗m, 2 encoding layers n∗100 and n∗75, 2 decoding layers n∗75 and n∗100, and 1 feature layer with 50 neurons.

For each dataset, we input all similarity measures listed in **Table 2** into three different MDADTI models whose layer configurations of two MDAs are different, and then we trained them to predict DTIs. The performance of MDADTI models for different layer configurations of two MDAs in four datasets under 5-repeats of 10-fold cross-validation is provided in **Table 3**; $n_\mathrm{d}$ and $n_\mathrm{t}$ are the number of drugs and targets in each of the four datasets, respectively. The AUC and AUPR values in bold are highest among three sets of evaluation indicator values corresponding tree different layer configurations of MDAs.

From **Table 3** we observed that for GPCR dataset, MDADTI achieved the highest AUC and AUPR when two MDAs have only one feature layer. Therefore, the MDA extracting the drug features is configured as $[\ n{*}n_d{,}50{,}n{*}n_d\ ]$, and the MDA extracting target features is configured as $[\ n{*}n_t{,}25{,}n{*}n_t\ ]$ when we applied MDADTI to predict DTIs in GPCR dataset. The AUC and AUPR of MDADTI are 0.980 and 0.978, respectively. For IC dataset, MDADTI achieved the highest AUC and AUPR when two MDAs have only one feature layer. Therefore, the MDA extracting drug features is configured as$[\ n{*}n_d{,}50{,}n{*}n_d\ ]$, and the MDA extracting target features is configured as $[\ n{*}n_t{,}50{,}n{*}n_t\ ]$ when we applied MDADTI to predict DTIs in IC dataset. The AUC and AUPR of MDADTI are 0.991 and 0.987, respectively. For E dataset, MDADTI achieved the highest AUC and AUPR when two MDAs have 1 encoding layer, 1 feature layer, and 1 decoding layer. Therefore, the MDA extracting drug features is configured as $[\ n * n_d, n * 200, 100, n * 200, n * n_d\ ]$ and the MDA extracting target features is configured as $[\ n * n_t, n * 200, 100, n * 200,\ n\ *\ n_t\ ]$ when we applied MDADTI to predict DTIs in E dataset. The AUC and AUPR of MDADTI are 0.983 and 0.980, respectively. For DrugBank_FDA dataset, MDADTI achieved the highest AUC and AUPR when two MDAs have 1 encoding layer, 1 feature layer, and 1 decoding layer. Therefore, the MDA extracting drug features are configured as $[\ n{*}n_d{,}n{*}200{,}100{,}n * 200, n * n_d\ ]$and the MDA extracting target features are configured as $[\ n\ *\ n_t,\ n\ *\ 200\ ,\ 100,\ n\ *\ 200,\ n\ *\ n_t\ ]$ when we applied MDADTI to predict DTIs in DrugBank_FDA dataset. The AUC and AUPR of MDADTI are 0.963 and 0.959, respectively.

For the small dataset NR, we applied transfer learning strategy to predict DTIs, and also applied 5-repeats of 10-fold cross-validation to evaluate MDADTI and obtain AUC and AUPR. Finally, we obtained AUC and AUPR of MDADTI in NR, GPCR, IC, E, and DrugBank_FDA datasets. AUC are 0.966, 0.980, 0.991, 0.983, and 0.963, respectively; AUPR are 0.959, 0.978, 0.987,

**TABLE 3 |** The comparison results of MDADTI models with different layer configurations of two MDAs under 5-repeats of 10-fold cross-validation on four datasets. The AUC and AUPR values in bold are highest among three sets of evaluation indicator values corresponding tree different layer configurations of MDAs in each dataset.

| Datasets | | Different layer configurations of MDAs | AUC | AUPR |
|---|---|---|---|---|
| GPCR | drug | $[n*n_d,50,n*n_d]$ | **0.980** | **0.978** |
| | target | $[n*n_t,25,n*n_t]$ | | |
| | drug | $[n*n_d,n*75,50,n*75,n*n_d]$ | 0.965 | 0.963 |
| | target | $[n*n_t,n*50,25,n*50,n*n_t]$ | | |
| | drug | $[\ n*n_d,n*150,n*75,50,n*75,n*150,n*n_d]$ | 0.930 | 0.925 |
| | target | $[\ n*n_t,n*75,n*50,25,n*50,n*75,n*n_t]$ | | |
| IC | drug | $[n*n_d,50,n*n_d]$ | **0.991** | **0.987** |
| | target | $[\ n*n_t,50,n*n_t]$ | | |
| | drug | $[n*n_d,n*75,50,n*75,n*n_d]$ | 0.944 | 0.923 |
| | target | $[n*n_t,n*75,50,n*75,n*n_t]$ | | |
| | drug | $[\ n*n_d,n*150,n*75,50,n*75,n*150,n*n_d]$ | 0.914 | 0.906 |
| | target | $n*n_t,n*150,n*75,50,n*75,n*150,n*n_t]$ | | |
| E | drug | $[n*n_d,100,n*n_d]$ | 0.956 | 0.947 |
| | target | $[n*n_t,100,n*n_t]$ | | |
| | drug | $[n*n_d,n*200,100,n*200,n*n_d]$ | **0.983** | **0.980** |
| | target | $[n*n_t,n*200,100,n*200,n*n_t]$ | | |
| | drug | $[n*n_d,n*300,n*200,100,n*200,n*300,n*n_d]$ | 0.893 | 0.886 |
| | target | $[n*n_t,n*300,n*200,100,n*200,n*300,n*n_t]$ | | |
| DrugBank_FDA | drug | $[n*n_d,100,n*n_d]$ | 0.925 | 0.912 |
| | target | $[n*n_t,100,n*n_t]$ | | |
| | drug | $[n*n_d,n*200,100,n*200,n*n_d]$ | **0.963** | **0.959** |
| | target | $[n*n_t,n*200,100,n*200,n*n_t]$ | | |
| | drug | $[n*n_d,n*300,n*200,100,n*200,n*300,n*n_d]$ | 0.946 | 0.938 |
| | target | $[n*n_t,n*300,n*200,100,n*200,n*300,n*n_t]$ | | |

0.980, and 0.959, respectively. **Figure 4** shows the ROC curve and precision-recall curve of the first repeat of 10-fold cross-validation in five datasets. The mean_AUC and mean_AUPR are the average AUC and average AUPR of MDADTI in the first repeat of 10-fold cross-validation. The train/valid accuracy-epoch and loss-epoch curves for each dataset are provided in **Figure S1** of **Supplementary Material** while selecting fold1 as the test set and the remaining as train set when we performed the first repeat of 10-fold cross-validation. From **Figure S1** we can observe that the change law of accuracy and loss of our model while validating is consistent with that while training, which demonstrates that overfitting has been effectively processed for each dataset. The hyperparameters of MDADTI model for each dataset under CVS1 setting are provided in **Table S1** of the **Supplementary Material**.

## Comparisons With Other Methods

In order to focus on the differences between MDADTI and other methods on NR, GPCR, IC, and E datasets under three different CV settings, we provided a detailed comparison with DDR, KronRLS-MKL, NRLMF, and BLM-NII methods. For DrugBank_FDA dataset, we only compared our method MDADTI with BLM-NII and NRLMF under three cross-validation settings because of the large amount of data in DrugBank_FDA dataset and the high-complexity of DDR and KronRLS-MKL methods, resulting in their longer runtime than our method.

DDR: First, it applied a similarity selection procedure to select a set of informative and less-redundant set of similarities for drugs and for target proteins. Then it manually extracted 12 different path-category-based feature matrices from the heterogeneous network, which consists of known drug-target interaction network and similarity networks for drugs and targets. Finally, it sent feature matrices to the Random Forest (RF) to predict DTIs.

KronRLS-MKL: First, it computed the weighted combination of multiple drug kernels and target kernels to get the final drug kernel and target kernel, then it computed the Kronecker product of final drug kernel and target kernel as the drug-target pairwise kernel. Finally, it applied Kronecker regularized least squares (KronRLS) to predict DTIs.

NRLMF: NRLMF represented the properties of a drug and a target as two latent vectors in the shared low dimensional latent space. For each drug-target pair, the interaction probability is modeled by a logistic function of the drug-specific and target-specific latent vectors. Moreover, the neighborhood regularization based on the drug similarities and target similarities is utilized to further improve the prediction ability of the model.

BLM-NII: BLM-NII integrated Bipartite Local Model (BLM) method with a neighbor-based interaction-profile inferring (NII) procedure to form a DTI prediction approach, where the RLS classifier with GIP kernel was used as the local model.

For comparison with these methods under CVS1 setting, we used 5-repeats of 10-fold cross-validation based on drug-target pairs to evaluate the predictive performance of DDR, KronRLS-MKL, NRLMF, and BLM-NII. **(Figure 5A)** shows the comparison of AUC and AUPR of MDADTI, DDR, KronRLS-MKL, NRLMF, and BLM-NII on five datasets under CVS1 setting. It can be seen from the figure that the performance of MDADTI has improved compared with the other methods. For NR, GPCR, IC, and E datasets, the growth rates of AUC of MDADTI compared to DDR, KronRLS-MKL, NRLMF, and BLM-NII are as follows: (NR: 4.43%, 9.65%, 1.79%, 6.74%),

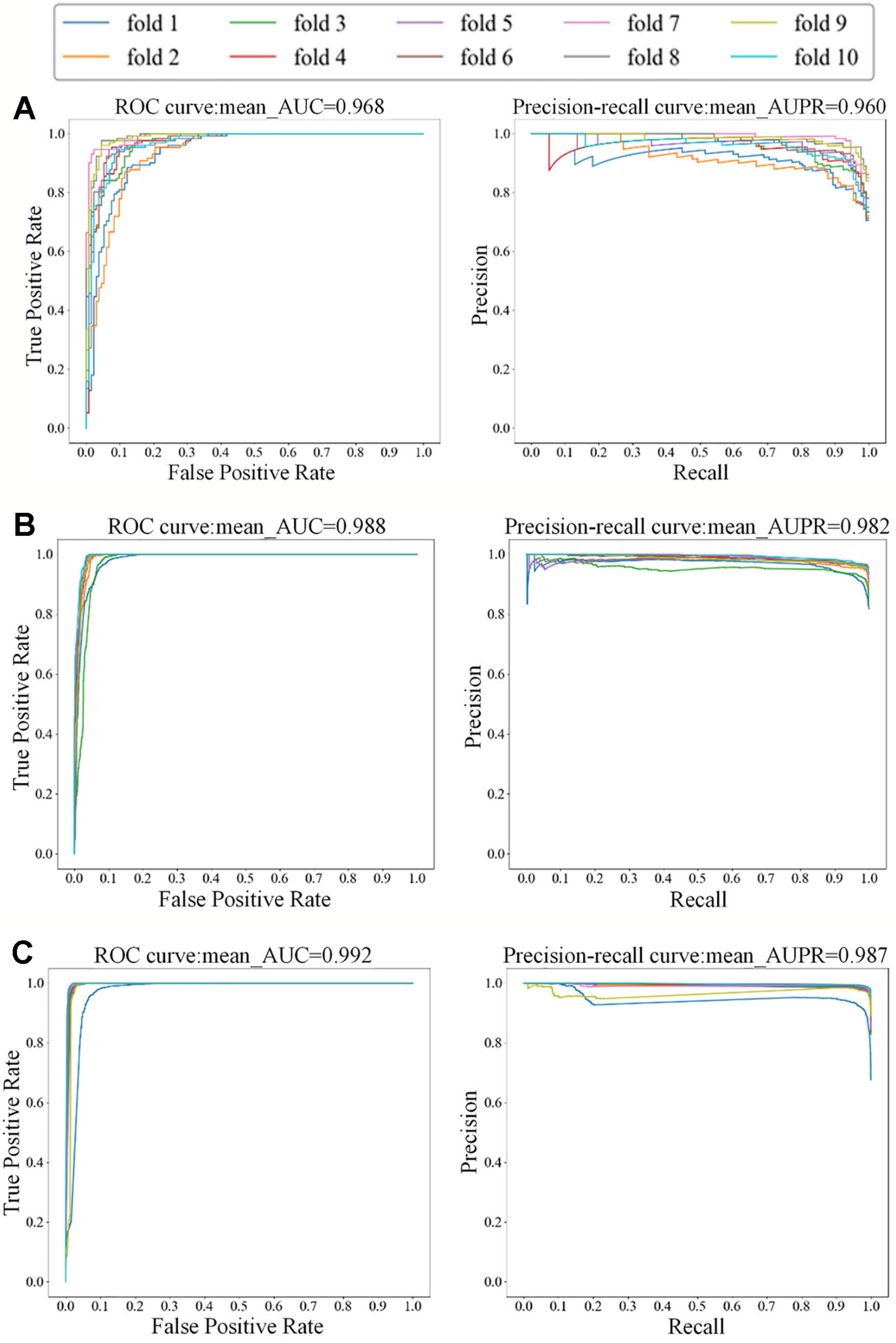


**FIGURE 4** | Continued

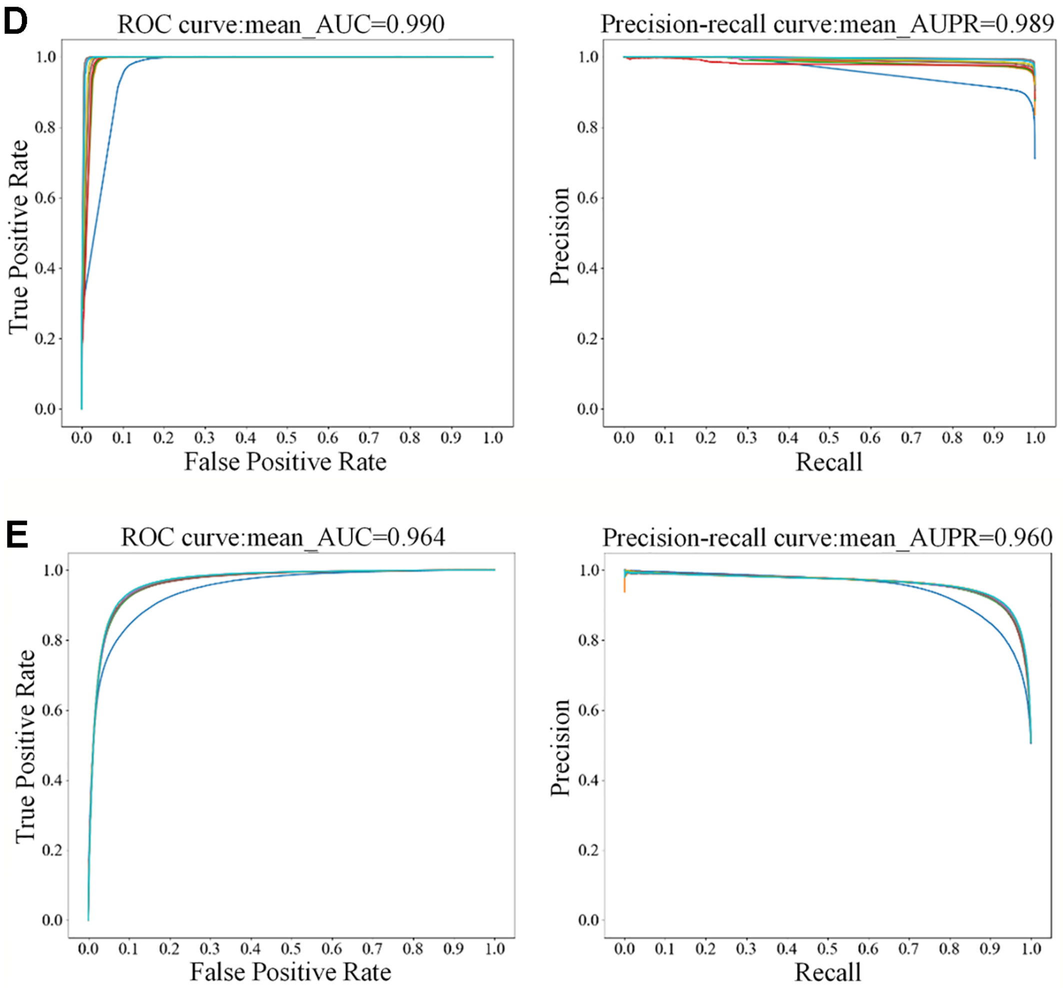


**FIGURE 4** | The ROC and precision-recall curves of the first repeat of 10-fold cross-validation for five datasets; the left is the ROC curve and the right is the precision-recall curve. **(A)** The ROC and precision-recall curves for NR dataset; **(B)** The ROC and precision-recall curves for GPCR dataset; **(C)** The ROC and precision-recall curves for IC dataset; **(D)** The ROC and precision-recall curves for E dataset; **(E)** The ROC and precision-recall curves for DrugBank_FDA dataset.

(GPCR: 1.77%, 3.16%, 2.08%, 3.81%), (IC: 0.20%, 0.92%, 0.71%, 1.02%), and (E: 1.03%, 0.61%, 0.72%, 1.34%). The growth rates of AUPR of MDADTI compared to DDR, KronRLS-MKL, NRLMF, and BLM-NII are as follows: (NR: 14.98%, 82.31%, 32.64%, 45.52%), (GPCR: 22.25%, 44.03%, 39.12%, 89.90%), (IC: 5.34%, 10.16%, 14.37%, 20.22%), and (E: 2.17%, 9.25%, 11.87%, 39.20%). For DrugBank_FDA dataset, we only compared our method with NRLMF and BLM-NII, and the growth rates of AUC of MDADTI compared to NRLMF and BLM-NII are 7.96% and 34.87%, respectively. In terms of AUPR, our method has improved by 213.40% than NRLMF that performs better between these two baselines methods.

The experimental results show that MDADTI is superior to DDR, KronRLS-MKL, NRLMF, and BLM-NII under CVS1 setting. The above comparison does not guarantee the efficacy and superiority of our proposed method. The possibility of getting good results by chance cannot be ignored. Thus, we performed paired t-test at significance level p = 0.05 to check if the differences between our method and the other methods are statistically significant or not under CVS1 setting. The specific details are as follows: we obtained 50 AUCs and 50 AUPRs for each method after performing five repeats of 10-fold cross-validation. In order to check if the differences between our method and each of baseline methods are statistically significant or not, i.e., check if mean AUCs (AUPRs) (mean AUC is the mean value of 50 AUCs) of them have significant differences, for each baseline method, we performed paired t-test based on 50 AUCs (AUPRs) of our method MDADTI and 50 AUCs (AUPRs) of the baseline method, respectively. We also combined bootstrap method to increase the sample size and used 2000 bootstrap samples for performing paired t-test.

The p-values of AUC and AUPR between our method and the other methods under CVS1 setting are reported in **Table S4(a)** of **Supplementary Material**, whereas p-values are less than 0.05 to

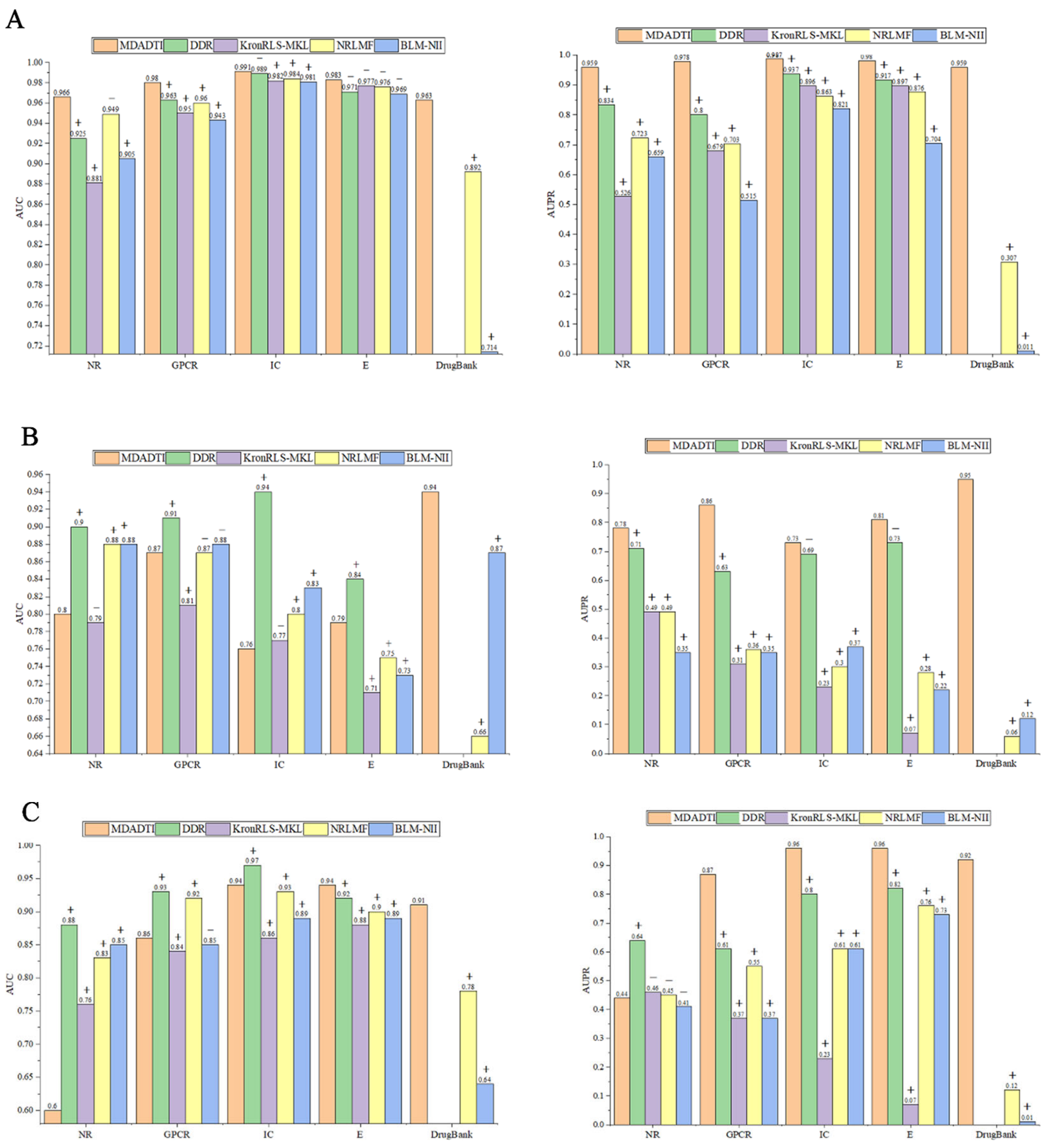


**FIGURE 5 |** Comparison of AUC and AUPR among MDADTI, DDR, KronRLS-MKL, NRLMF, and BLM-NII methods on NR, GPCR, IC, E, and Drugbank_FDA datasets under CVS1, CVS2, and CVS3 setting. **(A)** Comparison of AUC and AUPR under CVS1 setting; **(B)** Comparison of AUC and AUPR under CVS2 setting; **(C)** Comparison of AUC and AUPR under CVS3 setting. The symbols +/- denote if the differences between our method MDADTI and other methods are statistically significant (+) or not (-) at the significance level of 0.05.

demonstrate the statistical superiority of our method. For NR, GPCR, IC, and E dataset, in terms of AUC, from **Table S4(a)** we can observe that MDADTI outperforms the other baseline methods, being statistically significant in most cases at the significance level of 0.05, except one comparison case with DDR in IC dataset, one comparison case with NRLMF in NR dataset, and four comparison cases with the other four competing methods in E dataset. In terms of AUPR, we can observe that MDADTI outperforms other baseline methods, being statistically significant in all cases at the significance level of 0.05. For DrugBank_FDA dataset, in terms of AUC, we can see that our method performs best, and it outperforms NRLMF and BLM-NII methods, being statistically significant at the significance level of 0.05. In terms of AUPR, we can see that our method also performs best, and it outperforms NRLMF and BLM-NII methods, being statistically significant.

For comparison with these methods under CVS2 and CVS3 setting, we used 5-repeats of 10-fold cross-validation based on drugs and targets to evaluate the predictive performance of MDADTI, DDR, KronRLS-MKL, NRLMF and BLM-NII. The hyperparameters of MDADTI model for each dataset under CVS2 and CVS2 settings are provided in **Table S2** and **Table S3** of **Supplementary Material**, respectively. The comparison of AUC and AUPR among MDADTI, DDR, KronRLS-MKL, NRLMF, and

BLM-NII methods on NR, GPCR, IC, E, and DrugBank_FDA datasets under CVS2 and CVS3 settings are provided in **Figure 5B** and **Figure 5C**. The screenshot of MDADTI when it predicts DTIs in GPCR dataset under CVS2 setting is provided in **Figure S2** of **Supplementary Material**. The program flow chart of the code of MDADTI under CVS1,CVS2,and CVS3 settings is provided in **Figure S3** of **Supplementary Material**.

From **Figure 5B** we can see that under CVS2 setting in NR, GPCR, IC, and E datasets, for AUPR, the performance of MDADTI is improved on these four datasets, and the growth rates are 9.86%, 36.51%, 5.80%, 10.96%, respectively, compared with the best method DDR among the baseline methods. For AUC, compared with these four methods, MDADTI performed better on E dataset, which was 11.26%, 5.33%, and 8.22% higher than KronRLS-MKL, NRLMF, and BLM-NII, respectively; MDADTI performed moderately on the other three datasets. For DrugBank_FDA dataset, our method performed better than NRLMF and BLM-NII methods in both AUC and AUPR.

Under CVS3 setting, from **Figure 5C** we can see that for NR, GPCR, IC, and E datasets, the AUPR of MDADTI has a certain improvement in GPCR, IC, and E datasets. The AUPR has increased by 42.62%, 20%, and 17.07% in GPCR, IC, and E datasets, respectively, compared with the best performing method DDR; for AUC, the AUC of MDADTI is the highest in E dataset compared with the other methods, which was increased by 2.17% than the AUC of DDR. For IC dataset, the AUC of MDADTI was increased by 9.30%, 1.08%, and 5.62% compared with KronRLS-MKL, NRLMF, and BLM-NII, respectively. Our method performed moderately in the GPCR dataset, but our method performed poorly on NR dataset. After analysis, it is found that the data volume of the E dataset is 295480 and the large amount of samples make deep learning model perform better; however, the data volume of the NR dataset is only 1404, which is relatively small and does not meet the requirements for data volume of deep learning. Although we applied transfer learning method to predict DTIs of NR datasets under CVS3 setting, the train set of CVS3 setting contains relatively little information, which affects the effect of transfer learning and leads to poor prediction results. For DrugBank_FDA dataset, our method performed better than NRLMF and BLM-NII methods in both AUC and AUPR.

As a kind of data-driven method, deep learning methods are superior to traditional machine learning methods when the amount of data is quite large. By comparing the performance of our method on the five datasets, our method performed best on E and DrugBank_FDA datasets and performed worst on NR dataset, which is consistent with the theory of deep learning.

Similar to CVS1 setting, we performed paired t-test at significance level p = 0.05 to check if the differences between our method and the other methods are statistically significant or not under CVS2 and CVS3 settings. The p-values of AUC and AUPR between our method and the other methods under CVS2 and CVS3 settings are tabulated in **Table S4(b)** and **Table S4(c)** of **Supplementary Material**, respectively.

For CVS2 setting, in terms of AUC, from **Table S4(b)** we can see that our method outperforms KronRLS-MKL in GPCR dataset, being statistically significant at the significance level of 0.05, and it also outperforms KronRLS-MKL, NRLMF, and BLM-NII in E dataset. For DrugBank_FDA dataset, our method MDADTI performs best compared with NRLMF and BLM-NII methods, and it outperforms them, being statistically significant. In terms of AUPR, from **Table S4(b)** we can see that our method performs best in five datasets and it outperforms other baseline methods, being statistically significant in most cases at the significance level of 0.05, except two comparison cases with DDR in IC and E datasets.

For CVS3 setting, in terms of AUC, from **Table S4(c)** we can see that our method outperforms KronRLS-MKL method in GPCR dataset, being statistically significant at the significance level of 0.05. Our method also outperforms KronRLS-MKL, NRLMF, and BLM-NII methods in IC dataset, being statistically significant. For E and DrugBank_FDA datasets, our method outperforms all baseline methods, being statistically significant at the significance level of 0.05. In terms of AUPR, from **Table S4(c)** we can see that our method MDADTI performs best in GPCR, IC, E, and DrugBank_FDA datasets, and it outperforms all baseline methods, being statistically significant in all cases. The comparison of AUC and AUPR between MDADTI with transfer learning and MDADTI without transfer learning on NR dataset  is reported in  **Table S6** of **Supplementary Material**. The performance of MDADTI with SMOTE method and MDADTI without SMOTE method is reported in **Table S7** of **Supplementary Material**.

All above analyses demonstrate that MDADTI is superior to DDR, KronRLS-MKL, NRLMF, and BLM-NII. The main reason is that different from DDR, KronRLS-MKL, NRLMF, and BLM-NII, which directly took the original multiple similarity measures as input and manually extracted the features of the drug-target pairs, MDADTI applied RWR method and PPMI to capture the global structure information of the similarity measures, and applied the multi-layer nonlinear functions of MDA to capture the complex non-linear relationship among features, and automatically learned the deep feature representation of drugs and targets, which are helpful to improve the predictive performance.

For large datasets GPCR, IC, and E, MDA reduced the dimension of drug feature and target feature while automatically learning them. The dimension of drug feature in GPCR dataset is reduced from 223 to 50, and the dimension of target feature is reduced from 95 to 25. The dimension of drug feature and target feature in IC dataset are reduced from 210 and 204 to 50, respectively. The dimension of drug feature and target feature in E dataset are reduced from 1482 and 1408 to 100, respectively. Dimensionality reduction accelerates the training speed and saves the time costs running on large datasets of MDADTI model.

We observed that the predictive performance of MDADTI is greatly improved in NR dataset under CVS1 setting, which indicates that our transfer learning strategy helps MDADTI achieve superior performance with a small amount of labeled data. This is because we used DTIs in E datasets to pretrain MDADTI model, and froze all layers except the output layer of the pretrained model, that is, set the parameters of these frozen layers to be untrainable. These parameters contain the knowledge learned from the E dataset, which are also applicable to the NR dataset. Therefore, our transfer learning strategy can make MDADTI predict DTIs more accurately on small datasets.

According to statistics, about one-third of the small molecule drugs in the world drug market are activators or antagonists of GPCRs, which are related to many diseases, and GPCR is the target of about 40% of modern drugs (Marinissen and Gutkind, 2001). MDADTI has a significant improvement in predictive performance on GPCR datasets under CVS1 setting. Therefore, MDADTI can be used as an effective tool to predict GPCR target and has great significance for drug development and disease treatment.

## Effectiveness of Feature Learning for Drugs and Targets With MDA

In order to evaluate the effectiveness of feature learning for drugs and targets with MDA for improving the predictive performance of MDADTI, we designed another RWR_DNN method to be compared with MDADTI. Firstly, RWR_DNN takes multiple

similarity measures for drugs (targets) as original inputs. Then, it uses RWR and PPMI method to calculate multiple topological similarity matrices for drugs (targets). Next, it averages multiple topological similarity matrices of drugs (targets) to form the feature of drugs (targets). Finally, the features of drugs and targets are concatenated together and sent into DNN for predicting DTIs. Similarly, we applied 5-repeats of 10-fold cross-validation under CVS1, CVS2, and CVS3 setting to evaluate the performance of RWR_DNN. The hyperparameters of RWR_DNN method on NR, GPCR, IC, and E dataset under three settings are the same with that of MDADTI, which are reported in **Table S1–S3** of the **Supplementary Material**.

The comparison results of RWR_DNN and MDADTI on NR, GPCR, IC, and E datasets in 5-repeats of 10-fold cross-validation are shown in **Figure 6**, where **Figure 6 (A)** is the comparison of AUC and **Figure 6 (B)** is the comparison of AUPR. We can see the AUC and AUPR values of MDADTI are higher than that of RWR_DNN in all cases. The results demonstrate that MDA can automatically learn deep feature representations of drugs and targets from multiple topological similarity matrices and effectively improve the predictive performance of MDADTI method.

## Prediction and Validation of Unknown DTIs

In this paper, we used NR, GPCR, IC, E, and DrugBank_FDA datasets to evaluate the performance of our proposed method MDADTI, and for each dataset, we used 5-repeats of 10-fold cross-validation to evaluate the performance of MDADTI method. Since the negative samples in the NR, GPCR, IC, E and DrugBank_FDA datasets are unknown DTIs, we evaluated the practical ability of MDADTI model in predicting new (unknown) interactions. New interactions are predicted high-probability drug-target pairs, but they are unknown DTIs in NR, GPCR, IC, E, and DrugBank_FDA datasets.

In order to implement this, we used the trained model to predict unknown DTIs in each dataset and output the interaction probability of a drug-target pair. Then we ranked them in descending order according to the predicted probability. Finally, we selected the top 100 predicted unknown DTIs and validated them in six reference databases, i.e., to check if they are included in any of six reference databases: ChEMBL (Gaulton et al., 2011), DrugBank (Knox et al., 2010), KEGG (Kanehisa et al., 2011), Matador (Günther et al., 2007), CTD (Davis et al., 2016), and STITCH (Kuhn et al., 2007). These six reference databases are online databases that include a large number of proved known DTIs and they are used by related literature to evaluate the actual ability of their methods in predicting unknown DTIs (Liu et al., 2016; Nascimento et al., 2016; Olayan et al., 2017).

**Table 4** shows the top 30 unknown interactions predicted by the MDADTI model on E dataset. In this table, DTIs in bold indicate that they exist in one or more reference databases, and the third column shows their predicted probability. For each drug-target pair, the reference databases containing it are displayed in the last column of the table, where C is the abbreviation of ChEMBL, D is the abbreviation of DrugBank, M is the abbreviation of Matador, K is the abbreviation of KEGG, T is the abbreviation of CTD, and S is the abbreviation of STITCH. For example, the DTI ranking No. 1 is D00528, hsa1549 and its predicted interaction probability is 1.0, which is validated in the Matador database. It can be seen from the table that 21 out of 30 unknown interactions are validated in at least one of the six reference databases.

In order to visualize the validation of unknown DTIs more intuitively, we visualized 100 high-probability unknown DTIs in E dataset. **Figure 7** is the network visualization of the top 100 unknown DTIs in E dataset predicted by MDADTI model. Yellow and blue nodes represent drugs and targets, respectively. Solid lines represent verified interactions while dashed lines represent unverified interactions. It can be seen from the figure that there are potential interactions between a drug and multiple targets, and some of them have been verified in six reference databases. For example, 33.33% (3/10) of the potential targets of drug D00002 have been verified in reference databases; D00002 represents nicotinamide adenine dinucleotide (NADH), which is widely used in many diseases like tuberculosis, Alzheimer's, and Parkinson disease. 44.44% (4/9) of the potential targets of D00043 are validated in the reference databases, and D00043 represents isofluorphate, a powerful miotic used mainly in the treatment of glaucoma. 60% (3/5) of the potential targets of drug D00410 are validated in reference databases, and D00410 represents metyrapone, an inhibitor of the enzyme steroid 11-beta-monooxygenase, which is used as a test of the feedback hypothalamic-pituitary mechanism in the diagnosis of Cushing syndrome. 57.14% (4/7) of the potential targets of drug D00528 are verified in the reference database. We also observed that a target may interact with multiple drugs, and some of them are verified in six reference databases. For example,

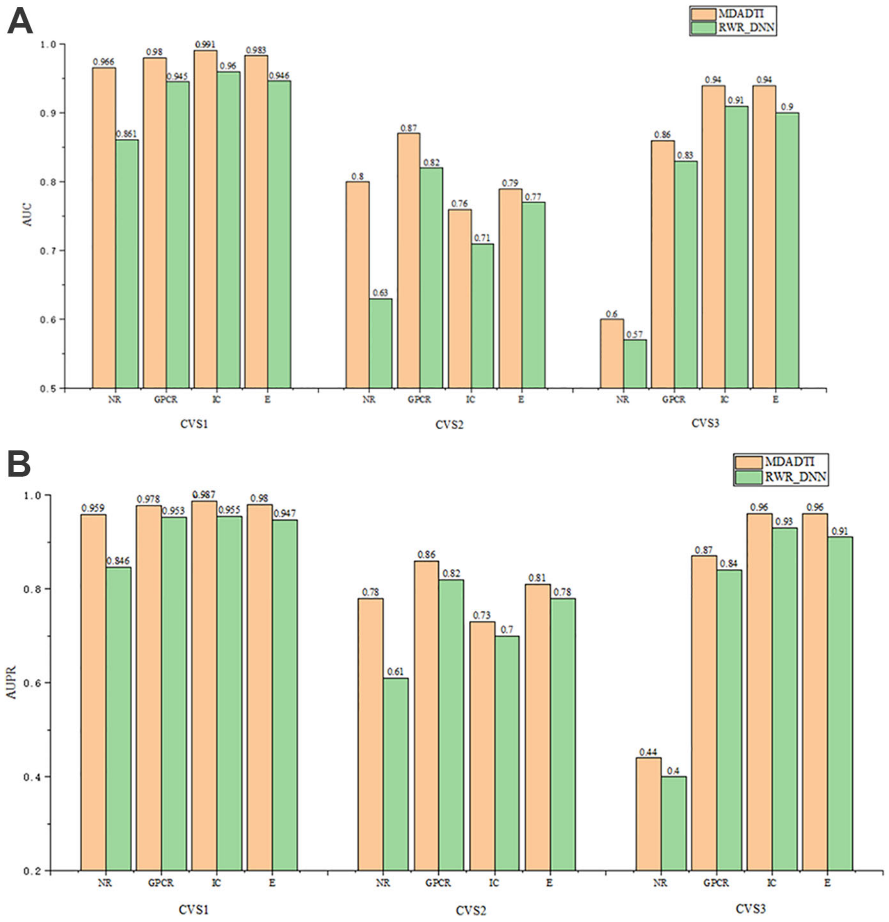


**FIGURE 6 |** The comparison of AUC and AUPR between MDADTI and RWR_DNN method on NR, GPCR, IC and E dataset under CVS1, CVS2 and CVS3 setting. **(A)** Comparison of AUC **(B)** Comparison of AUPR.

hsa1559 interacts with D00510, D00217, and D00437 at the same time, and all of them are verified in six reference databases. hsa5150 interacts with D00528, D00501, and D00691, and all of them are verified in six reference databases.

Finally, **Table 5** summarizes the validated proportion of top N unknown DTIs (N = 10, 30, 50, 100) on five datasets. The validated proportion of top 10 unknown DTIs are 50%, 80%, 80%, 100%, 80%, respectively. The fractions of validated DTIs of MDADTI, NRLMF, BLM-NII among the predicted Top N(N = 10, 30, 50) DTIs in NR,GPCR,IC, and E datasets are provided in **Table S5**. The fractions of validated DTIs of NRLMF, BLM-NII are provided by (Liu et al., 2016). Since these databases are still being updated, the proportion of new DTIs predicted by MDADTI model will increase in the future. All the above analyses proved that MDADTI can effectively predict unknown DTIs because MDADTI model integrated multiple similarity measures of drugs and targets, which provides abundant information for predicting DTIs. Moreover, MDADTI not only considered the original information of similarity measure but also captured the global structure information of similarity measures, which improved the prediction accuracy of DTIs. The most important reason is that MDADTI applied MDA to automatically learn the deep representation of drug feature and target feature from multiple topological similarity matrices of drugs and targets, which contributes to the effective prediction of unknown DTIs.

## DISCUSSION

We proposed a novel method MDADTI to predict DTIs based on MDA. Compared with existing methods, MDADTI applied RWR and PPMI to calculate the topological similarity matrices of drugs and targets, capturing the global structure information of the similarity measures. Then MDA was applied to fuse multiple topological similarity matrices and learn the feature of drugs and targets while capturing the non-linear relationship among features. In addition, MDA also reduced the dimension of the feature of drugs and targets, which speeded up the training of MDADTI. To evaluate the performance of MDADTI, we compared MDADTI with DDR, KronRLS-MKL, NRLMF, and BLM-NII under three different cross-validation settings. The

**TABLE 4 |** Top 30 unknown DTIs predicted by MDADTI model on E dataset. DTIs in bold indicate that they are validated in one or more reference databases.

| Rank | Drug | Target | Probability | Databases | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **D00528** | **hsa1549** | **1.0** | | | | **M** | | |
| **2** | **D00542** | **hsa1571** | **0.9997** | | | | **M** | | |
| **3** | **D00501** | **hsa5150** | **0.9997** | **C** | | | | | |
| **4** | **D00437** | **hsa1559** | **0.9997** | | | | **M** | | |
| **5** | **D00043** | **hsa11330** | **0.9995** | | | | **M** | | |
| **6** | **D00528** | **hsa5150** | **0.9992** | | **D** | **K** | | | |
| **7** | **D00410** | **hsa1543** | **0.9988** | | | | **M** | | |
| **8** | **D00691** | **hsa8564** | **0.9985** | | | | | | **S** |
| **9** | **D00437** | **hsa1585** | **0.9981** | | | | **M** | | |
| **10** | **D00410** | **hsa1585** | **0.9981** | | | | **M** | | |
| **11** | **D00139** | **hsa1543** | **0.9972** | | | | **M** | | |
| **12** | **D00043** | **hsa2147** | **0.9965** | | | | **M** | | |
| **13** | **D01441** | **hsa5594** | **0.9884** | | | | | **T** | |
| **14** | **D00126** | **hsa246** | **0.9869** | | | | **M** | | |
| 15 | D00043 | hsa1504 | 0.977 | | | | | | |
| **16** | **D00217** | **hsa1559** | **0.9683** | | | | | **T** | |
| **17** | **D01223** | **hsa3988** | **0.9644** | | | | **M** | | |
| **18** | **D00038** | **hsa5742** | **0.9640** | | | | | **T** | |
| 19 | D01223 | hsa5538 | 0.9616 | | | | | | |
| 20 | D00002 | hsa31 | 0.9553 | | | | | | |
| **21** | **D01441** | **hsa1021** | **0.9546** | | | | | **T** | |
| **22** | **D00528** | **hsa5743** | **0.9467** | | | | | **T** | |
| 23 | D00139 | hsa5742 | 0.9344 | | | | | | |
| **24** | **D00217** | **hsa1558** | **0.9338** | | | | | **T** | |
| **25** | **D00043** | **hsa1636** | **0.9326** | | | | **M** | | |
| 26 | D00002 | hsa7298 | 0.9207 | | | | | | |
| 27 | D03670 | hsa1579 | 0.8932 | | | | | | |
| 28 | D01441 | hsa3551 | 0.8806 | | | | | | |
| **29** | **D00097** | **hsa5743** | **0.8787** | | **D** | | **M** | | |
| 30 | D00043 | hsa686 | 0.8688 | | | | | | |

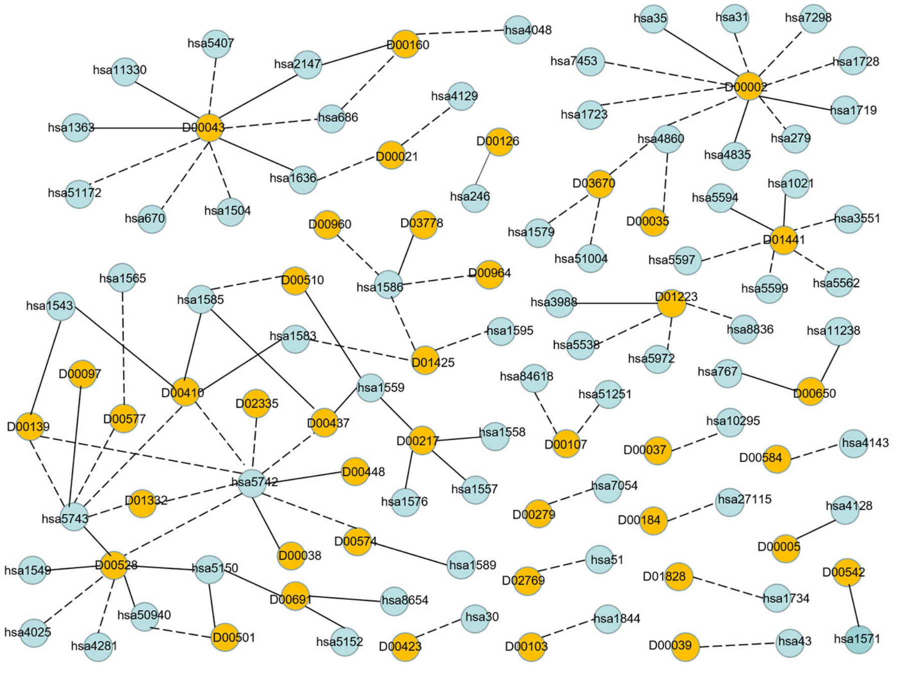


**FIGURE 7 |** Network visualization of the top 100 unknown DTIs in E dataset. Yellow and blue nodes represent drugs and targets, respectively. Solid lines represent verified interaction and dashed lines represent unverified interactions. There are 40 unknown DTIs that were verified.

**TABLE 5 |** The fractions of true DTIs among the predicted Top N (N = 10, 30, 50,100) unknown DTIs in five datasets.

| Dataset | Top N | Fraction |
|---|---|---|
| NR | Top10 | 50.00% |
| | Top30 | 43.33% |
| | Top50 | 28.00% |
| | Top100 | 20.00% |
| GPCR | Top10 | 80.00% |
| | Top30 | 66.67% |
| | Top50 | 60.00% |
| | Top100 | 40.00% |
| IC | Top10 | 80.00% |
| | Top30 | 50.00% |
| | Top50 | 40.00% |
| | Top100 | 32.00% |
| E | Top10 | 100.00% |
| | Top30 | 73.33% |
| | Top50 | 52.00% |
| | Top100 | 40.00% |
| DrugBank_FDA | Top10 | 80.00% |
| | Top30 | 66.67% |
| | Top50 | 68.00% |
| | Top100 | 46.00% |

results showed that MDADTI achieved higher AUC and AUPR in five datasets than the other four methods under CVS1 setting. The predictive performance of MDADTI was greatly improved especially in GPCR and NR datasets. For CVS2 and CVS3 settings, our method has a great improvement in AUPR in five datasets, and it performed better in large datasets, like E and DrugBank_FDA datasets. These results proved that MDADTI is better than the other four baseline methods in predicting DTIs.

In addition, we evaluated the actual ability of MDADTI method to predict new interactions. For each dataset, we applied the trained MDADTI model to predict unknown interactions and selected the top 100 predictions to validate them in the six reference databases: ChEMBL, DrugBank, KEGG, Matador, STITCH, and CTD. The results showed that MDADTI method can effectively identify unknown DTIs.

Since our method currently only predicts whether there is an interaction between a drug and a target, we plan to predict the binding affinity scores for drug-target pairs in the next step.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

## AUTHOR CONTRIBUTIONS

HW and JW performed the majority of the analysis and primarily wrote the manuscript. CD performed some analysis and provided biological expertise. ZY performed some analysis of data and helped conceive the project. YL and DL completed the drawing of the charts in the results analysis and the layout of the manuscripts. All authors edited and approved the manuscript.

## FUNDING

This study was supported by research grants from the National Key Research and Development Plan of China (2018YFD0100204), the National Natural Science Foundation of China (61672374), National Natural Science Foundation of China (61976150), Key Research and Development Plan of Shanxi Province (201903D121151) and the Scientific and Technological Project of Shanxi Province (No. 201603D22103-2).

## ACKNOWLEDGMENTS

We sincerely thank the other members of our team for the useful discussions and help.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.01592/full#supplementary-material

## REFERENCES

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2016). HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi: 10.1093/nar/gkw985

Ba-Alawi, W., Soufan, O., Essack, M., Kalnis, P., and Bajic, V. B. (2016). DASPfind: new efficient method to predict drug–target interactions. *J. Cheminformatics* 8 (1), 15. doi: 10.1186/s13321-016-0128-4

Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25 (18), 2397–2403. doi: 10.1093/bioinformatics/btp433

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al. (2016). "UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view," in *Plant Bioinformatics* (Springer), 23–54. doi: 10.1007/978-1-4939-3167-5_2

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321 (5886), 263–266. doi: 10.1126/science.1158140

Cao, S., Lu, W., and Xu, Q. (2016). Deep neural networks for learning graph representations. In *Thirtieth AAAI Conference on Artificial Intelligence* p. 1145–1152.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics* 32 (12), 1832–1839. doi: 10.1093/bioinformatics/btw074

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., et al. (2016). The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.* 45 (D1), D972–D978. doi: 10.1093/nar/gkw838

Deng, L., Fan, C., and Zeng, Z. (2017). A sparse autoencoder-based deep neural network for protein solvent accessibility and contact number prediction. *BMC Bioinf.* 18 (16), 569. doi: 10.1186/s12859-017-1971-7

Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., Zhu, K., and Lu, S. (2019). Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. *BMC Bioinf.* 20 (1), 87. doi: 10.1186/s12859-019-2675-y

Fang, H., and Gough, J. (2013). A disease-drug-phenotype matrix inferred by walking on a functional domain network. *Mol. Biosyst.* 9 (7), 1686–1696. doi: 10.1039/c3mb25495j

Fu, L., and Peng, Q. (2017). A deep ensemble model to predict miRNA-disease association. *Sci. Rep.* 7 (1), 14482. doi: 10.1038/s41598-017-15235-6

Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 28 (18), 2304–2310. doi: 10.1093/bioinformatics/bts360

Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., et al. (2007). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* 36 (suppl_1), D919–D922. doi: 10.1093/nar/gkm862

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40 (D1), D1100–D1107. doi: 10.1093/nar/gkr777

Gligorijević, V., Barot, M., and Bonneau, R. (2018). deepNF: deep network fusion for protein function prediction. *Bioinformatics* 34 (22), 3873–3881. doi: 10.1093/bioinformatics/bty440

Hao, M., Bryant, S. H., and Wang, Y. (2017). Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci. Rep.* 7, 40376. doi: 10.1038/srep40376

Hizukuri, Y., Sawada, R., and Yamanishi, Y. (2015). Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. *BMC Med. Genomics* 8 (1), 82. doi: 10.1186/s12920-015-0158-1

Jacob, L., and Vert, J.-P. (2008). Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 24 (19), 2149–2156. doi: 10.1093/bioinformatics/btn409

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82 (4), 949–958. doi: 10.1016/j.ajhg.2008.02.013

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40 (D1), D109–D114. doi: 10.1093/nar/gkr988

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., et al. (2009). Predicting new molecular targets for known drugs. *Nature* 462 (7270), 175. doi: 10.1038/nature08506

Khan, M., Hayat, M., Khan, S. A., and Iqbal, N. (2017). Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. *J. Theor. Biol.* 415, 13–19. doi: 10.1016/j.jtbi.2016.12.004

Klabunde, T. (2007). Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* 152 (1), 5–7. doi: 10.1038/sj.bjp.0707308

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2010). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39 (suppl_1), D1035–D1041. doi: 10.1093/nar/gkq1126

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2007). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 36 (suppl_1), D684–D688. doi: 10.1093/nar/gkm795

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44 (D1), D1075–D1079. doi: 10.1093/nar/gkv1075

Lim, H., Gray, P., Xie, L., and Poleksic, A. (2016). Improved genome-scale multi-target virtual screening *via a* novel collaborative filtering approach to cold-start problem. *Sci. Rep.* 6, 38860. doi: 10.1038/srep38860

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PloS Comput. Biol.* 12 (2), e1004760. doi: 10.1371/journal.pcbi.1004760

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486 (7403), 361. doi: 10.1038/nature11159

Lu, Y., Guo, Y., and Korhonen, A. (2017). Link prediction in drug-target interactions network using similarity indices. *BMC Bioinf.* 18 (1), 39. doi: 10.1186/s12859-017-1460-z

Marinissen, M. J., and Gutkind, J. S. (2001). G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends In Pharmacol. Sci.* 22 (7), 368–376. doi: 10.1016/S0165-6147(00)01678-3

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. (2012). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29 (2), 238–245. doi: 10.1093/bioinformatics/bts670

Núñez, S., Venhorst, J., and Kruse, C. G. (2012). Target–drug interactions: first principles and their application to drug discovery. *Drug Discovery Today* 17 (1-2), 10–22. doi: 10.1016/j.drudis.2011.06.013

Nascimento, A. C., Prudêncio, R. B., and Costa, I. G. (2016). A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinf.* 17 (1), 46. doi: 10.1186/s12859-016-0890-3

Olayan, R. S., Ashoor, H., and Bajic, V. B. (2017). DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 34 (7), 1164–1173. doi: 10.1093/bioinformatics/btx731

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22 (10), 1345–1359. doi: 10.1109/TKDE.2009.191

Pan, X., Fan, Y.-X., Yan, J., and Shen, H.-B. (2016). IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics* 17 (1), 582. doi: 10.1186/s12864-016-2931-8

Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., et al. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Syst. Biol.* 12 (2), 18. doi: 10.1186/s12918-018-0539-0

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.* 18 (2), 133–145. doi: 10.1089/cmb.2010.0213

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (1), 1929–1958. doi: 10.5555/2627435.2670313

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning In *International conference on machine learning*. p. 1139–1147.

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27 (21), 3036–3043. doi: 10.1093/bioinformatics/btr500

Vilar, S., and Hripcsak, G. (2016). The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions. *Briefings In Bioinf.* 18 (4), 670–681. doi: 10.1093/bib/bbw048

Waris, M., Ahmad, K., Kabir, M., and Hayat, M. (2016). Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* 199, 154–162. doi: 10.1016/j.neucom.2016.03.025

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2007). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36 (suppl_1), D901–D906. doi: 10.1093/nar/gkm958

Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC systems biology. 4 (2), S6. doi: 10.1186/1752-0509-4-S2-S6

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24 (13), i232–i240. doi: 10.1093/bioinformatics/btn162

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 1025–1033. doi: 10.1145/2487575.2487670

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

METHODS
published: 30 January 2020
doi: 10.3389/fphar.2019.01586

# Deep Learning Based Drug Metabolites Prediction

*Disha Wang*[1], *Wenjun Liu*[2], *Zihao Shen*[1], *Lei Jiang*[1], *Jie Wang*[1], *Shiliang Li*[1*] and *Honglin Li*[1*]

[1] *Shanghai Key Laboratory of New Drug Design, State Key Laboratory of Bioreactor Engineering, School of Pharmacy, East China University of Science and Technology, Shanghai, China,* [2] *Research and Development Department, Jiangzhong Pharmaceutical Co., Ltd., Nanchang, China*

Drug metabolism research plays a key role in the discovery and development of drugs. Based on the discovery of drug metabolites, new chemical entities can be identified and potential safety hazards caused by reactive or toxic metabolites can be minimized. Nowadays, computational methods are usually complementary tools for experiments. However, current metabolites prediction methods tend to have high false positive rates with low accuracy and are usually only used for specific enzyme systems. In order to overcome this difficulty, a method was developed in this paper by first establishing a database with broad coverage of SMARTS-coded metabolic reaction rule, and then extracting the molecular fingerprints of compounds to construct a classification model based on deep learning algorithms. The metabolic reaction rule database we built can supplement chemically reasonable negative reaction examples. Based on deep learning algorithms, the model could determine which reaction types are more likely to occur than the others. In the test set, our method can achieve the accuracy of 70% (Top-10), which is significantly higher than that of random guess and the rule-based method SyGMa. The results demonstrated that our method has a certain predictive ability and application value.

**Keywords: deep learning, drug metabolism, metabolites prediction, reaction rules, SMARTS**

OPEN ACCESS

***Edited by:***
*Alex Zhavoronkov,*
*Biogerontology Research Foundation,*
*United Kingdom*

***Reviewed by:***
*Cao Dongsheng,*
*Central South University, China*
*Mingyue Zheng,*
*Chinese Academy of Sciences, China*

***Correspondence:**
*Shiliang Li*
*slli403@163.com*
*Honglin Li*
*hlli@ecust.edu.cn*

***Specialty section:***
*This article was submitted to Experimental Pharmacology and Drug Discovery, a section of the journal Frontiers in Pharmacology*

***Received:*** *03 October 2019*
***Accepted:*** *09 December 2019*
***Published:*** *30 January 2020*

***Citation:***
*Wang D, Liu W, Shen Z, Jiang L, Wang J, Li S and Li H (2020) Deep Learning Based Drug Metabolites Prediction. Front. Pharmacol. 10:1586. doi: 10.3389/fphar.2019.01586*

## INTRODUCTION

The discovery of small molecule drugs is time-consuming, expensive and labor-intensive. (Dickson and Gagnon, 2004; Paul et al., 2010; Dimasi et al., 2015) It is resource intensive, and involves typical timelines of 10–20 years and costs that range from US$0.5 billion to US$2.6 billion (Paul et al., 2010; Avorn, 2015). In addition to economic and technical reasons, the main reason is that almost half of the candidate drugs failed in clinical trials. Up to 25% of compounds were withdrawn due to metabolic, pharmacokinetic, or toxic problems (Hwang et al., 2016). Drug metabolism can produce metabolites with physicochemical and pharmacological properties, which are significantly different from the physical and pharmacological properties of parent drugs (Kirchmair et al., 2013). As **Figure 1** shows, when drugs or other exogenous substances enter the human body, they are largely controlled by three stages of drug metabolism. In the first stage, reactive groups are introduced by oxidation, reduction, or hydrolysis. In the second stage, conjugation reactions with macromolecules occur *in vivo*. In the third stage, allogeneic and metabolites are removed from liver and intestinal

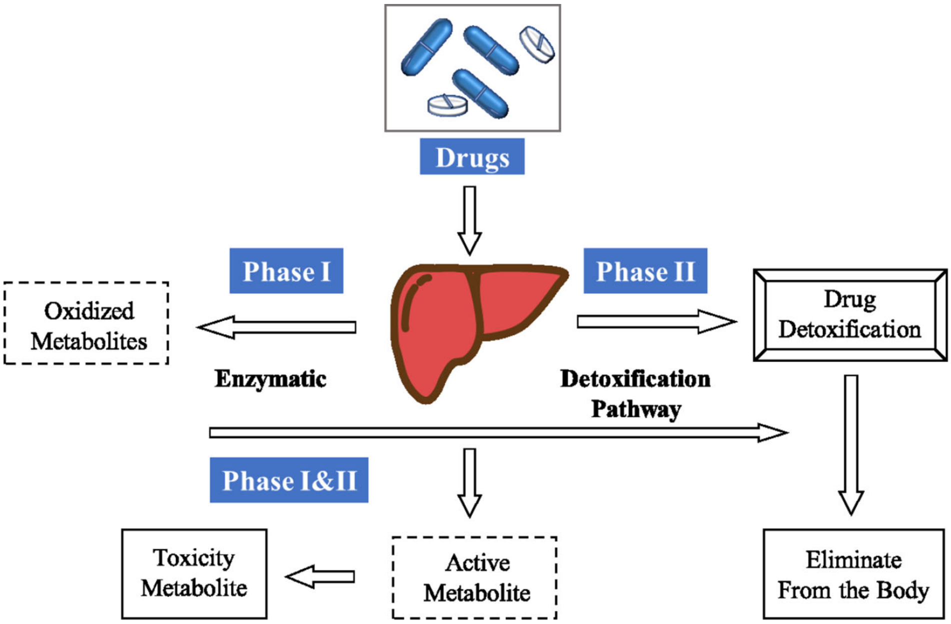


**FIGURE 1** | General pathway of drug metabolism.

cells. After these three stages, exogenous substances such as drugs may be transformed into non-physiological active substances or toxic metabolites. 70% clinical drugs are removed by the body's metabolic system, so as part of drug development, it is also necessary to conduct in-depth research on drug metabolism (Grant et al., 2001; Embrechts and Ekins, 2007; Lazar and Birnbaum, 2012; Damsky and Bosenberg, 2012; Di, 2014; Mackenzie et al., 2017; Kang et al., 2018). Understanding the metabolic process of drugs is essential for successful drug discovery and development, and helps to optimize the stability of drugs, so as to optimize the half-life *in vivo*.

In order to reduce the risk caused by metabolic characteristics of candidate drugs, effective andreliable methods are needed to predict drug metabolism *in vitro*. Many experimental methods can be used to explore the metabolic process of drugs (Diao et al., 2016; Mackenzie et al., 2017). For example, fast LC-MS scans can be carried out to specifically detect predicted metabolites. However, experimental methods are still highly demanding in terms of equipment, expertise, cost, and time (Kirchmair et al., 2013). Therefore, it has great prospects to develop computational tools for predicting drug metabolism with lower cost and higher throughput than existing tools. Many different methodologies to predict metabolites or sites of metabolism have been reported recently. Various methods in predicting drug metabolism using in silico approaches have been reviewed (Fox and Kriegl, 2006; Gleeson et al., 2011; Zhang et al., 2011; Tan et al., 2017). However, most of these methods are limited to P450 catalytic reactions and represent only unstable sites, rather than predicting the actual metabolites formed.

Metabolic sites (SOMs) and metabolite structure are two main research directions of computer-aided metabolic prediction methods, which can provide decisive support and guidance for experimenters. SOM prediction methods usually have high prediction accuracy. The program MetaSite estimate the possibility of metabolic reactions at an atomic site using protein structure information, GRID-derived MIFs of protein, and ligand and molecular orbital calculations (Gabriele et al., 2005). The program SMARTCyp contains a pre-calculated energy reaction analysis table for density functional theory activation, where a large number of ligand fragments pass through CYP3A4 or CYP2D6 mediated transformation (Rydberg et al., 2010a; Rydberg et al., 2010b). A method called cypscore, in which 2400 CYP-mediated transformations and 850 literature compounds are used as data bases (Hennemann et al., 2010). However, most of these methods are limited to CYP450 catalytic reactions and can only predict unstable sites rather than metabolite structures. Furthermore, predicted SOMs are not identical to identifying the correct bioinformations that will occur at an atomic location, and they do not provide information about the type of reaction that will occur. Therefore, these limitations make it difficult to draw any quantitative conclusions about the metabolic possibilities of a molecule.

So far, only a few computational methods have been developed for predicting the structure ofmetabolites. Existing methods can be divided into two categories: expert rule-based anddescriptor-based. Rule-based approaches use data mining techniques. Large databases with data onmetabolism are used to extract generalized rules to determine the part of a molecule that undergoes metabolic alteration (Cariello et al., 2002). The ligand-based approach relies on the assumption that the metabolic fate of compounds is entirely determined by their chemical structure and properties. These methods include

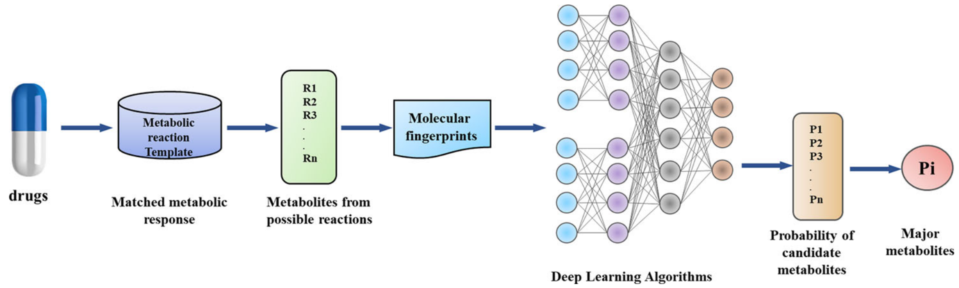

drugs
Metabolic reaction Template
Matched metabolic response
R1
R2
R3
.
.
.
Rn
Metabolites from possible reactions
Molecular fingerprints
Deep Learning Algorithms
P1
P2
P3
.
.
.
Pn
Probability of candidate metabolites
Pi
Major metabolites

**FIGURE 2 |** Metabolic reaction product prediction flow chart.

quantum mechanics methods. Descriptor-based methods to obtain an idea of the route of a drug through the metabolic system, the identification of the involved enzymes, and the reaction pathways is necessary (Livingstone, 2010). The program of Bioprint contains a database of most marketed drugs together with reference compounds and data from a wide variety of biological and *in-vitro* ADME assays, called the Biological fingerprint (Krejsa et al., 2003). Thus, the possible results of new compounds can be calculated by neighborhood relation and QSAR model. In the MetaDrug database (Ekins et al., 2006; Ekins et al., 2005b), metabolic reactions with substrates (including primary and secondary metabolites), xenobiotic reactions, and kinetic data on enzyme inhibition are stored. 317 molecules (parent drug and primary and secondary metabolite) were randomly selected from this database to build kernel-partial least squares models for metabolism rules (Embrechts and Ekins, 2007). Metabolite prediction is usually accomplished through a large set of transformation rules. Given the reactant, all rules are then matched to determine the site of metabolic instability. Expert systems such as METEOR (Testa et al., 2010; Button et al., 2015), META (Klopman et al., 1994; Talafous et al., 1994; Klopman et al., 1997), MetabolExpert (Darvas, 1988), RD-Metabolizer (Meng et al., 2017), MetaDrug (Korolev et al., 2003; Ekins et al., 2005a), and KnowItAll (Stouch et al., 2003) are based on these databases and provide a ranked list of most likely metabolites. In a study described by AstraZeneca (Scott et al., 2007), the substrates and reaction centers of the metabolite database were stored as fingerprints in two databases. Then the query molecule powders are compared with the two databases, and the proposed SOM is ranked by using the number of clicks as a weighted scheme. An approach called SyGMa based on the MDL metabolite database was developed (Ridder and Wagener, 2010). According to the corresponding rules of MDL metabolite data coding, the structure of possible metabolites is predicted, and probability scores are assigned to each metabolite, covering 70% of all known human metabolic reactions.

So far, one of the difficulties in predicting possible metabolites is that this task means identifying the reaction site (SOM) and the type of metabolic reaction correctly. Current methods for predicting metabolite structure tend to have high false-positive rates and can only be used for specific enzymes without covering all the metabolic enzymes involved in human reactions. In view of the above problems, we mainly designed a deep learning algorithm combined with drug metabolism characteristics.

In this work, by combining metabolic reaction template and Deep Learning, we have established a model to predict the main metabolites of drugs (**Figure 2**). Our method has the following innovations: (1) Data enhancement strategy, which provides chemically reliable examples of negative reactions through the metabolic reaction template library; (2) the implementation and validation of a neural network-based model, which can obtain that some reaction modes are more or less likely to occur than other potential modes.

# MATERIALS AND METHODS

## Data Collection and Processing

We collected metabolic reaction data from MDL Database (2011 edition). Here we used only human metabolic reactions with effective substrates and metabolites. The data were filtered to remove unreasonable structures, such as reactants and products containing R groups, free radicals, metal chelating, and structural errors, which could make it impossible to distinguish the reaction records of reaction sites. The pretreatment resulted in 7,380 reaction records, of which 74 reactions had only chiral changes. We randomly selected 300 response records from them as standby for external test sets.

## Generation of Metabolic Reactions Template Library

The process of constructing the metabolic reactions template library is shown in **Figure 3**. At present, the methods based on expert system mainly use the general metabolic rules deduced by experts to predict the structure of metabolites. However, this method has some drawbacks. The model needs to understand the influence of coding reaction functional groups. Such rules can not completely produce the desired response because the complete background of molecules is ignored. The remaining Non-coding functional groups of the molecule may affect or react competitively. So maybe even if the rules are matched, the ideal reaction product cannot be produced. Therefore, reaction rules need to be annotated with relevant information, such as functional groups, priority of reactions. However artificial code rules are time-consuming, laborious, and lack of internal ranking mechanism. Based on this, expert rules cannot be implemented

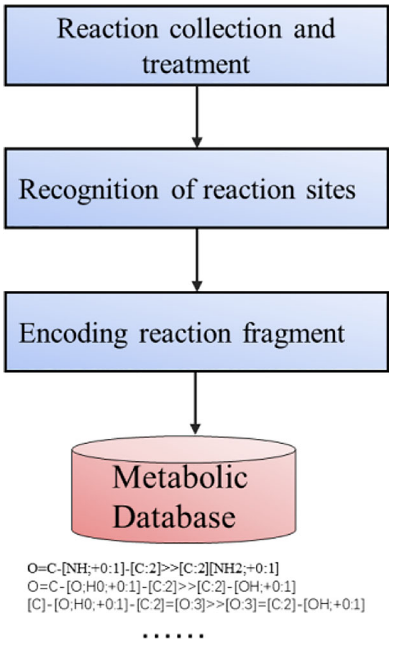

Reaction collection and treatment
Recognition of reaction sites
Encoding reaction fragment
Metabolic Database
O=C-[NH;+0:1]-[C:2]>>[C:2][NH2;+0:1]
O=C-[O;H0;+0:1]-[C:2]>>[C:2]-[OH;+0:1]
[C]-[O;H0;+0:1]-[C:2]=[O:3]>>[O:3]=[C:2]-[OH;+0:1]
......

**FIGURE 3 |** The generation of metabolic reactions template.

on a large scale. Marwin H.S. (Law et al., 2009; Segler et al., 2018) has proved that in predicting the products of inverse synthesis reaction, the artificial extraction of reaction rules based on expert rules is far less effective than the algorithm which automatically judges the type of reaction according to the reactants and products in the deep learning model training. In order to construct a database of metabolic templates, we also adopted the heuristic driving algorithm of Law (Law et al., 2009).

**Table 1** shows the most common reaction types in the database. It can be seen that the most common metabolic reactions are amide hydrolysis, carboxylic acid hydrolysis, and hydroxylation of N, O, S atoms.

## Producing Candidate Metabolites

The above-mentioned metabolic reaction templates are stored in the database for subsequent production of positive and negative potential metabolites. For each atom mapping reaction in the dataset, the reaction center is defined by determining which product atoms are different from the corresponding reactant atoms. The reaction center is expanded to include the surrounding environment, and then other factors that play a role in the reaction is found out. Adjacent atoms are defined as non-hydrogen substituents, where high coverage is achieved at the expense of low specificity. Metabolic Templates are defined with SMART format strings encoding reaction centers. The reaction template generated in this way does not depend on manual extraction, marking, or sorting.

As shown in **Figure 4**, we can match the reaction template of the metabolic database one by one and produce a large number of potential metabolic reaction products by using RDKit. Positive compounds are the products recorded in the database, and the rest are all negative products. This strategy can continuously produce negative products for later use.

## Model Training

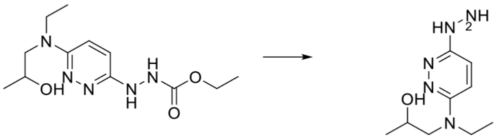
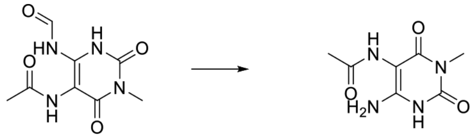
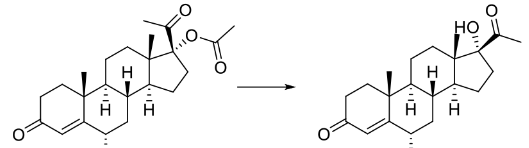
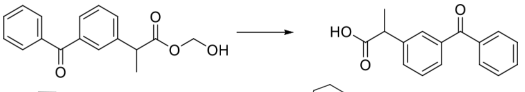
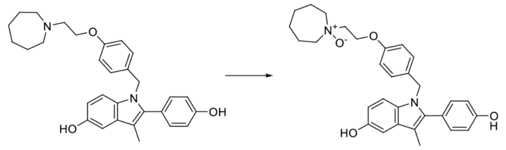
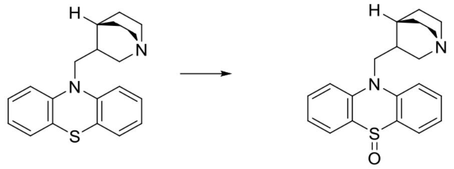
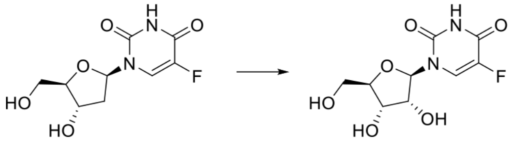
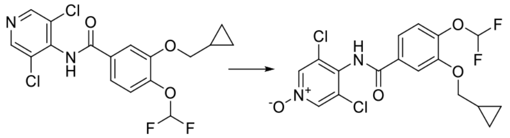
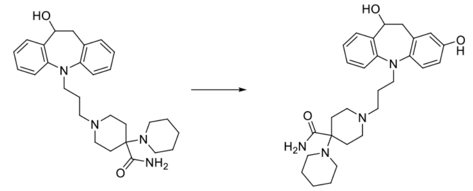
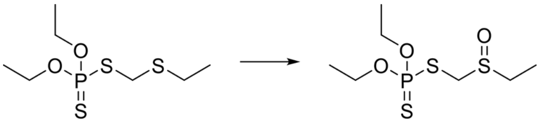
For deep learning and supervised learning, we need to input eigenvalues. What we need to consider is using what molecular descriptors to characterize the whole metabolic process. Here we choose molecular fingerprints to describe the atomic and functional characteristics of metabolic reactions. The abstract representation of molecular fingerprints, which encodes molecular transformation into a series of vectors, makes it easier for molecules to compare with each other. If two molecules are similar, there must be many common fragments between them. Then molecules with similar fingerprints will have a high probability of being similar in 2D structures. Here we use RDKit to generate 1,024-dimensional Morgan molecular fingerprints. Molecular fingerprint ECFP is suitable for machine learning because it contains more molecular structure details. The metabolites generated above through the metabolic reaction template will be scored separately by Deep Learning model. Here we use Python library of Keras. The input layer consists of molecular fingerprints of products and reactants, with a total of 2,048 dimensions. One reaction corresponds to multiple potential metabolites. Thus, the input layer generates a matrix of 2,048 dimensions with *n* vectors. We use keras wrapper to realize each vector, that is, each individual potential metabolite is fully connected independently, which increases the ability of the model to achieve one-to-many and many-to-many. The probability of all potential reaction products is finally mapped out by the output layer activation function softmax, so that the total probability of all potential metabolites is 1. According to the score of the output layer, it is most likely to describe which metabolites actually exist. Deep neural network models are trained here to solve problems similar to classification problems: given hundreds of possible classes (potential metabolites), predicting real classes (recording reaction products), each metabolic reaction may correspond to multiple classifications. We use cross-entropy as the loss function during training. This objective function can be understood as the negative logarithm of probability allocated to the true class (true metabolites). During the training period, we use five-fold cross-validation to divide all the data sets into five parts, one of them is taken as the validation set without repetition, and the other four are used as training model of training set. Cross-validation can avoid over-fitting and under-fitting, and the final results are more convincing.

# RESULTS AND DISCUSSION

## Accuracy of Prediction Results

Following the above steps, we cross-validated the model with five folds by using 200 epochs. The training set, validation set, test set segmentation was 7:1:2. The objective of the training period is to minimize the cross-entropy loss of classification, which is the

**TABLE 1 |** The most common type of reactions and SMART fragments in the dataset.

| Template SMART | Example |
|---|---|
| O = C-[NH;+0:1]-[C:2]> > [C:2][NH2;+0:1] |  |
| O = C-[NH;+0:1]-[c:2]> > [NH2;+0:1]-[c:2] |  |
| O = C-[O;H0;+0:1]-[C:2]> > [C:2]-[OH;+0:1] |  |
| [C]-[O;H0;+0:1]-[C:2] = [O:3]> > [O:3] = [C:2]-[OH;+0:1] |  |
| [C:1]-[N;H0;+0:2](-[C:3])-[C:4]> > [C:1]-[N+;H0:2](-[C:3])(-[C:4]) |  |
| [c:1]-[S;H0;+0:2] [c:3]> > O = [S;H0;+0:2](-[c:1])-[c:3] |  |
| [C:1]-[CH2;+0:2]-[C:3]> > O-[CH;+0:2](-[C:1])-[C:3] |  |
| [c:1]:[n;H0;+0:2]:[c:3]> > [O-]-[n+;H0:2](:[c:1]):[c:3] |  |
| [c:1]:[cH;+0:2]:[c:3]> > O-[c;H0;+0:2](:[c:1]):[c:3] |  |
| [C:1]-[S;H0;+0:2]-[C:3]> > O = [S;H0;+0:2](-[C:1])-[C:3] |  |

natural logarithm of probability allocated to real metabolites. Considering that there may be more than one metabolic reaction product for a drug, we believe that the top ten predicted products may have more reference value. As shown in **Table 2**, the model achieves an average test set accuracy of 70% for Top-10 in the five-fold cross-validation. In addition, we also calculated the accuracy of Top-1, Top-3, and Top-6 rankings. Since our metabolites are generated automatically by the metabolic template obtained by the algorithm, as long as the template is matched, the reaction products can be formed. There is a problem with the explosion of potential metabolite combinations. It is a great challenge for the model to hit the

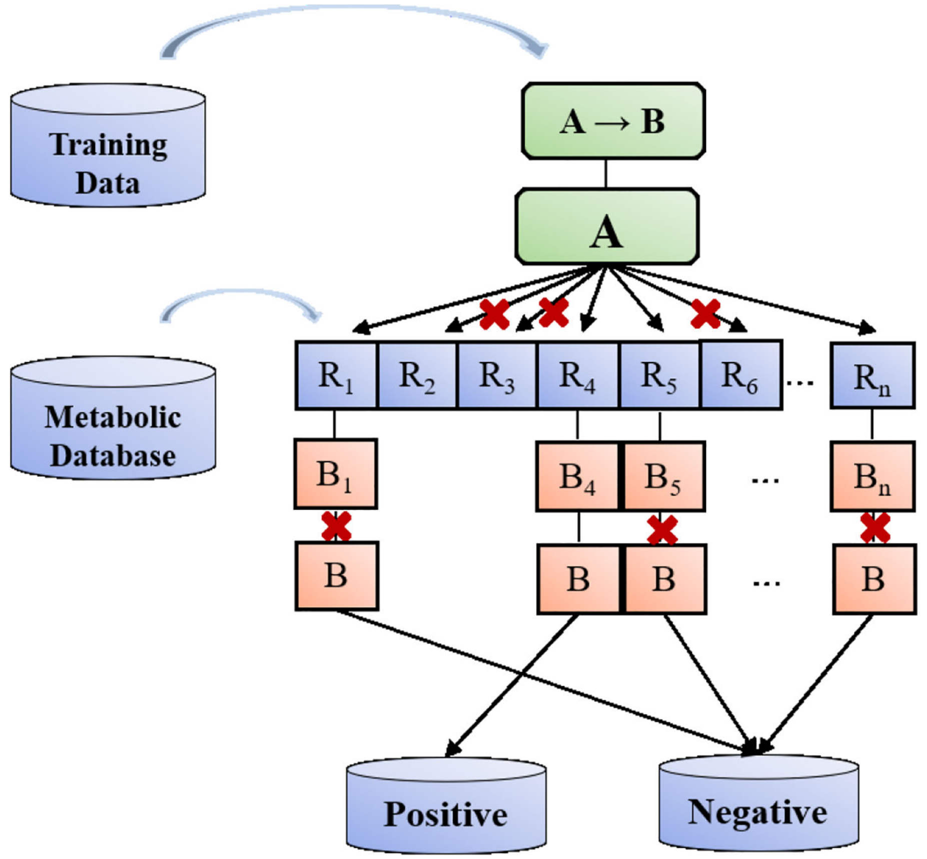


**FIGURE 4 |** Flow chart for potential metabolites production.

**TABLE 2 |** Prediction accuracies of the test set.

| | Accuracy |
|---|---|
| Top-1 | 34% |
| Top-3 | 51% |
| Top-6 | 68% |
| Top-10 | 70% |

product of the real reaction in the reaction product, but at the same time, the model can learn a lot of false product information because of the production of a large number of false metabolites, thus enhancing the learning ability.

Here, we conducted external tests on 300 reaction records that were not used for model training. It is also compared with the rule-based prediction method SyGMa. The accuracy rates of Top-1, Top-3, Top-6, and Top-10 are 35, 55, 67, and 78% respectively for our method (**Figure 5**). The accuracies of SyGMa for Top-1, Top-3, Top-6, and Top-10 are 20, 39, 50, and 70 respectively. The accuracy rates of our method are higher than SyGMa's. The main reason is that SyGMa does not produce the correct metabolites in some reactions.

As can be seen from **Figure 6**, correctly predicted metabolic reaction products by our method are common metabolic reactions, because these types of reaction samples account for the vast majority of the training set. Some of the metabolic reactions that cannot be correctly predicted are due to reaction types being uncommon with fewer occurrences in data sets, or because the reactants are too complicated and have multiple reaction sites. Furthermore, because usually only one metabolite of a compound is recorded in the reaction record, the Top-1 metabolite predicted by our method may not exactly be the recorded one, but it may still be one of the metabolites. Besides, in the reaction record involving multi-site and multi-step reactions, we can only predict a single-step metabolic reaction at one site. For our model, it is difficult for us to learn the changes in ring-opening and ring-closing reactions, because too much information is lost in those processes. It is difficult to characterize those metabolic processes only by the molecular fingerprints of reactants and products.

The amount of our data is small for Deep Learning to learn more information. The more reaction records that focused on a specific reaction, the more accurate the prediction of the reaction is. Thus we need to expand the data set for training. Next, we will collect more and more metabolic reactions from KEGG and other databases to train models, so as to improve the prediction accuracy of the models.

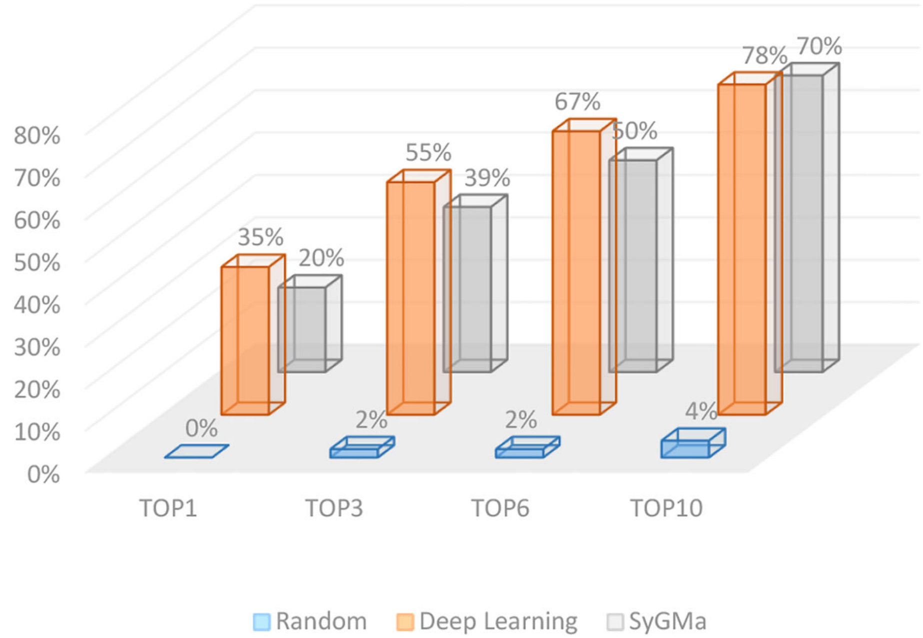


**FIGURE 5 |** Comparison results on external test set.

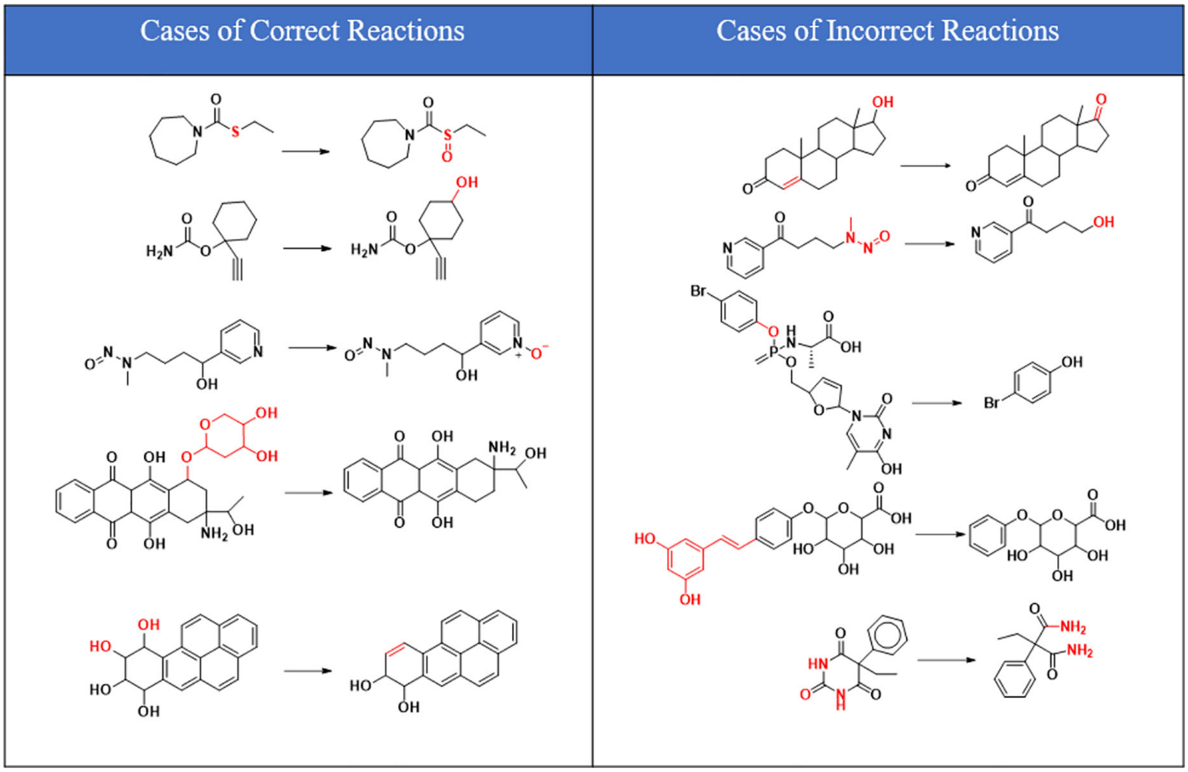


**FIGURE 6 |** Reaction cases for correct and incorrect predictions.

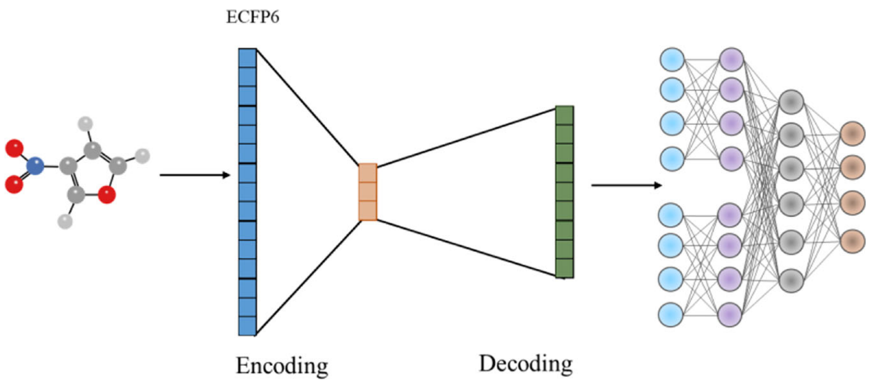


**FIGURE 7 |** Flow chart of AutoEncoder combined with molecular fingerprint.

**TABLE 3 |** Prediction accuracies at molecular fingerprint radius of 3.

| | Accuracy |
|---|---|
| Top-1 | 32% |
| Top-3 | 51% |
| Top-6 | 68% |
| Top-10 | 81% |

## Influences of Molecular Fingerprint Radius on the Results

We retrieved Morgan molecular fingerprints with radius 3 from potential metabolic reaction products in training set and retrained them with AutoEncoder algorithm (**Figure 7**). Morgan molecular fingerprint with a radius of 3 is equivalent to ECFP6, which will contain more information about molecular fragments.

As shown in **Table 3**, increasing fingerprint radius did not improve the prediction accuracy of Top-1 and Top-3, but did improve the prediction accuracy of Top-10. The results suggested that increasing fingerprint radius can improve the accuracy of the model to a certain extent, and AutoEncoder algorithm can help improve the prediction ability of the model as well.

Here we take Zileuton as an example to analyze its prediction results of metabolites. It is an inhibitor of 5-lipoxygenase for the maintenance treatment of asthma. The main metabolic pathways of Zileuton are hydroxylation of benzene ring, oxidation of sulfur atoms on sulfur-containing heterocycles, and hydrolysis of nitrogen atoms on amide groups (Joshi et al., 2004) (**Figure 8** and **Table 4**).

Here, three metabolites of zileuton were predicted correctly by our method, namely hydroxylation of the benzene ring and oxidation of sulfur atoms on sulfur-containing heterocycles. But our model has not predicted the hydrolysis of N atom of the side chain amide. The possible reason is that our training set has too little reactions to this type and the model has not adequately learned.

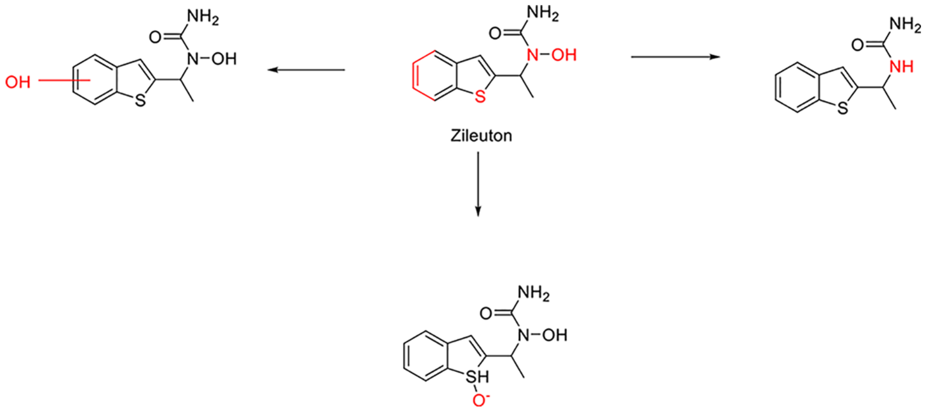
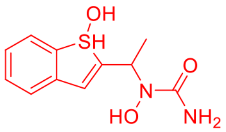
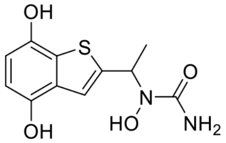
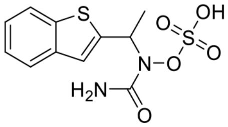
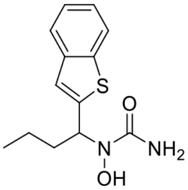
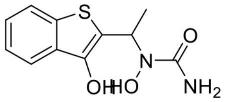
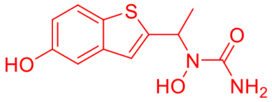
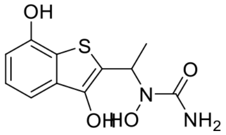
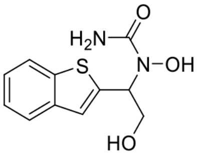
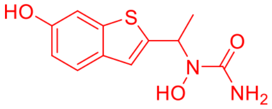
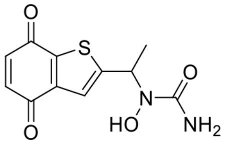


**FIGURE 8 |** Metabolic pathways of Zileuton.

**TABLE 4** | Top-10 predicted metabolites for Zileuton.

| Rank | Compounds |
|---|---|
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |

## CONCLUSION

In summary, we developed a deep learning based drug metabolites prediction algorithm to complement the experimental methods. By generating a broad coverage of metabolic reaction templates, we can generate a large number of potential metabolic reactants, and rank all metabolites by deep neural network algorithm to get the right metabolites ranked high. The accuracy of Top-1, Top-3, Top-6, and Top-10 in 300 external test sets with metabolic reactions is 35, 55, 67, and 78% respectively, which is significantly higher than that of random guess and the rule-based method SyGMa. Nevertheless, our method still has some limitations. It can rank the metabolites, but cannot give the probability of occurrence of metabolic sites. Besides, despite the relatively high prediction accuracy, it still has a high false-positive problem. To sum up, A approach of drug metabolites prediction based Deep learning was developed in this paper, which has certain predictive ability and can be used to provide some guidance information for researchers to improve the metabolic properties of lead compounds.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## AUTHOR CONTRIBUTIONS

DW conducted the research and wrote the paper. WL validated the model using known drugs and natural products. ZS helped collect the metabolic reaction data. LJ participated in building the metabolic reactions template library. JW tested the trained model. HL and SL designed and performed research, interpreted data, and approved the final manuscript.

## FUNDING

This work was supported by the National Key Research and Development Program (2016YFA0502304 to HL); the National Natural Science Foundation of China (grant 81825020 to HL, 81803437 to SL); the National Science & Technology Major Project “Key New Drug Creation and Manufacturing Program,” China (Number: 2018ZX09711002); the Fundamental Research Funds for the Central Universities, Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) under Grant No.U1501501. SL is also sponsored by Shanghai Sailing Program (No. 18YF1405100). HL is also sponsored by National Program for Special Supports of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

## REFERENCES

Avorn, J. (2015). The $2.6 Billion Pill — Methodologic and Policy Considerations. *New Engl. J. Med.* 372 (20), 1877–1879. doi: 10.1056/NEJMp1500848

Button, W. G., Judson, P. N., Long, A., and Vessey, J. D. (2015). Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *Cheminform* 34 (50), 1371–1377. doi: 10.1021/ci0202739

Cariello, N. F., Wilson, J. D., Britt, B. H., Wedd, D. J., Burlinson, B., Gombar, V., et al. (2002). Comparison of the computer programs DEREK and TOPKAT to predictbacterial mutagenicity. *Mutagenesis* 17 (4), 321–329. doi: 10.1093/mutage/17.4.3

Damsky, W. E., and Bosenberg, M. (2012). From bedding to bedside: genetically engineered mouse models of cancer inform concurrent clinical trials. *Pigment Cell Melanoma Res.* 25 (4), 404–405. doi: 10.1111/j.1755-148X.2012.01013.x

Darvas, F.(1988). Predicting metabolic pathways by logicprogramming. *J. Mol. Graphics* 6 (2), 80–86. doi: 10.1016/0263-7855(88)85004-5

Di, L. (2014). The role of drug metabolizing enzymes in clearance. *Expert Opin. Drug Metab. Toxicol.* 10 (3), 379–393. doi: 10.1517/17425255.2014.876006

Diao, X., Scheidweiler, K. B., Wohlfarth, A., Pang, S., Kronstrand, R., and Huestis, M. A. (2016). *In vitro* and *in vivo* Human Metabolism of Synthetic Cannabinoids FDU-PB-22 and FUB-PB-22. *AAPS J.* 18 (2), 455–464. doi: 10.1208/s12248-016-9867-4

Dickson, M., and Gagnon, J. P. (2004). The cost of new drug discovery and development. *Discovery Med.* 4 (22), 172–179.

Dimasi, J. A., Grabowski, H. G., and Hansen, R. W. (2015). The Cost of Drug Development. *New Engl. J. Med.* 372 (20), 1972. doi: 10.1056/NEJMc1503146

Ekins, S., Andreyev, S., Ryabov, A., Kirillov, E., Rakhmatulin, E. A., Bugrim, A., et al. (2005a). Computational prediction of human drugmetabolism. *Expert Opin. Drug Metab. Toxicol.* 1 (2), 303–324. doi: 10.1517/17425255.1.2.303

Ekins, S., Nikolsky, Y., and Nikolskaya, T. (2005b). Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol. Sci.* 26 (4), 202–209. doi: 10.1016/j.tips.2005.02.006

Ekins, S., Andreyev, S., Ryabov, A., Kirillov, E., Rakhmatulin, E. A., Sorokina, S., et al. (2006). A combined approach to drug metabolism and toxicity assessment. *Drug Metab. Disposition* 34 (3), 495–503. doi: 10.1124/dmd.105.008458

Embrechts, M. J., and Ekins, S. (2007). Classification of metabolites with kernel-partial least squares (K-PLS). *Drug Metab. Disposition* 35 (3), 325–327. doi: 10.1124/dmd.106.013185

Fox, T., and Kriegl, J. M. (2006). Machine learning techniques for *in silico* modeling of drug metabolism. *Curr. Topics Medicinal Chem.* 6 (15), 1579–1591. doi: 10.2174/156802606778108915

Gabriele, C., Carosati, E., De.Boeck., B., Ethirajulu., K., Mackie, C., Trevor Howe, A., et al. (2005). MetaSite: understanding metabolism in human cytochromes from the Perspective of the Chemist. *J. Medicinal Chem.* 48 (22), 6970–6979. doi: 10.1021/jm050529c

Gleeson, M. P., Hersey, A., and Hannongbua, S. (2011). In-Silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. *Curr. Topics Medicinal Chem.* 11 (4), 358–381. doi: 10.2174/156802611794480927

Grant, J. A., Pickup, B. T., and Nicholls, A. (2001). A smooth permittivity function for Poisson–Boltzmann solvation methods. *J. Comput. Chem.* 22 (6), 608–640. doi: 10.1002/jcc.1032

Hennemann, M., Friedl, A., Lobell, M., Keldenich, J., Hillisch, A., Clark, T., et al. (2010). CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *Chemmedchem* 4 (4), 657–669. doi: 10.1002/cmdc.200800384

Hwang, T. J., Carpenter, D., and Lauffenburger, J. C. (2016). Failure of Investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern Med.* 176 (12), 1826–1833. doi: 10.1001/jamainternmed.2016.6008

Joshi, E. M., Heasley, B. H., Chordia, M. D., and Macdonald, T. L. (2004). *In vitro* metabolism of 2-acetylbenzothiophene: relevance to zileuton hepatotoxicity. *Chem. Res. Toxicol.* 17 (2), 137–143. doi: 10.1021/tx0341409

Kang, Y. P., Ward, N. P., and Denicola, G. M. (2018). Recent advances in cancer metabolism: a technological perspective. *Exp. Mol. Med.* 50 (4), 31. doi: 10.1038/s12276-018-0027-z

Kirchmair, J., Howlett, A., Peironcely, J. E., Murrell, D. S., Williamson, M. J., Adams, S. E., et al. (2013). How do metabolites differ from their parent molecules and how are they excreted? *J. Chem. Inf. Modeling* 53 (2), 354–367. doi: 10.1021/ci300487z

Klopman, G., Dimayuga, M., and Talafous, J. (1994). META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.* 34 (6), 1320. doi: 10.1021/ci00022a014

Klopman, G., Tu, M., and Talafous, J. (1997). META. 3. A genetic algorithm for metabolic transform priorities optimization. *J. Chem. Inf Comput. Sci.* 37 (2), 329–334. doi: 10.1021/ci9601123

Korolev, D., Balakin, K. V., Nikolsky, Y., Kirillov, E., Ivanenkov, Y. A., Savchuk, N. P., et al. (2003). Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *J. Medicinal Chem.* 46 (17), 3631–3643. doi: 10.1021/jm030102a

Krejsa, C. M., Horvath, D., Rogalski, S. L., Penzotti, J. E., Mao, B., Barbosa, F., et al. (2003). Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discovery Devel* 6 (4), 470–480.

Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., et al. (2009). Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Modeling* 49 (3), 593. doi: 10.1021/ci800228y

Lazar, M. A., and Birnbaum, M. J. (2012). Physiology. *De-meaning Metab. Sci.* 336 (6089), 1651–1652. doi: 10.1126/science.1221834

Livingstone, D. J. (2010). The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf Comput. Sci.* 31 (23), 195–209. doi: 10.1021/ci990162i

Mackenzie, P. I., Somogyi, A. A., and Miners, J. O. (2017). Advances in Drug Metabolism and Pharmacogenetics Research in Australia. *Pharmacol. Res.* 116, 7–19. doi: 10.1016/j.phrs.2016.12.008

Meng, J., Li, S., Liu, X., Zheng, M., and Li, H. (2017). RD-Metabolizer: an integrated and reaction types extensive approach to predict metabolic sites and metabolites of drug-like molecules. *Chem. Cent. J.* 11 (1), 65. doi: 10.1186/s13065-017-0290-4

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* 9 (3), 203–214. doi: 10.1038/nrd3078

Ridder, L., and Wagener, M. (2010). SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *Chemmedchem* 3 (5), 821–832. doi: 10.1002/cmdc.200700312

Rydberg, P., Gloriam, D. E., and Olsen, L. (2010a). The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* 26 (23), 2988–2989. doi: 10.1093/bioinformatics/btq584

Rydberg, P., Gloriam, D. E., Zaretzki, J., Breneman, C., and Olsen, L. (2010b). SMARTCyp: a 2D method for prediction of cytochrome p450-mediated drug metabolism. *ACS Medicinal Chem. Lett.* 1 (3), 96–100. doi: 10.1021/ml100016x

Scott, B., Amby., C. H., Carlsson, L., Smith, J., Stein, V., and Glen, R. C. (2007). Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Modeling* 47 (2), 583. doi: 10.1021/ci600376q

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555 (7698), 604. doi: 10.1038/nature25978

Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X. Q., Doweyko, A., and Yi, L. (2003). *In silico* ADME/Tox: why models fail. *J. Comput. Aided Mol. Des.* 17 (2-4), 83–92. doi: 10.1023/A:1025358319677

Talafous, J., Sayre, L. M., Mieyal, J. J., and Klopman, G. (1994). META. 2. A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.* 34 (6), 1326. doi: 10.1021/ci00022a015

Tan, B. H., Pan, Y., Dong, A. N., and Ong, C. E. (2017). *In vitro* and *in silico* Approaches to Study Cytochrome P450-Mediated Interactions. *J. Pharm. Pharm Sci.* 20 (1), 319. doi: 10.18433/J3434R

Testa, B., Balmat, A. L., Long, A., and Judson, P. (2010). Predicting drug metabolism–an evaluation of the expert system METEOR. *Chem. Biodiversity* 2 (7), 872–885. doi: 10.1002/cbdv.200590064

Zhang, T., Chen, Q., Li, L., Angela Liu, L., and Wei, D. Q. (2011). In silico prediction of cytochrome P450-mediated drug metabolism. *Comb. Chem. High Throughput Screening* 14 (5), 388–395. doi: 10.2174/138620711795508412

**Conflict of Interest:** Author WL was employed by company Jiangzhong Pharmaceutical Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MZ declared a past co-authorship with one of the authors HL to the handling editor.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 31 January 2020
doi: 10.3389/fphar.2019.01665

# Patient-Level Effectiveness Prediction Modeling for Glioblastoma Using Classification Trees

Tine Geldof[1,2], Nancy Van Damme[3], Isabelle Huys[2†] and Walter Van Dyck[1,2*†]

[1] Healthcare Management Centre, Vlerick Business School, Ghent, Belgium, [2] Department of Pharmaceutical and Pharmacological Sciences, Research Centre for Pharmaceutical Care and Pharmaco-economics, KU Leuven, Leuven, Belgium, [3] Belgian Cancer Registry, Brussels, Belgium

**Objectives:** Little research has been done in pharmacoepidemiology on the use of machine learning for exploring medicinal treatment effectiveness in oncology. Therefore, the aim of this study was to explore the added value of machine learning methods to investigate individual treatment responses for glioblastoma patients treated with temozolomide.

**Methods:** Based on a retrospective observational registry covering 3090 patients with glioblastoma treated with temozolomide, we proposed the use of a two-step iterative exploratory learning process consisting of an initialization phase and a machine learning phase. For initialization, we defined a binary response variable as the target label using one-by-one nearest neighbor propensity score matching. Secondly, a classification tree algorithm was trained and validated for dividing individual patients into treatment response and non-response groups. Theorizing about treatment response was then done by evaluating the tree performance.

**Results:** The classification tree model has an area under the curve (AUC) classification performance of 67% corresponding to a sensitivity of 0.69 and a specificity of 0.51. This result in predicting patient-level response was slightly better than the logistic regression model featuring an AUC of 64% (0.63 sensitivity and 0.54 specificity). The tree confirms confounding by age and discovers further age-related stratification with chemotherapy-treatment dependency, both not revealed in preceding clinical studies. The model lacked genetic information confounding treatment response.

**Conclusions:** A classification tree was found to be suitable for understanding patient-level effectiveness for this glioblastoma–temozolomide case because of its high interpretability and capability to deal with covariate interdependencies, essential in a real-world environment. Possible improvements in the model's classification can be achieved by including genetic information and collecting primary data on treatment response. The model can be valuable in clinical practice for predicting personal treatment pathways.

**Keywords: real world evidence, oncology, exploratory study, propensity score modeling, decision tree, machine learning**

OPEN ACCESS

***Edited by:***
*Alex Zhavoronkov,*
*Biogerontology Research Foundation,*
*United Kingdom*

***Reviewed by:***
*Purna Mukherjee,*
*Boston College, United States*
*Evgeny Izumchenko,*
*University of Chicago, United States*

****Correspondence:***
*Walter Van Dyck*
*walter.vandyck@vlerick.com*

*†These authors share senior authorship*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *20 May 2019*
***Accepted:*** *19 December 2019*
***Published:*** *31 January 2020*

***Citation:***
*Geldof T, Van Damme N, Huys I and Van Dyck W (2020) Patient-Level Effectiveness Prediction Modeling For Glioblastoma Using Classification Trees. Front. Pharmacol. 10:1665. doi: 10.3389/fphar.2019.01665*

## INTRODUCTION

Glioblastoma is one of the most common and aggressive brain tumors in adults, with a median survival of less than one year from the time of diagnosis. Apart from the current standard of care treatment based on surgical resection and post-operative radiotherapy, there is only one medicinal product available for the treatment of glioblastoma patients. This temozolomide intervention has been shown to be efficacious in prolonging survival in Randomized Controlled Trials (RCTs) (Stupp et al., 2005; Stupp et al., 2009).

However, specific details on the mechanisms that drive individual response to temozolomide treatment in clinical practice, or on the drivers of real-world patient-level treatment effectiveness, are unknown (van Genugten et al., 2010; Eichler et al., 2011; Liu et al., 2016). To study these personal responses, traditional cohort-oriented methods, such as the Kaplan-Meier survival techniques currently used in pharmacoepidemiology (Strom and Kimmel, 2006) for investigating real-world evidence (RWE) data, have shown to be inadequate because of their difficulties to cope with heterogeneous patient populations; their restrictive assumptions regarding linear relationships among variables; their inability to provide patient-level predictions; and their inability to infer causality (Ankarfeldt et al., 2017; Arora et al., 2019).

For example, Kaplan-Meier methods provide (sub) population-level results, that is, they return the average or median treatment effect rather than patient-level results. Other statistical methods commonly used in the domain of medicine, such as logistic regression models, have hitherto focused mainly on investigating survival probability and their associated confounding factors when used in pharmacoepidemiology, as opposed to treatment effectiveness (Burke et al., 1995).

While currently primarily investigated for their application in drug discovery and development (Vamathevan et al., 2019), Onukwugha et al. (2017) suggested machine learning to be a valuable tool in pharmacoepidemiology as well aiming at studying this personal treatment's effectiveness (Onukwugha et al., 2017). Specifically, conducting exploratory treatment effectiveness studies using machine learning generates new knowledge on whether and how the treatment works in its specific real-world population and health care system context by accurately making individual predictions (Onukwugha et al., 2017; Berger et al., 2017; Puranam et al., 2018). These methods are increasingly being used by oncologists for cancer detection and prediction of risks, cancer recurrence, and survival (Lavrac, 1999; Kononenko and Kukar, 2001). Henceforward, machine learning develops as an alternative for traditional survival methods because it can be used for hypotheses generation on patient-level treatment effects in heterogenetic real-word patient populations, among others, through causal assessments (Vamathevan et al., 2019; Lavrac, 1999; Kononenko and Kukar, 2001; Cruz and Wishart, 2006; Onukwugha et al., 2017; Berger et al., 2017; Puranam et al., 2018). However, only little research has been done so far to explore the value of machine learning in pharmacoepidemiology (Crown, 2015).

In this paper, we present information-based machine learning methods – decision tree-based classification or classification trees (CT)—for use in a two-step iterative exploratory learning process to investigate the stratification factors of individual treatment response to temozolomide in glioblastoma patients using observational data. The well-known CT technique can then be used for patient-level effectiveness predictions of temozolomide.

## MATERIALS AND METHODS

To investigate the effects of real-world data (RWD) covariates on real-world treatment response on a patient-level basis and to be able to identify confounding factors influencing real-world treatment response, the methods that are used should allow for product performance-based data labeling if no primary data are available on real performance per patient. Hence, these models should use patient-level information and be able to handle personal treatment paths and/or genomic information. In this section, we will first describe the data collection process and provide a definition of the product's performance used to annotate the data set. Next, we will describe the classification models and exploratory learning process used for theorizing about personal treatment effectiveness.

### Data Setting

In this study, data were extracted from the Belgian Cancer Registry (BCR), including 4587 patients with glioblastoma (ICD-10 code C71.0-C71.9) diagnosed between 2004 and 2012, and vital status information updated until January 1, 2015. Variables for this study were taken from the full standard set of variables nationally collected by the BCR—including patient and tumor characteristics—and Inter Mutualistic Agency (IMA), including reimbursed therapeutic acts consisting of medical acts and medications administrated in hospitals and handed out in pharmacies. These variables were further limited by BCR oncologists for their potential relevance in the analysis.

The index date, or date of incidence of glioblastoma, was defined as the date of first microscopic confirmation of malignancy, first hospitalization for the cancer, first consultation for the malignancy, first clinical or technical diagnosis, start of treatment, or date of death, whichever date came first. Patients with incidence dates that were the same as the date of death as well as patients without a social security identification number were excluded.

Temozolomide therapy relevant for the treatment of glioblastoma was extracted from the IMA data set based on the medicines' anatomical therapeutic chemical (ATC) code (L01AX03) and treatment start data within −1 to 9 months from the date of incidence. Other chemotherapeutic interventions with possible interactive effects were extracted from the IMA data set based on the ATC code for chemotherapy (L01), starting −1 month from the date of incidence. Information on radiotherapeutic (RT) interventions, biopsy, and surgical resection were extracted from the IMA data set by BCR51 oncologists based on the relevant nomenclature codes used.

The final data set consisted of (a) the patient's overall survival (OS) period, a continuous variable calculated as the difference between the date of death or last confirmation that the patient was alive and date of incidence; (b) treatment path, that is, binary variables indicating biopsy and/or surgical resection and RT, and chemotherapeutic treatment; (c) five discrete covariates (age, tumor differentiation grade, topography, total number of tumors, and World Health Organization [WHO] performance score at diagnosis and recursive partitioning analysis [RPA] class), one binary covariate (sex), and one categorical covariate (tumor topography, specifying the location in the brain), confounding both the patient's OS and treatment path; and (d) OS binary observation status specifying whether the survival was censored, that is, whether the follow-up time was too short to observe the date of death. The final RWD set consisted of 4528 patients, of which 3090 treated with temozolomide (**Table 1**).

## Definitions

Because no primary data on treatment response was available for temozolomide, initialization was needed to label the data. For this purpose, a binary dependent variable with variables 1 and 0 representing individual-treatment response and non-response, respectively, was created based on the patients' gain in OS, that is, the number of months the patient gained in survival when being assigned to the temozolomide treatment. Here, OS was used as the main indicator of the treatment effect because this was the RCT's primary endpoint. Patients' gain in OS was calculated using nearest neighbor propensity score (PS) matching, a method commonly used on RWD to mitigate bias induced by the non-random assignment of treatments. Hence, let $T$ and $C$ be the set of treated (Z = 1) and control (Z = 0) patients, respectively. The PS = $\Pr(Z_i = 1|X_i)$ is defined as the probability of being assigned to the treatment of consideration conditional on the observed covariates X. Its value is estimated using a logit model (Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984) with the selected covariates X being the observed variables which significantly affect the survival time, because this variable selection approach is associated with better PS estimations (see supplementary materials for more details) (Austin et al., 2007). Following this nearest neighbor PS technique, each temozolomide-treated patient is matched to k control patients based on the smallest difference in estimated PSs, that is, $i \in T$ and $j \in C$ are matched if $dist\ (PS_i^T, PS_j^C)$ is minimal (Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984).

Here, we chose to set k equal to 10, given a set of 1438 control patients, to not average out possible covariate effects. This nearest neighbor PS matching algorithm was performed with the "MatchIt" 125 package within R (Ho et al., 2011).

Further, let $Y_T = OS_T$ and $Y_C = OS^C$ be the observed continuous outcomes of the treated and control units, respectively. Denote by $C\ (i)$ the set of k control patients $j \in C$ matched to the treated patient $i \in T$. Define the weights $w_{ij} = 1/k$ if $j \in (i)$ and $w_{ij} = 0$ otherwise. From the formula for the average treatment effect (Ho et al., 2011), we defined the treated patient's survival gain (SG): $SG_i = OS_i^T - \sum_{j \in C(i)} w_{ij} OS_j^C$. Following the guidelines of the European Society for Medical Oncology (ESMO) andMagnitude of Clinical Benefit scale (MCBS) and with the aim of maximizing treatment response rate(TRR) (Becker and Ichino, 2002), patients were labeled with "response" whenever their SG was longer than the threshold $\lambda$ equal to one month (Cherny et al., 2015).

**TABLE 1** | Main characteristics of the real-world study population.

| | Real-World | |
|---|---|---|
| | **Control group (n = 1438)** | **Treated group (n = 3090)** |
| **Age** | | |
| Range (median) | 0–94 (74) | 5–98 (61) |
| no. (%) < 50 | 98 (42%) | 582 (19%) |
| no. (%) >= 50 | 1,340 (58%) | 2,508 (81%) |
| **Sex – no. (%)** | | |
| Male | 814 (57%) | 1,847 (60%) |
| Female | 624 (43%) | 1243 (40%) |
| **WHO performance status—n (%)** | | |
| 0—asymptomatic | 253 (18%) | 415 (13%) |
| 1—symptomatic but completely ambulatory | 850 (59%) | 2265 (73%) |
| 2—symptomatic, up and about >50% walking hours | 197 (14%) | 313 (10%) |
| 3—symptomatic, confined to bed/chair > 50% walking | 84 (6%) | 61 (2%) |
| hours | 54 (4%) | 36 (1%) |
| 4—completely disabled; totally confined to bed/chair | | |
| **RPA—n (%)** | | |
| Class III† | 43 (3%) | 162 (5%) |
| Class IV‡ | 789 (55%) | 2,419 (78%) |
| Class V § | 606 (42%) | 509 (16%) |
| **Surgical procedure (biopsy/ debulking)—n (%)** | | |
| No | 169 (12%) | 23 (1%) |
| Yes | 1,269 (88%) | 3,067 (99%) |
| **Radiotherapy treatment—n (%)** | | |
| No | 899 (63%) | 130 (4%) |
| Yes | 539 (37%) | 2,960 (96%) |
| **Chemotherapy treatment—n (%)** | | |
| No | 1,342 (93%) | 2,277 (74%) |
| Yes | 96 (7%) | 813 (26%) |
| **Time from diagnosis to radiotherapy:** range (median) | 377.0–256.3 (Arora et al., 2019) | −313.6 to 186.9 (Vamathevan et al., 2019) |
| **Time from diagnosis to chemotherapy:** range (median) | −4.0 to 190.0 (Stupp et al., 2005) | −4.3 to 389.7 (Burke et al., 1995) |

*Patients were categorized according to recursive partitioning analysis (RPA) classes: †Age < 50 years and World Health Organization (WHO) status 0. ‡ Age < 50 years and WHO status > 0 or age ≥ 50 years and surgical resection. §Age ≥ 50 years and no surgical resection*

## Classification Model

We used classification techniques within machine learning to divide individual patients into treatment response and non-response groups, with the purpose to fully understand individual treatment response to temozolomide. For exploratory reasons, we used a CT to extract patterns from the data. CTs are highly interpretable and intuitive as well as well attuned to coping with missing data and heterogeneous data types (Kelleher et al., 2015). While recursively creating branches for different covariate values, ordered in function of their classification error minimization power, the CT algorithm (for details see **Supplementary Material**) gradually improves prediction accuracy. Missing data is handled by classifying these observations in branches based on surrogate variables, predicting the most likely missing variable value.

As pointed out by Puranam et al. (2018), we believe that our sample size of 3090 temozolomide-treated patients was sufficiently large to extract valuable evidence (Shaikhina et al., 2017). Although identification of the best classification model was not the main purpose of this research, we did compare this technique with a logistic regression model, one of the most commonly used statistical classification methods in the medicinal literature (Kononenko and Kukar, 2001).

The set of treated patients T was divided into a training set, comprising 80% (2472 units) of the temozolomide-treated patients sampled at random, and a test set, comprising the remaining 20% (618 units). The CT algorithm was trained and validated using 10-fold cross validation to obtain the most generalizable model using the "rpart" package within R, which implements the Classification and Regression Tree (CART) algorithm described by Breiman et al. (1984). Given that the difference between our defined binary response and predicted response by the classification model can be described by a confusion matrix, we can define the following properties: the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these properties, the true positive rate (TPR) and the true negative rate (TNR) are defined as $TPR = TP/(TP + FN)$ and $FPR = FP/(TN + FP)$, respectively. The CT and logistic regression model performance were then evaluated by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, mapping the models' sensitivity and specificity measured by the TPR and 1 −TNR, respectively (Fawcett, 2006; Cherny et al., 2015). The AUC and ROC curves were computed using the "pROC" package within R (Robin et al., 2011).

## Iterative Exploratory Learning Process

The focus of this study was on investigating the confounding factors and causal effects of individual treatment response to temozolomide. As classification methods within machine learning identify correlations but cannot by themselves reach causal inference (Puranam et al., 2018), further interpretation of the CT is required. We conducted a two-step iterative exploratory learning process, as depicted in **Figure 1**, which aids inductive theory building. This learning process consisted of the evaluation of (i) possible unobserved confounding variables, for example through expert consultation, and (ii) the redefinition of response as a target feature when not available as primary data, by changing TRR assumptions and/or using different response-identification algorithms. Iteration ended when no further improvements were obtained, giving the model's optimal AUC achievable in practice (see Appendix for pseudo-code).

# RESULTS

First, we will show results for the data labeling process for patients treated with temozolomide. Thereafter, the outcome of the trained and validated CT is given and evaluated. The training set for the CT model consisted out of 2472 temozolomide-treated patients. These CT results are finally compared to the results of the logistic regression model.

## Initialization: Binary Response Labeling

The observed covariates significantly affecting the survival time of temozolomide-treated patients included patients' age, RT, and chemotherapeutic treatment (p-value < 0.001), and WHO performance score (p-value < 0.01) (see supplementary materials for more details). Nearest neighbor PS matching based on these covariates resulted in 1063 control units matched once or multiple times to one treated unit. Following the ESMO-MCBS (Cherny et al., 2015), we obtained a TRR of 52%, meaning 1,607 of 3,090 temozolomide-treated patients showed SG > 1 month.

## Classification Results

The CART algorithm showed a maximal decrease in classification error when first dividing the treated patients according to their age (**Figure 2**, see supplementary materials for more details). Another covariate stratifying the training set included patients' chemotherapeutic treatment path, but such covariate interdependencies are currently not analyzed in RCT and treatment effectiveness studies (Stupp et al., 2005; Strom and Kimmel, 2006; Stupp et al., 2009; van Genugten et al., 2010). CT performance evaluation of the test data set resulted in an AUC of 0.6650 (**Figure 3A**). Compared to a model that is no better than a random classifier, featuring an AUC of 0.50, the CT performed better than chance but still showed poor prediction skills. Associated with this AUC was a sensitivity of 0.6850, meaning that 31% of patients who would benefit from the treatment were not recognized by the model, and a specificity of 0.5114, meaning that 49% of patients who would not benefit from the treatment were predicted to benefit by the model.

The logistic regression model achieved a slightly lower AUC of 0.6357 with a sensitivity of 0.6337 and specificity of 0.5420 (**Figure 3B**). Although they showed a better specificity than the CT, the results of the logistic regression model are still far too low.

## Iterative Exploratory Learning Process

With an AUC of 66.50% and 63.57% for the CT and logistic model, respectively, further interpretation of the model was done to obtain a higher sensitivity and lower specificity. In this temozolomide case, two learning steps were followed depicted in **Figure 1**: (i) theorization about possible unobserved confounding variables and (ii) redefinition of treatment response as a target feature. In the first case, a low AUC, which is associated with many misclassifications (false responders and non-responders), can result from the problem of spuriousness, suggesting that there may be some important

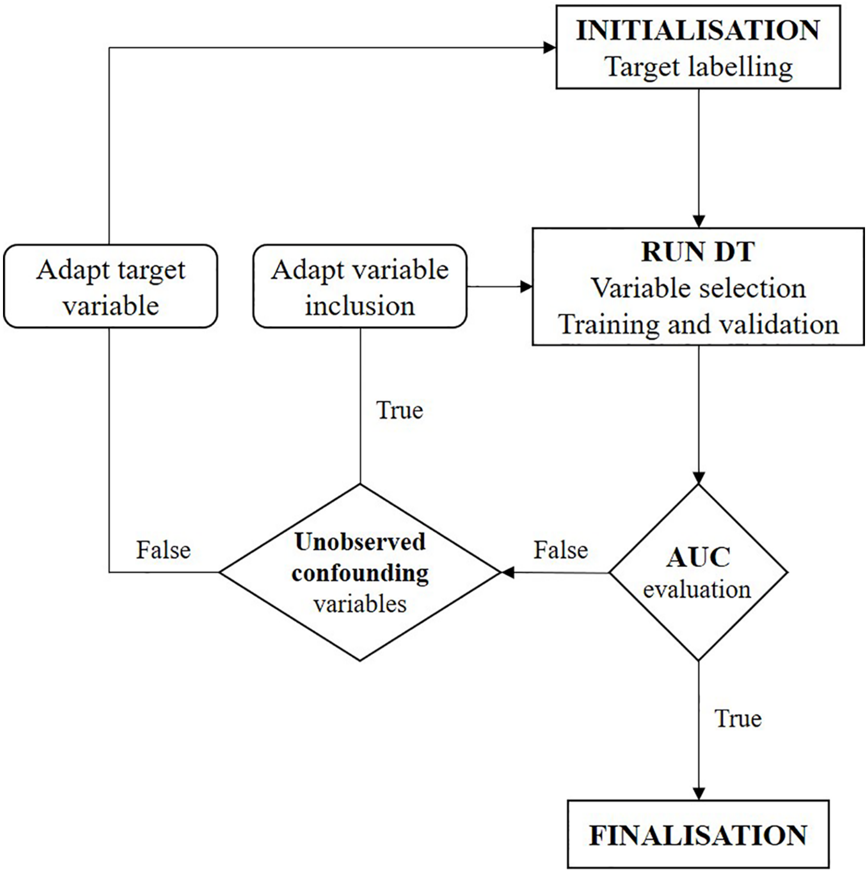


**FIGURE 1 |** Flowchart of two-step iterative exploratory learning process. The model is iterated until the area under the receiver operating characteristic curve (AUC) is satisfactory, i.e. until the highest achievable AUC in practice is found. Unobserved confounding variables are (unknown) variables currently not captured in real-world situations. (AUC, area under the receiver operating characteristic curve; CT, decision tree).

confounding variables that were omitted from the data set, that were not collected in the data source, or that were just unknown (i.e., not known from any translational research). As an example, from our case, the BCR does not dispose of genetic information such as the methylation of the promoter for the gene encoding O-6-methylguanine-DNA methyltransferase. However, based on clinical research literature, this appears to be associated with a higher survival benefit (Stupp et al., 2005; Stupp et al., 2009).

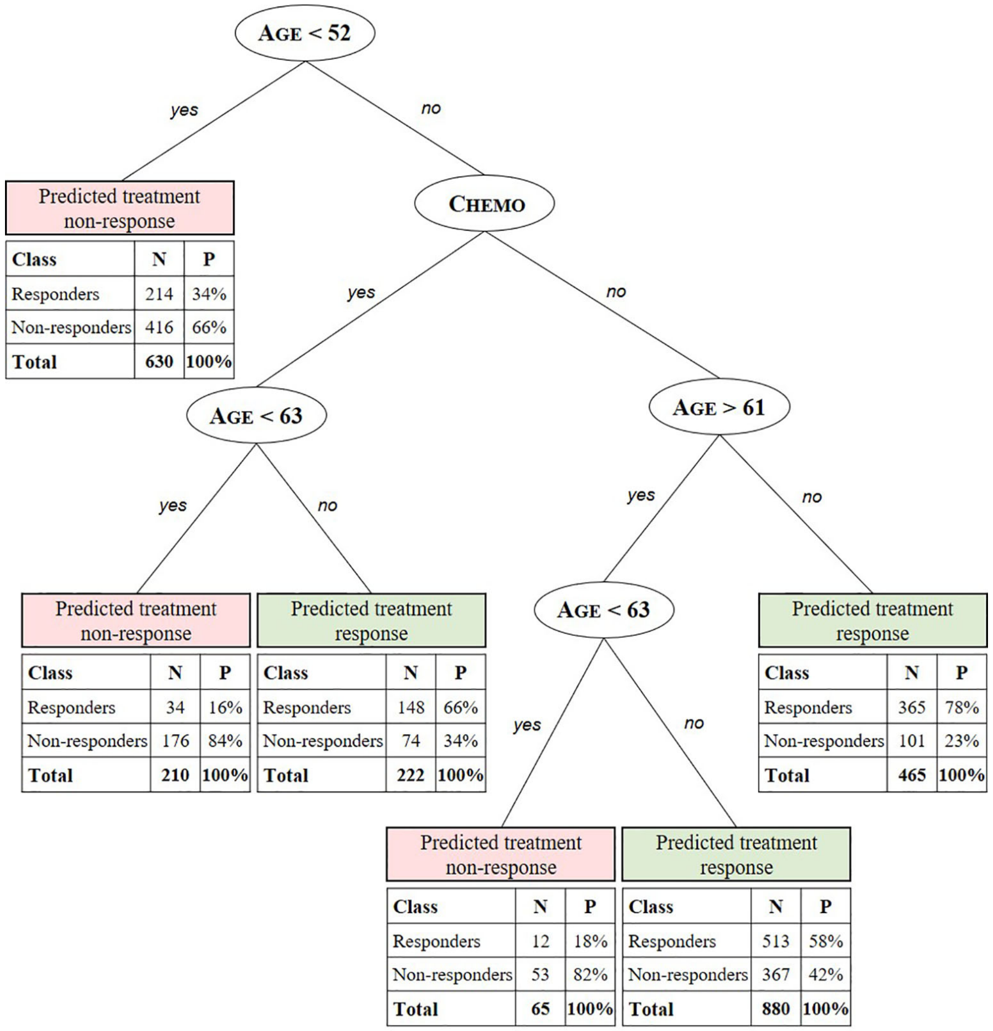
In the second case, one can modify the TRR definition. In our case, for example, modifying the threshold to 3 months (giving a TRR of 43%) in the algorithm led to a CT with a different structure and lower AUC of 0.6005 (see supplementary materials). Again, age and status of chemotherapeutic treatment were shown to be the main classification variables.

# DISCUSSION

Although the prediction structure induced by RWD confirms the importance of patient age, which was previously used as a stratification variable during RCT, the CT based on observational data reveals extra interdependencies of chemotherapy as a co-treatment effect, which was not found in preceding RCT-based studies. Such variable interdependencies cannot be investigated through current pharmacoepidemiology methods, including Kaplan Meier survival analysis techniques. In the following sections, we will discuss the causality assessment to generate hypotheses about personal treatment effectiveness and show the significance of this method. Next, we will discuss some limitations of the proposed method as well as possible issues with the data.

## Hypotheses Generation Through Exploratory Learning

Our CT model had an AUC of 67% with an associated sensitivity equal to 0.69 and specificity equal to 0.51. In the case of cancer treatments, a low specificity is undesirable because the treatment of false positives can be dangerous for the patient, depriving him or her of correct treatment, and can also be very costly, considering the high oncology drug prices during health care budget austerity. Therefore, theorizing about personal treatment


AGE < 52
yes
no
Predicted treatment non-response
CHEMO
yes
no
AGE < 63
AGE > 61
yes
no
yes
no
Predicted treatment non-response
Predicted treatment response
AGE < 63
Predicted treatment response
yes
no
Predicted treatment non-response
Predicted treatment response

| Class | N | P |
|---|---|---|
| Responders | 214 | 34% |
| Non-responders | 416 | 66% |
| **Total** | **630** | **100%** |

| Class | N | P |
|---|---|---|
| Responders | 34 | 16% |
| Non-responders | 176 | 84% |
| **Total** | **210** | **100%** |

| Class | N | P |
|---|---|---|
| Responders | 148 | 66% |
| Non-responders | 74 | 34% |
| **Total** | **222** | **100%** |

| Class | N | P |
|---|---|---|
| Responders | 365 | 78% |
| Non-responders | 101 | 23% |
| **Total** | **465** | **100%** |

| Class | N | P |
|---|---|---|
| Responders | 12 | 18% |
| Non-responders | 53 | 82% |
| **Total** | **65** | **100%** |

| Class | N | P |
|---|---|---|
| Responders | 513 | 58% |
| Non-responders | 367 | 42% |
| **Total** | **880** | **100%** |

**FIGURE 2 |** Summary predictive classification tree model after training and validation. Predicted stratification variables for TMZ in glioblastoma include age, RPA class, and chemotherapeutic (Chemo) and radiotherapeutic (RT) patient status. For each stratified patient class a confusion matrix indicates the number (N) and percentage (P) of treated patients from the test set for which the CT predicts treatment response correctly (responders to predicted treatment response and non-responders to predicted non-response) with respect to the labeled SG value. E.g. the CT model predicts the class of patients aged 52 to 61 years and >63 years not receiving concomitant or adjuvant chemotherapy to respond to the treatment with a true positive (TP) probability of 58%. For this class, with patients aged < 63 years 82% are correctly predicted (true negative [TN]) not to respond.

effectiveness was done following an iterative learning process. A starting point for the first learning step of the CT was to explain why false responders and non-responders were observed in the various groups because this could suggest that there are some essential variables not being collected in RWD, such as genomic information, or other unidentified factors confounding RWE that are not detected by cohort-oriented methods used in current efficacy and effectiveness studies. Mitigating this problem of spuriousness may be essential to avoid wrong causal conclusions. Thus, including known or yet unknown unobserved (depending on data set used) confounding variables, for example through expert-consolations or conducting translational research, may lead to a subsequent CART search to induce a CT with better prediction accuracy and possibly a higher specificity.

In the second learning step, one can experimentally modify the TRR definition (under the guidance of experts) and/or method. Ideally, this can be done by collecting a treatment response identifier as primary data from the data source, such as information on tumor growth. Here, the TRR was based on PS matching and a non-variable SG threshold of 1 month. Depending on the extent of the phenotype (e.g. blood pressure) and genotype (e.g. mutations) variable collection in RWD sources, advanced TRR identification algorithms can greatly improve the labeling.

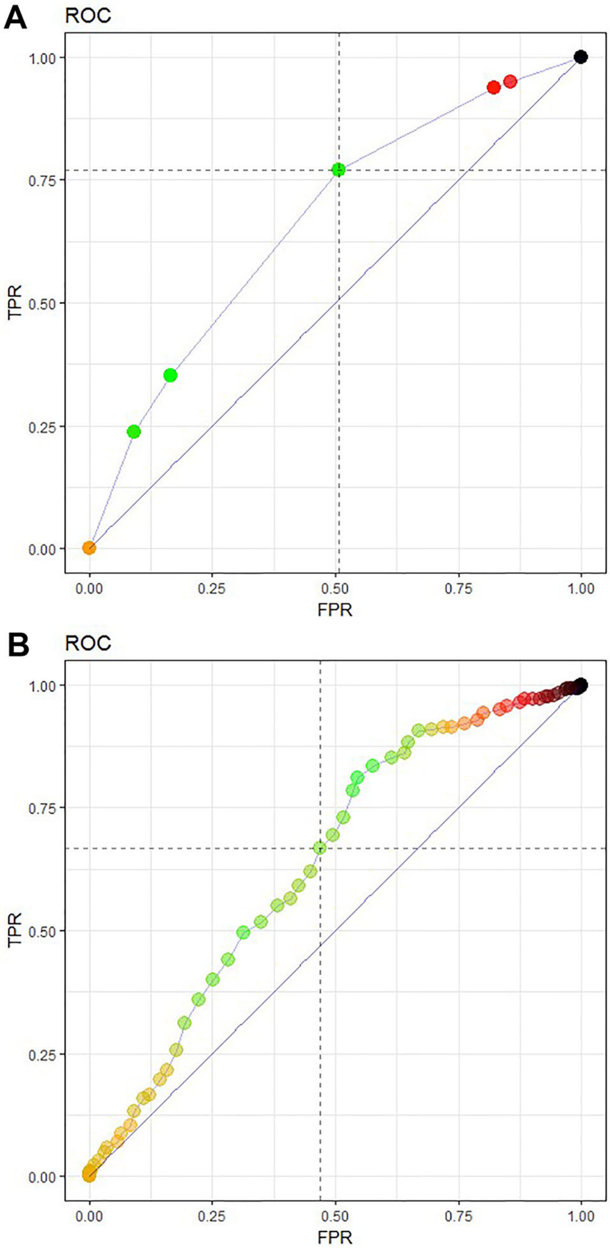


**FIGURE 3 |** Receiver operating characteristic (ROC) curve featuring model performance evaluation as an area under the curve (AUC), sensitivity TPR and FPR or (1-specificity) for **(A)** the CT prediction model 3 and **(B)** a logistic regression model of the test data set. The CT model **(A)** featured an AUC of 0.6650, a sensitivity of 0.6850, and a specificity of 0.5114, The logistic regression model **(B)** achieved a slightly lower AUC of 0.6357 with a sensitivity of 0.6337 and specificity of 0.5420.

## Patient-Level Effectiveness Prediction

We found a combination of age and chemotherapeutic treatment status to be the main stratification factors of real-world personal treatment response to temozolomide in glioblastoma. Additionally, further specifications of these factors not found in preceding RCT-based studies were discovered. For example, the CT predicts positive response to the treatment for patients being assigned to chemotherapeutic treatment and being older than 63 years with a probability of 66%. Additionally, patients aged 52 to 61 years and >63 years not receiving concomitant or adjuvant chemotherapy are predicted to respond to the treatment with a probability of 58%. Using the iterative learning process described in Hypotheses Generation Through Exploratory Learning section, a higher AUC and hence better predictions could be obtained when (un)known stratification factors are identified and included. As an example, in our case, the BCR does not yet dispose of genetic information, such as the methylated promoter for the gene encoding O-6-methylguanine-DNA methyltransferase, which is associated with a larger survival benefit (Stupp et al., 2005; Stupp et al., 2009; van Genugten et al., 2010). When the achieved AUC is satisfactory and thus treatment effectiveness is fully understood, that is, when all stratification and confounding variables are known, the model can be used for accurate patient-level effectiveness predictions.

## Significance of the Proposed Methodology for RWD

In this temozolomide-glioblastoma case, the CT was potentially useful for exploring covariate interdependencies and confounders of individual treatment responses. With this, the importance of factors yet unknown to previously conducted clinical research, such as phenotypical or genotypical variations, can easily be integrated and tested for their effects using this technique. Therefore, CTs may be valuable in terms of discovering variations in patient-level effectiveness of medicines, which might not be discovered otherwise. This confirms recent literature discussing the promise of machine learning techniques in pharmaceutical innovation and decision making (Reps et al., 2018; Beam and Kohane, 2019; Rajkomar et al., 2019). Therefore, we argue that RWE-based machine learning analysis can be used in exploratory treatment effectiveness studies (Berger et al., 2017; Puranam et al., 2018) for improving the understanding of TRR and the specification of treatment paths with a level of detail not previously achieved in pharmacoepidemiology studies of temozolomide. In practice, when considering cancers that are being treated following multiple sequences (e.g. first- to third-line treatments) with a range of different, possibly combined, interventions (as is the case for melanoma, colorectal, and breast cancer) in conjunction with a range of different diagnostic tools, the technique can also be useful for exploring and predicting optimal treatment sequences and therefore guide clinical decision making.

## Limitations of the Proposed Method

This study does not come without limitations. For the CT's predictive accuracy, the quality of the RWD is very important. Within health care, data sources may be of low veracity, that is, they may contain incomplete, imprecise, or inconsistent data. Data cleaning is an important step to mitigate this problem. Also, data sources may capture a low variety of information. Here, no primary data on treatment response was available, which required the use of PS matching to estimate personal treatment effect. Also, the BCR does not dispose of genetic information.

Additionally, we must note that the TRR definition did not consider survival censoring, that is, the OS of both treated and control patients were assumed to be uncensored. Fortunately, in this study, censoring was rarely observed given the severity of the disease; only 1% of matched cohort patients (13 of 1063) and 7% of treated patients (211 of 3090) had censored OS, and the latter was only of importance if the SG was less than one month because these would potentially be wrongfully classified as non-responsive. In such cases, the use of semi-supervised machine learning methods, where treatment response as the target feature is missing when the OS of either matched treated patient and/or matched control patient is censored, may improve these results.

Lastly, the used matching technique does not control for unobserved variables and does not consider early patient death before start of treatment. In our case, the latter may be important because of short patients' OS.

# CONCLUSIONS

Using machine learning, we showed an increased understanding of patient-level treatment responses and specification of individual treatment paths that were not be identified using cohort-oriented methods used in previous RCT studies. Through the iterative learning model, confounding factors can be identified to achieve the most optimal prediction model of patient-level effectiveness.

We believe that machine learning can be effective in the observational phase following "initial" licensing in an adaptive licensing approach, as suggested by Eichler et al. (2012), or in the pilot phase after licensing following Phase III pre-approval studies in the sequential study design suggested by Franklin et al. (2014). In both cases, machine learning can be used for exploratory treatment effectiveness studies where hypotheses are generated to further guide efficient designs of large-scale confirmatory observational trials, both in disease database and pragmatic RCTs.

The CT method was found to be the suitable for this case because of its high interpretability and capability to deal with covariate interdependencies. However, the CT is suitable up to a maximum level of complexity characterized by the number of baseline variables, amount of possible treatment pathways and their combinations, and extent of OS censoring. Thus, when considering medicinal products such as cetuximab or panitumumab for colorectal cancer, CTs become inadequate because more patients will have censored OS while receiving multiple and more combined treatments in different sequences depending on their genetic expression, resulting in a smaller sample-to-feature ratio. As a result, methods should account for label uncertainty, for example, by including the likelihood of the treatment response measure. Further studies involving predictive data analytics used for real-world effectiveness exploration are needed to determine whether more advanced techniques within machine learning should be considered to deal with the higher complexity in these cases. These methods include probability-based Bayesian classification, support vector machines, and neural networks conducted through supervised or semi-supervised learning.

# DATA AVAILABILITY STATEMENT

The data sets that support the findings of this study are available from the Belgian Cancer Registry but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Belgian Cancer Registry.

## AUTHOR CONTRIBUTIONS

TG and WD participated in the design of the research, interpretation of the results and writing the manuscript. TG performed the data retrieval, statistical analysis and made substantial contributions to the writing of the manuscript. ND provided the materials and assisted in data cleaning. TG, WD, IH, and ND read and amended the manuscript and approved the final version for publication.

## FUNDING

This work was supported by the Vlerick Business School Academic Research Fund. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and publishing the report.

## ACKNOWLEDGMENTS

This study benefitted from the contribution of Yves Moreau and the review of Edward De Brouwer, Department of Electrical Engineering (ESAT) STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics Department, University of Leuven, Belgium. We would also like to thank the BCR for providing us with unique access to historical observational data and for their research assistance.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found onlineat: https://www.frontiersin.org/articles/10.3389/fphar.2019.01665/full#supplementary-material

## REFERENCES

Ankarfeldt, M. Z., Adalsteinsson, E., Groenwold, R., Ali, S., and Klungel, O. (2017). A systematic literature review on the efficacy– effectiveness gap: comparison of randomized controlled trials and observational studies of glucose-lowering drugs. *Clin. Epidemiol.* 9, 41–51. doi: 10.2147/CLEP.S121991

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., and Druzdzel, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value Health* 22 (4), 439–445. doi: 10.1016/j.jval.2019.01.006

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat. Med.* 26 (4), 734–753. doi: 10.1002/sim.2580

Beam, A. L., and Kohane, I. (2019). Big data and machine learning in health care. *J. Am. Med. Assoc.* 319 (13), 1317–1318. doi: 10.1001/jama.2017.18391

Becker, S. O., and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata J. 4th Quarter* 2 (4), 358–3770. doi: 10.1177/1536867x0200200403

Berger, M. L., Sox, H., Willke, R. J., Brixner, D. L., Eichler, H. G., Goettsch, W., et al. (2017). Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Value Health* 6 (9), 1003–1008. doi: 10.1002/pds.4297

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. I. (1984). *Classification and Regression Trees* (Boca Raton: FL CRC Press), 18–55.

Burke, H. B., Rosen, D. B., and Goodman, P. H. (1995). Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. *Adv. Neural Inf. Process. Syst.* (Cambridge, MA;), 1064–1067.

Cherny, N. I., Sullivan, R., Dafni, U., Kerst, J. M., Sobrero, A., Zielinski, C., et al. (2015). A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European society for medical oncology magnitude of clinical benefit scale (ESMO-MCBS). *Ann. Oncol.* 26 (8), 1547–1573. doi: 10.1093/annonc/mdv249

Crown, W. H. (2015). Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health* 18 (2), 137–140. doi: 10.1016/j.jval.2014.12.005

Cruz, J., and Wishart, D. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77. doi: 10.1177/117693510600200030

Eichler, H. G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., et al. (2011). Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Clin. Pharmacol. Ther.* 97, 234–246. doi: 10.1038/nrd3501

Eichler, H. G., Oye, K., Baird, L. G., Abadie, E., Brown, J., Drum, C. L., et al. (2012). Adaptive licensing: taking the next step in the evolution of drug approval. *Clin. Pharmacol. Ther.* 91 (3), 426–437. doi: 10.1038/clpt.2011.345

Fawcett, T. (2006). *An introduction to ROC analysis, Pattern Recognition Letters.* (Elsevier), 27 (8), 861–874. doi: 10.1016/j.patrec.2005.10.010

Franklin, J., Rassen, J., and Bartels, D.Schneeweiss, S. (2014). Prospective cohort studies of newly marketed medications: using covariate data to inform the design of large-scale studies. *Epidemiol.* 44625 (1), 126–133. doi: 10.1097/EDE.0000000000000020

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* 42 (8), 1–28. doi: 10.18637/jss.v042.i08

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (Cambridge MA: The MIT Press).

Kononenko, I., and Kukar, M. (2001). Machine learning for medical diagnosis: history, state of the art, and perspective. *Artif. Intell. Med.* 23 (1), 89–109. doi: 10.1016/S0933-4023657(01)00077-X

Lavrac, L. (1999). Selected techniques for data mining in medicine. *Artif. Intell. Med.* 16 (1), 3–23. doi: 10.1016/S0933-3657(98)00062-1

Liu, L., Hummel, N., Mauer, M., Morais, E., and Olivares, R. (2016). PCN23 – A systematic literature review on the drivers of effectiveness and the efficacy-effectiveness gap in hematological malignancies with a focus on 376Hodgkin's Lymphoma. *Value Health* 19 (7), A712. doi: 10.1016/j.jval.2016.09.2095

Onukwugha, E., Bjarnadottir, M., Zhou, S., and Czerwinski, D. (2017). Visualizing data for hypothesis generation using large-volume claims data. *Value Outcomes Spotlight* 3 (1), 6–10.

Puranam, P., Shrestha, Y. R., He, V. F., and von Krogh, G. (2018). Algorithmic induction through machine learning: using predictions to theorize. *INSEAD Working Paper*. doi: 10.2139/ssrn.3140617

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England J. Med.* 380 (14), 1347–1358. doi: 10.1056/NEJMra1814259

Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B., and Rijnbeek, P. R. (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.* 25 (8), 969–975. doi: 10.1093/jamia/ocy032

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12, 77. doi: 10.1186/1471-2105-12-77

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biom.* 70 (1), 41–55. doi: 10.1093/biomet/70.1.41

Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using sub classification on the propensity score. *J. Am. Stat. Assoc.* 79 (387), 516–524. doi: 10.1080/01621459.1984.10478078

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., et al. (2017). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control.*, 426. doi: 10.1016/j.bspc.2017.01.012

Strom, B., and Kimmel, S. E. (2006). *Textbook of Pharmacoepidemiology* (Chichester, UK: John Wiley & Sons, Ltd.).

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J. B., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Stupp, R., Hegi, M. E., Mason, W. P., van den Bent, M. J., Taphoorn, M. J., Taphoorn, M. J., et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 10, 459–466. doi: 10.1016/S1470-2045(09)70025-7

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Zhao, S., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 2019, 1. doi: 10.1038/s41573-019-0024-5

van Genugten, J. A. B., Leffers, P., Baumert, B. G., Tjon-a-Fat, H., and Twijnstra, A. (2010). Effectiveness of temozolomide for primary glioblastoma multiforme in routine clinical practice. *J. Neurooncol.* 96, 249– 370. doi: 10.1007/s11060-009-9956-7

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 19 February 2020
doi: 10.3389/fphar.2019.01675

# Applying Machine Learning to Ultrafast Shape Recognition in Ligand-Based Virtual Screening

*Etienne Bonanno*[1] *and Jean-Paul Ebejer*[2]*

[1] *Department of Artificial Intelligence, University of Malta, Msida, Malta,* [2] *Centre for Molecular Medicine and Biobanking, University of Malta, Msida, Malta*

Ultrafast Shape Recognition (USR), along with its derivatives, are Ligand-Based Virtual Screening (LBVS) methods that condense 3-dimensional information about molecular shape, as well as other properties, into a small set of numeric descriptors. These can be used to efficiently compute a measure of similarity between pairs of molecules using a simple inverse Manhattan Distance metric. In this study we explore the use of suitable Machine Learning techniques that can be trained using USR descriptors, so as to improve the similarity detection of potential new leads. We use molecules from the Directory for Useful Decoys-Enhanced to construct machine learning models based on three different algorithms: Gaussian Mixture Models (GMMs), Isolation Forests and Artificial Neural Networks (ANNs). We train models based on full molecule conformer models, as well as the Lowest Energy Conformations (LECs) only. We also investigate the performance of our models when trained on smaller datasets so as to model virtual screening scenarios when only a small number of actives are known *a priori*. Our results indicate significant performance gains over a state of the art USR-derived method, ElectroShape 5D, with GMMs obtaining a mean performance up to 430% better than that of ElectroShape 5D in terms of Enrichment Factor with a maximum improvement of up to 940%. Additionally, we demonstrate that our models are capable of maintaining their performance, in terms of enrichment factor, within 10% of the mean as the size of the training dataset is successively reduced. Furthermore, we also demonstrate that running times for retrospective screening using the machine learning models we selected are faster than standard USR, on average by a factor of 10, including the time required for training. Our results show that machine learning techniques can significantly improve the virtual screening performance and efficiency of the USR family of methods.

**Keywords: virtual screening, machine learning, ultrafast shape recognition, ligand based virtual screening, ligand similarity, ElectroShape**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Tingjun Hou,*
*Zhejiang University, China*
*Honglin Li,*
*East China University of Science and Technology, China*

***Correspondence:***
*Jean-Paul Ebejer*
*jean.p.ebejer@um.edu.mt*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *02 October 2019*
***Accepted:*** *23 December 2019*
***Published:*** *19 February 2020*

***Citation:***
*Bonanno E and Ebejer J-P (2020) Applying Machine Learning to Ultrafast Shape Recognition in Ligand-Based Virtual Screening. Front. Pharmacol. 10:1675. doi: 10.3389/fphar.2019.01675*

**Abbreviations:** ANN, artificial neutral Nework; AUC, area under curve; CSR, chiral shape recognition; DG, distance geometry; DUD, directory of useful decoys; DUD-E, directory of useful decoys-enchanced; EF, enrichment factor; ETKDG, experimental-torsion knowledge distance geometry; GMM, gaussian mixture model; HTS, high throughput screening; LBVS, ligand-based virtual screening; LEC, lowest energy conformation; ROC, receiver operator characteristic; SBVS, structure-based virtual screening; SMILES, simplified molecular input line entry specification; USR, ultrafast shape recognition; USRCAT, ultrafast shape recognition with CREDO atom types; VS, virtual screening.

## INTRODUCTION

The discovery and development of a new drug is a time-consuming process that can take 14 years to complete successfully, incurring a cost of about 2.5 billion US dollars (DiMasi et al., 2016). Virtual Screening (VS) is a search approach that leverages electronic databases of chemical compounds and modern computing resources to streamline this process. The aim of this process is to computationally pre-screen molecules to find those that are most likely to exhibit affinity for binding to a given target protein. In this way, laboratory time and resources associated with High Throughput Screening (HTS) can be drastically reduced by preferentially testing only the compounds that are more likely to become successful leads (Leach and Gillet, 2007). Advances in processing power and high-capacity storage as well as development of Big-Data techniques has made this process of molecular screening feasible, resulting in significant savings of time and cost and significantly streamlining the drug discovery cycle (Leach and Gillet, 2007; Lavecchia and Giovanni, 2013).

Ligand-Based Virtual Screening (LBVS) is underpinned by the concept of similarity as defined in the Similarity Property Principle, which simply states that similar molecules tend to exhibit similar properties (Johnson and Maggiora, 1990). Many LBVS methods exist, but in essence they all require two steps. First, is the generation of a descriptor which represents a molecule. Second, is the search for a quantitative distance function which given two descriptors pertaining to different molecules computes the similarity between these. Descriptors for a library of molecules are compared to a query molecule's descriptor, which typically exhibits bioactivity. The result is a similarity ranking of all the molecules in the library. The top molecules from this list, *i.e.* the most similar to the bioactive one, are moved forward for physical testing.

There are many different types of LBVS methods such as fingerprints, pharmacophore modelling, Quantitative Structure-Activity Relationship modelling (QSAR), Ultrafast Shape Recognition (USR), *etc.* LBVS methods may use physicochemical properties, 2D topology, 3D molecular shape, and other dimensions such as electrostatics, lipophilicity, *etc.* in their descriptor generation stage. Some methods use a combination of these features (*e.g.* SHAFTS uses both pharmacophores and 3D structure information (Liu et al., 2011). In the case of LBVS methods that use shape information, these may be broadly divided into alignment and alignment-free methods. Alignment methods build a 3D model of the query and target molecules which are then superimposed. A common metric is to calculate volume overlap between the aligned (superpositioned) models. Alignment-free methods do not require an alignment for the descriptor comparison and are generally more efficient. For a review of shape-based similarity methods please refer to Finn and Morris (2013).

Ultrafast Shape Recognition (USR) is an alignment-free LBVS technique (Ballester and Richards, 2007a; Ballester and Richards, 2007b) that distils molecular shape into a rotation-invariant descriptor vector made up of 12 real numbers. These descriptors are then compared directly using a modified Manhattan Distance metric in order to obtain a measure of similarity.

The greatest advantage of this method is the exceedingly concise way in which the shape of a molecule is condensed into a small 12-element descriptor. The comparison of such small descriptors is fast to compute and efficient to store. This significant feature of USR made it orders of magnitude faster than any other shape-based similarity method that existed at the time (Ballester and Richards, 2007a).

This method was developed in 2007, however, extensions to this algorithm have since been proposed that extend the purely shape-based descriptors of USR with other physicochemical properties of the molecule, examples of which are ElectroShape 4D (Armstrong et al., 2010), ElectroShape 5D (Armstrong et al., 2011) and USRCAT (Schreyer and Blundell, 2012), which respectively add atomic partial charges, lipophilicity, and atomic types to pure USR descriptors, obtaisning better virtual screening scores than the original USR algorithm.

Even though extensive research has been carried out in the application of machine learning techniques to structure-based as well as ligand-based virtual screening, to the best of our knowledge there has not been a study systematically applying machine learning to USR and USR-based descriptors. The aim is to improve virtual screening performance with respect to the standard USR method.

In this study, we use the datasets provided in Directory of Useful Decoys-Enhanced (DUD-E) to train machine learning models based on Gaussian Mixture Models, Isolation Forests, and Artificial Neural Networks using USR and ElectroShape 5D descriptors in order to explore the performance improvement achievable by abandoning the standard USR similarity metric based on the inverse Manhattan Distance function in favour of a full machine learning approach.

GMMs and Isolation Forests were chosen because they are unsupervised, one-class learning methods that can be trained only on positive examples, in a sense, mimicking the standard USR method of using actives as search templates. GMMs and Isolation Forests take different approaches to this one-class learning problem. The former is a generative model, aiming to learn the probability distribution governing the training examples, whilst the latter is an outlier detection model, which rather than find clusters in the training data, detects outlying points. Further to these two algorithms, we chose to explore the use of ANNs in this study. This is a supervised method in wide use that gives excellent performance in a varied range of domains. We chose this algorithm because it enabled us to compare the performances of the two unsupervised methods with a supervised model. One-class learning methods are interesting in virtual screening since DUD-E contains real active molecules but only putative inactives (hence termed decoys).

Ballester et al. (2009) determined that using the LECs as active search templates provides a good performance-speed balance when evaluating compound databases using USR. We, therefore train alternative models using full active molecule conformers as training data as well as using only the active LECs in order to determine the performance differences between the two approaches.

Additionally, we also train similar models based on successively smaller fractions of the available training dataset so as to gauge the performance degradation of our models with respect to training dataset size. A good performance achieved even with a small number of active training examples is desirable because often, only a small number of actives are known *a priori* at the commencement of a prospective virtual screening exercise.

Through this study we demonstrate the potential of these techniques in significantly improving their retrospective screening performance. Our models obtain performance improvements over the state-ofthe-art ElectroShape 5D algorithm of a similar magnitude to those obtained by ElectroShape 5D itself over the original USR method, which were on the order of a maximum improvement of 738% and mean improvement of 253% for full conformers and a maximum of 755% and mean of 283% for LECs.

## Ultrafast Shape Recognition

The USR technique was ideated by Ballester and Richards (2007a; 2007b) wherein they proposed a novel nonsuperpositional shape-based virtual screening technique meant to preserve the virtual screening performance of superpositional algorithms while obtaining the speed benefits of non-superpositional methods.

Ballester et al. point out that the 3D shape of a molecule can be encoded by taking the Euclidean distance of each atom to a predetermined number of centroids located within the space occupied by the molecule. The number and position of the centroids can be arbitrary, however, while pointing out that their selection had not been validated to be the optimal one, the authors chose four well-defined centroids as follows:

1. The molecular centroid (*ctd*)
2. The closest atom to *ctd* (*cst*)
3. The furthest atom from *ctd* (*fct*)
4. The furthest atom to *fct* (*ftf*).

Centroids computed for an example molecule are shown in **Figure 1**. Computing the Euclidean distances of all the atoms in the conformer to each of these four centroids yields four separate distance distributions of size proportional to the number of atoms making up the molecule.

As Ballester et al. indicate, however, there are several reasons these distributions are problematic to work with for the purposes of similarity searching. Most importantly, making use of these distributions as-is, it would not be possible to compare molecules having differing numbers of atoms because the distributions yielded by molecules of different sizes would also be of different sizes. In addition to this, distributions are normally represented as histograms, however this would still leave open the question of finding an optimal bin size given distributions of wildly differing sizes and characteristics generated from a database of molecules, not to mention the storage volume and processing power required for their processing.

They solve these problems by pointing out that a distribution is completely determined by its statistical moments (Hall, 1983), and condensing the four distributions into their respective first three moments, corresponding to the mean, the variance and the skewness of the distribution (Ballester and Richards, 2007a). This

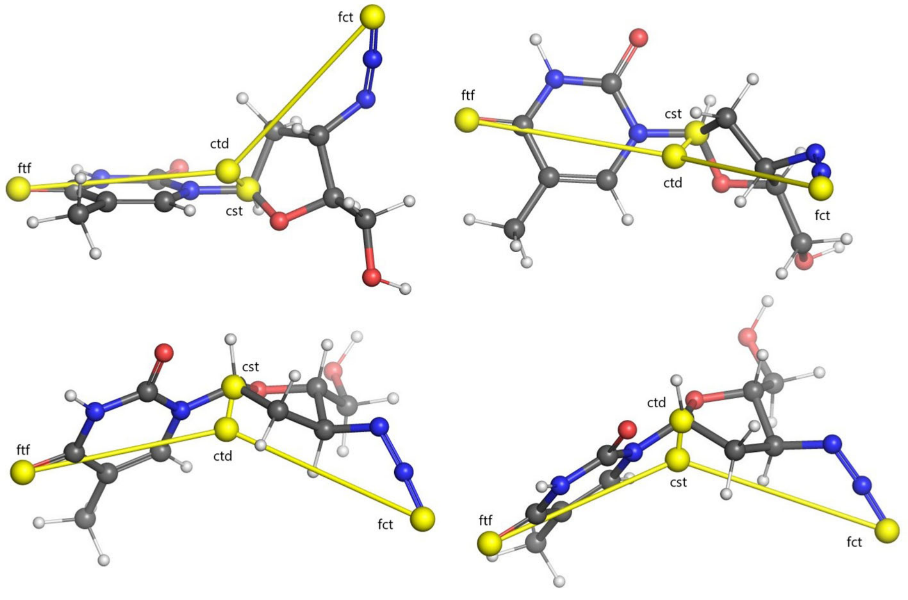


**FIGURE 1 |** Illustration of USR centroids computed for a sample conformer of the Zidovoudine molecule. Centroids are indicated with yellow spheres. Lines between every centroid and the molecular centre are displayed for clarity. Four different rotations of the molecule are illustrated. Legend: ctd, molecular centroid; cst, closest atom to ctd; fct, furthest atom to ctd; ftf, furthest atom from fct.

results in a vector of 12 decimal values making up a descriptor encapsulating shape information for a given conformer. The authors propose using this vector as a stand-in for the molecule's 3D structure in similarity comparisons. Ballester et al. (2009) modify this process by taking the square root and cube root of the second and third moments respectively, thus normalising them to a scale comparable to that of the first moment and resulting in better similarity matching performance.

The resulting descriptors could, in theory, be compared to each other using any similarity measure, however Ballester et al. chose to use a metric based on the Manhattan distance according to Equation 1.

$$S_{qi} = (1 + \frac{1}{12}\sum_{l=1}^{12}|M_l^q - M_l^i|)^{-1} \tag{1}$$

where $S_{qi}$ gives a similarity value between the query conformer $q$ and the conformer $i$ being screened and $\vec{M}^q$ and $\vec{M}^i$ are the descriptor vectors for the query conformer and the conformer being screened, respectively. Here the sum is normalised by dividing it by the number of elements in the USR descriptor.

Ballester et al. (2009) formally evaluated the USR method comparing it to ESshape3D in terms of Enrichment Factor (EF) finding it to offer, on average, significantly better ranking performance. They furthermore pointed out that the ideal active conformers to use as search templates are those experimentally observed in their bound state *via* X-ray crystallography or MRI. When this is not available, however, they show that using the LECs is a good, but obviously not perfect, approximation. When using LECs they obtained retrospective virtual screening performance that is only slightly worse than the maximum possible enrichment.

As they point out, the method can be easily extended by incorporating into the descriptors other, nonspatial, atomic-centred information (Ballester et al., 2009). This was achieved by Armstrong et al. in a series of three papers—Armstrong et al. (2009); Armstrong et al. (2010) and Armstrong et al. (2011).

Armstrong's first effort at extending USR (Armstrong et al., 2009) was in the development of the Chiral Shape Recognition (CSR) method, aimed at overcoming the shortcoming of USR that enantiomers, i.e. molecules that are mirror images of each other, generate identical descriptors, however do not necessarily bind equally to a protein, causing false positives. Armstrong et al. modified the USR method to account for chirality in the descriptor calculation, thus eliminating this source of error and obtaining enrichment factor improvements of 121%, 113%, and 106% at 0.25%, 0.5%, and 1% EF respectively.

Subsequently, Armstrong et al. (2010) again modified CSR by incorporating atomic partial charges into its descriptors, resulting in a new method they called ElectroShape. They did this by adding an extra dimension to the descriptors, consisting of the partial charge pertaining to each atom scaled by a constant quantity $Q$ so as to give them a magnitude comparable to the other spatial dimensions. This method resulted in a near doubling in performance over USR.

Armstrong et al. further extended their ElectroShape method in 2011 by adding lipophilicity in the form of ALogP to the ElectroShape descriptors in a similar manner as they had done for electrostatics, obtaining a further mean performance improvement of 110% over ElectroShape (Armstrong et al., 2011). This method shall hereafter by referred to as ElectroShape 5D.

Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) is a further method that extends USR. Proposed by Schreyer and Blundell (2012), this method incorporates the atom types maintained in the CREDO Structural Interatomics Database (Schreyer and Blundell, 2009), these being hydrophobic, aromatic, hydrogen bond donor and hydrogen bond acceptor. It does this by computing separate distributions for each atom type, joining the resulting distribution moments into a single descriptor vector with 60 elements. USRCAT, on average, obtained a slightly higher average performance score than ElectroShape in retrospective screening on the DUD-E database with an $EF_{0.25\%}$ of 15.64 as opposed to 8.84 for USR and 14.48 for ElectroShape, however the exact performance depended on the target under consideration, with some targets scoring better than ElectroShape and others worse.

Other extensions to USR have also been proposed with a variety of modifications, ranging from the combination of USR descriptors with 2D fingerprints, incorporating atomic types and applying graph theory to the USR centroid concept (Cannon et al., 2008; Shave et al., 2015).

## Machine Learning Methods

Machine learning techniques have been applied extensively to virtual screening; both in Structure-Based Virtual Screening (SBVS) (Betzi et al., 2006; Ain et al., 2015; Wojcikowski et al., 2017) as well as LBVS where 2D fingerprints are naturally suited to be used as training data for machine learning algorithms (Stahura and Bajorath, 2004; Hert et al., 2006; Chen et al., 2007; Geppert et al., 2010; Kurczab et al., 2011; Lavecchia, 2015). This has, however, not been the case with USR, where to our knowledge, only Cannon et al. (2008) have applied machine learning to USR descriptors, and even then, in combination with 2D fingerprints.

In the work presented in this paper, we make an initial effort to fill this lacuna in current research related to USR, obtaining significant performance improvements over one of the highest performing USR-derived methods, ElectroShape 5D, by training several machine learning models on ElectroShape 5D descriptors.

LBVS can be considered as a ranking problem, where the objective is to sort molecules by similarity to one or more ligands that are used as search templates. We have chosen three machine learning algorithms to explore in this study, that are well suited to model this problem—GMMs, Isolation Forests, and ANNs.

A Gaussian Mixture Model (Reynolds, 2015) is a generative machine-learning model that models a distribution of data points using a combination of weighted Gaussian distributions. It can be considered to be a clustering algorithm similar to k-means (Hartigan and Wong, 1979); however, in a GMM, cluster membership of a data point is not absolute but instead is influenced probabilistically by several centroids. A GMM is described mathematically by Equation 2 below:

$$f(x|\mu,\Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} exp\left[\left(x-\mu_k\right)^T \Sigma_k^{-1} \left(x-\mu_k\right)\right] \quad (2)$$

where $M$ is the number of Gaussians, also known as components, making up the GMM; $\mu_k$ is the mean for component $k$; $\Sigma_k$ is the covariance matrix for component $k$, giving the co-variance between every pair of dimensions; and $c_k$ is the weight for component $k$. These number of components is a hyperparameter of the algorithm as is usually tuned through an iterative cross-validation process. The GMM is trained using the Expectation Maximization algorithm (Dempster et al., 1977).

GMMs have wide-ranging applications in machine learning. They have been used in speech recognition (Stuttle, 2003), audio speech classification (Siegler et al., 1997), for language and speaker identification (Reynolds, 1995; Reynolds and Rose, 1995), as well as in visual object tracking (Santosh et al., 2013) and image enhancement applications (Celik and Tjahjadi, 2011). They have also been used in virtual screening and, in particular, protein-ligand docking (Grant and Pickup, 1995; Grant et al., 1996; Jahn et al., 2010; Jahn et al., 2011).

Isolation forests (Liu et al., 2008) are a class of machine learning models known as ensemble models. Ensemble models make use of a collection of simpler models to improve their predictions over those that would have been obtained by any single one model. Isolation Forests are similar to the Random Forest algorithm (Ho, 1995) in that they create a number of Decision Trees (Breiman, 2017) based on the training data and averages the predictions from each decision tree to arrive at a final result. While Random Forests are a supervised algorithm used to perform classification tasks, Isolation Forests are unsupervised and are meant to be used to perform anomaly detection in a set of observations.

Contrary to other clustering algorithms which attempt to identify similar samples within the input dataset, Isolation Forests explicitly identify anomalies in the data. They do so by exploiting the fact that, averaged over a number of Decision Trees, the path length that will be needed to generate a prediction for an outlier will be, on average, significantly shorter than that required for an inlier observation.

The rationale for using Isolation Forests as an algorithm for ranking USR descriptors is by extension of the formal evaluation of the USR method by Ballester et al. (2009). Herein it was shown that upon clustering the conformers of the active molecules for a given protein, several cluster centroids emerge, corresponding to shapes matching the one or more binding modes presented by the target protein.

By definition, a large number of actives will fall on, or close to a given centroid, since most active molecules will have at least one conformer that matches a binding mode of the target protein. This means that, taking all the active conformers as a set, high-density zones should be apparent and centred around the cluster centroids. Non-binding conformers, on the other hand, will fall outside these high-density zones, making them into outliers or anomalies. Training an Isolation Forest using the descriptors for the active molecules and ranking these points by their anomaly score should yield results with good predictive power.

The third machine learning algorithm that we explored along the course of this study is the Artificial Neural Network (ANN). ANNs are models loosely inspired by the structure of the brain, being made up of several successive layers of nodes (neurons), each output of one layer of nodes feeding in to the inputs of the next.

The neural net is usually set up with an input layer having the same number of nodes as the number of features in the input data. The output of the input layer is then routed through one or more hidden layers and into an output layer which gives the result predicted by the network.

A single node $j$ in layer $i$ of a neural network consists of a vector of weights $W_{i,j}$ equal in length to the number of nodes in layer $i - 1$ and an activation function, which computes an output value for the neuron $a_{i,j}$ by taking into account the outputs of the previous layer $a_{i-i}$ and the corresponding weights $W_i$.

There are a variety of activation functions that may be used in a neural network layer and it is possible to use different activation functions in different layers of a single network. Common ones include linear, sigmoid, and Rectified Linear Unit (RelU).

ANNs can be used for both classification as well as regression problems. For regression tasks, the output layer normally consists of one node with a linear activation function giving a real-valued output. For a classification network, the output layer is normally set up with one node for each class. The Softmax function, also called the Normalised Exponential Function, is applied to the outputs resulting a set of probabilities over the output classes.

In the context of molecule similarity ranking, regression networks are clearly the type of neural network that are the most suitable and the type of network used in this study. In our experiments, we used RelU activation for our hidden layer and linear activation on the output layer. The RelU activation is simple and is described by Equation 3 below:

$$f(x) = \begin{cases} 0, \text{ if } & x \leq 0 \\ x, \text{ if } & x \geq 0 \end{cases} \quad (3)$$

The linear activation function is also simple: $f(x) = x$.

Our intention in the selection of these three particular machine learning algorithms for our study was primarily to explore one-class learning models. Additionally, the “traditional” virtual screening process only involves using the known actives as “templates” against which to compare candidate molecules and not any decoys. Translating this into the machine-learning domain, this could be compared to one-class learning methods that, unlike supervised binary classifiers, do not make use of negative examples, but only positive ones. For these reason, we focussed most of our resources on exploring one-class learning algorithms, as we believed they would be better suited to the LBVS problem. However, we selected ANN as a general-purpose, widely-used supervised algorithm against which to compare the performance of the other one-class learning algorithms.

## METHODS

Most of the previous literature involving USR has been evaluated on the Directory of Useful Decoys (DUD) database of compounds (Huang et al., 2006), however shortcomings have since been identified in DUD (Mysinger et al., 2012). Actives in the dataset were not diverse enough to ensure unbiased results from virtual screening algorithms. Decoy selection was also not optimal as significant imbalance existed between the net charges of actives and decoys with 42% of the actives having a net charge versus only 15% of the decoys. In 2012, Mysinger et al. released a new and updated database named DUD-E which tackled these shortcomings (Mysinger et al., 2012). DUD-E provides active and decoy datasets for 102 protein targets with an average active/decoy ratio of 1:50. To our knowledge, only the USRCAT method has been evaluated on DUD-E. We, therefore, made the choice of using the DUD-E the purposes of training and evaluating our models.

As previous work was evaluated on the DUD database, for ease of comparison, we selected the DUD38 subset of targets provided by DUD-E which consists of 38 of the 40 targets in DUD. The protein targets we considered together with the respective number of actives, decoys and resulting conformers are shown in **Table 1**. We have also provided the dataset sizes on disk for the 3D conformers that we generated from the SMILES representations of the molecule datasets as well as the sizes of the descriptors generated from said conformer data. These can be seen in **Table S2** in the **Supplementary Material**.

As with many virtual screening methods that depend on molecular 3D shape, a sufficient number of conformers have to be generated to adequately sample the molecules' conformational space in order to produce effective results in USR. We generated conformers from the Simplified Molecular Input Line Entry Specification (SMILES) strings provided in DUD-E using the RDKit open-source cheminformatics library (Landrum and Others, 2013) following the protocol devised by Ebejer et al. (2012).

**TABLE 1 |** The list of 38 protein targets that we considered in this study along with the number of active and decoy molecules that were available for each protein target, and the respective number of active and decoy conformers we generated. These targets correspond to the "Dud38" subset in DUD-E.

| Target | Description | Active Mols. | Decoy Mols. | Active Confs. | Decoy Confs. | Confs./mol (Actives) | Confs./mol (Decoys) |
|---|---|---|---|---|---|---|---|
| ACE | Angiotensin-converting enzyme | 282 | 16,900 | 31,947 | 1,266,730 | 113 | 74 |
| ACES | Acetylcholinesterase | 453 | 26,250 | 55,549 | 2,153,887 | 122 | 82 |
| ADA | Adenosine deaminase | 93 | 5,450 | 7,786 | 332,177 | 83 | 60 |
| ALDR | Aldose reductase | 159 | 9,000 | 4,797 | 375,355 | 30 | 41 |
| AMPC | Beta-lactamase | 48 | 2,850 | 1,351 | 99,431 | 28 | 34 |
| ANDR | Androgen Receptor | 269 | 14,350 | 12,068 | 543,761 | 44 | 37 |
| CDK2 | Cyclin-dependent kinase 2 | 474 | 27,850 | 21,273 | 1,371,687 | 44 | 49 |
| COMT | Catechol O-methyltransferase | 41 | 3,850 | 1,262 | 147,125 | 30 | 38 |
| DYR | Dihydrofolate reductase | 231 | 17,200 | 16,679 | 873,009 | 72 | 50 |
| EGFR | Epidermal growth factor receptor erbB1 | 542 | 35,050 | 41,580 | 2,405,525 | 76 | 68 |
| ESR1 | Estrogen receptor alpha | 383 | 20,685 | 21,024 | 1,212,349 | 54 | 58 |
| FA10 | Coagulation factor X | 537 | 28,325 | 38,757 | 2,087,845 | 72 | 73 |
| FGFR1 | Fibroblast growth factor receptor 1 | 139 | 8,700 | 9,232 | 535,529 | 66 | 61 |
| GCR | Glucocorticoid receptor | 258 | 15,000 | 12,111 | 652,595 | 46 | 43 |
| HIVPR | Human immunodeficiency virus type 1 protease | 536 | 35,750 | 67,552 | 3,436,686 | 126 | 96 |
| HIVRT | Human immunodeficiency virus type 1 reverse transcriptase | 338 | 18,891 | 16,576 | 836,334 | 49 | 44 |
| HMDH | HMG-CoA reductase | 170 | 8,750 | 22,037 | 827,459 | 129 | 94 |
| HS90A | Heat shock protein HSP 90-alpha | 88 | 4,850 | 4,918 | 235,367 | 55 | 48 |
| INHA | Enoyl-[acyl-carrier-protein] reductase | 43 | 2,300 | 3,900 | 118,362 | 90 | 51 |
| KITH | Thymidine kinase | 57 | 2,850 | 3,168 | 150,295 | 55 | 52 |
| MCR | Mineralocorticoid receptor | 94 | 5,150 | 3,960 | 215,697 | 42 | 41 |
| MK14 | MAP kinase p38 alpha | 578 | 35,850 | 34,310 | 2,096,198 | 59 | 58 |
| NRAM | Neuraminidase | 98 | 6,200 | 6,030 | 325,337 | 61 | 52 |
| PARP1 | Poly [ADP-ribose] polymerase-1 | 508 | 30,050 | 18,925 | 1,242,760 | 37 | 41 |
| PDE5A | Phosphodiesterase 5A | 398 | 27,550 | 32,657 | 1,876,746 | 82 | 68 |
| PGH1 | Cyclooxygenase-1 | 195 | 10,800 | 8,123 | 410,263 | 41 | 37 |
| PGH2 | Cyclooxygenase-2 | 435 | 23,150 | 19,598 | 960,837 | 45 | 41 |
| PNPH | Purine nucleoside phosphorylase | 103 | 6,950 | 3,277 | 284,801 | 31 | 40 |
| PPARG | Peroxisome proliferator-activated receptor gamma | 484 | 25,300 | 71,166 | 2,527,881 | 147 | 99 |
| PRGR | Progesterone receptor | 293 | 15,650 | 13,041 | 578,492 | 44 | 36 |
| PUR2 | GAR transformylase | 50 | 2,700 | 7,931 | 195,987 | 158 | 72 |
| PYGM | Muscle glycogen phosphorylase | 77 | 3,950 | 3,300 | 212,652 | 42 | 53 |
| RXRA | Retinoid X receptor alpha | 131 | 6,950 | 8,008 | 316,919 | 61 | 45 |
| SAHH | Adenosylhomocysteinase | 63 | 3,450 | 1,883 | 118,691 | 29 | 34 |
| SRC | Tyrosine-protein kinase SRC | 524 | 34,500 | 39,561 | 2,313,655 | 75 | 67 |
| THRB | Thrombin | 461 | 27,004 | 57,028 | 2,131,048 | 123 | 78 |
| TRY1 | Trypsin I | 449 | 25,980 | 47,961 | 1,933,063 | 106 | 74 |
| VGFR2 | Vascular endothelial growth factor receptor 2 | 409 | 24,950 | 25,349 | 1,518,622 | 61 | 60 |

Conformer generation is performed using open-source code by Steven Kearnes[1] which follows the protocol laid out by (Ebejer et al., 2012). We modified this code in two ways:

- Use of ETKDG. We modified the code to use Experimental Torsion Knowledge Distance Geometry (ETKDG) as the conformer generation algorithm (Riniker and Landrum, 2015). ETKDG is a stochastic conformer generation method which builds upon the existing Distance Geometry (DG) algorithm (Blaney and Dixon, 1994) by using experimental knowledge about preferential torsional-angles. The major advantage in using ETKDG as opposed to DG is that the output of DG is not optimal and the resulting conformers may be in a distorted state (e.g., aromatic rings which are not planar). In order to remedy this, a second energy minimisation step is usually performed on these conformers in which inter-atomic force-field calculations are used to relax the molecule into a stable, energy-minimized state. This computationally expensive step is avoided by Experimental-Torsion Knowledge Distance Geometry (ETKDG) as the embedded knowledge in the algorithm produces conformers that are already energy minimized.
- Maximum energy cutoff. We removed all conformers which had a total energy higher than that of the LEC by 5 kcal/mol or more. This ensures that conformers with high energy (typically unsound structures) are discarded.

Prior to conformer generation, we validated and standardized the molecules using the MolVS tool[2]. This tool has been now integrated into RDKit.

Once we had generated a sufficient number of conformers for the compounds pertaining to our chosen protein targets, we calculated USR descriptors as well as descriptors for CSR, ElectroShape, and ElectroShape 5D for all the generated conformers. Note, however, that for reasons of time and resource availability, we chose to perform our machine learning experiments exclusively on the descriptors for USR and those for ElectroShape 5D. ElectroShape 5D was chosen because it is the highest performing USR-like method among those we evaluated.

The processes of conformer and descriptor generation resulted in excess of 300 GB of data. In order to generate and process this in a feasible amount of time, we used a Python 3.6/Spark 2.3.0 cluster on Amazon Web Services consisting of 3 compute-optimised c5.2xlarge instances having 8 cores and 16 GB of memory each. Cheminformatics analysis was performed using RDKit (version 2018.09.1). We also used the machine learning algorithms supplied with version 0.20.2 of the Scikit-learn library as well as Keras v.2.2.4/Tensorflow v.1.14.

## Experiments

The first experiments that we conducted were retrospective virtual screening using both USR and ElectroShape 5D over all the DUD38 protein targets in DUD-E. This gave us a baseline performance level against which to compare the results of the machine learning experiments.

For both USR as well as ElectroShape 5D, two versions of the experiments were performed.

The first used the full molecule conformer models of the actives as search templates for the similarity matching, comparing each conformer of each unknown molecule to each conformer of the template, taking the maximum similarity as the similarity score between the two molecules.

The second used only the LEC for each active as the search templates rather than all the active conformers in order to replicate the results of Ballester et al. (2009).

Having obtained baseline performance measures for the standard Manhattan distance-based USR and ElectroShape 5D screening processes, we proceeded to train the three types of machine learning models described previously.

Our training protocol was similar for all three algorithms and is described as follows:

1. Partition the training set into test set $T$ (20%) and training set $L$ (80%).
2. Partition $L$ set into 5 folds, $L_1 \ldots L_5$.
3. For every choice of hyperparameter (grid search), perform 5-fold cross validation on $L$, i.e. perform training and testing over $j = 1 \ldots 5$ iterations, each time taking $L_{j=x}$ as a test set and the 4 folds $L_{j \neq x}$ together as training set.
4. Select highest scoring grid search hyperparameter value combination averaged over the 5 iterations.
5. Train model using highest performing hyperparameter combination using $L$ as the training set and $T$ as the test set to evaluate final model. This ensures that the final test set is completely disjoint from the training data and avoids bias in the final results.

This process was repeated for every protein target at successively smaller portions of the entire dataset available in DUD-E equivalent to 100%, 80%, 60%, 50%, 30%, 10%, 5%, and 10 molecules, selected at random. All this is furthermore repeated for models trained using full molecule conformer models and for LECs models, running the training/testing cycle for a total of 16 times per protein target.

## Evaluation

For every model trained, we evaluated the performance using two criteria—the Receiver Operator Characteristic (ROC) Area Under Curve (AUC) and the EF. EF is a measure used specifically in retrospective virtual screening studies. EF at a given percentage of a dataset is defined as the ratio of the fraction of actives correctly found within the first x% of the ranked dataset to the fraction of actives that would be found by chance. This is defined formally in Equation 4.

$$EF_{x\,\%} = \frac{a_{x\,\%}/c_{x\,\%}}{a_{100\,\%}/c_{100\,\%}}$$

where $EF_{x\%}$ is the enrichment factor at x%, $a_{x\%}$ is the number of actives found in the top x% of the sorted dataset and $c_{x\%}$ is the total number of compounds in x% of the dataset. This measure, however, depends on the ratio of decoys to actives that are

[1] https://github.com/pandegroup/vs-utils [Last Accessed 4 May 2019]

[2] https://github.com/mcs07/MolVS [Last Accessed: 5 August 2019]

present in the dataset, and therefore is problematic to use when comparing results across different studies. For this reason, we also evaluate our models based on the ROC AUC.

The disadvantage to using ROC AUC performance metric, in the context of retrospective virtual screening, is that they give a picture of the performance of the method across the entire dataset, however in virtual screening only the top-ranked molecules are of interest. This is because in a prospective screening scenario, it is not possible to physically test all the compounds in the dataset and the available resources for testing in the laboratory would be invested only on the best-ranked compounds.

Unlike the EF, the ROC AUC does not depend on the structure of the dataset, making it more suitable and robust when used for comparison across studies using different benchmark datasets.

## RESULTS

The first stage in our experiments was to implement and evaluate the standard USR and ElectroShape 5D methods. Evaluation of our results with those of Ballester et al. (2009) and Armstrong et al. (2011) show them to be comparable albeit with differences, since they are evaluated on different datasets with a different decoy selection. Our results are shown in **Figures 2** and **3**. As can be seen, ElectroShape 5D obtains better performance than standard USR in all the protein targets being considered. The corresponding ROC AUC measures can be seen in the **Supplementary Material**.

We observed that, in general, our results show a similar trend to those presented by Armstrong et al. (2011) (reproduced in **Figure 4**), i.e., most targets that show a high enrichment in our results also show a high enrichment in Armstrong's results and vice versa, but there are differences. The Pearson productmoment correlation coefficient for the two sets of data is 0.35, indicating a mild positive correlation. Given the differences in decoy selection in DUD-E in comparison with DUD (Mysinger et al., 2012), it is not surprising that our results differ from those obtained by Armstrong. This relatively low, albeit positive, correlation coefficient, indicates that differences in dataset selection can have a significant impact on virtual screening results.

Once we generated results for our baseline methods, we trained and evaluated our machine learning models as described in the section *Experiments*. The results obtained from our machine learning experiments are visualised as follows. For each machine learning model, we have graphed the $EF_{1\%}$ as well as the ROC AUC achieved by the model along with the corresponding evaluation result achieved by ElectroShape 5D. Along with these we graph the improvement ratio between the performance of the model and the performance of ElectroShape 5D so as to indicate immediately the advantage in performance afforded by the use of the machine learning

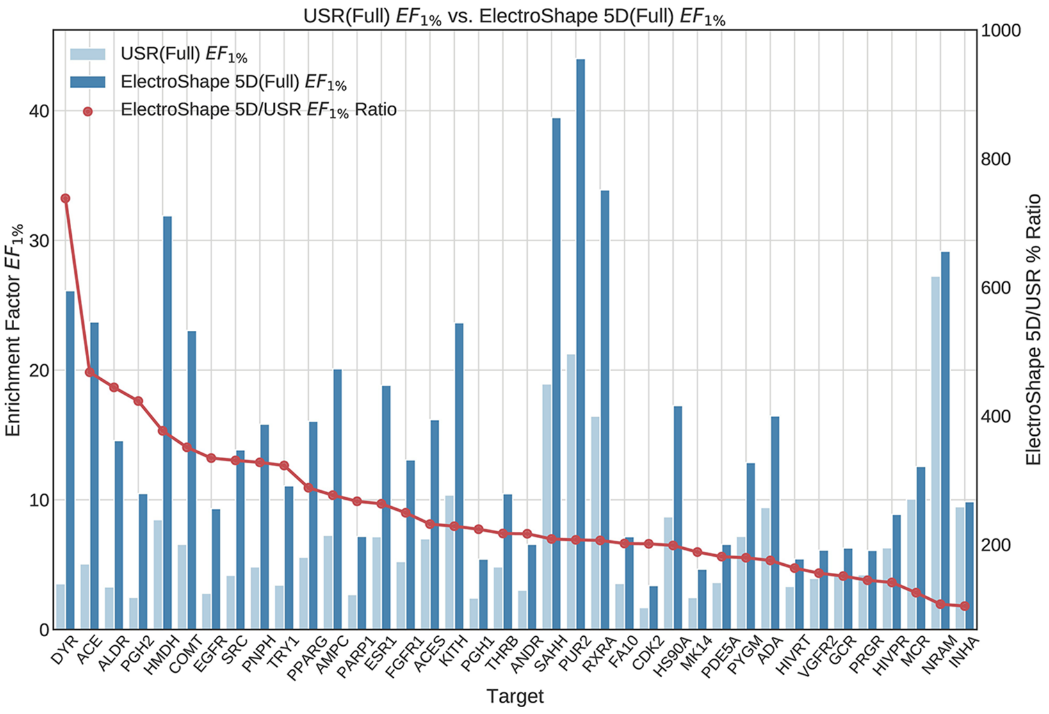


**FIGURE 2 |** Comparison of Enrichment Factor at 1% ($EF_{1\%}$) obtained by USR with that obtained by ElectroShape 5D using full conformer models. Also plotted is the percentage ratio of the Enrichment Factor score of ElectroShape 5D compared to Ultrafast Shape Recognition (USR). Mean ratio = 253% ± 122%, max = 738%, min = 104%.

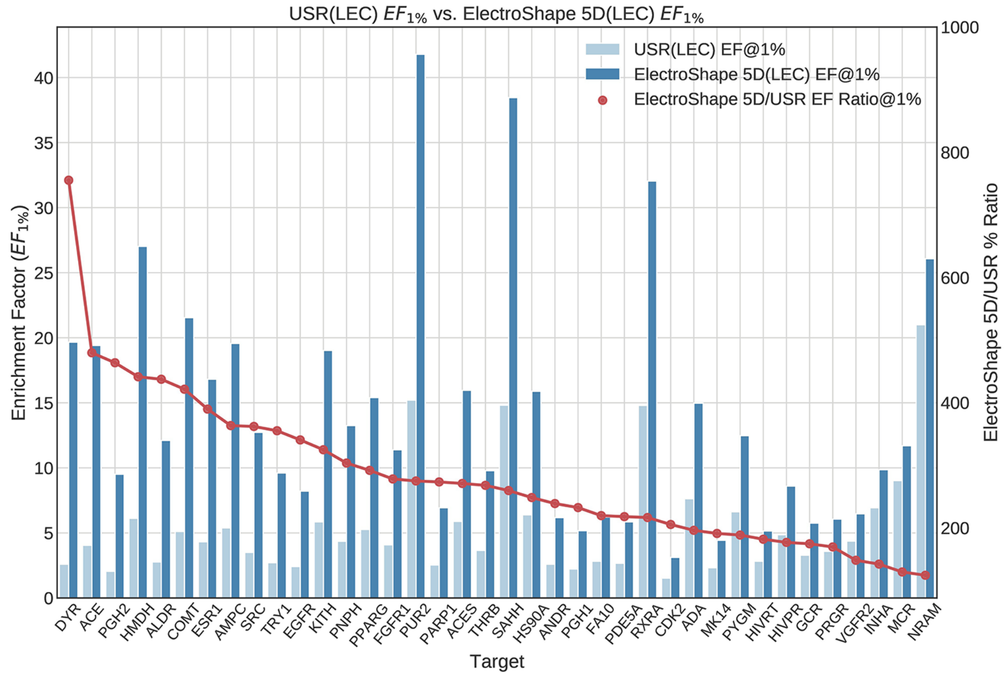


**FIGURE 3 |** Comparison of Enrichment Factor at 1% obtained by USR with that obtained by ElectroShape 5D using Lowest Energy Conformers. Also plotted is the percentage ratio of the Enrichment Factor score of ElectroShape 5D compared to Ultrafast Shape Recognition (USR). Mean ratio = 283% ± 125%, max = 755%, min = 124%.

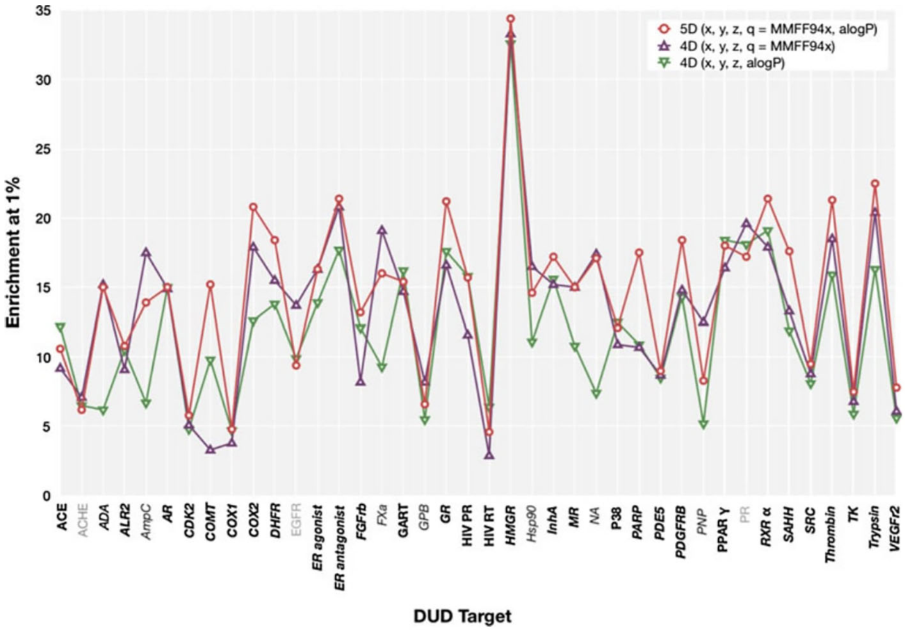


**FIGURE 4 |** ElectroShape 5D $EF_{1\%}$ calculated on the DUD dataset as reported in Armstrong et al. (2011). Legend: 5D(x,y,z,q = MMFF94x,aLogP)—ElectroShape 5D with partial charge and lipophilicity as the 4th and 5th dimensions, 4D(x,y,z,q = MMFF94x)—ElectroShape 4D using partial charge as the 4th dimension, 4D(x,y,z, q = aLogP)—ElectroShape 4D using lipophilicity as the 4th dimension. Reproduced from Armstrong et al., 2011.

method over ElectroShape 5D for every protein target. We do this for models trained on full conformer models as well as for those trained on LECs. Due to space constraints, we only present the $EF_{1\%}$ results. These can be seen in **Figures 5–11**. A complete set of visualisations is made available in the **Supplementary Material** (**Figures S1–S15**).

All the results obtained by the machine learning models we trained are presented in tabular form in **Table 2**.

Note that when training the ANNs, we expected to see a performance drop in the LEC model with respect to the full conformer-trained model, as for the other models, however, training both using a hidden layer size of 100 nodes, this did not materialise and the performance obtained for the LEC-trained model, in terms of mean $EF_{1\%}$ improvement ratio over ElectroShape 5D, was virtually the same for the same hidden layer size (255% ± 106% vs. 256% ± 129%). Upon increasing the hidden layer size to 500 nodes, this situation did not change (333% ± 128% vs. 327% ± 148%). It is also interesting that the ANN performance did not surpass that of the full-conformer GMM. Based on these results, the ANN model does not perform as well as GMMs.

It is also important to note that the imbalance in the training datasets, i.e., the ~1:50 active/decoy ratio, can cause some supervised machine learning models such as ANNs to give misleading test results by adapting their response to the distribution of labels in the training data rather than to the structure of the data itself. We verified the effect of the DUD-E unbalanced datasets on our ANN models by training alternative models using oversampling of the active conformers to balance the active/decoy ratio. Through these experiments we saw that the results obtained by balancing the datasets were comparable to those obtained from the unbalanced ones (mean unbalanced ROC AUC = 0.937 ± 0.037 vs. balanced ROC AUC = 0.955 ± 0.33, mean unbalanced $EF_{1\%} = 38.2 \pm 11.7$ vs. mean balanced $EF_{1\%} = 37.3 \pm 14.7$). Balancing the datasets in this way, however results in almost twice the training data for each model that is trained, and therefore a correspondingly longer training time. Given the marginal differences in results obtained through these experiments, therefore, we stuck to using the original unbalanced data to train our ANNs. Note that dataset balance is not an issue with either GMMs or Isolation Forests since decoys are not used when training these models.

## Varying the Size of the Training Dataset

We have repeated our experiments for every machine learning algorithm multiple times using successively smaller portions of the available dataset so as to explore the manner in which the performance given by each model degrades with dataset size and to understand how the performance of machine learning models degrades with a reduced dataset.

**Figures 12** and **13** contain plots illustrating the performance variation with number of known actives of our GMM models, the best performing models in our tests. The complete set of figures illustrating the performance change with dataset size for all our trained models can be found in the **Supplementary Material** (**Figures S1–S15**).

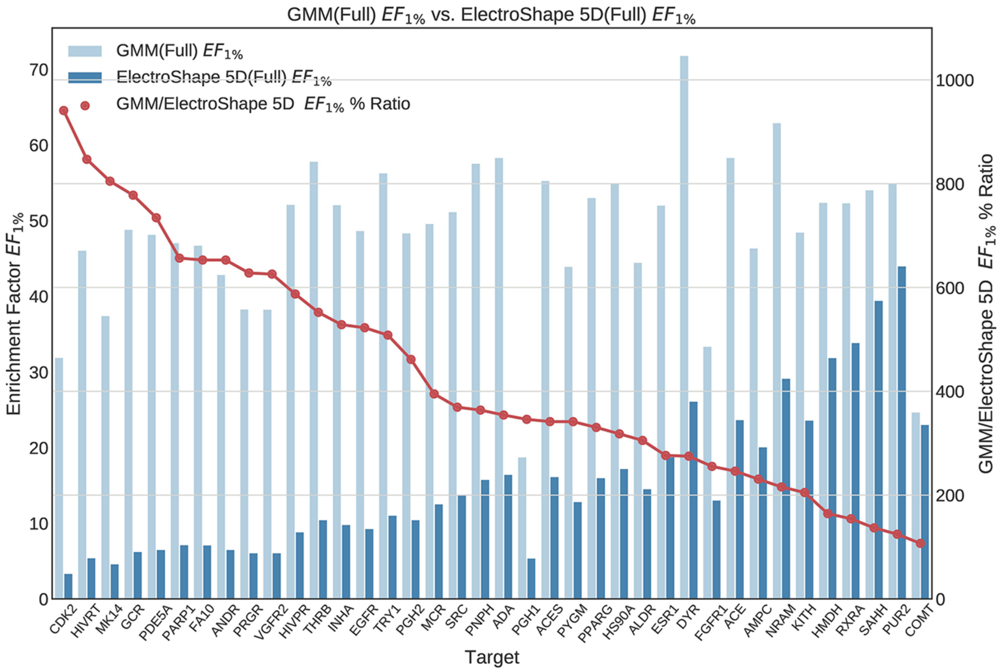


**FIGURE 5** | Comparison of Enrichment Factor at 1% obtained by Gaussian Mixture Models with that obtained by ElectroShape 5D using full conformer model. Also plotted is the percentage ratio of the Enrichment Factor score of Gaussian Mixture Model (GMM) compared to ElectroShape 5D. Mean ratio = 430% ± 223%, max = 941%, min = 107%.

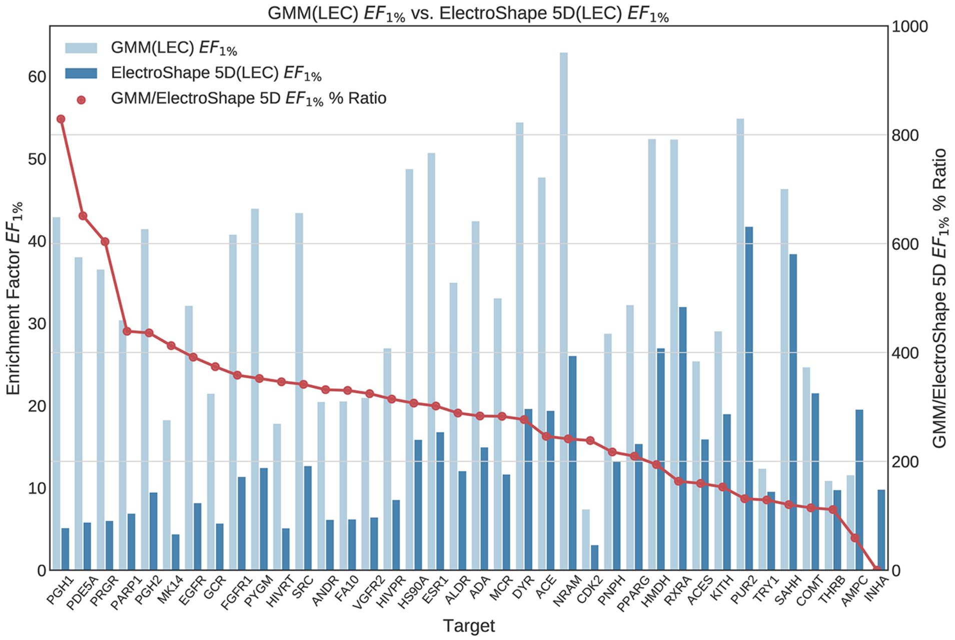


**FIGURE 6 |** Comparison of Enrichment Factor at 1% obtained by Gaussian Mixture Model with that obtained by ElectroShape 5D using Lowest Energy Conformers. Also plotted is the percentage ratio of the Enrichment Factor score of Gaussian Mixture Model (GMM) compared to ElectroShape 5D. Mean ratio = 291% ± 162%, max = 829%, min = 0%.

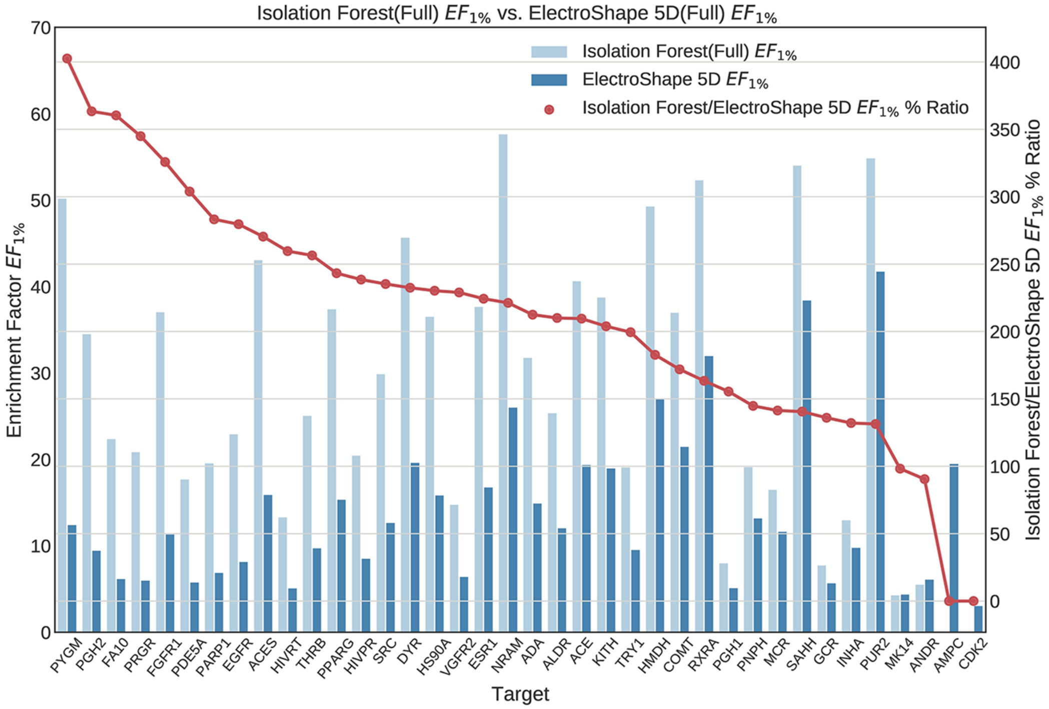


**FIGURE 7 |** Comparison of Enrichment Factor at 1% obtained by Isolation Forest with that obtained by ElectroShape 5D using full conformer model. Also plotted is the percentage ratio of the Enrichment Factor score of Isolation Forest compared to ElectroShape 5D. Mean ratio = 211% ± 90%, max = 941%, min = 107%.

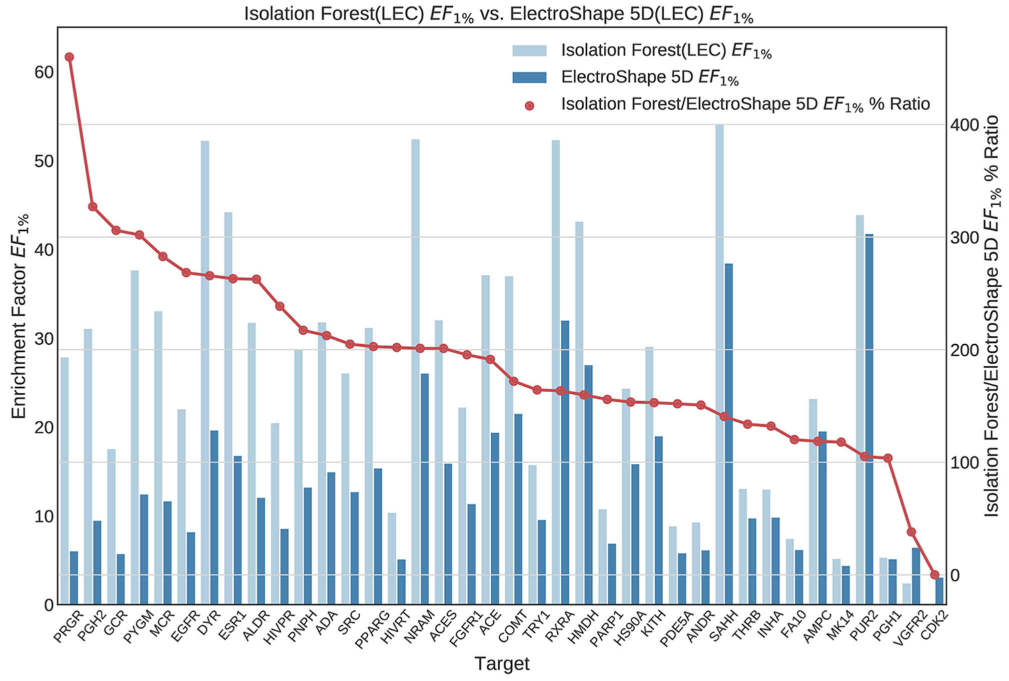


**FIGURE 8 |** Comparison of Enrichment Factor at 1% obtained by Isolation Forest with that obtained by ElectroShape 5D using Lowest Energy Conformers. Also plotted is the percentage ratio of the Enrichment Factor score of Isolation Forest compared to ElectroShape 5D. Mean ratio = 190% ± 84%, max = 460%, min = 0%.

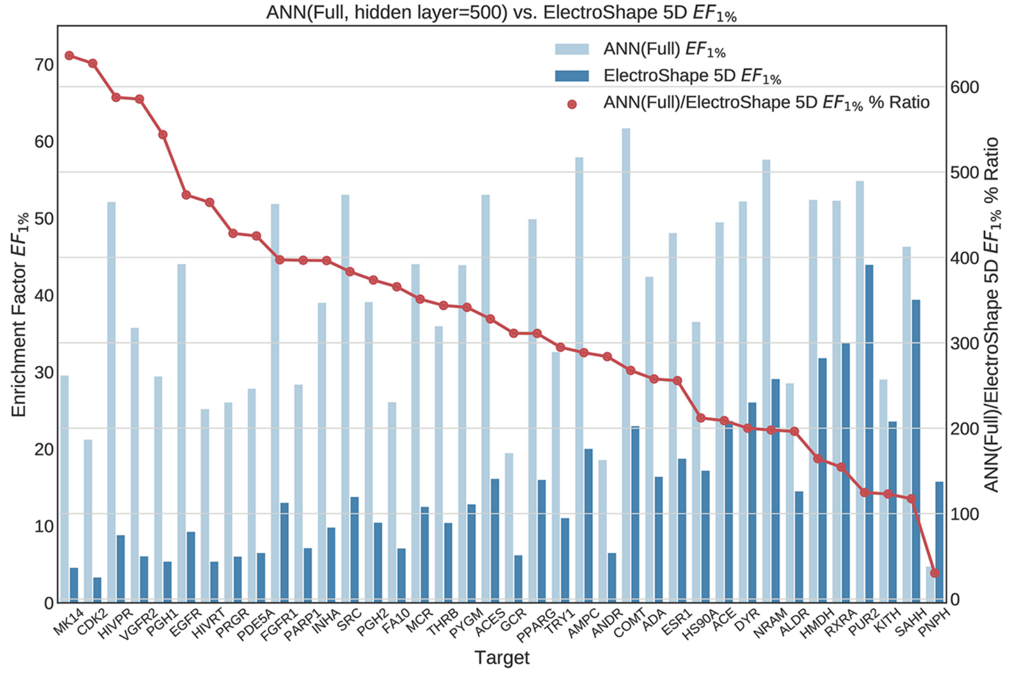


**FIGURE 9 |** Comparison of Enrichment Factor at 1% obtained by Artifical Neural Networks with 500-node hidden layer with that obtained by ElectroShape 5D using full conformer models. Also plotted is the percentage ratio of the Enrichment Factor score of Artificial Neural Network (ANN) compared to ElectroShape 5D. Mean ratio = 328% ± 149%, max = 636%, min = 30%.

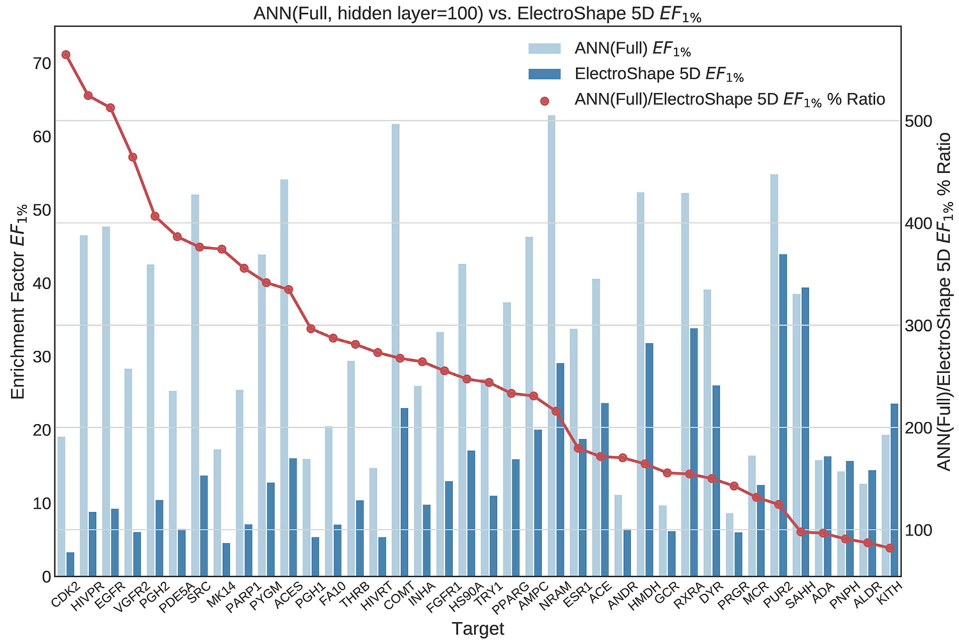


**FIGURE 10 |** Comparison of Enrichment Factor at 1% obtained by Artificial Neural Networks with 100-node hidden layer with that obtained by ElectroShape 5D using full conformer models. Also plotted is the percentage ratio of the Enrichment Factor score of Artificial Neural Network (ANN) compared to ElectroShape 5D. Mean ratio = 256% ± 129%, max = 565%, min = 82%.

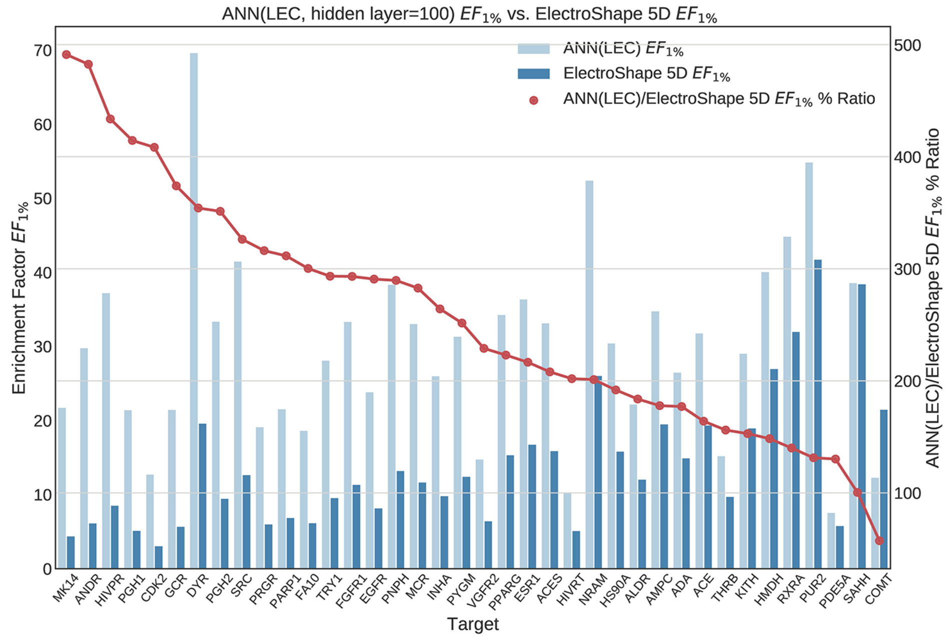


**FIGURE 11 |** Comparison of Enrichment Factor at 1% obtained by Artifical Neural Networks with 100-node hidden layer with that obtained by ElectroShape 5D using Lowest Energy Conformers. Also plotted is the percentage ratio of the Enrichment Factor score of Artificial Neural Network (ANN) compared to ElectroShape 5D. Mean ratio = 256% ± 107%, max = 491%, min = 57%.

**TABLE 2 |** Summary of machine learning results expressed as percentage ratios over ElectroShape 5D. A value of 100% indicates that the same performance as ElectroShape 5D was obtained.

| LEC | GMM | Isolation Forest | ANN (Hidden layer = 100) | ANN (Hidden layer = 500) |
|---|---|---|---|---|
| Mean $EF_{1\%}$( ± std) | 291%( ± 162%) | 191%( ± 084%) | 256%( ± 107%) | 333%( ± 130%) |
| Max $EF_{1\%}$ | 829% | 450% | 491% | 618% |
| Min $EF_{1\%}$ | 0% | 0% | 57% | 121% |
| Mean AUC( ± std) | 133%( ± 17%) | 126%( ± 15%) | 139%( ± 17%) | 144%( ± 21%) |
| Max AUC | 171% | 155% | 175% | 179% |
| Min AUC | 104% | 99% | 104% | 105% |
| Full Conformers | | | | |
| Mean $EF_{1\%}$( ± std) | 430%( ± 223%) | 211%( ± 90%) | 256%( ± 129%) | 328%( ± 149%) |
| Max $EF_{1\%}$ | 941% | 403% | 565% | 636% |
| Min $EF_{1\%}$ | 107% | 0% | 82% | 30% |
| Mean AUC( ± std) | 137%( ± 19%) | 124%( ± 14%) | 136%( ± 20%) | 143%( ± 20%) |
| Max AUC | 173% | 153% | 173% | 177% |
| Min AUC | 105% | 99% | 103% | 104% |

The statistical significance annotations were computed using the Wilcoxon rank-sum test (Mann and Whitney, 1947). This is a non-parametric test and therefore does not assume normality in the data. We have visually checked the distribution for each bin using histograms and found that they were not normal. It also assumed that the groups being compared are independent and not paired, which is the case with our box plots. The Wilcocon rank-sum test tests the null hypothesis that for any two observations *a* and *b* drawn from group *A* and group *B* respectively, the probability of *a* being greater that *b* is the same as that for *b* being greater that *a*. This test is used to investigate whether two sampling distributions are the same.

It is apparent from these figures that performance is better maintained for low number of actives by using full conformer models than by LECs. This is most pronounced for Neural Networks as well as GMMs, however it is also apparent for Isolation Forests, albeit more weakly. Nevertheless, even for small active training sets for which the mean performance is low, outliers are apparent with high enrichment factors. This shows that the performance of the methods we have explored is highly dependent on the protein target that is being considered and it is difficult to know *a-priori*, how well a method will perform given the number of available actives.

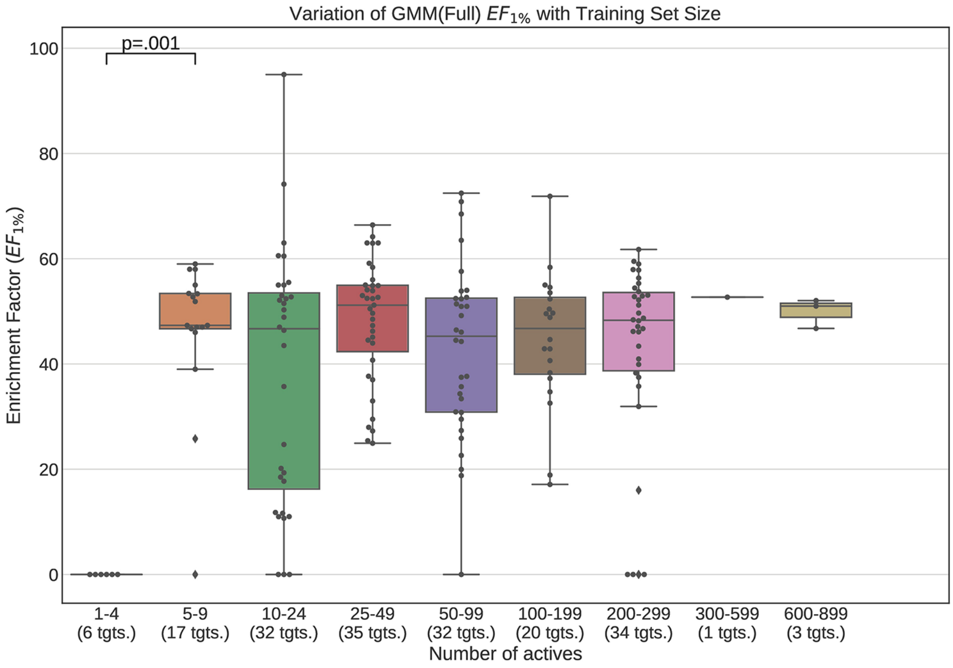


**FIGURE 12 |** Performance variation of full-conformer model Gaussian Mixture Models with number of actives. Scatter plot indicates one point per template within the given range. The number of templates captured within the range is indicated in the axis labels. Note that multiple points belonging to the same target could fall within a single range due to the binning thresholds used.

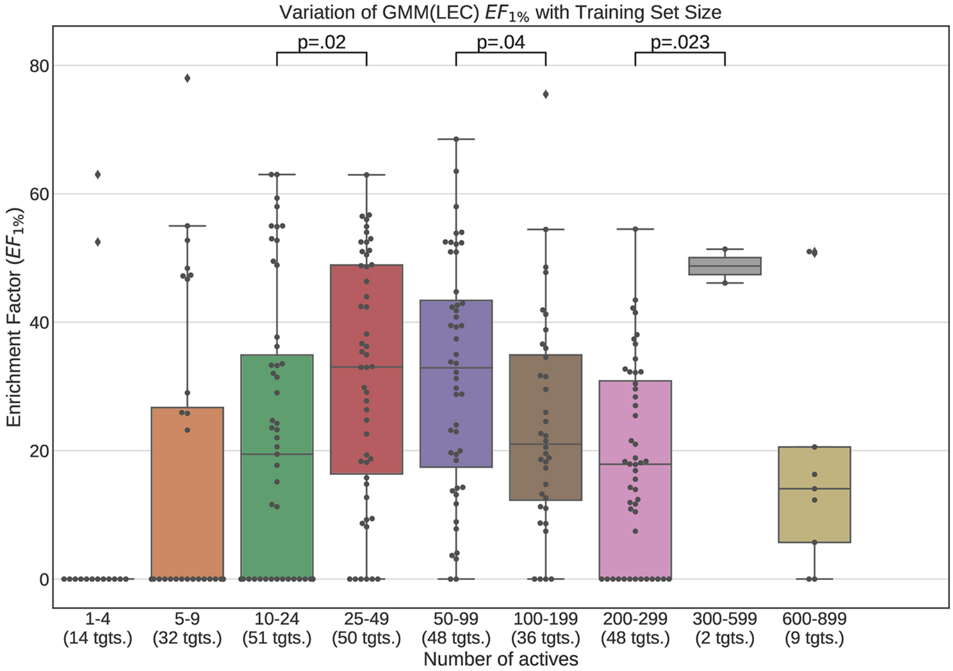


**FIGURE 13 |** Performance variation of Lowest Energy Conformation (LEC) model Gaussian Mixture Model with number of actives. Scatter plot indicates one point per template within the given range. The number of templates captured within the range is indicated in the axis labels. Note that multiple points belonging to the same target could fall within a single range due to the binning thresholds used.

For LEC models a performance peak is apparent at around 25–49 actives, beyond which performance degrades again. We observed this effect on GMMs and Isolation Forest models, but not on Neural Networks. It is possible that implementing a more comprehensive parameter sweep during the tuning of these models could eliminate or reduce this effect. For example, in the case of GMMs, allowing a larger number of Gaussian components would probably resolve the active clusters better and improve performance for larger numbers of actives.

A general observation in our results is that, across the machine learning models that we trained, those trained on full conformers preserve good performance when trained with as little as 5–9 actives, while with those trained on LECs, the cutoff is in the 10–24 actives range. These results indicate that for small datasets, models should be trained using full conformer models.

## Running Times

In order to understand how the time required to train and perform a retrospective virtual screening run varies with dataset size, we plotted the time taken to perform our experiments against the corresponding dataset portion used as training set using box plots, with separate boxes representing the run-time for each machine learning algorithm. The timings include the time taken to train the final, tuned model and evaluate the molecules under test. This does not include the time required to generate the conformers and the USR and ElectroShape 5D descriptors. These plots can be found in the **Supplementary Material** (**Figures S16–S19**). Additionally we have also presented running-time statistics in **Table S1**.

Note that, if used in a prospective screening scenario, a machine learning model would have been pre-trained from the available training data, therefore the time required for training would not be a factor when measuring the running time for such a study. In this case, however, since a retrospective experiment was being carried out we considered the total time required for training as well as testing/evaluation to be an important consideration.

It is apparent from the plots supplied in the **Supplementary Material** that GMMs were the quickest models overall for LEC models (8s ± 11s mean time) and the second quickest for the full conformer models (787s ± 868s mean time). For full conformer-trained models, GMMs were quicker for dataset fractions up to 60% of the full dataset, however, were slower than Isolation Forest for dataset fractions larger than 60%. At the 30% fraction the GMM running time increased. This could have been caused by transient resource contention on the machine on which the experiments were being run.

Isolation Forest speed performance compared favourably to GMMs for large datasets when using full conformer models (397s ± 373s mean time for isolation forest vs. 787s for GMMs), however, for smaller datasets using LECs it was considerably slower than the other algorithms, including ANNs (453s ± 423s for Isolation Forest vs. 131s ± 89s for ANNs). This is quite surprising and is likely due to the fact that no matter the size of the training data, an ensemble of decision trees of comparable size need to be created by the algorithm. Tweaking the hyperparameters to use smaller ensembles for LECs would probably make this model faster, however, this was not attempted in this study.

Neural Networks appear to be the most consistent with respect to speed performance. In general, it is the slowest algorithm (1855s ± 1659s mean time for full conformers and 131s ± 89s mean time for LECs), except for Isolation Forest in the LEC scenario.

It is worth noting that, notwithstanding the necessity to train the machine learning models before running the virtual screening procedure, the total time required to perform our retrospective screening on each target took, on average, a much shorter time to complete than the standard USR algorithms which took, on average 10 times more time to complete. Part of this discrepancy is likely the efficiency of our Python implementation of USR, which must necessarily be slower than the C-based implementations of the algorithms in the scikit-learn library. The magnitude of the difference, however, makes it unlikely for this to be the entire explanation. A large part of the discrepancy also comes from the fact that, in USR, all the conformers in the test set of molecules must be compared to every conformer of every active template. Over the course of an entire retrospective screening cycle, this adds up to a large amount of computation.

With machine learning algorithms, however, this is not necessary. The bulk of the running-time when using machine learning methods is the training of the model, however this, in general, does not require the repeated comparison of all the data points with all the active data points in a Cartesian product fashion. Additionally, once a model is trained, classifying new data points is generally a fast process because it does not involve comparing the new point with the training data directly, but only requires that the new data be evaluated according to the model built during training. All this, clearly depending on which particular machine learning algorithm is being used, implies a much smaller amount of computation than the "brute force" approach inherent in standard USR.

## DISCUSSION

Throughout this study we sought to answer two research questions, namely:

- Can machine learning techniques replace the naïve Manhattan distance in USR and USR-like methods to improve Virtual Screening performance?
- What is the minimal amount of data required to adequately train USR and USR-like machine learning models?

In pursuit of the first question, we used the datasets provided in DUD-E to generate a suitable number of conformers to adequately sample the conformational space of the molecules from which we generated corresponding USR and ElectroShape 5D descriptors.

We then selected three suitable machine learning algorithms, namely Gaussian Mixture Models, Isolation Forests, and Artificial Neural Networks and we trained and evaluated these models using the descriptors we had previously generated. In doing so, we obtained results that significantly outperformed USR as well as ElectroShape 5D when using both the full conformer models of the active molecules as training data, as well as when using only the Lowest Energy Conformations (LECs). Concretely, in terms of $EF_{1\%}$ the best mean improvement over ElectroShape 5D was that of 430% obtained using GMMs trained on full conformers, the same models having obtained a maximum improvement of 941% over ElectroShape 5D. This was followed by a mean improvement of 328% with a maximum of 636%, obtained by ANNs, again trained on full conformer models. When using LECs as training data, GMMs obtained a mean performance improvement of 291% and a maximum of 829%, outperforming ANNs with a hidden layer size of 100, which obtained a mean improvement of 256% with a maximum of 613%. It is clear, however, that some targets are more responsive to screening by USR descriptors, there being a relatively large variance in the mean performance figures. This is also reflected in the literature (Armstrong et al., 2009; Ballester et al., 2009; Armstrong et al., 2010; Armstrong et al., 2011) and is, therefore, expected.

These improvements over ElectroShape 5D are of a similar magnitude to the performance increase afforded by ElectroShape 5D itself over USR and are, therefore, highly significant. Machine learning algorithms assimilate the features of all the active molecules into a single model, in contrast to the naïve USR-based algorithms which can only consider one molecule at a time as a search query. This feature of machine-learning algorithms appears to make a large difference to the similarity matching performance in the LBVS context when compared with the standard algorithm for the USR family of methods.

In order to explore our second research question, we trained the machine learning models on progressively smaller fractions of the selected DUD-E targets so as to explore the manner in which the performance of the models varied whilst decreasing training dataset size. Our results demonstrate that when using full conformers to train the models, better performance is obtained when the number of actives is low. In general a performance peak is observed when training with 25–49 actives. With the LEC models, this peak is more pronounced, indicating that for small active training sets it is more advantageous to train with full conformers than LECs.

We also observed that performance of our models was preserved when only 5–9 actives are used for training when using full conformer models while, for the LEC-trained models, the performance remained acceptable down to the 10–24 actives level.

Taking into account all the results obtained, in terms of VS performance as well as running times, and we come to the conclusion that GMMs were, overall, the most efficient models that we tested, achieving excellent performance in the shortest time (except for the largest datasets; see **Figures S16** and **S17** in **Supplementary Information**) and while also exhibiting good stability with decreasing dataset size.

## CONCLUSION

To the best of our knowledge, this research project constitutes the first study to explore the viability of several machine learning algorithms in their application to LBVS using USR and USR-like descriptors.

We have demonstrated the utility of applying machine learning methods to the LBVS scenario when using USR-like descriptors, managing to obtain significant performance improvements over both the USR and the ElectroShape 5D algorithms using the Gaussian Mixture Model (GMM), Isolation Forest and Artificial Neural Network (ANN) algorithms. The GMM models were found to achieve the best performance improvement over ElectroShape 5D in terms of enrichment factor, giving an improvement of 291% for LEC-trained models and 430% for full conformer trained models with maximum improvements of 829% and 940%, respectively. These results clearly represent non-trivial improvements over the classical, non-machine learning, USR family of methods.

Furthermore we demonstrated that these trained models maintain stable performance when trained with drastically smaller quantities of training data, especially when full conformer molecule models are used, maintaining statistically similar performance from full dataset down to the 5–9 active range for full conformer models.

We also demonstrated the significant advantages in terms of running times, where retrospective screening took, on average 10 times less time to complete using our machine learning models than for USR and ElectroShape 5D.

Due to the sheer magnitude of the options available when it comes to machine learning methods, this work must be considered as a starting point for further research into the topic of machine learning on USR, however, we believe that it makes a valid contribution to the field, as it demonstrates significant performance improvements over current state-of-the-art methods that do not use machine learning.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Database For Useful Decoys-Enhanced (DUD-E).

## AUTHOR CONTRIBUTIONS

J-PE contributed to the conception and design of the study and guided and supervised the research. EB implemented and carried out the experiments and drafted the manuscript. J-PE revised and submitted the manuscript. All the authors have read and approved the final version of the manuscript.

## FUNDING

This research was partially funded by the Research Support Services Directorate at the University of Malta (Grant number MMERP04-18), as well as Amazon Web Services through their AWS Educate initiative. These funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

## ACKNOWLEDGMENTS

The authors would like to thank the Centre for Molecular Medicine and Biobanking at the University of Malta for providing computational facilities. We would also like to thank the Research Support Services Directorate at the University of Malta and Amazon Web Services for funding the study.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.01675/full#supplementary-material

## REFERENCES

Ain, Q. U., Aleksandrova, A., Roessler, F. D., and Ballester, P. J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 405–424. doi: 10.1002/wcms.1225

Armstrong, S. M., Morris, G. M., Finn, P. W., Sharma, R., and Richards, W. G. (2009). Molecular similarity including chirality. *J. Mol. Graph. Model.* 28, 368–370. doi: 10.1016/j.jmgm.2009.09.002

Armstrong, S. M., Morris, G. M., Finn, P. W., Sharma, R., Moretti, L., Cooper, R. I., et al. (2010). ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided. Mol. Des.* 24, 789–801. doi: 10.1007/s10822-010-9374-0

Armstrong, S. M., Finn, P. W., Morris, G. M., and Richards, W. G. (2011). Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J. Comput. Aided. Mol. Des.* 25, 785–790. doi: 10.1007/s10822-011-9463-8

Ballester, P. J., and Richards, W. G. (2007a). Ultrafast shape recognition for similarity search in molecular databases. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 463, 1307–1321. doi: 10.1098/rspa.2007.1823

Ballester, P. J., and Richards, W. G. (2007b). Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* 28, 1711–1723. doi: 10.1002/jcc.20681

Ballester, P. J., Finn, P. W., and Richards, W. G. (2009). Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J. Mol. Graph. Model.* 27, 836–845. doi: 10.1016/j.jmgm.2009.01.001

Betzi, S., Suhre, K., Chetrit, B., Guerlesquin, F., and Morelli, X. (2006). GFscore: a general nonlinear consensus scoring function for high-throughput docking. *J. Chem. Inf. Modeling* 46, 1704–1712. doi: 10.1021/ci0600758

Blaney, J. M., and Dixon, J. S. (1994). Distance geometry in molecular modeling. *Rev. Comput. Chem.* 5, 299–335. doi: 10.1002/9780470125823.ch6

Breiman, L. (2017). Classification and regression trees (Routledge). doi: 10.1201/9781315139470

Cannon, E. O., Nigsch, F., and Mitchell, J. B. (2008). A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chem. Cent. J.* 2, 1–9. doi: 10.1186/1752-153X-2-3

Celik, T., and Tjahjadi, T. (2011). Automatic image equalization and contrast enhancement using Gaussian mixture modeling. *IEEE Trans. Image Process.* 21, 145–156. doi: 10.1109/TIP.2011.2162419

Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., et al. (2007). Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided. Mol. Des.* 21, 53–62. doi: 10.1007/s10822-006-9096-5

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.: Ser. B (Methodological)* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* 47, 20–33. doi: 10.1016/j.jhealeco.2016.01.012

Ebejer, J. P., Morris, G. M., and Deane, C. M. (2012). Freely available conformer generation methods: how good are they? *J. Chem. Inf. Model.* 52, 1146–1158. doi: 10.1021/ci2004658

Finn, P. W., and Morris, G. M. (2013). Shape-based similarity searching in chemical databases. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 226–241. doi: 10.1002/wcms.1128

Geppert, H., Vogt, M., and Bajorath, J. (2010). Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* 50, 205–216. doi: 10.1021/ci900419k

Grant, J. A., and Pickup, B. T. (1995). A Gaussian description of molecular shape. *J. Phys. Chem.* 99, 3503–3510. doi: 10.1021/j100011a016

Grant, J. A., Gallardo, M. A., and Pickup, B. T. (1996). A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* 17, 1653–1666. doi: 10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K

Hall, P. (1983). A distribution is completely determined by its translated moments. *Z. für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 62, 355–359. doi: 10.1007/BF00535259

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Applied Statistics)* 28, 100–108. doi: 10.2307/2346830

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2006). New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* 46, 462–470. doi: 10.1021/ci050348j

Ho, T. K. (1995). "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. (Montreal, Quebec, Canada: IEEE), 278–282.

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Jahn, A., Hinselmann, G., Fechner, N., Henneges, C., and Zell, A. (2010). Probabilistic modeling of conformational space for 3D machine learning approaches. *Mol. Inform.* 29, 441–455. doi: 10.1002/minf.201000036

Jahn, A., Rosenbaum, L., Hinselmann, G., and Zell, A. (2011). 4D flexible atom-pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening. *J. Cheminform.* 3, 23. doi: 10.1186/1758-2946-3-23

Johnson, M. A., and Maggiora, G. M. (1990). Concepts and applications of molecular similarity (Wiley).

Kurczab, R., Smusz, S., and Bojarski, A. (2011). Evaluation of different machine learning methods for ligand-based virtual screening. *J. Cheminform.* 3, P41. doi: 10.1186/1758-2946-3-S1-P41

Landrum, G.Others (2013). RDKit: cheminformatics and machine learning software. RDKIT. ORG.

Lavecchia, A., and Giovanni, C. D. (2013). Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* 20, 2839–2860. doi: 10.2174/09298673113209990001

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331. doi: 10.1016/j.drudis.2014.10.012

Leach, A. R., and Gillet, V. J. (2007). *An Introduction to Cheminformatics*. *Revised ed edn.* (Sheffield: Springer). doi: 10.1007/978-1-4020-6291-9

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). "Isolation forest," in *2008 Eighth IEEE Int. Conf. Data Min.* (Pisa: IEEE), 413–422. doi: 10.1109/ICDM.2008.17

Liu, X., Jiang, H., and Li, H. (2011). SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model.* 51, 2372–2385. doi: 10.1021/ci200060s

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18 (1), 50–60. doi: 10.1214/aoms/1177730491

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Reynolds, D. A., and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* 3, 72–83. doi: 10.1109/89.365379

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Commun.* 17, 91–108. doi: 10.1016/0167-6393(95)00009-D

Reynolds, D. (2015). Gaussian mixture models. *Encyclopedia Biometrics*, 827–832. doi: 10.1007/978-1-4899-7488-4_196

Riniker, S., and Landrum, G. A. (2015). Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* 55, 2562–2574. doi: 10.1021/acs.jcim.5b00654

Santosh, D. H. H., Venkatesh, P., Poornesh, P., Rao, L. N., and Kumar, N. A. (2013). Tracking multiple moving objects using gaussian mixture model. *Int. J. Soft Computing Eng. (IJSCE)* 3, 114–119.

Schreyer, A., and Blundell, T. (2009). CREDO: a protein–ligand interaction database for drug discovery. *Chem. Biol. Drug Des.* 73, 157–167. doi: 10.1111/j.1747-0285.2008.00762.x

Schreyer, A. M., and Blundell, T. (2012). USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* 4. doi: 10.1186/1758-2946-4-27

Shave, S., Blackburn, E. A., Adie, J., Houston, D. R., Auer, M., Webster, S. P., et al. (2015). UFSRAT: ultra-fast shape recognition with atom types – the discovery of novel bioactive small molecular scaffolds for FKBP12 and 11βHSD1. *PLoS One* 10, 1–15. doi: 10.1371/journal.pone.0116570

Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. (1997). "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, vol. 1997 (Morgan Kaufmann).

Stahura, F., and Bajorath, J. (2004). Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen.* 7, 259–269. doi: 10.2174/1386207043328706

Stuttle, M. N. (2003). *A Gaussian mixture model spectral representation for speech recognition* [PhD thesis] (Cambridge, U. K.: University of Cambridge).

Wojcikowski, M., Ballester, P. J., and Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* doi: 10.1038/srep46710

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 25 February 2020
doi: 10.3389/fphar.2020.00069

# Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets

*Jincai Yang[1,2], Cheng Shen[2,3] and Niu Huang[2,4]**

[1] School of Life Sciences, Peking University, Beijing, China, [2] National Institute of Biological Sciences, Beijing, China, [3] Graduate School of Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [4] Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, China

Predicting protein-ligand interactions using artificial intelligence (AI) models has attracted great interest in recent years. However, data-driven AI models unequivocally suffer from a lack of sufficiently large and unbiased datasets. Here, we systematically investigated the data biases on the PDBbind and DUD-E datasets. We examined the model performance of atomic convolutional neural network (ACNN) on the PDBbind core set and achieved a Pearson $R^2$ of 0.73 between experimental and predicted binding affinities. Strikingly, the ACNN models did not require learning the essential protein-ligand interactions in complex structures and achieved similar performance even on datasets containing only ligand structures or only protein structures, while data splitting based on similarity clustering (protein sequence or ligand scaffold) significantly reduced the model performance. We also identified the property and topology biases in the DUD-E dataset which led to the artificially increased enrichment performance of virtual screening. The property bias in DUD-E was reduced by enforcing the more stringent ligand property matching rules, while the topology bias still exists due to the use of molecular fingerprint similarity as a decoy selection criterion. Therefore, we believe that sufficiently large and unbiased datasets are desirable for training robust AI models to accurately predict protein-ligand interactions.

**Keywords: artificial intelligence, convolutional neural network, protein-ligand interaction, virtual screening, molecular docking, scoring function, topology fingerprint**

OPEN ACCESS

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Chi Zhang,*
*University of Nebraska-Lincoln,*
*United States*
*Shanfeng Zhu,*
*Fudan University, China*

****Correspondence:***
*Niu Huang*
*huangniu@nibs.ac.cn*

***Specialty section:***
*This article was submitted to*
*Translational Pharmacology,*
*a section of the journal*
*Frontiers in Pharmacology*

***Received:*** *11 November 2019*
***Accepted:*** *24 January 2020*
***Published:*** *25 February 2020*

***Citation:***
*Yang J, Shen C and Huang N (2020)*
*Predicting or Pretending: Artificial*
*Intelligence for Protein-Ligand*
*Interactions Lack of Sufficiently Large*
*and Unbiased Datasets.*
*Front. Pharmacol. 11:69.*
*doi: 10.3389/fphar.2020.00069*

## INTRODUCTION

Structure-based virtual screening (molecular docking) has been widely used to discover new ligands based on target structures (Kitchen et al., 2004; Shoichet, 2004; Irwin and Shoichet, 2016; Zhou et al., 2016; Wang et al., 2017; Lyu et al., 2019; Peng et al., 2019). The molecular docking approach is designed to identify small molecules from a large chemical library that possess complementary to a protein binding site. The heart of molecular docking is the scoring function for estimation of binding affinities of protein-ligand complexes. Large research efforts in the field have been dedicated to the development of scoring functions in terms of their abilities to reproduce crystal ligand binding poses, to prioritize the known active compounds in a large compound database, and to predict the relative binding affinities (Stahl and Rarey, 2001; Halgren et al., 2004; Huang et al., 2006a; Wang et al., 2016; Liu et al., 2017; Guedes et al., 2018; Su et al., 2019). Despite some success, it is still very challenging to predict protein-ligand interactions accurately and efficiently using molecular docking.

In the retrospective studies, the performance of virtual screening was evaluated on several public available benchmarking datasets, including the Community Structure-Activity Resource (CSAR) (Dunbar et al., 2011), the PDBbind (Liu et al., 2017), the Directory of Useful Decoys (DUD) (Huang et al., 2006b), and the Directory of Useful Decoys - Enhanced (DUD-E) (Mysinger et al., 2012). The CSAR and PDBbind datasets were compiled to facilitate the prediction of the binding affinities based on experimental complex structures. The availability of experimental protein-ligand complex structures allows the structure-based featurization to correlate the protein-ligand binding interactions and the binding affinities. The DUD and DUD-E datasets were originally designed to assess docking enrichment performance by distinguishing the annotated actives from among a large database of computationally generated non-binding decoy molecules.

In recent years, deep learning (DL) technologies in the field of artificial intelligence (AI) have rapidly developed, and have been quickly introduced into the different aspects of drug discovery and development process (Chen et al., 2018; Ching et al., 2018; Hu et al., 2018; Ivanenkov et al., 2019; Xu et al., 2019; Zhavoronkov et al., 2019). However, DL relies on large and high-quality annotated datasets, and this approach is only in the early stages of applicability for protein-ligand binding prediction (Shen et al., 2019). Two types of representations have been applied in studying protein-ligand interactions (Ching et al., 2018). One is three-dimensional (3D) grid, which discretize protein-ligand complex structure into a 3D grid with features stored at the grid point (Wallach et al., 2015; Ragoza et al., 2017; Jiménez et al., 2018; Stepniewska-Dziubinska et al., 2018). For example, a 3D convolutional neural network (CNN) model was shown to outperform the AutoDock Vina in enrichment performance by achieving a mean area under the curve (AUC) of 0.86 on the DUD-E dataset (Ragoza et al., 2017). Another model (named Pafnucy) was tested for binding affinity prediction on the PDBbind v2013 core set with a Pearson $R^2$ of 0.49 (Stepniewska-Dziubinska et al., 2018).

The other representation is graph neural network (Battaglia et al., 2018), every atom is a vertex and the atomic features (including atom type, charge, distances, and neighbors) in molecule are stored at the atom (Pereira et al., 2016; Gomes et al., 2017; Cang et al., 2018; Feinberg et al., 2018). For example, DeepVS was reported to achieve a mean AUC of 0.81 for cross-target cross validation (CV) on the DUD dataset (Pereira et al., 2016). The atomic convolutional neural network (ACNN) was developed for binding affinity prediction but did not outperform random forest (RF) on the PDBbind datasets (Gomes et al., 2017). Cang et al. (2018) achieved a Pearson $R^2$ of 0.66 on the PDBbind v2013 core set using the model trained on the refined set.

However, Sieg et al. (2019) recently reported that the AI models were heavily biased by 1D properties and 2D topology trained on the DUD and DUD-E datasets. Only with the use of six physicochemical properties, RF classifiers achieved mean AUCs up to 1.0 for intra-target CV, while for cross-target CV on DUD and DUD-E, maximum mean AUCs of 0.78 and 0.80 were able to obtain, individually. Only using topology information of compounds, RF and DeepVS achieved a mean AUC of 0.78 for cross-target CV on DUD, and grid-based CNN model yielded a mean AUC of 0.84 for cross-target CV on DUD-E. Similarly, Chen et al. (2019) also reported the bias on topology in DUD-E. These studies demonstrate that AI models trained on ligand properties or ligand topology have comparable enrichment performance as those trained on docked complexes.

In the present work, we systematically investigated the data biases in the PDBbind and DUD-E datasets, including different data splitting methods, featurization, models, and metrics. We trained ACNN models (Gomes et al., 2017) on the protein-ligand complex structures, as well as on the ligand structures without the presence of proteins or on the protein structures by removing the ligand information. Strikingly, all these models performed comparably well in predicting binding affinities in test subsets, which strongly suggests that the ACNN models did not require learning essential protein-ligand interactions. Furthermore, we visualized the individual atomic contributions decomposed from the ACNN scores and found that the ACNN models may actually rely on the similarity of atomic features that exist in the training and test subsets to predict binding affinities. These results indicate that PDBbind has data biases in both proteins and ligands for building reliable AI models. Finally, we demonstrated that model learned the topology bias in DUD-E even after reducing the property bias by carefully designed CV experiments. We expect that our study will provide a useful guideline to assess the model performance in predicting protein-ligand interactions using state-of-the-art AI approaches.

# METHODS

## Datasets

The PDBbind is a comprehensive collection of protein-ligand complexes in the Protein Data Bank (PDB) with experimentally measured binding affinities, which contains core, refined, and general sets (**Table 1**) (Li et al., 2014). For clarification, the PDBbind v2013 core set is identical to the v2015 core set. At present study, we only report the results obtained from the PDBbind v2015. The general set contains a total of 11,987 protein-ligand complexes in PDB with experimentally measured binding affinity data. The refined set contains 3,796 complex structures chosen from the general dataset to enforce higher quality protein-ligand complex structures and binding affinities.

**TABLE 1** | The PDBbind and DUD-E datasets.

| Name | Task type | Sets | Crystal structures | #Actives | #Decoys |
|---|---|---|---|---|---|
| **PDBbind** | Regression | Core | 195 | 195 | 0 |
| | | Refined | 3,706 | 3,706 | 0 |
| | | General | 11,987 | 11,987 | 0 |
| **DUD-E** | Classification | Original | 102 | 22,886 | 1,411,214 |
| | | MW ≤ 500 | 102 | 19,374 | 1,182,039 |

The core set consists of 195 high-quality complexes clustered in 65 structural groups, each containing three complexes with low, medium, and high binding affinities. In addition, Wan et al. (2013) modeled 2,431 binding interactions of 17 kinase inhibitors against 143 protein kinases using physics-based approach. We also tested the kinase inhibitor selectivity prediction on this dataset using ACNN models trained on the PDBbind refined set.

The DUD and DUD-E datasets were designed for benchmarking molecular docking enrichment power by providing challenging decoys. For each annotated active, 50 decoys with six similar physicochemical properties, including molecular weight (MW) and cLogP, but dissimilar topology (fingerprint) were selected from the ZINC12 database (Irwin et al., 2012). The DUD-E dataset consists of 22,886 actives and 1,411,214 decoys against 102 targets. We compiled a variation of DUD-E, named DUD-E$_{(MW \leq 500)}$ by simply removing actives with MW (only accounting for all heavy atoms) greater than 500 and the same fraction of decoys (**Table 1**).

## Dataset Splitting

Each PDBbind set was split into the training, validation, and test subsets following an 80/10/10 ratio. We trained models on the training subset by using early stopping to avoid overfitting, tuned hyperparameters on the validation subset to select the best model, and subsequently evaluated model performance on the test subset. We applied three types of dataset splitting methods, including random, ligand scaffold-based, and protein sequence-based splitting. Scaffold-based splitting was based on ligand scaffold similarity, where the ligand 2D scaffolds (Bemis and Murcko, 1996) were extracted using RDKit software (Landrum, 2006) and clustered using Extended-Connectivity Fingerprints (ECFP) (Rogers and Hahn, 2010) with Tanimoto coefficient (Tc) cutoff value of 0.8. The obtained large, medium, and small clusters were assigned into the training, validation, and test subsets, respectively. The test subset contained the smallest clusters to create a greater challenge for AI models. The sequence-based splitting was performed by using the UCLUST (Edgar, 2010) program with sequence identity cutoff of 0.4.

To stay consistent with a previous report, we trained models on the refined and general sets, and tested on the core set. To avoid the same protein-ligand complex used in training and testing simultaneously, we removed samples in the refined and general sets overlapping with the core set. In addition, we removed analogs or homologs based on ligand scaffold or protein sequence similarity when we applied scaffold-based and sequence-based splitting in training. Nevertheless, we subsampled the same number of samples (2,036 samples accounting for 55% of the refined set, 7,792 samples accounting for 65% of the general set) from the rest of samples in the refined or general sets, respectively, and split them into the training and validation subsets following a 90/10 ratio.

We split DUD-E into three folds based on target classes to perform the cross-class CV study. There are 26 kinases in the first fold, 31 targets in the second fold (including 15 proteases, 11 nuclear receptors, and five G-protein coupled receptors), and the rest of 45 targets in the third fold. We also applied a random CV on DUD-E by randomly splitting the targets into three folds with the same fold sizes as the cross-class CV.

## Models

### ACNN

We applied the graph-based model ACNN implemented in the open source DeepChem package (Ramsundar et al., 2019) for predicting protein-ligand interactions in PDBbind. The ACNN model only requires atomic numbers and Cartesian coordinates of protein-ligand complexes as input to predict binding affinities. First, the ACNN model applies three independent atomic convolution blocks to extract atomic features from the ligand, protein, and protein-ligand complex, individually. In an atomic convolution block, the maximum number of closest neighbors (M) is used to represent the atomic environment for each atom. To represent the pairwise interaction, a radial basis function kernel is applied to map the distance between the atom and its each neighbor into a vector. And the atomic feature (a vector) is obtained by element-wise sum of M pairwise vectors. The atomic convolution blocks share the same initial parameters but will be changed after training. Secondly, one weight-sharing atomistic fully connected layer predicts atomic energies from all the atomic features. Thirdly, the ACNN model sums up the atomic energies to predict the energies of protein, ligand, and complex, individually, and then obtains the binding energy by subtracting the energies of protein and ligand from the energy of the binding complex. For analysis of bias in PDBbind, we modified ACNN to model only protein structures (protein alone), and only ligand structures (ligand alone) (**Supplementary Figure 1**). For protein alone, two independent atomic convolution blocks were used to extract atomic features from the same protein, and led to two different protein energies calculated from the same fully connected layer. The predicted "binding affinity" was the difference between two protein energies. The same strategy was applied for ligand alone as well. This strategy decouples the correlation of molecule size (number of atoms) and binding energy (sum of atomic energies), which enforces the ACNN model with the ability to learn atomic features.

All models were trained with an early-stopping strategy by stopping training if the performance on the validation subset did not improve in five epochs. The maximum number of neighbors of each atom was set to 4 at present study. We used a batch size of 16 and grouped samples with similar binding affinities into batches without changing the samples in one batch from the first to the last epochs. This training strategy is similar to the "curriculum learning" strategy (Bengio et al., 2009) because it reduces the difficulty of learning *via* training on the organized data.

### Random Forest

Two feature sets for decoy selection were used to build the RF models (Breiman, 2001) to evaluate the bias in the DUD-E dataset. The first feature set consisted of six physicochemical properties, including MW (only accounting all heavy atoms), cLogP, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, and net charge. The second feature set was ECFP (Morgan fingerprint with a radius of 2 and 2,048 bits in RDKit), which has been widely applied to encode molecular 2D topology into fixed length binary vector. We computed the properties and ECFP using the open source RDKit package.

The RF classifier from scikit-learn (Pedregosa et al., 2011) version 0.21.3 was used. The default parameters were used except that the number of estimators was set to 100 and the seed of random state was set to 0 for deterministic behavior during fitting. The AUC value was used to evaluate the classification performance of the RF. The enrichment factor was calculated as $EF_{subset} = (Actives_{subset}/N_{subset})/(Actives_{total}/N_{total})$. The higher the percentage of known actives found at a given percentage of the ranked database, the better the enrichment performance of the virtual screening. Since the practical value of virtual screening is to find active compounds as early as possible, we chose the enrichment factor at the top 1% of the ranked dataset ($EF_1$) to evaluate the early enrichment performance in the present study. In kinase inhibitor selectivity prediction, we used predictive index (PI) as a semi-quantitative measurement of the power of the target ranking order, where PI value (ranging from 1 to −1) of 1 indicates the perfect prediction, and 0 is completely random (Pearlman and Charifson, 2001).

# RESULTS

## High Performance Achieved on the PDBbind Datasets Using Random Splitting

We evaluated the performance of ACNN model to predict protein-ligand binding affinities on the PDBbind datasets using different data splitting approaches. The Pearson $R^2$ values on test subsets are reported in **Supplementary Table 1**. Firstly, we used a random splitting approach to split each PDBbind dataset into the training, validation, and test subsets five times with different random seeds. The increased number of protein-ligand complexes in the refined and general sets improved the ACNN model performance significantly (**Figure 1A**). The core set had the lowest mean $R^2$ value of 0.04, the refined and general sets with more samples were shown much higher performance with $R^2$ values of 0.80 and 0.70, respectively. We also trained the models on the refined and general sets, and tested the models on the core set, individually. The results were also promising, outperformed previously reported results of $R^2$ value of 0.66 using model trained on the refined set (Cang et al., 2018; Shen et al., 2019), with $R^2$ values of 0.70 and 0.73 using models trained on the refined and general sets, individually (**Figure 1B** and **Supplementary Table 2**).

Since PDBbind contains large number of kinase targets (309 kinase structures accounting 9.76% of the refined set), we wanted to test the performance of ACNN model on a benchmarking dataset for kinase inhibitor selectivity modeling (Wan et al., 2013). Using the models trained on the PDBbind refined set, the calculated mean $EF_{20}$ value of 1.12 and PI value of 0.01 indicate that such ACNN models cannot be used to predict the ranking order of the kinase targets for a given inhibitor (**Supplementary Table 3**).

To study the prediction power of the ACNN model, it is critical to decompose the contributions of the ligands and protein from the complex structure. Therefore, we generated two extra datasets by

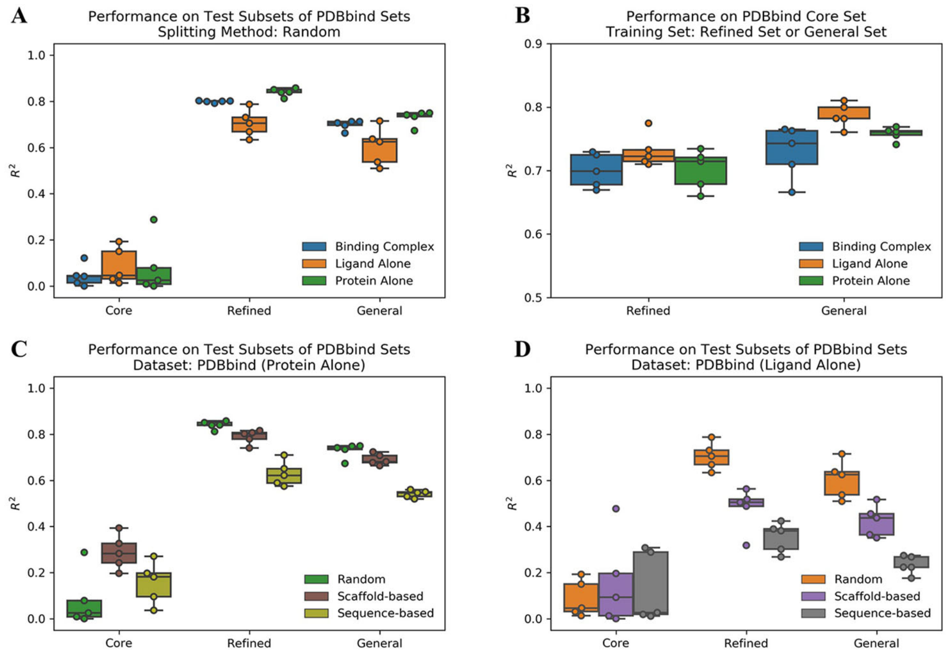


**FIGURE 1 |** Atomic convolutional neural network performance measured by the Pearson $R^2$ values obtained from the different PDBbind datasets using different splitting approaches. Each dataset was split into the training, validation, and test subsets five times with different random seeds following an 80/10/10 ratio, and studied on three different binding components, including protein-ligand complex structure (binding complex), only ligand structure (ligand alone), and only protein structure (protein alone), individually. **(A)** Models trained and tested within the same set. **(B)** Models trained on randomly selected subsets of the refined and the general sets (removing the core set structures) and tested on the core set. Models trained on the PDBbind datasets **(C)** (protein alone) and **(D)** (ligand alone) using different splitting methods.

dividing the protein-ligand complex structure (binding complex) into ligand structure (ligand alone) and protein structure (protein alone), individually. Strikingly, the model performance did not change significantly on datasets of ligand alone or protein alone in both the refined and general sets (**Figure 1A, B** and **Supplementary Table 1**). These results indicate that the ACNN model does not require learning protein-ligand interactions to achieve high performance, and suggest that data biases exist in PDBbind, both with proteins and with ligands.

## Protein and Ligand Similarity Biases in PDBbind

Li et al. reported that the protein similarity impacts the performance of AI models (Li and Yang, 2017). Therefore, we applied sequence-based splitting to reduce the impact of the protein similarity between the training and test subsets. When trained on protein alone, the $R^2$ value was reduced from 0.84 (random splitting) to 0.63 (sequence-based splitting) in the refined set; while it was reduced from 0.73 to 0.54 in the general set (**Figure 1C** and **Supplementary Table 1**). In addition, we guessed that ACNN learned the bias on ligand similarity. Therefore, we split the PDBbind datasets based on ligand scaffold similarity, and the performance of ACNN models was reduced significantly. When trained on ligand alone, the $R^2$ value was reduced from 0.71 (random splitting) to 0.48 (scaffold-based splitting) in the refined set, and from 0.60 to 0.42 in the general set. Since similar targets bind similar ligands, it is not surprising that protein sequence-based splitting also significantly reduced the model performance compared to random splitting. The $R^2$ values were reduced to 0.35 and 0.23 in the refined and general sets, individually (**Figure 1D** and **Supplementary Table 1**).

To further investigate what the ACNN model exactly learned from the ligand structures, we derived the atomic contributions from the ACNN models (ligand alone) trained on the PDBbind refined set (with structures in the core set removed) (**Figure 2**). Three representative systems were chosen from the core set to illustrate the atomic contributions of the ligands. Two protein tyrosine phosphatase 1B (PTP1B) inhibitors had similar atomic scores in Br atoms but different scores in S atoms, which suggests that the ACNN model could predict atomic contributions based on local atomic features (**Figure 2A**). However, the derived atomic contributions differed significantly in models trained with different random seeds, as demonstrated by the scores of the same Br atom changing from 0.55 to −0.04 in different models (**Supplementary Figure 2**). Atomic scores on the ligands bound to the antibody Fab showed that the model could predict one ligand (1zea) with larger molecular size but lower affinity by assigning negative scores on atoms with potentially unfavorable binding contributions (**Figure 2B**). For two acetylcholinesterase (AChE) inhibitors with similar size, the model correctly predicted the more potent inhibitor by identifying the presence of specific functional groups, such as Cl atom and ethyl group (**Figure 2C**). Combing the observations from those representative systems, the ACNN model is able to learn the correlation between atomic features and binding affinities. However, this correlation does not have to relate to protein-ligand interactions and may only represent the similarity of the ligands in PDBbind.

## Property Bias in DUD-E

Although the accurate prediction of ligand binding affinities is the ultimate goal of molecular docking, the practical value of structure-based virtual screening is to enrich the active compounds in the top ranked subset. Generally, the success of a virtual screening method is evaluated by its capacity to discriminate known active compounds from a background of decoy molecules. However, Sieg et al. (2019) reported that the distributions of MW beyond 500 Da between actives and decoys in DUD-E were mismatched (**Supplementary Figure 3**). Indeed, only using six properties as features, RF achieved a mean $EF_1$ of 22.2 and a mean AUC of 0.73 in random CV on DUD-E (**Figure 3A**). Therefore, we compiled the DUD-$E_{(MW \leq 500)}$ dataset to remove this specific MW bias (**Supplementary Figure 4**). A mean $EF_1$ of 15.4 and a mean AUC of 0.71 was achieved in random CV on DUD-$E_{(MW \leq 500)}$, more importantly, a mean $EF_1$ of 5.14 and a mean AUC of 0.66 was achieved in cross-class CV, which indicates that the model cannot use property bias to achieve high performance in cross-class CV on the DUD-$E_{(MW \leq 500)}$.

## Topology Bias in DUD-E

In DUD and DUD-E, the actives and decoys against the same target are dissimilar on topology and can be easily differentiated based on fingerprint (von Korff et al., 2009; Venkatraman et al., 2010; Hu et al., 2012; Lagarde et al., 2015; Kearnes et al., 2016; Sieg et al., 2019). However, whether the actives and decoys can be differentiated in cross-target CV based on fingerprint remains unclear, due to the mixed property bias and topology bias. By avoiding the use of property bias, we may study the independent contribution of topology bias on DUD-E. As shown in **Figure 3B**, using RF with molecular fingerprint (FP) as features, a mean AUC of 0.91 and a mean $EF_1$ of 32.75 in random CV was obtained on DUD-E. The model achieved a mean AUC of 0.86 and a mean $EF_1$ of 15.33 in cross-class CV on the DUD-$E_{(MW \leq 500)}$. These results indicate that the model can still use topology bias in DUD-E even after avoiding the property bias.

To investigate the topology bias in the DUD-E dataset, we calculated the relative frequency of bit set on each bit (2,048 bits) for actives and decoys in DUD-$E_{(MW \leq 500)}$ and the bit frequencies of ZINC12 compounds as reference (**Supplementary Figure 5**). Eighty-four bits with absolute log2 fold change ≥ 1 and mean relative frequency ≥ 0.03 were selected as representative bits (**Supplementary Figure 6**). About half of bit frequencies of actives and decoys are located on the opposite side of the bit frequencies of ZINC12 compounds, for example, the most populated bit 1,452 representing an aryl-alkyl ether group (**Figure 4**). This indicates that the topology distribution of decoys is strikingly different to actives. The rest of representative bits have relatively close frequencies between decoys and ZINC12 compounds, while larger differences between actives and ZINC12 compounds exist, such as bit 235 (representing six-membered aromatic ring) and bit 352 (representing aromatic ring with a sp2-hybridized carbon substituent). This further demonstrates that topology bias is not only caused by using

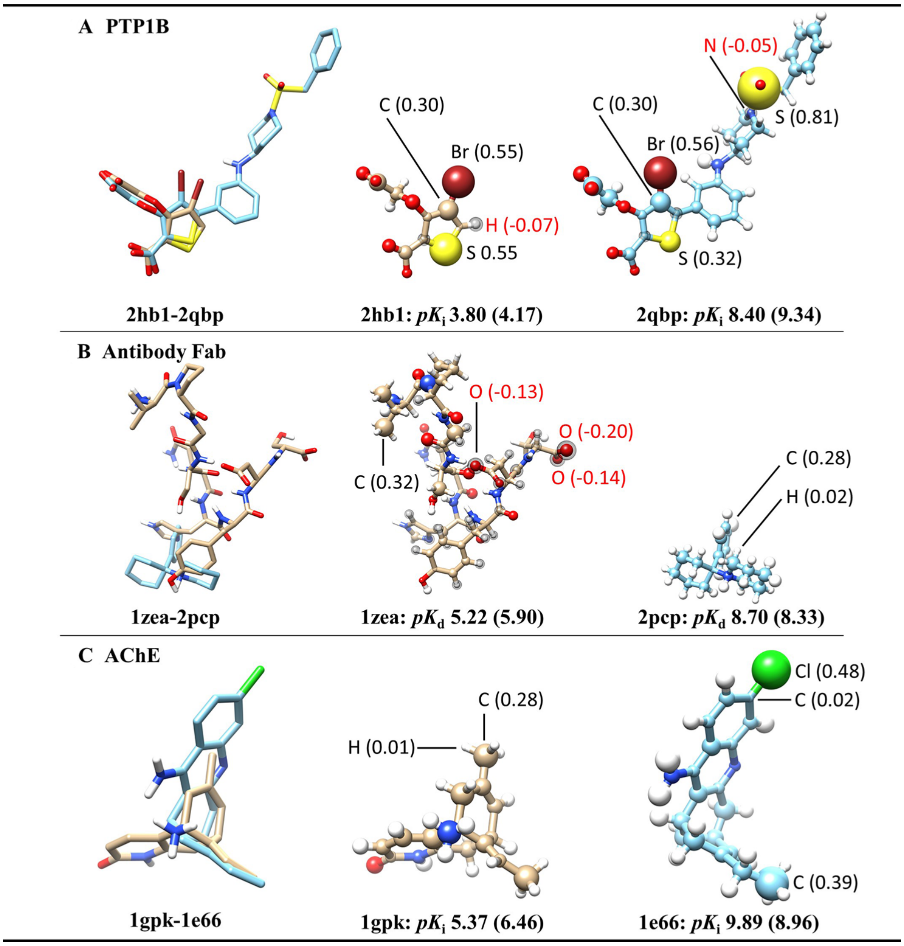


**FIGURE 2 |** Atomic contributions derived from the ACNN model (ligand alone) on three representative systems chosen from the PDBbind core set, including **(A)** protein tyrosine phosphatase 1B (PTP1B) inhibitors, **(B)** ligands bound to the antibody Fab and **(C)** acetylcholinesterase (AChE) inhibitors. The ACNN model (ligand alone) was trained on the refined set (removing the core set structures) and tested on the core set. Each row shows two ligands from the same protein target with different binding affinities ($pK_i$ or $pK_d$) (predictive values included inside the parentheses). The first column shows the superimposed ligand structures using the binding pocket alignment approach. The second and third columns show atomic contributions of each ligand. The size of the balls represents the absolute values of atomic scores. The atomic scores of selected atoms are labeled explicitly. The atoms with black spheres have negative scores. The molecular images were generated using UCSF Chimera (Pettersen et al., 2004).

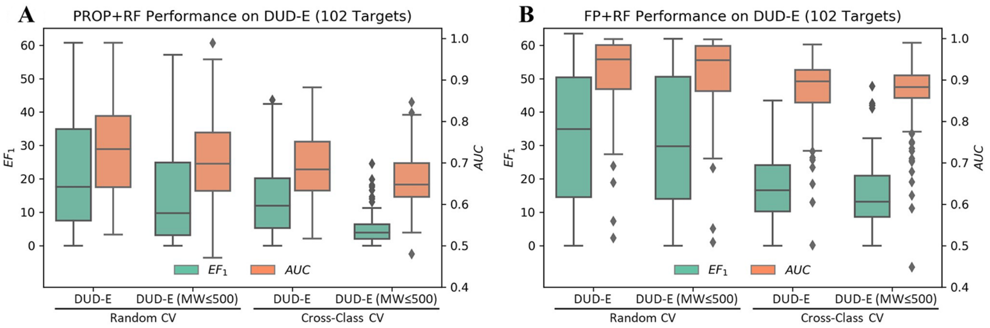


**FIGURE 3 |** Performance of RF on the DUD-E datasets using **(A)** six properties or **(B)** topology fingerprints. Note that the DUD-$E_{(MW \leq 500)}$ dataset was compiled by removing actives with MW (only including heavy atoms) greater than 500 and their associated decoys. The cross-class CV split the dataset into three folds based on target classes, and the random CV randomly split targets with the same fold sizes as in cross-class CV.

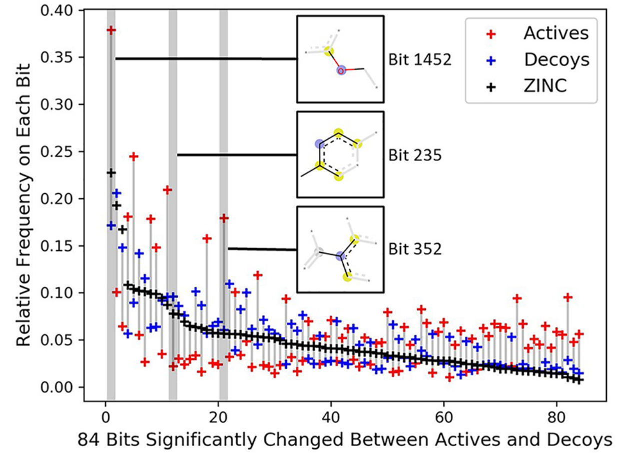


**FIGURE 4 |** Significantly changed bits between actives and decoys on DUD-E $_{(MW \le 500)}$. Eighty-four bits with absolute log2 fold change ≥ 1 between the actives and decoys and mean relative frequency ≥ 0.03 were selected as representative bits from the Morgan fingerprints (2,048 bits). The bits were sorted by frequencies of ZINC12 compounds. The chemical features of three selected bits are presented, and the chemical features of all 84 bits are summarized in **Supplementary Table 4**.

fingerprint as a decoy filter, but also resulted from the different topology distribution between actives and ZINC12 compounds. Therefore, the DUD and DUD-E datasets are not suitable for training models which directly or indirectly utilize the compound topological information.

## DISCUSSION AND CONCLUSIONS

State-of-the-art AI technologies represent a new paradigm in virtual screening with both opportunities and challenges for future improvement. The differences in different AI models mainly come from two aspects: one is the training dataset, and the other is the characterization method. At present work, we focused on analyzing the biases in two widely applied datasets for protein-ligand interactions. The former is represented by PDBbind, a collection of experimentally determined protein-ligand complex structures with known binding affinities, which is reliable, but the amount of data is small and arguably suffers from the data redundancy caused by the protein and ligand similarity. Our systematic investigation of ACNN models on the PDBbind datasets led to a surprising observation that the model performance was not correlated with learning essential protein-ligand interactions. Even the models trained on ligands or proteins performed as well as trained on complexes, while data splitting based on the similarity (protein sequence or ligand scaffold) clustering reduced the performance significantly. This suggests that the model performance may rely on the similarity of atomic features existing in the training and test subsets. It is expected that the rapidly increased amount of protein-ligand binding and structural data will improve the generality of the models by sampling the much larger and diverse chemical space.

DUD-E has become a common dataset for evaluating structure-based virtual screening methods, which were designed to benchmark enrichment performance by prioritizing the actives among a large amount of property-match but topology-dissimilar decoy molecules. As evidenced at present study, the topology bias is difficult to avoid when train on DUD-E. Therefore, care must be taken when using DUD-E for training AI models to predict protein-ligand interactions. However, DUD-E can still serve as an independent dataset to test the prediction power of AI models without using it for training. The use of fingerprint for selecting topological dissimilar decoys in the DUD and DUD-E datasets introduces topology bias in cross-target, and even cross-class CV. If we want to perform cross-target CV on DUD-like datasets for benchmarking AI models, the decoys shall be selected not only dissimilar to actives of a specific target, but also similar to actives of the other targets. Therefore, it is desirable to develop a more sophisticated approach for DUD-like decoy selection by depleting the topology bias, and such dataset may serve as a general-purpose benchmarking dataset to assess the enrichment performance of different virtual screening approaches (including AI models).

Nevertheless, it is encouraging that ACNN models have shown powerful capability for learning correlations hidden in structural data. Using the same neural network structure, ACNN was able to learn the structural similarities between ligands and between proteins. Even after protein sequence similarity clustering, ACNN still performed well in predicting ligand binding affinities. It is likely that ACNN model is well suitable for analysis of protein binding pocket, and it can be applied in protein pocket similarity analysis and protein pocket druggability prediction.

In summary, sufficiently large and unbiased datasets are desirable to fully exploit the potential of AI models for protein-ligand interactions. In addition to the guidelines proposed by Sieg et al. (2019), we can envision extra practical guidelines in developing and applying AI-based models. First of all, target structure-based methods do not guarantee that the performance of predicting ligand binding affinities is correlated with the learning of protein-ligand interactions. Vice versa, we demonstrated that ACNN models trained on the PDBbind datasets did not learn the essential protein-ligand interactions. Therefore, control experiments of training on the free ligands (ligand alone) and the free proteins (protein alone) can facilitate our understanding of what the AI models learned from the complex structures. Secondly, PDBbind is probably still going to be the best quality and the most accessible dataset for benchmarking protein-ligand interactions. However, it is necessary to evaluate the model performance by splitting datasets based on protein sequence and ligand scaffold similarity. Redundancy reduction increases the level of difficulty in model training, but will definitely improve the robustness of model transferability. Lastly, protein-ligand binding follows the laws of physics. The interpretability of AI models is critical for studying protein-ligand binding interactions, and visualization of atomic contributions decomposed from the models shall be engaged in extracting human understandable insights.

## DATA AVAILABILITY STATEMENT

All scripts and user tutorial are available at https://github.com/hnlab/can-ai-do. The kinase inhibitors dataset is available at http://www.huanglab.org.cn/kinome/kinome-ligand.tgz.

## AUTHOR CONTRIBUTIONS

NH and JY designed the project. JY and CS performed the computational studies. JY and NH analyzed the data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

## FUNDING

Computational support was provided by the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) under Grant No. U1501501 to NH.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2020.00069/full#supplementary-material

## REFERENCES

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv180601261 Cs Stat*.

Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi: 10.1021/jm9602928

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). “Curriculum Learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning ICML ‘09* (New York, NY, USA: ACM), 41–48. doi: 10.1145/1553374.1553380

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cang, Z., Mu, L., and Wei, G.-W. (2018). Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PloS Comput. Biol.* 14, e1005929. doi: 10.1371/journal.pcbi.1005929

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Chen, L., Cruz, A., Ramsey, S., Dickson, C. J., Duca, J. S., Hornak, V., et al. (2019). Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS One* 14, e0220113. doi: 10.1371/journal.pone.0220113

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc Interface* 15, 20170387. doi: 10.1098/rsif.2017.0387

Dunbar, J. B.Jr., Smith, R. D., Yang, C. Y., Ung, P. M., Lexa, K. W., Khazanov, N. A., et al. (2011). CSAR benchmark exercise of 2010: selection of the protein-ligand complexes. *J. Chem. Inf Model* 51, 2036–2046. doi: 10.1021/ci200082t

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). PotentialNet for molecular property prediction. *ACS Cent. Sci.* 4, 1520–1530. doi: 10.1021/acscentsci.8b00507

Gomes, J., Ramsundar, B., Feinberg, E. N., and Pande, V. S. (2017). Atomic convolutional networks for predicting protein-ligand binding affinity. *ArXiv170310603 Phys. Stat*.

Guedes, I. A., Pereira, F. S. S., and Dardenne, L. E. (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9, 1089. doi: 10.3389/fphar.2018.01089

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., and Tang, Y. (2012). Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *J. Chem. Inf. Model.* 52, 1103–1113. doi: 10.1021/ci300030u

Hu, Q., Feng, M., Lai, L., and Pei, J. (2018). Prediction of drug-likeness using deep autoencoder neural networks. *Front. Genet.* 9, 585. doi: 10.3389/fgene.2018.00585

Huang, N., Kalyanaraman, C., Bernacki, K., and Jacobson, M. P. (2006a). Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.* 8, 5166–5177. doi: 10.1039/B608269F

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006b). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Irwin, J. J., and Shoichet, B. K. (2016). Docking screens for novel ligands conferring new biology. *J. Med. Chem.* 59, 4103–4120. doi: 10.1021/acs.jmedchem.5b02008

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf Model* 52, 1757–1768. doi: 10.1021/ci3001277

Ivanenkov, Y. A., Zhavoronkov, A., Yamidanov, R. S., Osterman, I. A., Sergiev, P. V., Aladinskiy, V. A., et al. (2019). Identification of novel antibacterials using machine learning techniques. *Front. Pharmacol.* 10, 913. doi: 10.3389/fphar.2019.00913

Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). KDEEP: protein–ligand absolute binding affinity prediction *via* 3d-convolutional neural networks. *J. Chem. Inf. Model.* 58, 287–296. doi: 10.1021/acs.jcim.7b00650

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* 30, 595–608. doi: 10.1007/s10822-016-9938-8

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* 3, 935–949. doi: 10.1038/nrd1549

Lagarde, N., Zagury, J.-F., and Montes, M. (2015). Benchmarking data sets for the evaluation of virtual ligand screening methods: review and perspectives. *J. Chem. Inf. Model.* 55, 1297–1307. doi: 10.1021/acs.jcim.5b00090

Landrum, G. (2006). RDKit: Open-source cheminformatics.

Li, Y., and Yang, J. (2017). Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *J. Chem. Inf. Model.* 57, 1007–1012. doi: 10.1021/acs.jcim.7b00049

Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., et al. (2014). Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *J. Chem. Inf. Model.* 54, 1700–1716. doi: 10.1021/ci500080q

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., et al. (2017). Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.* 50, 302–309. doi: 10.1021/acs.accounts.6b00491

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., et al. (2019). Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229. doi: 10.1038/s41586-019-0917-9

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Pearlman, D. A., and Charifson, P. S. (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 map kinase protein system. *J. Med. Chem.* 44, 3417–3423. doi: 10.1021/jm0100279

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peng, S., Xiao, W., Ju, D., Sun, B., Hou, N., Liu, Q., et al. (2019). Identification of entacapone as a chemical inhibitor of FTO mediating metabolic regulation through FOXO1. *Sci. Transl. Med.* 11, eaau7116. doi: 10.1126/scitranslmed.aau7116

Pereira, J. C., Caffarena, E. R., and dos Santos, C. N. (2016). Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506. doi: 10.1021/acs.jcim.6b00355

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* 57, 942–957. doi: 10.1021/acs.jcim.6b00740

Ramsudar, B., Eastman, P., Walters, P., and Pande, V. (2019). Deep learning for the life sciences : applying deep learning to genomics, microscopy, drug discovery and more. *First edition*. (Sebastopol, CA: O'Reilly Media).

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., and Hou, T. (2019). From machine learning to deep learning: advances in scoring functions for protein–ligand docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 0, e1429. doi: 10.1002/wcms.1429

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature* 432, 862–865. doi: 10.1038/nature03197

Sieg, J., Flachsenberg, F., and Rarey, M. (2019). In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* 59, 947–961. doi: 10.1021/acs.jcim.8b00712

Stahl, M., and Rarey, M. (2001). Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* 44, 1035–1042. doi: 10.1021/jm0003992

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., Siedlecki, P., and Valencia, A. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 34, 3666–3674. doi: 10.1093/bioinformatics/bty374

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., et al. (2019). Comparative assessment of scoring functions: the CASF-2016 Update. *J. Chem. Inf. Model.* 59, 895–913. doi: 10.1021/acs.jcim.8b00545

Venkatraman, V., Pérez-Nueno, V. I., Mavridis, L., and Ritchie, D. W. (2010). Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* 50, 2079–2093. doi: 10.1021/ci100263p

von Korff, M., Freyss, J., and Sander, T. (2009). Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* 49, 209–231. doi: 10.1021/ci800303k

Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *ArXiv151002855 Cs Q-Bio Stat*.

Wan, X., Zhang, W., Li, L., Xie, Y., Li, W., and Huang, N. (2013). A new target for an old drug: identifying mitoxantrone as a nanomolar inhibitor of PIM1 kinase *via* kinome-wide selectivity modeling. *J. Med. Chem.* 56, 2619–2629. doi: 10.1021/jm400045y

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., et al. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* 18, 12964–12975. doi: 10.1039/C6CP01555G

Wang, Y., Sun, Y., Cao, R., Liu, D., Xie, Y., Li, L., et al. (2017). In silico identification of a novel hinge-binding scaffold for kinase inhibitor discovery. *J. Med. Chem.* 60, 8552–8564. doi: 10.1021/acs.jmedchem.7b01075

Xu, Y., Lin, K., Wang, S., Wang, L., Cai, C., Song, C., et al. (2019). Deep learning for molecular generation. *Future Med. Chem.* 11, 567–597. doi: 10.4155/fmc-2018-0358

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040. doi: 10.1038/s41587-019-0224-x

Zhou, Y., Ma, J., Lin, X., Huang, X.-P., Wu, K., and Huang, N. (2016). Structure-based discovery of novel and selective 5-hydroxytryptamine 2B receptor antagonists for the treatment of irritable bowel syndrome. *J. Med. Chem.* 59, 707–720. doi: 10.1021/acs.jmedchem.5b01631

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/fphar.2020.00112

# EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction

*Fangping Wan*[1†], *Shuya Li*[1†], *Tingzhong Tian*[1], *Yipin Lei*[2], *Dan Zhao*[1*] *and Jianyang Zeng*[1*]

*[1] Institute of Interdisciplinary Information Science, Tsinghua University, Beijing, China, [2] Machine Learning Department, Silexon AI Technology Co. Ltd., Nanjing, China*

Synthetic lethality (SL), an important type of genetic interaction, can provide useful insight into the target identification process for the development of anticancer therapeutics. Although several well-established SL gene pairs have been verified to be conserved in humans, most SL interactions remain cell-line specific. Here, we demonstrated that the cell-line-specific gene expression profiles derived from the shRNA perturbation experiments performed in the LINCS L1000 project can provide useful features for predicting SL interactions in human. In this paper, we developed a semi-supervised neural network-based method called EXP2SL to accurately identify SL interactions from the L1000 gene expression profiles. Through a systematic evaluation on the SL datasets of three different cell lines, we demonstrated that our model achieved better performance than the baseline methods and verified the effectiveness of using the L1000 gene expression features and the semi-supervise training technique in SL prediction.

**Keywords: synthetic lethality, L1000 gene expression profiles, machine learning, semi-supervised neural network, target identification**

OPEN ACCESS

**Edited by:**
*Alex Zhavoronkov,
Biogerontology Research Foundation,
United Kingdom*

**Reviewed by:**
*Feng ZHU,
Zhejiang University, China
Vasileios Stathias,
University of Miami, United States
Qi Zhao,
Shenyang Aerospace University,
China
Jihye Kim,
University of Colorado Anschutz
Medical Campus,
United States
Bhaskar Roy,
Beijing Genomics Institute (BGI),
China*

***Correspondence:**
*Dan Zhao
zhaodan2018@tsinghua.edu.cn
Jianyang Zeng
zengjy321@tsinghua.edu.cn*

*†These authors have contributed equally to this work*

**Specialty section:**
*This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology*

***Received:*** *27 October 2019*
***Accepted:*** *28 January 2020*
***Published:*** *28 February 2020*

**Citation:**
*Wan F, Li S, Tian T, Lei Y, Zhao D and Zeng J (2020) EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction.
Front. Pharmacol. 11:112.
doi: 10.3389/fphar.2020.00112*

## INTRODUCTION

Two genes are considered a synthetic lethal (SL) pair if perturbation of both genes induces a defect in cell viability, while perturbation of either gene is not harmful to cell survival (Boone et al., 2007). Different types of perturbations were considered to trigger SL in previous studies, including knockdown, knockout, mutation, aberrant gene expression, copy number variation, and drug treatment (Whitehurst et al., 2007; Jerby-Arnon et al., 2014; Han et al., 2017; Sinha et al., 2017). Studying synthetic lethal interactions may help gain novel insights into target identification. Many cancer cells carry specific mutations in one gene (e.g., a tumor suppressor gene) of a synthetic lethal pair, and thus its synthetic lethal partner becomes a promising drug target (O'Neil et al., 2017). For example, the known synthetic lethal interactions between the tumor suppressor gene *BRCA1/2* and the drug target gene *PARP1* can be used to selectively kill cancer cells by triggering fatal DNA damages (Bryant et al., 2005; Farmer et al., 2005). To this end, PARP1 inhibitors have been approved to treat certain types of *BRCA*-mutated cancers (Fong et al., 2009).

SL gene pairs can be experimentally screened by developing double-knockout strains in model organisms and human cell lines. The synthetic lethality network in yeast has been well constructed using synthetic genetic arrays (SGA) (Tong et al., 2001) and diploid synthetic lethality analysis with microarrays (dSLAM) (Pan et al., 2007). Nearly one million gene pairs covering 90% of the whole yeast genome were screened in a recent study (Costanzo et al., 2016). Compared to yeast strains, which can undergo sexual reproduction to generate double-knockout offspring from parents bearing different single knockouts, it is more challenging to develop double-knockout human cell lines in an efficient manner. Thus, a relatively low number of human gene pairs (about hundreds or thousands) can be screened by RNA interference (Whitehurst et al., 2007; Barbie et al., 2009) and CRISPR-Cas9 (Shen et al., 2017; Han et al., 2017) based double-knockout experiments. Due to the difficulty in the establishment of large-scale double-knockout systems in human cell lines, the currently screened gene pairs only account for a small fraction of all possible combinations of human genes.

To overcome the current difficulty in experimental screen and generate more SL interactions in human, computational methods have recently been proposed to predict novel human SL pairs recently. The most direct idea is to leverage the abundant SL pairs characterized in yeast to infer human SLs through ortholog mapping (Deshpande et al., 2013; Wu et al., 2013; Srivas et al., 2016). The application of these methods was limited, as a large number of human genes do not have evolutionarily close yeast orthologs. Network-based methods predict human SLs through analyzing the protein-protein interaction (PPI) networks, metabolic networks, or signaling pathways (Folger et al., 2011; Kranthi et al., 2013; Zhang et al., 2015; Apaolaza et al., 2017). Statistical methods were also developed to identify SL gene pairs from human cancer cells based on the principle that the perturbations (*e.g.*, mutation, aberrant gene expression, and copy number variation) of both SL genes should be subject to negative selection and exhibit a mutually exclusive pattern (Jerby-Arnon et al., 2014; Srihari et al., 2015; Jacunski et al., 2015; Sinha et al., 2017; Lee et al., 2018). Besides, there exist several machine-learning-based approaches for predicting SL gene pairs. Most of these approaches learn from the adequate amount of supervised information of yeast (Wong et al., 2004; Pandey et al., 2010; Li et al., 2011). Only a few machine learning methods for predicting human SLs were developed. For example, Das et al. used a Random Forest classifier with multi-omics features (*e.g.*, differential expression, expression correlation, mutual exclusivity and shared pathways) to predict SL pairs in human cancer (Das et al., 2018); and Liu et al. proposed a logistic matrix factorization model regularized by the PPI similarity network and the gene ontology (GO) semantic similarity network to predict SL pairs (Liu et al., 2019).

Although a number of SL interactions are conserved in humans, most of them are only observed in specific cell lines or tissues (Ryan et al., 2018). A recent study detected SL pairs in three cell lines and found that only about 10% of SL interactions were shared by two cell lines, and no SL pair was identified in all the three cell lines (Shen et al., 2017). Despite the extensive applications of the above computational methods in SL prediction, most of them make predictions for the human genetic network without considering the cell line or tissue context. Although one of the aforementioned methods (Das et al., 2018) can predict SL in different human cancer types, it is difficult to directly apply this method to cell lines, as the homogenous genetic background of cell lines cannot provide enough mutation-related omics data. To provide a feasible tool for capturing the unique SL interaction networks for individual cell types, we aim to develop a computational method to learn from the experimentally measured SL interactions through considering the cell-line specific genetic information.

In this paper, we have proposed a novel computational method, EXP2SL, to predict cell-line specific SL interactions in human. The cell-line specific gene expression profiles resulting from the shRNA knockdown experiments in the LINCS L1000 project (Subramanian et al., 2017) were used to capture the information of cell-line specific genetic background. Since the available labeled data in single cell lines are limited, a semi-supervised objective function is used to exploit the large amount of unlabeled data. Tested on the combinatorial CRISPR-Cas9 perturbation-based SL datasets in three different cell lines, our model showed competitive prediction ability compared to the baseline methods. We also verified the effectiveness of the features derived from the L1000 gene expression profiles and the semi-supervised objective function. Furthermore, we evaluated the importance of each gene included in the L1000 gene expression profiles and found that the cell viability related functions were enriched among the top attributing genes.

# METHODS

## Data Processing

### The L1000 Gene Expression Profiles

The LINCS L1000 project (Subramanian et al., 2017) measured the expression levels of 978 landmark genes under different perturbations (*i.e.*, shRNA or compounds) and control conditions (*i.e.*, empty vectors or solvents) in different human cell lines. Here, we used the gene expression profiles resulting from shRNA perturbations to construct the features of the corresponding shRNA target genes, which were 978-dimensional vectors.

Specifically, the raw data from the LINCS L1000 project were preprocessed based on the pipeline in the original paper (Subramanian et al., 2017) with minor modifications; We first directly obtained the Level 3 data from L1000, which contained the quantile normalized gene expression profiles. The shRNA profiles perturbed after 96 hours were used, as the data amount for this time point was the largest. Based on this dataset, we calculated the z-score for each dimension of a shRNA perturbed profile $x \in \mathbb{R}^{978}$ by

$$\boldsymbol{z} = \frac{\boldsymbol{x} - median(\boldsymbol{V})}{1.4826 * MAD(\boldsymbol{V})}, \tag{1}$$

where $\boldsymbol{z}$ is a 978-dimensional z-score of the shRNA perturbation profile $\boldsymbol{x}$, $\boldsymbol{V}$ is the set of vector control profiles from the same plate, *median(**V**)* and *MAD(**V**)* stand for the median value and the median absolute deviation of $\boldsymbol{V}$, and 1.4826 is a scaling factor to make the resulted z-scores close to normal distribution. Notably, in the original L1000 preprocessing pipeline (Subramanian et al., 2017), the control profiles were replaced by all the profiles on the plate, called population control. Here, we argue that this data preprocessing scheme may cause a biased control distribution due to the specific perturbation design. Thus, we use the expression levels treated with empty vectors as the control for the shRNA perturbed profiles.

For each gene, typically more than one types of shRNA were designed to knock down the expression of the corresponding gene product. To eliminate the off-target effects of shRNAs and obtain a robust signature for each single gene, the z-scores obtained from the replicated trials of the same shRNA were first processed using an algorithm with L1000 Level 5 data (Subramanian et al., 2017), then the same protocol was used to reduce the shRNAs targeting the same gene. More specifically, the z-scores were weighted and averaged according to the Spearman correlations to obtain a final 978-dimensional L1000 gene expression profile for each gene, which was then used as the input gene features for our model and other baseline models.

## SL Labels

The SL labels in our datasets were constructed from the CRISPR double-knockout experiments performed in human cell lines (Shen et al., 2017; Zhao et al., 2018; Najm et al., 2018). A recently proposed computational approach called GEMINI (Zamanighomi et al., 2019) was used to identify SL interactions from the combinatorial CRISPR perturbation based cell viability studies. We adopted the GEMINI scores to select the positive and negative SL pairs for constructing our datasets. In particular, for each cell line, positive SL pairs were selected from gene pairs satisfying two criteria: 1) GEMINI "strong" scores larger than zero, which indicates the existence of the synergic lethal effect, and 2) GEMINI "strong" scores ranking among top 5%, to reduce the potential false positives. The main reason for choosing this threshold is that the top 5% gene pairs were considered as "the most significant hits in each screen" in the GEMINI paper (Zamanighomi et al., 2019). To more thoroughly evaluate the performance of our method, we also tested another threshold (*i.e.*, 10%) for choosing the positive SL pairs (**Tables S1-S2**). Negative SL pairs were those gene pairs satisfying 1) a GEMINI "strong" score less than zero, which means that there exists no synergic lethal effect between these two genes, and 2) a GEMINI "strong" score among the bottom 50%, to remove the potential false negatives. The gene pairs that were not selected as positive or negative SL pairs were considered as unknown pairs. Finally, cell lines with adequate numbers (>100) of gene pairs with both SL labels and L1000 gene expression profiles, including A549, A375, and HT29, were used in our study. The numbers of training samples for the cell lines are summarized in **Table 1**.

# The Workflow of EXP2SL

The basic idea of our EXP2SL model is to extract useful information from the L1000 expression profiles to accurately predict cell-line specific SL interactions. To achieve this goal, a semi-supervised objective function was designed to fully exploit the large amount of unlabeled data (**Figure 1**).

**TABLE 1** | Number of labeled training samples for each cell line.

| | A549 | A375 | HT29 |
|---|---|---|---|
| Positive SL gene pairs | 126 | 18 | 18 |
| Negative SL gene pairs | 1106 | 44 | 123 |
| Total | 1232 | 62 | 141 |

## The Network Architecture of EXP2SL

For a given cell line, suppose that there are $N$ genes (marked as the indices 1, 2,…, $N$) with measured shRNA data from the LINCS L1000 project (Subramanian et al., 2017). The corresponding L1000 gene expression profiles can be represented as a set of feature vectors $\{f_i \in \mathbb{R}^{978}\}_{i=1}^{N}$.

For a given cell line, our model first encodes the gene features through $E$ sequential fully-connected layers, that is,

$$\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{e}} = ReLU\left(\boldsymbol{W}_{encoder}^{e}\boldsymbol{h}_i^{e-1} + \boldsymbol{b}_{encoder}^{e}\right), \tag{2}$$

$$e = 1, 2, \ldots, E, i = 1, 2, \ldots, N,$$

where $\boldsymbol{h}_i^0 = \boldsymbol{f}_i$, $ReLU(x)$ stands for the rectifier linear activation function $ReLU(x) = max(0,x)$, $\boldsymbol{W}_{encoder}^{1} \in \mathbb{R}^{d\times 978}$, $\boldsymbol{W}_{encoder}^{e} \in \mathbb{R}^{d\times d}(e = 2, \ldots, E)$, and $\boldsymbol{b}_{encoder}^{e} \in \mathbb{R}^{d}(e = 1, \ldots, E)$ denote the learnable parameters ($d$ is the dimension of the hidden layers).

After $E$ encoding layers, the updated gene features $\{\boldsymbol{h}_i^E\}_{i=1}^{N}$ are then used to predict SL interactions. More specifically, for a gene pair $(i, j)$, $i, j = 1,2,\ldots, N$ and $i \neq j$, a confidence score is calculated through a linear layer to predict the potential of SL interaction between this gene pair, that is,

$$s_{i,j} = \frac{1}{2}\left(\boldsymbol{W}_{out}\left[\boldsymbol{h}_i^E, \boldsymbol{h}_j^E\right] + \boldsymbol{W}_{out}\left[\boldsymbol{h}_j^E, \boldsymbol{h}_i^E\right]\right) + \boldsymbol{b}_{out}, \tag{3}$$

where $\boldsymbol{W}_{out}\in\mathbb{R}^{1\times 2d}$ and $\boldsymbol{b}_{out}\in\mathbb{R}$ stand for learnable parameters. Note that the pairs $(i, j)$ and $(j, i)$ are equivalent to each other, so we calculate the average prediction scores of concatenations of $[\boldsymbol{h}_i^E, \boldsymbol{h}_j^E]$ and $[\boldsymbol{h}_j^E, \boldsymbol{h}_i^E]$ to obtain the equivalent prediction results for input pairs $(i, j)$ and $(j, i)$.

## The Semi-Supervised Objective Function

As described in SL Labels, the gene pairs with different SL labels can be classified into positive, negative, and unknown sets, denoted as $P$, $N$, and $U$, respectively. Here, we designed a semi-supervised loss function that utilizes information from all three sets to optimize the parameters of our model. More specifically, our loss consisted of three parts:

The first part of our objective function is the mean squared error (MSE) of positive and negative samples, calculated as

$$L_{MSE} = \sum_{(i,j)\in P\cup\mathcal{N}} \left(\hat{s}_{i,j} - s_{i,j}\right)^2, \tag{4}$$

where $\hat{s}_{i,\,j} = 1$ if $(i, j) \in P$, $\hat{s}_{i,\,j} = -1$ if $(i, j) \in N$, and $s_{i,\,j}$ stands for the potential score of gene pair $(i, j)$ predicted by EXP2SL.

The second part of the objective function is inspired by the semi-supervised Bayesian personalized ranking (BPR) loss (Rendle et al., 2009), which uses the unknown labels to boost the prediction performance. In particular, the BPR loss is defined as

$$L_{BPR} = \sum_{(a,b)\in P,(c,d)\in\mathcal{U}} \log\,\sigma\left(s_{a,b} - s_{c,d}\right)$$

$$+ \sum_{(c,d)\in\mathcal{U},(e,f)\in\mathcal{N}} \log\,\sigma\left(s_{c,d} - s_{e,f}\right), \tag{5}$$

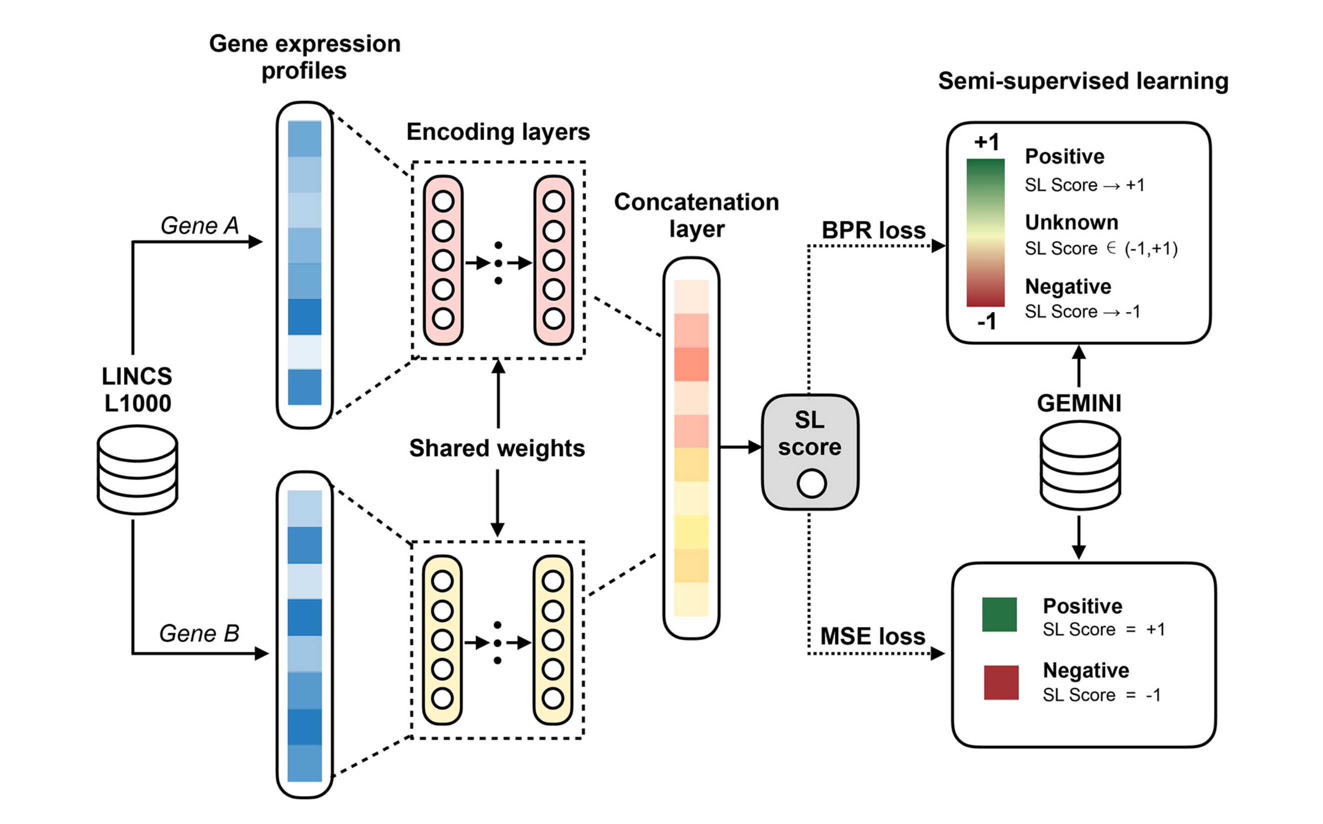


**FIGURE 1 |** Workflow of the EXP2SL model. For a pair of gene, their L1000 gene expression profiles derived from knockdown conditions are the inputs of the encoding layers. Then, the updated features for both genes in a given pair are concatenated to predict the confidence score of being an SL pair by a linear combination. In addition, a semi-supervised objective function is used to train the model parameters, which aims to utilize the information from both known (positive and negative) and unknown SL gene pairs.

where $\sigma$ stands for the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. This objective function aims to enlarge the margins of the predicted scores between positive SL and unknown pairs, as well as those between the unknown and negative SL pairs. To calculate this loss, we sample the negative and unknown pairs with the sample number equal to the positive pairs during model training.

The above MSE and BPR objective functions are further combined with an L2 regularizier over all the learnable model parameters to construct the final objective function of our EXP2SL model, that is,

$$L(\theta) = L_{MSE} + \lambda_1 L_{BPR} + \lambda_2 ||\theta||^2, \tag{6}$$

where $\theta$ denotes the model parameters, and $\lambda_1$ and $\lambda_2$ stand for the weight parameters controlling the contributions of the BPR loss and the L2 regularization term, respectively.

To train the EXP2SL model, we used the Adam optimizer (Kingma and Ba, 2014) with the default learning rate 0.001 and the number of training epochs 1,000. We also clipped the gradient if it was larger than 5 to stabilize the training process. We implemented our model with PyTorch 1.0.1 (Paszke et al., 2017).

## Hyper-Parameters

The hyper-parameters of our model include the weight of the BPR loss $\lambda_1$ from [16, 32, 64, 128], the weight of the L2 regularization $\lambda_2$ from [0.1, 0.05, 0.01, 0.005, 0.0001], the number of encoding layers from [0, 1, 2, 3, 4], and the dimension of hidden features $d$ from [32, 64, 128, 256]. For each cell line, a grid search was performed to select the best combination of hyper-parameter settings from the above mentioned ranges, according to the AUC scores achieved by five repeats of 5-fold cross validations under the “split pair” setting (*i.e.*, gene pairs were randomly split into training and test sets). Details about the cross-validation settings can be found in *Performance Evaluation*. The baseline models were tuned using the same strategy, and the ranges for hyper-parameters in each baseline model are described in the *Baseline Models*.

## Extraction of Feature Importance

Here, we used the saliency map-based approach proposed in (Simonyan et al., 2013) to evaluate the importance of each position along the 978-dimensional input features $\{f_i\}_{i=1}^{N}$. The basic idea of this method is to calculate the gradients of the output score with respect the to the input features, and the larger absolute values of gradients would suggest the more importance of the corresponding feature dimension. After the training process, the positive and negative SL pairs of each cell line are fed into the EXP2SL model, and the corresponding importance for each input feature dimension is calculated by

$$\boldsymbol{w} = \sum_{(i,j)\in P \cup \mathcal{N}} \left|\frac{\partial s_{i,j}}{\partial \boldsymbol{f}_i}\right| + \left|\frac{\partial s_{i,j}}{\partial \boldsymbol{f}_j}\right|, \tag{7}$$

where $s_{i,\,j}$ is the predicted confidence score of gene pair $(i, j)$, and $\boldsymbol{w}$ is a 978-dimensional vector containing the importance score of each dimension of the input L1000 gene expression profiles. To reduce the variance caused by random initialization of network parameters and random sampling of the unknown and negative gene pairs for calculating the BPR loss during the training process, we also take the summation of $\boldsymbol{w}$ vectors from 10 trained EXP2SL models to obtain the final importance scores for the 978 feature dimensions. The top 50 ranked features are then selected for each cell line. We examined the overlaps of the selected features between cell lines and calculated the over-representations of functional gene sets and pathways using the WebGestalt server (Liao et al., 2019).

## Baseline Models

### Logistic Regression

We used the logistic regression (LR) model implemented based on scikit-learn (Buitinck et al., 2013). The L1000 expression profiles were used as input to the LR model. For each pair of input genes $(i,j)$, the features of genes $i$ and $j$ (denoted as $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, respectively) were concatenated before being fed into the LR model. Since LR may produce different results for pairs $(i, j)$ and $(j, i)$, each of the two pairs were treated as an individual input with the same label in the training phase. In the test phase, the prediction values from both inputs were then averaged to obtain the final prediction score. The inverse of regularization strength (a hyper-parameter) was chosen from [10, 1, 0.5, 0.1, 0.05, 0.01].

### Random Forest

We used the random forest (RF) classifier implemented based on scikit-learn (Buitinck et al., 2013). The input and output of RF were the same as those of LR described above. The number of trees was selected from [32, 64, 128] and the maximum depth of the trees was selected from [8, 16, None], where "None" means that the trees will keep expanding until no node can be split.

### Support Vector Machine

We used the support vector machine (SVM) classifier implemented based on scikit-learn (Buitinck et al., 2013). The input and output of SVM were the same as those of LR and RF described above. The only hyper-parameter, the inverse of regularization strength, was selected from [100, 50, 10, 5, 1, 0.5, 0.1].

### Gradient Boosting Decision Tree

We used the gradient-boosting decision tree (GBDT) classifier implemented by the XGBoost project (Chen and Guestrin, 2016). The input and output of GBDT were the same as other classifiers described above. The number of trees was selected from [32, 64, 128] and the maximum depth of the trees was selected from [4, 8, 16].

### NetLapRLS

NetLapRLS (Xia et al., 2010) (a semi-supervised regressor) was implemented based on pyDTI (https://github.com/stephenliu0423/PyDTI). As NetLapRLS treats symmetric gene pairs $(i, j)$ and $(j, i)$ in the same way, there is no need to average the predictions of both pairs. Three types of similarity matrices were used as the input to NetLapRLS: 1) The protein-protein interaction (PPI) similarity matrix $S_p$, *i.e.*, the pairwise PPI similarities between all pairwise genes used in the cell line. The human PPI data were obtained from the STRING database v11 (Szklarczyk et al., 2014). Protein pairs marked with STRING scores larger than 0.8 were considered positive interaction pairs in the PPI network. The PPI similarity between two proteins $(i, j)$ were calculated as the Jaccard similarity of their interaction partners in the PPI network, that is,

$$S_p(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}, \tag{8}$$

where $N(x)$ stands for the neighbors of protein $x$ in the PPI network. 2) The L1000 profile similarity matrix $S_l$, *i.e.*, the absolute values of the pairwise L1000 profile similarities between all the genes used in the cell line. The L1000 profile similarity between two genes were calculated as the Pearson correlation between their L1000 gene expression profiles. 3) The combination of both PPI and L1000 similarities, calculated as $1 - (1 - S_p)(1 - S_l)$. The best hyper-parameter settings were selected from all the combinations over $\gamma_d = \gamma_t$ from [0.0001, 0.001, 0.01, 0.1, 1] and $\beta_d = \beta_t$ from [0.003, 0.03, 0.3,3, 30].

# RESULTS

## Cell-Line Specificity of SL Interactions

To demonstrate the cell-line specificity of SL interactions, we examined 378 CRISPR knockout pairs screened in different cell lines from the Big Papi SynLet library (Najm et al., 2018). Their SL scores were calculated by GEMINI (Zamanighomi et al., 2019), a computational tool for identifying SL interactions from pairwise CRISPR knockout screens. Three cell lines were used in our performance evaluation, including A549, A375, and HT29. Among these three cell lines, A549 and A375 exhibited relatively high correlation (Pearson correlation 0.71, **Figure 2A**) in GEMINI scores, which measure the strength of the SL interactions. Meanwhile, the correlations between HT29 and the other two cell lines are relatively low (Pearson correlations 0.36 and 0.28, **Figure 2A**). These results indicate that the SL interaction patterns between the same gene pairs in different cell lines can be quite different.

Next, we examined the positive and negative SL samples selected from the Big Papi dataset according to the criteria described in *SL Labels*. By comparing the SL labels of the same gene pairs in the three cell lines, we found that most gene pairs have inconsistent labels cross different cell lines (**Figure 2B**). There are 38 gene pairs with at least one positive label in the three cell lines, but only one of them (*i.e.*, the *BRCA1-PARP1* gene pair) is always labeled as a positive SL. Among these 38 gene pairs, 16 have negative labels in one cell line but positive labels in another one.

Based on the above observation that most SL pairs were not conserved across different cell lines, we built prediction models for each cell line separately. In addition to the Big Papi dataset, we also included the data from other literature (Shen et al., 2017;

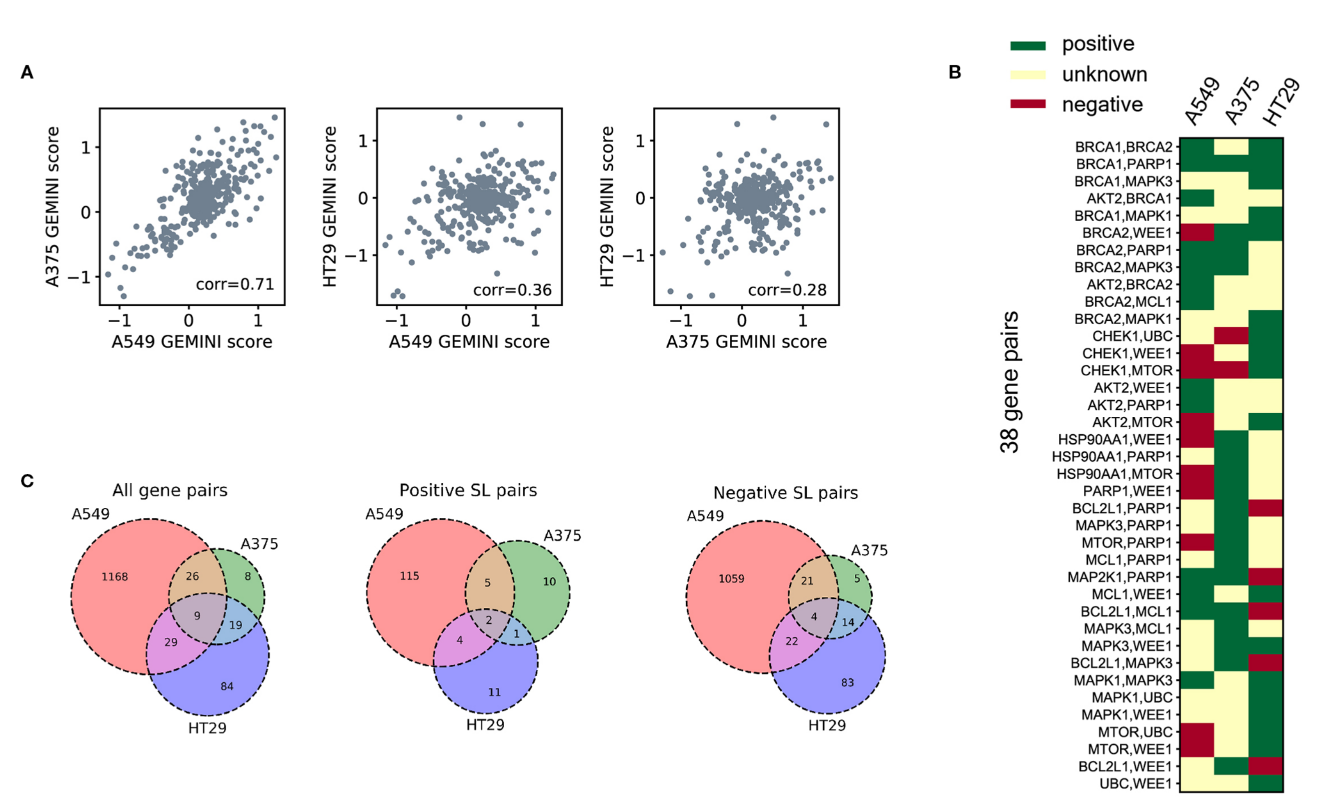


**FIGURE 2 |** SL datasets for three human cell lines. **(A)** Correlations of the GEMINI scores between three different cell lines for the same gene pairs measured in the Big Papi dataset. **(B)** The binary SL labels for the gene pairs in the Big Papi dataset. The 38 gene pairs measured in all the three cell lines and with at least one positive SL label are included in the figure. **(C)** The Venn diagrams of all labeled SL pairs, positive SL pairs, and negative SL pairs used in our dataset, which were constructed from the Big Papi dataset and other available CRISPR-Cas9 based experimental screens in the literature.

Zhao et al., 2018), which further enlarged the SL data of cell line A549. The overlaps of gene pairs used as labeled training samples between the three cell lines are shown in **Figure 2C**.

## Performance Evaluation

We compared the performance of our model to that of several baseline methods through cross-validation on the aforementioned datasets for the three cell lines. LR, RF, SVM, and GBDT were selected as the baseline methods because they are the machine learning baseline models and accept vector input, which is suitable for our case. NetLapRLS is also used as a baseline model, as it is a well-established semi-supervised method that accepts network input and which can be used to test the effectiveness of other features, such as the PPI network. Two settings were used to split the training and test samples. The first one was called "split pair" in which gene pairs were randomly split into training and test sets. The second one was called "split gene" in which, for each test gene pair, at least one gene is not seen in training data. The "split gene" setting was mainly used to test whether the prediction can be generalized to unseen genes, which is more challenging. Note that the splitting was performed over positive and negative SL pairs, and our model also utilized the unknown pairs during the training process.

Area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPR), F1 score, accuracy, precision, sensitivity and selectivity were used to evaluate the classification performance (**Tables 2** and **3**). The receiver operating characteristic (ROC) and precision-recall (PR) curves achieved by EXP2SL and the baseline models are shown in **Figures S2–S3**. Under the "split pair" setting, all the models achieved relatively high performance, which indicates that the prediction problem defined under this setting was relatively easy. The performance of our model was comparable with the top-performing baseline methods under this setting. However, under the more practical "split gene" setting in which we wished to predict SL pairs containing novel genes without experimental screen data (due to the limited existing experimental data), the SL prediction task became difficult as all the models achieved relatively lower AUC and AUPR scores than those under the "split pair" setting. However, our model exhibited a significantly better performance than that of all the baseline models under this "split gene" setting. EXP2SL achieved the best performance in at least 6/7 metrics for all the three cell lines (**Table 3**). We also tested our model and the baseline methods with a less strict threshold for defining the positive SL pairs (*i.e.*, 10%), and our model also achieved a better performance than that of the baseline methods (**Tables S1–S2**).

## Ablation Study and Feature Comparison

To evaluate the contribution of the semi-supervised objective function to the final prediction, we tested our EXP2SL model without the BPR loss. That is, we modified the objective function

**TABLE 2 |** Performance evaluation in three different cell lines under the "split pair" setting. The mean and standard deviation (in brackets) of metrics over 10 repeats of 5-fold cross-validations are shown. The best results for each cell line and each metric are marked in bold.

| Dataset | Model name | AUC | AUPR | F1 | Accuracy | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| A549 | LR | 0.863 (0.041) | 0.556 (0.089) | 0.577 (0.068) | 0.913 (0.030) | 0.622 (0.109) | 0.573 (0.033) | 0.952 (0.032) |
| | RF | 0.854 (0.039) | 0.552 (0.076) | 0.567 (0.069) | 0.912 (0.027) | 0.600 (0.104) | 0.559 (0.032) | 0.952 (0.026) |
| | SVM | 0.809 (0.038) | 0.505 (0.084) | 0.555 (0.060) | 0.914 (0.019) | 0.610 (0.104) | 0.523 (0.037) | **0.958 (0.019)** |
| | GBDT | 0.847 (0.039) | 0.520 (0.086) | 0.552 (0.065) | 0.908 (0.029) | 0.573 (0.120) | 0.552 (0.037) | 0.948 (0.033) |
| | NetLapRLS(L1000)[1] | 0.760 (0.044) | 0.344 (0.088) | 0.407 (0.068) | 0.845 (0.034) | 0.357 (0.119) | 0.512 (0.039) | 0.883 (0.038) |
| | NetLapRLS(PPI) [2] | 0.760 (0.045) | 0.344 (0.090) | 0.407 (0.079) | 0.845 (0.034) | 0.357 (0.130) | 0.512 (0.032) | 0.883 (0.037) |
| | NetLapRLS(combined) [3] | 0.827 (0.042) | 0.488 (0.091) | 0.519 (0.061) | 0.898 (0.025) | 0.523 (0.100) | 0.539 (0.017) | 0.938 (0.027) |
| | EXP2SL(no BPR loss) [4] | 0.866 (0.038) | **0.576 (0.086)** | **0.583 (0.071)** | **0.916 (0.032)** | **0.638 (0.135)** | 0.565 (0.036) | 0.955 (0.035) |
| | EXP2SL(PPI) [5] | 0.870 (0.041) | 0.574 (0.078) | 0.583 (0.055) | 0.915 (0.020) | 0.636 (0.081) | 0.573 (0.039) | 0.954 (0.020) |
| | EXP2SL | **0.871 (0.044)** | 0.573 (0.083) | 0.582 (0.070) | 0.914 (0.024) | 0.634 (0.084) | **0.579 (0.063)** | 0.952 (0.023) |
| A375 | LR | 0.994 (0.004) | 0.983 (0.006) | 0.981 (0.011) | 0.989 (0.007) | 0.967 (0.018) | 1.000 (0.015) | 0.984 (0.011) |
| | RF | 0.997 (0.004) | 0.990 (0.015) | 0.987 (0.016) | 0.993 (0.007) | 0.977 (0.028) | 1.000 (0.010) | 0.990 (0.010) |
| | SVM | 0.991 (0.004) | 0.978 (0.017) | 0.972 (0.020) | 0.984 (0.008) | 0.962 (0.033) | 0.991 (0.000) | 0.983 (0.009) |
| | GBDT | 0.999 (0.009) | 0.997 (0.013) | 0.993 (0.019) | 0.996 (0.013) | 0.993 (0.020) | 0.994 (0.022) | 0.997 (0.012) |
| | NetLapRLS(L1000) [1] | 0.989 (0.005) | 0.983 (0.006) | 0.969 (0.014) | 0.976 (0.013) | 0.956 (0.026) | 0.990 (0.012) | 0.966 (0.022) |
| | NetLapRLS(PPI) [2] | 0.990 (0.002) | 0.985 (0.003) | 0.972 (0.012) | 0.978 (0.010) | 0.956 (0.021) | 0.995 (0.000) | 0.966 (0.017) |
| | NetLapRLS(combined) [3] | 0.994 (0.007) | 0.990 (0.007) | 0.983 (0.016) | 0.987 (0.018) | 0.971 (0.026) | 1.000 (0.000) | 0.979 (0.033) |
| | EXP2SL(no BPR loss) [4] | 1.000 (0.003) | 1.000 (0.011) | 1.000 (0.013) | 1.000 (0.008) | 1.000 (0.023) | 1.000 (0.000) | 1.000 (0.012) |
| | EXP2SL(PPI)[5] | 1.000 (0.008) | 1.000 (0.010) | 1.000 (0.015) | 1.000 (0.014) | 1.000 (0.026) | 1.000 (0.000) | 1.000 (0.023) |
| | EXP2SL | **1.000 (0.012)** | **1.000 (0.029)** | **1.000 (0.026)** | **1.000 (0.016)** | **1.000 (0.043)** | **1.000 (0.000)** | **1.000 (0.021)** |
| HT29 | LR | 0.967 (0.015) | 0.861 (0.049) | 0.851 (0.032) | 0.958 (0.012) | 0.855 (0.053) | 0.895 (0.048) | 0.968 (0.017) |
| | RF | 0.955 (0.020) | 0.821 (0.067) | 0.824 (0.030) | 0.947 (0.005) | 0.792 (0.039) | 0.899 (0.073) | 0.955 (0.005) |
| | SVM | 0.949 (0.017) | 0.765 (0.079) | 0.808 (0.065) | 0.943 (0.015) | 0.744 (0.069) | **0.942 (0.100)** | 0.941 (0.018) |
| | GBDT | **0.973 (0.016)** | 0.880 (0.061) | 0.855 (0.029) | **0.960 (0.015)** | 0.861 (0.065) | 0.897 (0.040) | **0.969 (0.021)** |
| | NetLapRLS(L1000) [1] | 0.935 (0.017) | 0.738 (0.094) | 0.778 (0.064) | 0.941 (0.025) | 0.786 (0.139) | 0.836 (0.053) | 0.954 (0.034) |
| | NetLapRLS(PPI) [2] | 0.927 (0.024) | 0.729 (0.086) | 0.772 (0.053) | 0.939 (0.008) | 0.787 (0.048) | 0.822 (0.056) | 0.953 (0.009) |
| | NetLapRLS(combined) [3] | 0.939 (0.019) | 0.764 (0.094) | 0.784 (0.054) | 0.939 (0.020) | 0.778 (0.107) | 0.850 (0.035) | 0.949 (0.026) |
| | EXP2SL(no BPR loss[4] | 0.957 (0.026) | 0.834 (0.071) | 0.826 (0.043) | 0.943 (0.017) | 0.779 (0.088) | 0.926 (0.051) | 0.946 (0.023) |
| | EXP2SL(PPI)[5] | 0.967 (0.018) | 0.869 (0.033) | 0.851 (0.026) | 0.956 (0.011) | 0.838 (0.067) | 0.912 (0.084) | 0.962 (0.022) |
| | EXP2SL | 0.969 (0.008) | **0.880 (0.027)** | **0.866 (0.027)** | 0.959 (0.012) | **0.872 (0.055)** | 0.903 (0.049) | 0.968 (0.018) |

[1] *The NetLapRLS method using only the L1000 similarity.*

[2] *The NetLapRLS method using only the PPI similarity.*

[3] *The NetLapRLS method using the combination of L1000 and PPI similarities.*

[4] *The EXP2SL model without the BPR loss.*

[5] *The EXP2SL model with additional PPI information incorporated by a graph convolution module.*

in Equation 6 and used only the MSE loss and the L2 regularization term; our model can thus be trained in a supervised manner. An obvious decrease in performance under the "split gene" setting could be observed when we removed the BPR loss (see the "EXP2SL (no BPR loss)" row in **Table 3**). Therefore, the results demonstrated that the semi-supervised objective function had an important contribution to the prediction performance of our model.

One of the baseline models, NetLapRLS, can also incorporate different similarity matrices (*i.e.*, the L1000 profile similarities, the PPI similarities, and the combined similarities, as described in *NetLapRLS*), thus allowing the comparison between different settings using different input information. The NetLapRLS models with L1000 profile similarities and with PPI similarities as the input features achieved similar performance, and the combination of both features only led to a slight increase in performance in most cases. In general, the performance of NetLapRLS was worse than EXP2SL.

We also incorporated the PPI network into our EXP2SL framework (denoted as EXP2SL (PPI) in **Tables 2** and **3**) using a graph convolution network (Lei et al., 2017), as described in **Supporting Material** and **Figure S1**. In this case, no significant improvement in AUC and AUPR scores was observed after adding the PPI network information ($p$ values larger than 0.1 for all the cell lines in both conditions, Wilcoxon rank-sum test). These results indicate that using only the L1000 gene expression profiles is adequate to enable the models to capture useful features for accurately predicting SL interactions.

## Feature Importance Analysis

We used the scheme described in *Extraction of Feature Importance* to extract the important features based on the saliency map approach (Simonyan et al., 2013). Those features (*i.e.*, the corresponding expression levels of 978 genes) ranked among the top 50 (about 5% from the 978-dimensional features) were selected as the important features for each cell line. Among the selected feature sets, there is only one gene shared across all the three cell lines, that is, *AKT1*. AKT1 is known as a serine/threonine protein kinase, which regulates many viability related cellular processes, including proliferation, apoptosis, and cell survival (Chen et al., 2001; Lee et al., 2011). Most features were considered as the top 50 important features only in one cell line (47, 46, and 46 unique important features for A549, A375, and HT29, respectively), which suggests that the prediction may rely on the specific gene expression landscapes in different cell lines.

We also checked the over-representation of functional gene sets and pathways among the selected important features of the three

**TABLE 3 |** Performance evaluation in three different cell lines under the "split gene" setting. The mean and standard deviation (in brackets) of metrics over 10 repeats of 5-fold cross-validations are shown. The best results for each cell line and each metric are marked in bold.

| Dataset | Model name | AUC | AUPR | F1 | Accuracy | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| A549 | LR | 0.709 (0.039) | 0.328 (0.050) | 0.373 (0.039) | 0.816 (0.044) | 0.404 (0.070) | 0.435 (0.059) | 0.853 (0.058) |
| | RF | 0.715 (0.037) | 0.348 (0.052) | 0.379 (0.038) | 0.850 (0.024) | 0.461 (0.058) | 0.394 (0.038) | 0.896 (0.027) |
| | SVM | 0.708 (0.026) | 0.340 (0.051) | 0.380 (0.032) | 0.838 (0.020) | 0.433 (0.037) | 0.432 (0.060) | 0.876 (0.030) |
| | GBDT | 0.715 (0.030) | 0.333 (0.051) | 0.363 (0.032) | 0.841 (0.043) | 0.401 (0.094) | 0.399 (0.057) | 0.888 (0.054) |
| | NetLapRLS(L1000) [1] | 0.668 (0.024) | 0.252 (0.038) | 0.321 (0.021) | 0.815 (0.016) | 0.294 (0.057) | 0.407 (0.029) | 0.858 (0.018) |
| | NetLapRLS(PPI) [2] | 0.668 (0.030) | 0.252 (0.048) | 0.321 (0.041) | 0.815 (0.016) | 0.294 (0.070) | 0.407 (0.036) | 0.858 (0.019) |
| | NetLapRLS(combined) [3] | 0.685 (0.032) | 0.331 (0.043) | 0.371 (0.035) | 0.863 (0.021) | 0.426 (0.083) | 0.368 (0.046) | **0.918 (0.027)** |
| | EXP2SL(no BPR loss) [4] | 0.699 (0.032) | 0.358 (0.053) | 0.389 (0.035) | 0.857 (0.033) | 0.450 (0.083) | 0.401 (0.043) | 0.906 (0.042) |
| | EXP2SL(PPI) [5] | 0.755 (0.024) | 0.390 (0.044) | 0.419 (0.034) | 0.861 (0.041) | **0.465 (0.079)** | **0.450 (0.047)** | 0.903 (0.054) |
| | EXP2SL | **0.756 (0.030)** | **0.392 (0.043)** | **0.419 (0.024)** | **0.863 (0.048)** | 0.458 (0.073) | 0.448 (0.050) | 0.907 (0.061) |
| A375 | LR | 0.945 (0.026) | 0.884 (0.050) | 0.874 (0.046) | 0.930 (0.034) | 0.866 (0.054) | 0.897 (0.031) | 0.925 (0.033) |
| | RF | 0.947 (0.028) | 0.886 (0.045) | 0.891 (0.038) | 0.934 (0.032) | 0.865 (0.039) | 0.938 (0.025) | 0.917 (0.027) |
| | SVM | 0.924 (0.027) | 0.860 (0.047) | 0.873 (0.035) | 0.916 (0.026) | 0.864 (0.044) | 0.915 (0.032) | 0.905 (0.030) |
| | GBDT | 0.923 (0.019) | 0.852 (0.056) | 0.875 (0.048) | 0.920 (0.022) | 0.862 (0.047) | 0.926 (0.040) | 0.909 (0.047) |
| | NetLapRLS(L1000) [1] | 0.915 (0.050) | 0.822 (0.054) | 0.821 (0.085) | 0.895 (0.052) | 0.827 (0.020) | 0.889 (0.112) | 0.933 (0.069) |
| | NetLapRLS(PPI) [2] | 0.915 (0.033) | 0.823 (0.063) | 0.821 (0.046) | 0.895 (0.036) | 0.827 (0.047) | 0.889 (0.029) | 0.933 (0.025) |
| | NetLapRLS(combined) [3] | 0.921 (0.022) | 0.837 (0.054) | 0.840 (0.045) | 0.912 (0.030) | 0.858 (0.063) | 0.869 (0.024) | 0.955 (0.025) |
| | EXP2SL(no BPR loss) [4] | 0.952 (0.035) | 0.895 (0.052) | 0.905 (0.042) | 0.943 (0.031) | 0.873 (0.045) | **0.967 (0.032)** | 0.922 (0.033) |
| | EXP2SL(PPI) [5] | **0.976 (0.028)** | **0.936 (0.028)** | **0.932 (0.022)** | **0.966 (0.024)** | **0.919 (0.046)** | 0.959 (0.062) | **0.961 (0.055)** |
| | EXP2SL | 0.976 (0.023) | 0.935 (0.055) | 0.926 (0.046) | 0.964 (0.030) | 0.902 (0.045) | 0.965 (0.038) | 0.960 (0.025) |
| HT29 | LR | 0.754 (0.056) | 0.417 (0.075) | 0.531 (0.041) | 0.823 (0.050) | 0.505 (0.059) | 0.709 (0.048) | 0.841 (0.067) |
| | RF | 0.846 (0.030) | 0.494 (0.062) | 0.587 (0.037) | 0.858 (0.028) | 0.524 (0.057) | 0.763 (0.057) | 0.869 (0.026) |
| | SVM | 0.827 (0.034) | 0.465 (0.044) | 0.595 (0.043) | 0.857 (0.032) | 0.539 (0.066) | **0.792 (0.056)** | 0.863 (0.036) |
| | GBDT | 0.823 (0.057) | 0.452 (0.071) | 0.546 (0.044) | 0.822 (0.046) | 0.495 (0.055) | 0.758 (0.026) | 0.839 (0.057) |
| | NetLapRLS(L1000) [1] | 0.801 (0.043) | 0.441 (0.056) | 0.542 (0.042) | 0.826 (0.042) | 0.475 (0.079) | 0.755 (0.070) | 0.837 (0.055) |
| | NetLapRLS(PPI) [2] | 0.794 (0.026) | 0.423 (0.047) | 0.525 (0.030) | 0.818 (0.022) | 0.458 (0.069) | 0.761 (0.040) | 0.828 (0.034) |
| | NetLapRLS(combined) [3] | 0.814 (0.029) | 0.464 (0.081) | 0.550 (0.045) | 0.840 (0.043) | 0.479 (0.062) | 0.758 (0.073) | 0.853 (0.055) |
| | EXP2SL(no BPR loss) [4] | 0.788 (0.035) | 0.481 (0.040) | 0.577 (0.059) | 0.830 (0.037) | 0.531 (0.086) | 0.752 (0.040) | 0.835 (0.048) |
| | EXP2SL(PPI) [5] | 0.865 (0.032) | 0.553 (0.038) | 0.612 (0.024) | 0.872 (0.012) | 0.563 (0.049) | 0.766 (0.046) | 0.882 (0.018) |
| | EXP2SL | **0.866 (0.039)** | **0.558 (0.066)** | **0.620 (0.046)** | **0.877 (0.028)** | **0.577 (0.065)** | 0.756 (0.065) | **0.890 (0.035)** |

[1] *The NetLapRLS method using only the L1000 similarity.*

[2] *The NetLapRLS method using only the PPI similarity.*

[3] *The NetLapRLS method using the combination of L1000 and PPI similarities.*

[4] *The EXP2SL model without the BPR loss.*

[5] *The EXP2SL model with additional PPI information incorporated by a graph convolution module.*

cell lines using the WebGestalt server (Liao et al., 2019). The gene ontology (GO) related to biological processes was first used to examine the enriched functional annotations of the selected feature sets (**Tables S3–S5**). The enriched GO terms were ranked according to the false discovery rate (FDR) scores and *p* values. As a result, the top 10 enriched functional annotations for the selected features of HT29 contains the regulation of cell death, proliferation, and apoptosis (*p* values < $10^{-6}$ and FDRs < $10^{-3}$), which are cell viability related functions. Then, we also checked the over-representation of selected genes among the KEGG pathways using the WebGestalt server (Liao et al., 2019) (**Tables S6–S8**). Among the top 10 enriched pathways ranked according to the FDR scores and *p* values, we found multiple cancer-related pathways for cell line HT29 and also cell cycle or cancer-regulatory pathways for A375 and A549, *e.g.*, the *p53* and *ERBB* signaling pathways. All these results indicated that the selected features are probably related to the regulation of cell viability.

## CONCLUSION

In this paper, we proposed a semi-supervised neural network based method, EXP2SL, to accurately predict cell-line specific SL interactions. Our method exploits the L1000 expression profiles measured from the shRNA knockdown experiments performed in different cell lines to learn the cell-line specific SL interactions from the labeled data generated by CRISPR-Cas9 double-knockout based screens. In addition, a semi-supervised objective function is designed to make use of the large amount of unlabeled data. Tests on three datasets corresponding to three different cell lines showed that our model achieved better performance than the baseline models. At the same time, we verified that the L1000 gene expression profiles and the semi-supervised objective function are useful in SL prediction. Moreover, we analyzed the most important genes among the whole L1000 gene expression profiles, and found that the top attributing genes are related to the regulation of cell viability, which suggested that our model may pay more attention to such meaningful components of the whole gene expression profiles.

The major contributions of our work are the demonstration of L1000 expression profiles as effective features for SL prediction, and a novel semi-supervised neural network algorithm to accurately capture SL interactions. To our best knowledge, our model is the *first* computational approach for predicting cell-line specific synthetic lethal interactions, which may potentially benefit the target identification for specific tissue or cancer types. However, the application of our model may be limited in certain cancer types with high heterogeneity. Another limitation of our model is the dependence of the available L1000 gene expression profiles as input to EXP2SL. Although the L1000 expression profiles of more than 3,500 genes have been measured by shRNA knockdown experiments in the three cell lines analyzed in this work, there exist some cell lines with a paucity of data, which may thus limit the applications of our model on such cell lines.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the L1000 datasets GSE92742 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742) and the GEMINI datasets (Additional file 2 in https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1745-9#additional-information). Codes and processed data for this study can be found in https://github.com/FangpingWan/EXP2SL.

## AUTHOR CONTRIBUTIONS

JZ, DZ, and FW conceived the project. FW, SL, and TT designed the method. FW, SL, YL, and DZ performed the analyses. All the authors contributed to the writing of the manuscript.

## FUNDING

This work was supported in part by the National Natural Science Foundation of China [61872216, 81630103, 31900862]. The authors declare that this study received funding from the Turing AI Institute of Nanjing and the Zhongguancun Haihua Institute for Frontier Information Technology. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## ACKNOWLEDGMENTS

The authors are grateful to Ms. Jitong Cai for her helpful discussions about this work.

## SUPPLEMENTARY MATERIALS

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2020.00112/full#supplementary-material

## REFERENCES

Apaolaza, I., San José-Eneriz, E., Tobalina, L., Miranda, E., Garate, L., Agirre, X., et al. (2017). An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. *Nat. Commun.* 8, 459. doi: 10.1038/s41467-017-00555-y

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108. doi: 10.1038/nature08460

Boone, C., Bussey, H., and Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* 8, 437. doi: 10.1038/nrg2085

Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., et al. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase. *Nature* 434, 913. doi: 10.1038/nature03443

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238.*

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. KDD '16. (New York, NY, USA: Association for Computing Machinery), 785–794. ACM. doi: 10.1145/2939672.2939785

Chen, W. S., Xu, P.-Z., Gottlob, K., Chen, M.-L., Sokol, K., Shiyanova, T., et al. (2001). Growth retardation and increased apoptosis in mice with homozygous disruption of the akt1 gene. *Genes Dev.* 15, 2203–2208. doi: 10.1101/gad.913901

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353, aaf1420. doi: 10.1126/science.aaf1420

Das, S., Deng, X., Camphausen, K., and Shankavaram, U. (2018). DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers. *Bioinformatics* 35, 701–702. doi: 10.1093/bioinformatics/bty673

Deshpande, R., Asiedu, M. K., Klebig, M., Sutor, S., Kuzmin, E., Nelson, J., et al. (2013). A comparative genomic approach for identifying synthetic lethal interactions in human cancer. *Cancer Res.* 73, 6128–6136. doi: 10.1158/0008-5472.CAN-12-3956

Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N., Johnson, D. A., Richardson, T. B., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434, 917. doi: 10.1038/nature03445

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* 7. doi: 10.1038/msb.2011.35

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., et al. (2009). Inhibition of poly (ADP-ribose) polymerase in tumors from BRCA mutation carriers. *New Engl. J. Med.* 361, 123–134. doi: 10.1056/NEJMoa0900212

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., and Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.* 35, 463. doi: 10.1038/nbt.3834

Jacunski, A., Dixon, S. J., and Tatonetti, N. P. (2015). Connectivity homology enables inter-species network models of synthetic lethality. *PloS Comput. Biol.* 11, e1004506. doi: 10.1371/journal.pcbi.1004506

Jerby-Arnon, L., Pfetzer, N., Waldman, Y. Y., McGarry, L., James, D., Shanks, E., et al. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* 158, 1199–1209. doi: 10.1016/j.cell.2014.07.027

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Kranthi, T., Rao, S., and Manimaran, P. (2013). Identification of synthetic lethal pairs in biological systems through network information centrality. *Mol. Biosyst.* 9, 2163–2167. doi: 10.1039/c3mb25589a

Lee, M. W., Kim, D. S., Lee, J. H., Lee, B. S., Lee, S. H., Jung, H. L., et al. (2011). Roles of akt1 and akt2 in non-small cell lung cancer cell survival, growth, and migration. *Cancer Sci.* 102, 1822–1828. doi: 10.1111/j.1349-7006.2011.02025.x

Lee, J. S., Das, A., Jerby-Arnon, L., Arafeh, R., Auslander, N., Davidson, M., et al. (2018). Harnessing synthetic lethality to predict the response to cancer treatment. *Nat. Commun.* 9, 2546. doi: 10.1038/s41467-018-04647-1

Lei, T., Jin, W., Barzilay, R., and Jaakkola, T. (2017). "Deriving neural architectures from sequence and graph kernels," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (Sydney, NSW, Australia: JMLR.org), 2024–2033.

Li, B., Cao, W., Zhou, J., and Luo, F. (2011). Understanding and predicting synthetic lethal genetic interactions in saccharomyces cerevisiae using domain genetic interactions. *BMC Syst. Biol.* 5, 73. doi: 10.1186/1752-0509-5-73

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). Webgestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47 (W1), W199–W205. doi: 10.1093/nar/gkz401

Liu, Y., Wu, M., Liu, C., Li, X., and Zheng, J. (2019). SL2MF: Predicting synthetic lethality in human cancers via logistic matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2909908

Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., et al. (2018). Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nat. Biotechnol.* 36, 179. doi: 10.1038/nbt.4048

O'Neil, N. J., Bailey, M. L., and Hieter, P. (2017). Synthetic lethality and cancer. *Nat. Rev. Genet.* 18, 613. doi: 10.1038/nrg.2017.47

Pan, X., Yuan, D. S., Ooi, S.-L., Wang, X., Sookhai-Mahadeo, S., Meluh, P., et al. (2007). dslam analysis of genome-wide genetic interactions in saccharomyces cerevisiae. *Methods* 41, 206–221. doi: 10.1016/j.ymeth.2006.07.033

Pandey, G., Zhang, B., Chang, A. N., Myers, C. L., Zhu, J., Kumar, V., et al. (2010). An integrative multi-network and multi-classifier approach to predict genetic interactions. *PloS Comput. Biol.* 6, e1000928. doi: 10.1371/journal.pcbi.1000928

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). *Automatic differentiation in PyTorch*.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (Montreal, Quebec, Canada: AUAI Press), 452–461.

Ryan, C. J., Bajrami, I., and Lord, C. J. (2018). Synthetic lethality and cancer–penetrance as the major barrier. *Trends In Cancer* 4, 671–683. doi: 10.1016/j.trecan.2018.08.003

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., et al. (2017). Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* 14, 573. doi: 10.1038/nmeth.4225

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sinha, S., Thomas, D., Chan, S., Gao, Y., Brunen, D., Torabi, D., et al. (2017). Systematic discovery of mutation-specific synthetic lethals by mining pan-cancer human primary tumor data. *Nat. Commun.* 8, 15580. doi: 10.1038/ncomms15580

Srihari, S., Singla, J., Wong, L., and Ragan, M. A. (2015). Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biol. Direct* 10, 57. doi: 10.1186/s13062-015-0086-1

Srivas, R., Shen, J. P., Yang, C. C., Sun, S. M., Li, J., Gross, A. M., et al. (2016). A network of conserved synthetic lethal interactions for exploration of precision cancer therapy. *Mol. Cell* 63, 514–525. doi: 10.1016/j.molcel.2016.06.022

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452. doi: 10.1016/j.cell.2017.10.049

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368. doi: 10.1126/science.1065810

Whitehurst, A. W., Bodemann, B. O., Cardenas, J., Ferguson, D., Girard, L., Peyton, M., et al. (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature* 446, 815. doi: 10.1038/nature05697

Wong, S. L., Zhang, L. V., Tong, A. H., Li, Z., Goldberg, D. S., King, O. D., et al. (2004). Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci.* 101, 15682–15687. doi: 10.1073/pnas.0406614101

Wu, M., Li, X., Zhang, F., Li, X., Kwoh, C.-K., and Zheng, J. (2013). "Meta-analysis of genomic and proteomic features to predict synthetic lethality of yeast and human cancer," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics* (Washington, DC, USA: ACM), 384.

Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol. (BioMed Central).* 4, S6. doi: 10.1186/1752-0509-4-S2-S6

Zamanighomi, M., Jain, S. S., Ito, T., Pal, D., Daley, T. P., and Sellers, W. R. (2019). GEMINI: a variational bayesian approach to identify genetic interactions from combinatorial CRISPR screens. *Genome Biol.* 20, 137. doi: 10.1186/s13059-019-1745-9

Zhang, F., Wu, M., Li, X.-J., Li, X.-L., Kwoh, C. K., and Zheng, J. (2015). Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *J. Bioinf. Comput. Biol.* 13, 1541002. doi: 10.1142/S0219720015410024

Zhao, D., Badur, M. G., Luebeck, J., Magaña, J. H., Birmingham, A., Sasik, R., et al. (2018). Combinatorial CRISPR-Cas9 metabolic screens reveal critical redox control points dependent on the KEAP1-NRF2 regulatory axis. *Mol. Cell* 69, 699–708. doi: 10.1016/j.molcel.2018.01.017

**Conflict of Interest:** YL was employed by company Silexon AI Technology Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 17 April 2020
doi: 10.3389/fphar.2020.00269

# Molecular Generation for Desired Transcriptome Changes With Adversarial Autoencoders

*Rim Shayakhmetov[1†], Maksim Kuznetsov[1†], Alexander Zhebrak[1], Artur Kadurin[1], Sergey Nikolenko[1,2], Alexander Aliper[1] and Daniil Polykovskiy[1*]*

*[1] Insilico Medicine, Hong Kong, Hong Kong, [2] Neuromation OU, Tallinn, Estonia*

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Youjun Xu,*
*Peking University, China*
*Chittaranjan Tripathy,*
*Walmart Labs, United States*

****Correspondence:***
*Daniil Polykovskiy*
*daniil@insilico.com*

*[†]These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *11 November 2019*
***Accepted:*** *25 February 2020*
***Published:*** *17 April 2020*

***Citation:***
*Shayakhmetov R, Kuznetsov M, Zhebrak A, Kadurin A, Nikolenko S, Aliper A and Polykovskiy D (2020) Molecular Generation for Desired Transcriptome Changes With Adversarial Autoencoders. Front. Pharmacol. 11:269. doi: 10.3389/fphar.2020.00269*

Gene expression profiles are useful for assessing the efficacy and side effects of drugs. In this paper, we propose a new generative model that infers drug molecules that could induce a desired change in gene expression. Our model—the Bidirectional Adversarial Autoencoder—explicitly separates cellular processes captured in gene expression changes into two feature sets: those *related* and *unrelated* to the drug incubation. The model uses *related* features to produce a drug hypothesis. We have validated our model on the LINCS L1000 dataset by generating molecular structures in the SMILES format for the desired transcriptional response. In the experiments, we have shown that the proposed model can generate novel molecular structures that could induce a given gene expression change or predict a gene expression difference after incubation of a given molecular structure. The code of the model is available at https://github.com/insilicomedicine/BiAAE.

**Keywords: deep learning, generative models, adversarial autoencoders, conditional generation, representation learning, drug discovery, gene expression**

## INTRODUCTION

Following the recent advances in machine learning, deep generative models found many applications in biomedicine, including drug discovery, biomarker development, and drug repurposing (Mamoshina et al., 2016; Zhavoronkov, 2018). A promising approach to drug discovery is conditional generation, where a machine learning model learns a distribution $p(x \mid y)$ of molecular structures $x$ with given property $y$. Such models can generate molecules with a given synthetic accessibility, binding energy, or even activity against a given protein target (Kadurin et al., 2016; Polykovskiy et al., 2018a).

In this paper, we studied how conditional models scale to a more complex biological property; specifically, we have studied how drug incubation influences gene expression profiles. Using the LINCS L1000 (Duan et al., 2014) dataset, we build a joint model $p(x, y)$ on molecular structures $x$ and induced gene expression changes $y$.

In many conditional generation tasks, $x$ completely defines $y$. For example, molecular structure completely defines its synthetic accessibility score. For our task, however, some transcriptome changes are unrelated to the drug effect on cells, and we cannot infer them only from an incubated drug.

We propose a new model—the Bidirectional Adversarial Autoencoder—that learns a joint distribution $p(x, y)$ of objects and conditions. The model decomposes objects and their properties into three feature parts: shared features $s$ common to both $x$ and $y$; exclusive features $z_x$ relevant only to $x$ and not $y$; and exclusive features $z_y$ relevant only to $y$ and not $x$: $p(x, y) = p(s, z_x, z_y)$. For the transcriptomes and drugs, shared features $s$ may contain pharmacophore properties, target protein information, binding energy, and inhibition level; exclusive features $z_x$ may describe the remaining structural information; and $z_y$ may represent unrelated cellular processes. As features $s$ are common to both $x$ and $y$, the model can extract them from both $x$ and $y$.

The paper is organized into sections: *Related Work* surveys related work; *Models* presents the proposed Bidirectional Adversarial Autoencoder; *Experimental Evaluation* compares and validates the models on two datasets: the toy Noisy MNIST dataset of hand-written digits and LINCS L1000 dataset of small molecules with corresponding gene expression changes; and *Conclusion* concludes the paper.

# RELATED WORK

Conditional generative models generate objects $x$ from a conditional distribution $p(x \mid y)$, with $y$ usually being limited to class labels. The Adversarial Autoencoder (AAE) (Makhzani et al., 2015) consists of an autoencoder with a discriminator on the latent representation $z$ that tries to make the latent space distribution indistinguishable from a prior distribution $p(z)$; its conditional extension—Supervised AAE (Makhzani et al., 2015)—works well for simple conditions but can violate the conditions in other cases (Polykovskiy et al., 2018b). Conditional Generative Adversarial Networks (CGAN) (Mirza and Osindero, 2014) supplied the condition as an auxiliary input to both generator and discriminator. Perarnau et al. (2016) inverted CGANs, allowing us to edit images by changing the labels $y$. In FusedGAN (Bodla et al., 2018), a GAN generated a generic "structure prior" with no supervision, and a CGAN generated an object $x$ from condition $y$ and the latent representation learned by the unconditional GAN. Other papers explored applications of Conditional AAE models to the task of image modification (Antipov et al., 2017; Lample et al., 2017; Zhang et al., 2017).

CausalGAN (Kocaoglu et al., 2018) allowed components of the condition to have a dependency structure in the form of a causal model making conditions more complex. The Bayesian counterpart of AAE, the Variational Autoencoder (VAE) (Kingma and Welling, 2013), also had a conditional version (Sohn et al., 2015a), where conditions improved structured output prediction. CycleGAN (Zhu et al., 2017) examined a related task of object-to-object translation.

Multimodal learning models (Ngiam et al., 2011) and multi-view representation models (Wang et al., 2016a) explored translations between different modalities, such as image to text. Wang et al. (2016b) presented a VAE-based generative multi-view model. Our Bidirectional Adversarial Autoencoder provided explicit decoupling of latent representations and brought the multi-view approach into the AAE framework, where the basic Supervised AAE-like models (Makhzani et al., 2015) did not yield correct representations for sampling (Polykovskiy et al., 2018b).

Information decoupling ideas have been previously applied in other contexts: Yang et al. (2015) disentangled identity and pose factors of a 3D object; adversarial architecture from Mathieu et al. (2016) decoupled different factors in latent representations to transfer attributes between objects; Creswell et al. (2017) used VAE architecture with separate encoders for class label $y$ and latent representation $z$, forcing $z$ to exclude information about $y$; InfoVAE (Zhao et al., 2017) maximized mutual information between input and latent features; and Li et al. (2019) proposed a VAE modification that explicitly learns a "disentangled" representation $s$ to predict the class label and a "non-interpretable" representation $z$ that contains the rest of the information used for decoding.

InfoGAN (Chen X. et al., 2016) maximized mutual information between a subset of latent factors and the generator distribution. FusedGAN (Bodla et al., 2018) generated objects from two components, where only one component contains all object-relevant information. Hu et al. (2018) explicitly disentangles different factors in the latent representation and maps a part of the latent code to a particular external information.

## Conditional Generation for Biomedicine

Machine learning has numerous applications in biomedicine and drug discovery (Gawehn et al., 2016; Mamoshina et al., 2016; Ching et al., 2018). Deep neural networks demonstrated positive results in various tasks, such as prediction of biological age (Putin et al., 2016; Mamoshina et al., 2018a; Mamoshina et al., 2019), prediction of targets and side effects Aliper et al., 2017; Mamoshina et al., 2018b; West et al., 2018), and applications in medicinal chemistry (Lusci et al., 2013; Ma et al., 2015).

Alongside large-scale studies that measure cellular processes, deep learning applications explore transcriptomics (Aliper et al., 2016b; Chen Y. et al., 2016); these works study cellular processes and their change following molecular perturbations. Deep learning has also been applied to pathway analysis (Ozerov et al., 2016), the prediction of protein functions (Liu, 2017), the discovery of RNA binding proteins (Zheng et al., 2017), the discovery of binding patterns of transcription factors (Qin and Feng, 2017), medical diagnostics based on omics data (Chaudhary et al., 2017), and the analysis of DNA and RNA sequences (Budach and Marsico, 2018).

In drug discovery, apart from predicting pharmacological properties and learning useful representations of small molecules (Duvenaud et al., 2015; Aliper et al., 2016a; Kuzminykh et al., 2018), deep learning is being widely applied to the generation of molecules (Sanchez and Aspuru-Guzik, 2018). Multiple authors have published models that generate new molecules that are similar to the training data or molecules with predefined properties (Kadurin et al., 2017a; Kadurin et al., 2017b; Segler et al., 2017 Gómez-Bombarelli et al., 2018). AI-generated molecules have also been tested *in vitro* (Polykovskiy et al., 2018b). Reinforcement learning and generative models further enabled the generation of complex non-differentiable objectives, such as novelty (Guimaraes et al., 2017; Putin et al., 2018a; Putin et al., 2018b). Generative models aim to eliminate the bottleneck of traditional drug development pipelines by providing promising new lead molecules for a specific target and automating the initial proposal of lead molecules with desired properties. Recently, Zhavoronkov et al. (2019) developed a model GENTRL to discover potent inhibitors of discoidin domain receptor 1 (DDR1) in 21 days.

# MODELS

In this section, we introduce Unidirectional and a Bidirectional Adversarial Autoencoders and discuss their applications to conditional modeling. While we have focused on an example of molecular generation for transcriptome changes, in general, our model is not limited to these data types and can be used for generation in other domains.

## Supervised Adversarial Autoencoder

Our model for conditional generation is based on a Supervised Adversarial Autoencoder (Supervised AAE, SAAE) (Makhzani et al., 2015) shown in **Figure 1**. The Supervised AAE learns three neural networks—an encoder $E_x$, a generator (decoder) $G_x$, and a discriminator $D$. The encoder maps a molecule $x$ onto a latent representation $z = E_x(x)$, and a generator reconstructs the molecule back from $z$ and gene expression changes $y$: $G_x(z, y)$. We trained a discriminator $D$ to distinguish latent codes from samples of the prior distribution $p(z)$ and modified the encoder to make the discriminator believe that encoder's outputs are samples from the prior distribution:

$$\begin{aligned}\min_{E_x,G_x}\max_{D}\ &\lambda_1\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{y})}l^{x}_{\mathrm{rec}}(\boldsymbol{x},G_{\boldsymbol{x}}(E_{\boldsymbol{x}}(\boldsymbol{x}),\boldsymbol{y}))\\ &+\mathbb{E}_{z\sim p(z)}\log\ D(\boldsymbol{z})+\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{d}}(\boldsymbol{x})}\log\ (1-D(E_{\boldsymbol{x}}(\boldsymbol{x}))),\end{aligned}\tag{1}$$

where $l^x_{\mathrm{rec}}$ is a similarity measure between the original and reconstructed molecule, and $p_{\mathrm{d}}(x, y)$ is the data distribution. Hyperparameter $\lambda_1$ balances reconstruction and adversarial losses. We trained the model by alternately maximizing the loss in Equation 1 with respect to the parameters of $D$ and minimizing it with respect to the parameters of $E_x$ and $G_x$ (Goodfellow et al., 2014).

Besides passing gene expression changes $y$ directly to the generator, we could also train an autoencoder ($E_y$, $G_y$) on $y$ and pass its latent codes to the molecular decoder $G_x$ (**Figure 2**). We call this model a Latent Supervised Adversarial Autoencoder (Latent SAAE). Its optimization problem is:

$$\begin{aligned}\min_{E_x,E_y,G_x,G_y}\max_{D}\ &\lambda_1\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{y})}l^{x}_{\mathrm{rec}}\left(\boldsymbol{x},G_{\boldsymbol{x}}\left(E_{\boldsymbol{x}}(\boldsymbol{x}),E_{\boldsymbol{y}}(\boldsymbol{y})\right)\right)\\ &+\lambda_2\mathbb{E}_{\boldsymbol{y}\sim p_{\mathrm{d}}(\boldsymbol{y})}l^{y}_{\mathrm{rec}}\left(\boldsymbol{y},G_{\boldsymbol{y}}\left(E_{\boldsymbol{x}}(\boldsymbol{y})\right)\right)+\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\log\ D(\boldsymbol{z})\\ &+\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{d}}(\boldsymbol{x})}\log\ (1-D(E_{\boldsymbol{x}}(\boldsymbol{x})))\,.\end{aligned}\tag{2}$$

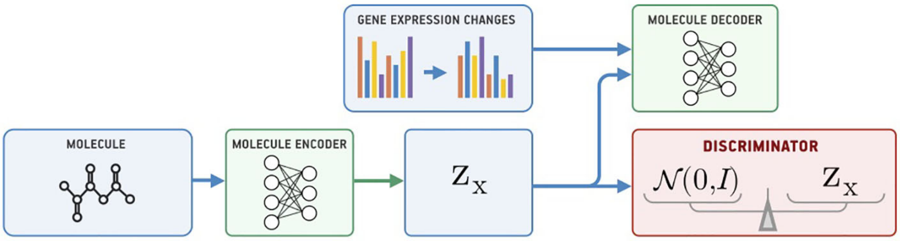


**FIGURE 1 |** The Supervised Adversarial Autoencoder model (SAAE).

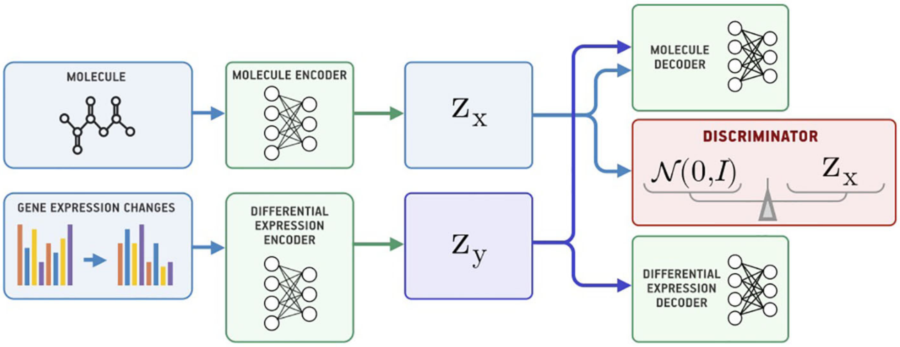


**FIGURE 2 |** The Latent Supervised Adversarial Autoencoder model (Latent SAAE).

Hyperparameters $\lambda_1$ and $\lambda_2$ balance object and condition reconstruction losses as well as the adversarial loss.

## Bidirectional Adversarial Autoencoder

Both SAAE and Latent SAAE models learn conditional distribution $p(x \mid y)$ of molecules for specific transcriptome changes. In this paper, we learned a joint distribution $p(x, y)$ instead. Our model is symmetric in that it can generate both $x$ for a given $y$ and $y$ for a given $x$. We assume that the data are generated with a graphical model shown in **Figure 3**. Latent variables $z_x$ and $z_y$ are exclusive parts that represent features specific only to molecules or transcriptome changes. Latent variable $s$ represents a shared part that describes features significant for both molecules and expression changes. To produce a new data point, we sampled exclusive ($z_x$, $z_y$) and shared ($s$) parts independently and used generative distributions $G_x$ ($x \mid s$, $z_x$) and $G_y$ ($y \mid s$, $z_y$) to produce $x$ and $y$.

To train a model, we used inference networks that predict values of $s$, $z_x$, and $z_y$: $E_x(z_x \mid x)$, $E_y(z_y \mid y)$, and $E(s \mid x, y) = E_x(s \mid x) = E_y(s \mid y)$. Note that we used two separate networks for inference of $s$ from one of $x$ and $y$ to perform conditional sampling (when only one of $x$ or $y$ is known). For example, to sample $p(x \mid y)$, we would do the following steps:

$$\boldsymbol{s} \sim E_y(\boldsymbol{s}|\boldsymbol{y}), \quad \boldsymbol{z_x} \sim p(\boldsymbol{z_x}), \quad \boldsymbol{x} \sim G_{\boldsymbol{x}}(\boldsymbol{s}, \boldsymbol{z_x}) \,. \tag{3}$$

For the molecule, $s$ may describe its pharmacophore—binding points that are recognized by macromolecules. For the gene expression, $s$ may describe affected proteins. Note that we can infer pharmacophore from a list of affected genes and vice versa. The exclusive part $z_x$ of a molecule describes the remaining structural parts besides the pharmacophore points. The exclusive part $z_y$ of a transcriptome describes cellular processes that influence the expression but are not caused by the drug.

**Figure 4** shows the proposed Bidirectional AAE architecture. We used two deterministic encoders $E_x$ and $E_y$ that infer latent codes from molecules and transcriptomes:

$$(\boldsymbol{z_x}, \boldsymbol{s_x}) = E_x(\boldsymbol{x}), \quad (\boldsymbol{z_y}, \boldsymbol{s_y}) = E_y(\boldsymbol{y}) \,. \tag{4}$$

Two deterministic decoders (generators) $G_x$ and $G_y$ reconstruct molecules $x$ and gene expression changes $y$ back from the latent codes:

$$\boldsymbol{x} = G_x(\boldsymbol{z_x}, \boldsymbol{s_x}), \quad \boldsymbol{y} = G_y(\boldsymbol{z_y}, \boldsymbol{s_y}) \tag{5}$$

The objective function consists of three parts, each capturing restrictions from the graphical model—the structure of the shared representation, reconstruction quality, and independence of shared and exclusive representations.

**Shared loss** ensures that shared representations extracted from the molecule $s_x$ and gene expression $s_y$ are close to each other, as suggested by the graphical model:

$$\min_{E_x, E_y} \mathcal{L}_{\text{shared}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y} \sim Pd(\boldsymbol{x},\boldsymbol{y})} \; \| \boldsymbol{s_x} - \boldsymbol{s_y} \|_2^2 \,. \tag{6}$$

**Reconstruction loss** ensures that decoders reconstruct molecules and gene expressions back from the latent codes produced by the encoders. We also use a cross-reconstruction loss, where molecular decoder $E_x$ uses shared part $s_y$ from a gene expression encoder $E_y$ for reconstruction and vice versa:

$$\min_{E_x, E_y, G_x} \mathcal{L}^x_{\text{rec}} = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{d}}(\boldsymbol{x})} l^{\boldsymbol{x}}_{\text{rec}}(\boldsymbol{x}, G_{\boldsymbol{x}}(\boldsymbol{z_x}, \boldsymbol{s_x})) + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y} \sim p_{\text{d}}(\boldsymbol{x},\boldsymbol{y})} l^{\boldsymbol{x}}_{\text{rec}}\left(\boldsymbol{x}, G_{\boldsymbol{x}}\left(\boldsymbol{z_x}, \boldsymbol{s_y}\right)\right) \tag{7}$$

$$\min_{E_x, E_y, G_y} \mathcal{L}^y_{\text{rec}} = \mathbb{E}_{\boldsymbol{y} \sim p_{\text{d}}(\boldsymbol{y})} l^{\boldsymbol{y}}_{\text{rec}}\left(\boldsymbol{y}, G_{\boldsymbol{x}}\left(\boldsymbol{z_y}, \boldsymbol{s_y}\right)\right) + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y} \sim p_{\text{d}}(\boldsymbol{x},\boldsymbol{y})} l^{\boldsymbol{y}}_{\text{rec}}\left(\boldsymbol{y}, G_{\boldsymbol{y}}\left(\boldsymbol{z_x}, \boldsymbol{s_y}\right)\right) \tag{8}$$

where $l^{\boldsymbol{x}}_{\text{rec}}$ and $l^{\boldsymbol{y}}_{\text{rec}}$ are some distance measures in the molecules and gene expression space.

**Discriminator loss** is an objective that encourages distributions $p(s)$, $p(z_x)$, and $p(z_y)$ to be independent, which means that shared and exclusive parts must learn different features. This restriction comes from a graphical model. It also encourages $p(s)$, $p(z_x)$, and $p(z_y)$ to be standard Gaussian distributions $N(0, I)$ to perform a sampling scheme from Equation 3. We optimized the discriminator in an adversarial manner (Goodfellow et al., 2014) similar to SAAE:

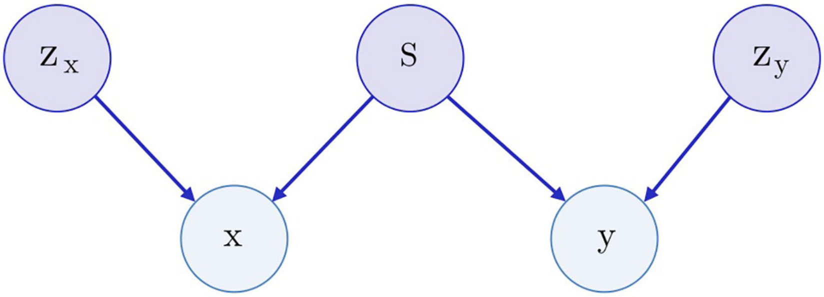


**FIGURE 3** | The underlying graphical model of the data: molecules $x$, gene expression changes $y$, three latent variables correspond to the exclusive ($z_x$, $z_y$) and shared ($s$) features between $x$ and $y$.

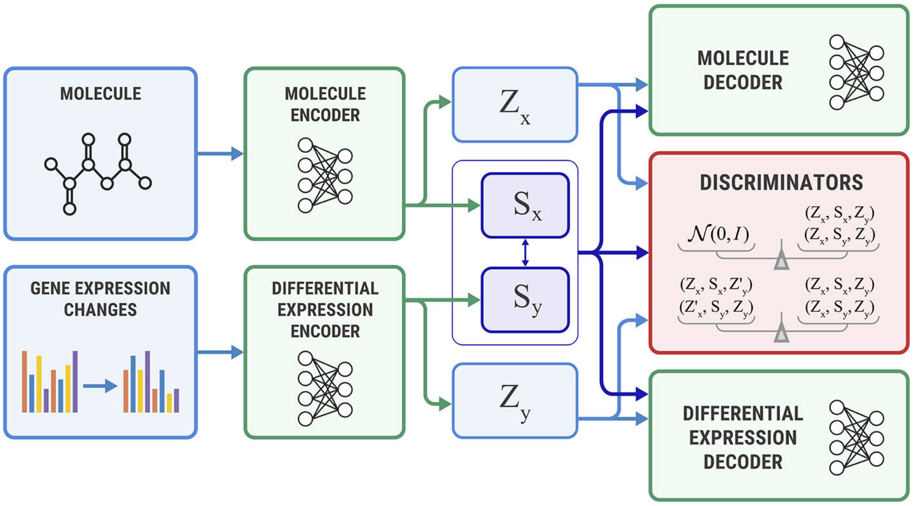


**FIGURE 4 |** The Bidirectional Adversarial Autoencoders model. The discriminators ensure that three latent code components are independent and indistinguishable from the prior distribution.

$$\min_{E_x,E_y,G_x,G_y} \max_{D} \mathcal{L}_{\text{adv}} = \mathbb{E}_{s',z'_x,z'_y \sim p(s)p(z_x)p(z_y)} \log D(z'_x, s', z'_y)$$
$$+ \tfrac{1}{2}\mathbb{E}_{x,y\sim p_{\text{d}}(x,y)} \log\left(1 - D(z_x, s_x, z_y)\right) \tag{9}$$
$$+ \tfrac{1}{2}\mathbb{E}_{x,y\sim p_{\text{d}}(x,y)} \log\left(1 - D(z_x, s_y, z_y)\right)$$

Note that since the target distribution for adversarial training is factorized, we expected that the trained model would learn independence of $s$, $z_x$, and $z_y$.

**Additional discriminator losses** We also added additional discrimination objective to explicitly encourage independence of $z_x$ from $(s_y, z_y)$ and $z_y$ from $(s_x, z_x)$:

$$\min_{E_x,E_y,G_x,G_y} \max_{D} \mathcal{L}_{\text{info}} = \mathbb{E}_{x,y\sim p_{\text{d}}(x,y)} \mathbb{E}_{y'\sim p_{\text{d}}(y)}$$
$$\left[\log D(z_x, s_x, z_y) + \log\left(1 - D(z_x, s_x, z'_y)\right)\right]$$
$$+\mathbb{E}_{x,y\sim p_{\text{d}}(x,y)} \mathbb{E}_{x'\sim p_{\text{d}}(x)} \left[\log D(z_x, s_y, z_y) + \log\left(1 - D(z'_x, s_y, z_y)\right)\right], \tag{10}$$

where $z'_x$ is an exclusive latent code of $x'$, and $z'_y$ is an exclusive latent code of $y'$. In practice, we obtain $z_x'$ and $z_y'$ by shuffling $z_x$ and $z_y$ in each batch.

Combining these objectives, the final optimization problem becomes a minimax problem that can be solved by alternating gradient descent with respect to encoder and decoder parameters, and gradient ascent with respect to the discriminator parameters:

$$\min_{E_x,E_y,G_x,G_y} \max_{D} \lambda_1 \mathcal{L}_{\text{shared}} + \lambda_2 \mathcal{L}^{x}_{\text{rec}} + \lambda_3 \mathcal{L}^{y}_{\text{rec}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{info}} \,. \tag{11}$$

The hyperparameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ balance different objectives. In general, we optimize lambdas based on the performance of BiAAE on the holdout set in terms of the target metrics, such as estimated negative conditional log-likelihood. In practice, we found that optimal values of lambdas yielded the gradients of loss components on a similar scale.

## Unidirectional Adversarial Autoencoder

The Bidirectional AAE can generate molecules that cause given transcriptome changes and transcriptome changes caused by a given molecule. However, if we only need conditional generation of molecules $p(x \mid y)$, we simplify the model by removing the encoder of $s_x$. The encoder $E_x$ returns only an exclusive part: $z_x = E_x(x)$. For this model, we derived the objective from Equation 11 by setting $s_x$ equal to $s_y$ (**Figure 5**).

# EXPERIMENTAL EVALUATION

In this section, we have described the experimental setup and presented numerical results on the toy Noisy MNIST dataset and a LINCS L1000 dataset (Duan et al., 2014) of gene expression data.

## Noisy MNIST

We start by validating our models on the Noisy MNIST (Wang et al., 2015) dataset of image pairs $(x, y)$, for which we know the correct features in the shared representation $s$. The image $x$ is a handwritten digit randomly rotated by an angle in $[-\pi/4, \pi/4]$. The image $y$ is also a randomly rotated version of another image containing the same digit as $x$ but with strong additive Gaussian noise. As a result, the only common feature between $x$ and $y$ is the digit. Bidirectional and Unidirectional AAEs should learn to store only the information about the digit in $s$.

The train-validation-test splits contain 50,000, 10,000, and 10,000 samples respectively. We set the batch size to 128 and the learning rate to 0.0003, and we used the Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.9$ for models with

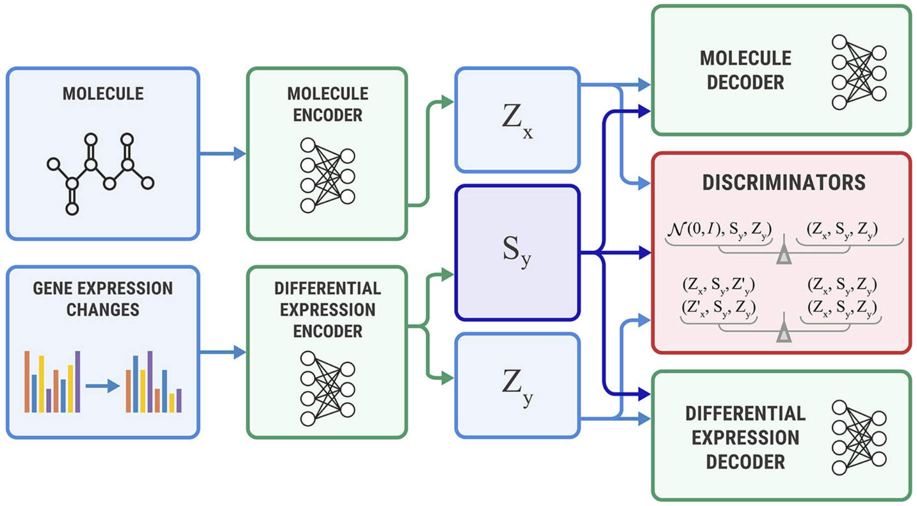


**FIGURE 5 |** The Unidirectional Adversarial Autoencoder: a simplified version of a Bidirectional Adversarial Autoencoder for generating from $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{y})$. The discriminator part ensures that the three latent code components are independent, and the object's exclusive latent code is indistinguishable from the prior distribution.

adversarial training and $\beta_1 = 0.99$ and $\beta_1 = 0.999$ for others with a single update of autoencoders per a single update of the discriminator. Encoder and decoder architectures were the same for all models, with 12-dimensional $z_x$, $z_y$ and 4-dimensional $s$. The encoder had 2 convolutional layers with a number of channels $1 \rightarrow 32 \rightarrow 16$ with 2D dropout rate 0.2 followed by three fully-connected layers of size $64 \rightarrow 128 \rightarrow 128 \rightarrow 16$ with batch normalization. The decoder consisted of 2 fully connected layers followed by 3 transposed convolution layers; the discriminators have two hidden layers with $1024 \rightarrow 512$ units. We set the weights for $\mathcal{L}_{\text{rec}}$ to 10 and 0.1 for $\mathcal{L}_{\text{shared}}$. Other $\lambda$ were set to 1. For Unidirectional AAE, we increased weight for $\mathcal{L}_{\text{info}}$ to 100. For baseline models we used similar architectures. Please refer to the **Supplementary Material** for additional hyperparameters.

Conditional generative model $p(x \mid y)$ should produce images with the same digit as image $\boldsymbol{y}$, which we evaluate by training a separate convolutional neural network to predict the digit from $x$ and comparing the most probable digit to the actual digit of $y$ known from the dataset. We also estimated a conditional mutual information $\mathcal{MI}(\boldsymbol{x},\boldsymbol{s_y}|\boldsymbol{y})$ using a Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) algorithm for BiAAE, UniAAE, JMVAE, and VCCA models. For SAAE, LatentSAAE, CVAE, and VIB we estimated $\mathcal{MI}(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{y})$ since these models do not separate embeddings into shared and exclusive parts explicitly. Models with high mutual information extract relevant information from $y$. A neural network for MINE consisted of a convolutional encoder for $x$ and fully-connected encoder for $s_y$. We then passed a concatenated embedding through a fully-connected neural network to get a final estimate of mutual information. Results in **Table 1** suggest that the BiAAE model extracted relevant mutual information which, besides all, contained information about the digit of $y$. In **Figure 6**, we show example samples from the model.

## Differential Gene Expression

In this section, we have validated Bidirectional AAE on a gene expression profiles dataset with 978 genes. We use a dataset of transcriptomes from the Library of Integrated Network-based Cellular Signatures (LINCS) L1000 project (Duan et al., 2014). The database contains measurements of gene expressions before and after cells react with a molecule at a given concentration.

For each cell line, the training set contains experiments characterized by the control ($ge_b \in \mathbb{R}^{978}$) and perturbation-induced ($ge_a \in \mathbb{R}^{978}$) gene expression profiles. We represented molecular structures in the SMILES format (Weininger, 1988; Weininger et al., 1989). We augmented the dataset by randomly matching control and perturbation-induced measurements from the same plate.

We preprocessed the training dataset by removing molecules with a molecular weight less than 250 and more than 550 Da. We then removed molecules that did not contain any oxygen or nitrogen atoms or contained atoms besides C, N, S, O, F, Cl, Br, and H. Finally, we removed molecules that contained rings with

**TABLE 1 |** Quantitative results for a Noisy MNIST experiment. Conditional Generation section evaluates how often the model produced a correct digit. Latent Codes section estimates the Mutual Information between $z_x$ and $s$ ($y$ for SAAE).

| Model | Accuracy, % | MI($x$,$s_y$\|$y$) | MI($x$,$s$\|$y$) |
|---|---|---|---|
| SAAE (Makhzani et al., 2015) | 43.68 | – | 1.665 |
| Latent SAAE | 34.76 | – | **1.681** |
| CVAE (Sohn et al., 2015b) | 0.4583 | – | 0.3074 |
| JMVAE (Suzuki et al., 2017) | 5.38 | 0.9515 | – |
| VIB (Alemi et al., 2017) | 43.6 | – | 1.121 |
| VCCA (Wang et al., 2016b) | 23.35 | 1.239 | – |
| BiAAE (our) | **49.21** | 1.432 | – |
| UniAAE (our) | 47.61 | **1.627** | – |

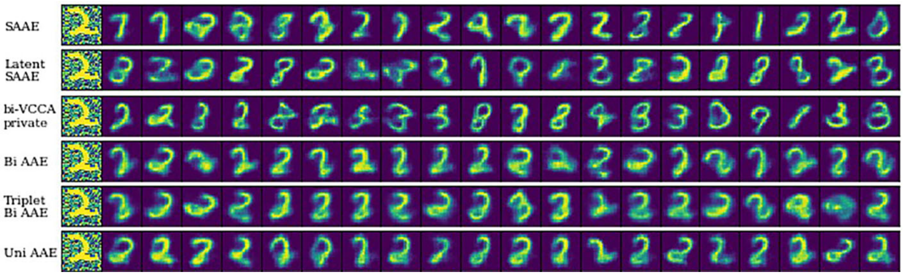


**FIGURE 6 |** Qualitative results on a Noisy MNIST dataset. The figure shows generated images *x* for a noisy image *y* (left column) as a condition. Generated images must have the same digit as *y*.

more than eight atoms or tetracyclines. The resulting dataset contained 5,216 unique SMILES. Since the dataset is small, we pretrained an autoencoder on the MOSES (Polykovskiy et al., 2018a) dataset and used its encoder and decoder as initial weights in all models.

For all baseline models on differential gene expressions, we used similar hyperparameters shown in **Table 2** (please refer to the **Supplementary Material** for the exact hyperparameters). In all experiments, we split our dataset into train, validation, and test sets, all containing different drugs. To construct a training example, we sampled a drug-dose pair, a perturbation for this drug and dose, and a control expression from the same plate as the perturbed expression.

We used a two-step encoder for $y = (\eta, \Delta\text{ge})$ shown in **Figure 7**, where $\Delta\text{ge}=\text{ge}_a-\text{ge}_b$. We first embedded $\Delta$ge with a fully-connected neural network, and then concatenated the obtained representation with a logarithm of concentration $\eta$. We passed the resulting vector through a final encoder. The decoder has a symmetric architecture.

## Generating Molecular Structures for Gene Expression Profiles

The proposed BiAAE model can generate molecules for given gene expression changes and vice versa. We started by experimenting with the molecular generation (**Table 3**). In the experiment, we reported a negative log-probability of generating the exact incubated drug $x$ given the dose and gene expression change averaged over tokens $log\ p(\boldsymbol{x}|\Delta\text{ge},\eta)$. We also estimated a Mutual Information $\mathcal{MI}(\boldsymbol{x}, s_y|\Delta\text{ge},\eta)$ similar to the MNIST experiment described above. For each $\eta$ and $\Delta$ge, we generated a set of molecules $G$ and estimated a fraction of valid molecules and internal diversity of $G$:

$$\text{IntDiv}(G) = 1 - \frac{1}{|G|(|G|-1)} \sum_{\substack{m_1, m_2 \in G \\ m_1 \neq m_2}} T(m_1, m_2), \qquad (12)$$

where $T$ is a Tanimoto similarity on Morgan fingerprints. This metric shows whether a model can produce multiple candidates for a given gene expression or collapses to a single molecule.

**TABLE 2 |** Hyperparameters for neural networks training on gene expression data. All neural networks are fully connected, and decoders have an architecture symmetric to the encoders.

| Hyperparameter | Value |
|---|---|
| Molecular Encoder | GRU; hidden size 128; 2 layers |
| Expression Encoder | IN(978)→256→OUT(128) |
| Difference Encoder | IN(129)→128→OUT(10 + 10) |
| Discriminator | IN→1024→512→OUT(1) |
| Batch Normalization | After each linear layer in encoders |
| Activation Function | LeakyReLU |
| Learning Rate | 0.0003 |

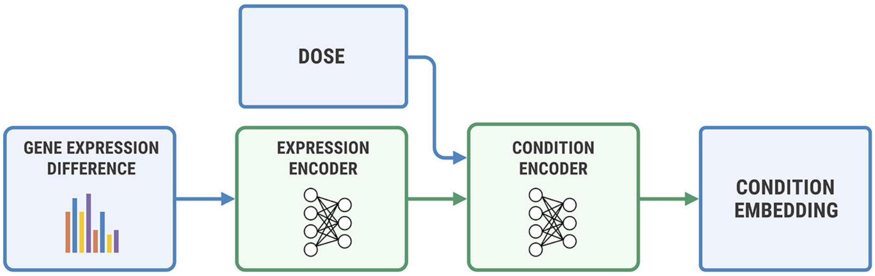


**FIGURE 7 |** The architecture of the condition encoder for changes in the transcriptome. The input to the expression encoder is the difference between the control and perturbed expressions. We passed the dose to the last layers of the encoder.

The proposed BiAAE and UniAAE architectures show the ability to capture the dependencies in the training set and generalize to new objects from the validation set. The BiAAE model provides better mutual information while preserving valid diverse molecules.

## Comparing Generated Molecular Structures to Known Active Molecules

In this experiment, we show that the proposed generative model (BiAAE) can produce biologically meaningful results. We used a manually curated database of bioactive molecules ChEMBL 24.1 (Gaulton et al., 2016) and additional profiles of gene expression knockdown from LINCS L1000 (Duan et al., 2014).

The first experiment evaluates molecular generation given a transcriptome change of a small molecule inhibitor of a specific protein. The ChEMBL dataset has experimental data on molecules that inhibit a certain human protein. We chose template molecules that are present in both LINCS molecule perturbation dataset and ChEMBL dataset. We used molecules that had inhibition concentration less than 10 $\mu$M IC50 for only one protein.

The condition for molecular generation is a transcriptome change and a dose of a template molecule. Specifically, the condition is a shared part $s_y$ of the gene expression and dose embedding. The model is expected to generate molecules that are similar to known drugs. In **Figure 8**, for several protein targets, we show a known inhibitor and generated molecules that could induce similar transcriptome profile changes.

The second experiment evaluates molecular generation given a transcriptome change of a specific gene knockdown. The LINCS dataset contains gene knockdown transcriptomes that the model was not trained on. For each gene knockdown, we found a corresponding human protein in the ChEMBL dataset. We chose template molecules that had a proven IC50 less than 10$\mu$M for only one protein. The condition for molecular generation is a transcriptome change of a gene knockdown and the most common dose 10 $\mu$M in LINCS. The model is expected to generate molecules that produce the same transcriptome change of gene knockdowns.

The condition is different compared to the previous experiment in a way that the gene knockdown expression profile is not induced by a small molecule but rather shows the desired behavior of the potential drug. In **Figure 9**, we show generated molecules and compare them to known inhibitors of a protein corresponding to a knocked down gene. We expect these molecules to produce similar effects in gene expression to gene knockdown.

**TABLE 3 |** Validation results of conditional generation $p(x|\Delta \text{ge}, \eta)$.

| Model | NLL | MI<br>$(x,s_y|\Delta\text{ge},\eta)$ | MI<br>$(x,s|\Delta\text{ge},\eta)$ | Internal Diversity | Validity |
|---|---|---|---|---|---|
| SAAE | 0.55 | – | **0.11** | **0.85** | 0.64 |
| Latent SAAE | 0.55 | – | 0.00 | **0.85** | 0.62 |
| CVAE | 1.22 | – | 0.00 | 0.84 | 0.58 |
| JMVAE | 1.42 | 0.00 | – | 0.61 | **0.82** |
| VIB | 1.46 | – | 0.00 | 0.17 | 0.29 |
| VCCA | 1.36 | 0.00 | – | 0.53 | 0.71 |
| BiAAE | 0.77 | **0.32** | – | **0.85** | 0.76 |
| UniAAE | **0.53** | 0.00 | – | **0.85** | 0.61 |

## Predicting Gene Expression Profiles for an Incubated Drug

We experimented with predicting gene expression changes after drug incubation (**Table 4**). First, we report estimated mutual information $\mathcal{MI}(\Delta\text{ge},\eta,\boldsymbol{s}_x|\boldsymbol{x})$ similar to the previous experiments. We also report the $R^2$ metric, which measures the determination coefficient between the real and predicted ($\Delta$ge, $\eta$) for a given molecule. Finally, we report a top-1 precision metric, which shows the fraction of samples for which the largest absolute change in real and predicted $\Delta$ge matched.

To compute $R^2$ and top-1 precision, we only used drugs that were administered at $\eta = 10\ \mu$M concentration. Since we are only interested in a certain concentration, we discarded generated ($\Delta$ge, $\eta$) tuples if $\eta$ was far from 10 $\mu$M (outside the range [−6.5, −5.5] in $\log_{10}$ scale). Note that VIB was not able to generate any gene expression changes near 10 $\mu$M.

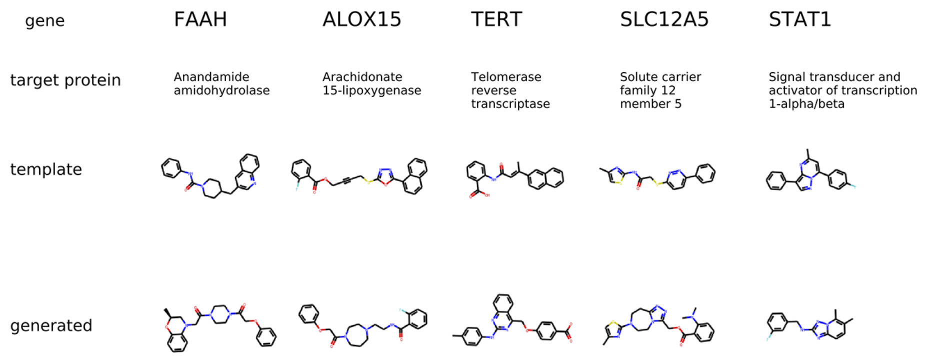

gene
FAAH
ALOX15
TERT
SLC12A5
STAT1
target protein
Anandamide amidohydrolase
Arachidonate 15-lipoxygenase
Telomerase reverse transcriptase
Solute carrier family 12 member 5
Signal transducer and activator of transcription 1-alpha/beta
template
generated

**FIGURE 8 |** The examples of generated molecules conditioned on gene expression changes from a protein inhibitor; Real most similar inhibitors from ChEMBL are provided for comparison.

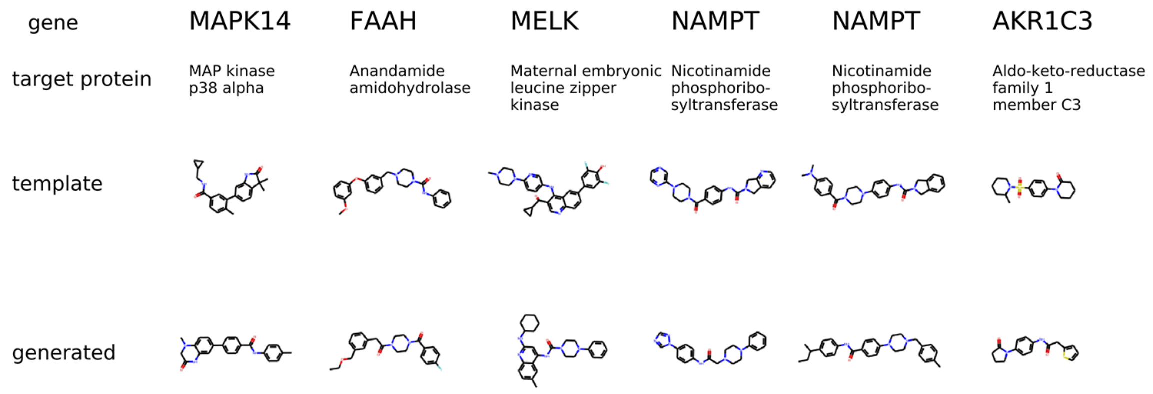

gene
MAPK14
FAAH
MELK
NAMPT
NAMPT
AKR1C3
target protein
MAP kinase p38 alpha
Anandamide amidohydrolase
Maternal embryonic leucine zipper kinase
Nicotinamide phosphoribo-syltransferase
Nicotinamide phosphoribo-syltransferase
Aldo-keto-reductase family 1 member C3
template
generated

**FIGURE 9 |** The examples of generated molecules conditioned on gene expression changes from a gene knockdown; Real most similar inhibitors of a knocked down gene are provided for comparison.

**TABLE 4 |** Validation results of conditional generation $\boldsymbol{p(\Delta ge,\eta|x)}$.

| Model | MI($\Delta$ge,$\eta$,$s_y$\|$x$) | MI($\Delta$ge,$\eta$,$s$\|$x$) | Top-1 precision | $R^2$ score |
|---|---|---|---|---|
| SAAE | — | 0.00 | 0.58 | 0.26 |
| Latent SAAE | — | **0.23** | 0.74 | 0.28 |
| CVAE | — | 0.01 | 0.29 | **0.33** |
| JMVAE | 0.00 | — | 0.0 | 0.03 |
| VIB | — | 0.00 | — | — |
| VCCA | 0.00 | — | 0.0 | 0.03 |
| BiAAE | 0.20 | — | 0.74 | 0.32 |
| UniAAE | **0.21** | — | **0.77** | 0.27 |

The experiment demonstrates that proposed UniAAE, BiAAE, and LatentSAAE models generalize well the symmetric task and show good metrics on predicting gene expression changes.

## DISCUSSION

The key advantage of the proposed model compared to the previous works is the joint adversarial learning of latent representations of paired objects. This representation improves conditional generation metrics and shows promising results in molecular generation for desired transcriptome changes.

Three discriminator neural networks ensure that the latent representations divided into shared and exclusive parts are more meaningful and useful for the conditional generation. Two additional discriminator losses help the model learn a more expressive shared part and make sure that all three parts are mutually independent.

However, adversarial training slightly complicates the training procedure for the BiAAE model. In comparison with other baseline models, the training loss contains more terms, each with a coefficient to tune. In general, we tune these coefficients using grid search, and we select the best coefficients according to the generative metrics on the validation set. In practice, we simplify the grid search and use the same coefficient for the adversarial terms $\lambda_1=\lambda_4=\lambda_5$ since the corresponding losses have values on the same scale. We choose the search space for coefficients $\lambda_2,\lambda_3$ in a way that the second and third terms provide the gradient in the same scale as the other terms.

Another problem that arises when we use the adversarial approach is the instability of training. The instability is the consequence of the minimax nature of adversarial training (Goodfellow et al., 2014). To overcome the instability, we use approaches described in (Bang and Shim, 2018), i.e., we use shallow discriminators and Adam optimizer with parameters $\beta_1=0.5,\beta_2=0.9$.

## CONCLUSION

In this work, we proposed a Bidirectional Adversarial Autoencoder model for the generation of molecular structures for given gene expression changes. Our AAE-based architecture extracts shared information between molecule and gene expression changes and separates it from the remaining exclusive information. We showed that our model outperforms baseline conditional generative models on the Noisy MNIST dataset and the generation of molecular structures for the desired transcriptome changes.

## DATA AVAILABILITY STATEMENT

The code and datasets for this study are available at https://github.com/insilicomedicine/BiAAE.

## AUTHOR CONTRIBUTIONS

RS and MK implemented the BiAAE and baseline models and conducted the experiments. RS, AK, and AA prepared the datasets. RS, MK, AK, and DP derived the BiAAE and

UniAAE models. RS, AZ, AK, SN, and DP wrote the manuscript. AK and DP supervised the project.

## ACKNOWLEDGMENTS

The original idea for molecular generation for a specific transcriptional or proteomic profile, a technology used broadly at Insilico Medicine, was proposed in 2016 by Dr. Alex Zhavoronkov, who is the co-author of the patent covering this technology.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2020.00269/full#supplementary-material.

## REFERENCES

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2017). Deep Variational Information Bottleneck. *Int. Conf. Learn. Representations*.

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016a). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol. Pharm.* 13, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Aliper, A. M., Plis, S. M., Artemov, A. V., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016b). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13 7, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Aliper, A., Jellen, L., Cortese, F., Artemov, A., Karpinsky-Semper, D., Moskalev, A., et al. (2017). Towards Natural Mimetics of Metformin and Rapamycin. *Aging (Albany NY)* 9, 2245–2268. doi: 10.18632/aging.101319

Antipov, G., Baccouche, M., and Dugelay, J. (2017). Face aging with conditional generative adversarial networks. In. *2017 IEEE Int. Conf. Image Process. (ICIP).*, 2089–2093. doi: 10.1109/ICIP.2017.8296650

Bang, D., and Shim, H. (2018). "Improved training of generative adversarial networks using representative features," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. Eds. J. Dy and A. Krause (Stockholmsmässan, Stockholm Sweden: PMLR), 433–442.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Hjelm, D., et al. (2018). "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, JMLR.org vol. 80. Eds. J. Dy and A. Krause(Stockholmsmässan, Stockholm Sweden: PMLR)), 531–540.

Bodla, N., Hua, G., and Chellappa, R. (2018). "Semi-supervised FusedGAN for conditional image generation," in *Proceedings of the European Conference on Computer Vision* (ECCV), Springer, Cham 669–683.

Budach, S., and Marsico, A. (2018). pysster: Learning Sequence and Structure Motifs in DNA and RNA Sequences using Convolutional Neural Networks. *bioRxiv*. 34, 3035–3037 doi: 10.1093/bioinformatics/bty222

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2017). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* 24, 1248–1259. doi: 10.1158/1078-0432.ccr-17-0853

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 29 . Eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (Curran Associates, Inc. in Red Hook, NY) 2172–2180.

Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics* 32, 1832–1839. doi: 10.1093/bioinformatics/btw074

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and Obstacles for Deep Learning in Biology and Medicine. *J. R Soc. Interface* 15, 141, 1–47. doi: 10.1098/rsif.2017.0387

Creswell, A., Bharath, A. A., and Sengupta, B. (2017). Conditional autoencoders with adversarial information factorization. *CoRR*. abs/1711.05175.

Duan, Q., Flynn, C., Niepel, M., Hafner, M., Muhlich, J. L., Fernandez, N. F., et al. (2014). LINCS canvas browser: Interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* 42, W449–W460. doi: 10.1093/nar/gku476

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, vol. 28. Eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (Curran Associates, Inc. in Red Hook, NY) 2224–2232.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2016). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi: 10.1093/nar/gkw1074

Gawehn, E., Hiss, J. A., and Schneider, G. (2016). Deep learning in drug discovery. *Mol. Inf.* 35, 3–14. doi: 10.1002/minf.201501008

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc. vol. 27. , 2672–2680.

Guimaraes, G. L., Sanchez-Lengeling, B., Farias, P. L. C., and Aspuru-Guzik, A. (2017). Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *CoRR*. abs/1705.10843.

Hu, Q., SzabÃ, A., Portenier, T., Favaro, P., and Zwicker, M. (2018). "Disentangling factors of variation by mixing them," in *The IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2797. (CVPR).

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2016). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8, 10883–10890. doi: 10.18632/oncotarget.14073

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2017a). The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* 8, 10883–10890. doi: 10.18632/oncotarget.14073

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017b). druGAN: An advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* 14, 3098–3104. doi: 10.1021/acs.molpharmaceut.7b00346

Kingma, D. P., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Representations*.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*. abs/1312.6114.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2018). CausalGAN: Learning causal implicit generative models with adversarial training. *Int. Conf. Learn. Representations*.

Kuzminykh, D., Polykovskiy, D., Kadurin, A., Zhebrak, A., Baskov, I., Nikolenko, S., et al. (2018). 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharm.* 15, 4378–4385. doi: 10.1021/acs.molpharmaceut.7b01134

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. A. (2017). "Fader networks: Manipulating images by sliding attributes," in *Advances in Neural Information Processing Systems 30*. Eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, Inc.), 5967–5976.

Li, Y., Pan, Q., Wang, S., Peng, H., Yang, T., and Cambria, E. (2019). Disentangled variational auto-encoder for semi-supervised learning. *Inf. Sci.* 482, 73–85. doi: 10.1016/j.ins.2018.12.057

Liu, X. L. (2017). Deep recurrent neural network for protein function prediction from sequence. *bioRxiv*. doi: 10.1101/103994

Lusci, A., Pollastri, G., and Baldi, P. (2013). Deep Architectures and Deep Learning in Chemoinformatics: the Prediction of Aqueous Solubility for Drug-like Molecules. *J. Chem. Inf. Model* 53, 1563–1575. doi: 10.1021/ci400187y

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* 55, 263–274. doi: 10.1021/ci500747n. PMID: 25635324.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. J. (2015). Adversarial Autoencoders. *CoRR.* abs/1511.05644. 73, 1482–1490.

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Mol. Pharma.* 13, 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982

Mamoshina, P., Kochetov, K., Putin, E., Cortese, F., Aliper, A., Lee, W.-S., et al. (2018a). Population specific biomarkers of human aging: A big data study using south korean, canadian, and eastern european patient populations. *Journals Gerontology: Ser. A.* 73, 1482–1490. doi: 10.1093/gerona/gly005

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., et al. (2018b). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* 9, 242. doi: 10.3389/fgene.2018.00242

Mamoshina, P., Kochetov, K., Cortese, F., Kovalchuk, A., Aliper, A., Putin, E., et al. (2019). Blood biochemistry analysis to detect smoking status and quantify accelerated aging in smokers. *Sci. Rep.* 9, 142. doi: 10.1038/s41598-018-35704-w

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. (2016). "Disentangling factors of variation in deep representation using adversarial training," in *Advances in Neural Information Processing Systems*, vol. 29. Eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (Curran Associates, Inc. in Red Hook, NY) 5040–5048.

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *CoRR.* abs/1411.1784.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, vol. 11. (Omnipress, ICML: 2600 Anderson St, Madison, WI, United States), 689–696.

Ozerov, I. V., Lezhnina, K. V., Izumchenko, E., Artemov, A. V., Medintsev, S., Vanhaelen, Q., et al. (2016). In Silico Pathway Activation Network Decomposition Analysis (iPANDA) as a Method for Biomarker Development. *Nat. Commun.* 7, 13427. doi: 10.1038/ncomms13427

Perarnau, G., van de Weijer, J., Raducanu, B., and Álvarez, J. M. (2016). Invertible conditional gans for image editing. *Neural Inf. Process. Syst. Workshop Adversarial Training.*

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al. (2018a). Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv preprint arXiv:1811.12823.*

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Bozdaganyan, M., et al. (2018b). Entangled conditional adversarial autoencoder for de-novo drug discovery. *Mol. Pharm.* 15, 4398–4405. doi: 10.1021/acs.molpharmaceut.8b00839

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., et al. (2016). Deep Biomarkers of Human Aging: Application of Deep Neural Networks to Biomarker Development. *Aging (Albany NY)* 8, 1021–1033. doi: 10.18632/aging.100968

Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sanchez-Lengeling, B., Aspuru-Guzik, A., et al. (2018a). Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model* 58, 1194–1204. doi: 10.1021/acs.jcim.7b00690

Putin, E., Asadulaev, A., Vanhaelen, Q., Ivanenkov, Y., Aladinskaya, A. V., Aliper, A., et al. (2018b). Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol. Pharm.* 15, 4386–4397. doi: 10.1021/acs.molpharmaceut.7b01137

Qin, Q., and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PloS Comput. Biol.* 13, e1005403+. doi: 10.1371/journal.pcbi.1005403

Sanchez, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science* 361, 360–365. doi: 10.1126/science.aat2663

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2017). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi: 10.1021/acscentsci.7b00512

Sohn, K., Lee, H., and Yan, X. (2015a). "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems* , vol. 28 . Eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (Curran Associates, Inc. in Red Hook, NY) 3483–3491.

Sohn, K., Lee, H., and Yan, X. (2015b). "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, vol. 28 . Eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (Curran Associates, Inc.), 3483–3491.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2017). "Joint Multimodal Learning with Deep Generative Models," in *International Conference on Learning Representations Workshop*.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015). "On deep multi-view representation learning," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. (JMLR.org), ICML'15. 1083–1092.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. A. (2016a). On deep multi-view representation learning: Objectives and optimization. *CoRR.* abs/1602.01024.

Wang, W., Lee, H., and Livescu, K. (2016b). Deep variational canonical correlation analysis. *CoRR.* abs/1610.03454.

Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* 29, 97–101. doi: 10.1021/ci00062a008

Weininger, D. (1988). SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36. doi: 10.1021/ci00057a005

West, M. D., Labat, I., Sternberg, H., Larocca, D., Nasonkin, I., Chapman, K. B., et al. (2018). Use of Deep Neural Network Ensembles to Identify Embryonic-fetal Transition Markers: Repression of COX7A1 in Embryonic and Cancer Cells. *Oncotarget* 9, 7796–7811. doi: 10.18632/oncotarget.23748

Yang, J., Reed, S., Yang, M.-H., and Lee, H. (2015). "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1. (Cambridge, MA, USA: MIT Press), 1099–1107. NIPS'15.

Zhang, Z., Song, Y., and Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. *IEEE Conf. Comput. Vision Pattern Recogn. (CVPR).* 4352–4360. doi: 10.1109/CVPR.2017.463

Zhao, S., Song, J., and Ermon, S. (2017). InfoVAE: Information maximizing variational autoencoders. *CoRR.* abs/1706.02262. doi: 10.1609/aaai.v33i01.33015885

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nat. Biotechnol.*, 1–4. doi: 10.1038/s41587-019-0224-x

Zhavoronkov, A. (2018). Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Mol. Pharm.* 15, 4311–4313. doi: 10.1021/acs.molpharmaceut.8b00930

Zheng, J., Zhang, X., Zhao, X., Tong, X., Hong, X., Xie, J., et al. (2017). Deep-RBPPred: Predicting RNA binding proteins in the proteome scale based on deep learning. *bioRxiv.* 8, 15264 doi: 10.1101/210153

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image \ translation using cycle-consistent adversarial networks. *2017 IEEE Int. Conf. Comput. Vision (ICCV)*, 2242–2251. doi: 10.1109/ICCV.2017.244

**Conflict of Interest:** RS, MK, AZ, AK, AA, and DP work for Insilico Medicine, a commercial artificial intelligence company. SN works for Neuromation OU, a company engaged in AI development through synthetic data and generative models.

ORIGINAL RESEARCH
published: 21 May 2020
doi: 10.3389/fphar.2020.00639

# Dual Transcriptomic and Molecular Machine Learning Predicts all Major Clinical Forms of Drug Cardiotoxicity

*Polina Mamoshina*[1,2*], *Alfonso Bueno-Orovio*[1] *and Blanca Rodriguez*[1]

*[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom, [2] Insilico Medicine Hong Kong Ltd, Hong Kong, Hong Kong*

Computational methods can increase productivity of drug discovery pipelines, through overcoming challenges such as cardiotoxicity identification. We demonstrate prediction and preservation of cardiotoxic relationships for six drug-induced cardiotoxicity types using a machine learning approach on a large collected and curated dataset of transcriptional and molecular profiles (1,131 drugs, 35% with known cardiotoxicities, and 9,933 samples). The algorithm generality is demonstrated through validation in an independent drug dataset, in addition to cross-validation. The best prediction attains an average accuracy of 79% in area under the curve (AUC) for safe versus risky drugs, across all six cardiotoxicity types on validation and 66% on the unseen set of drugs. Individual cardiotoxicities for specific drug types are also predicted with high accuracy, including cardiac disorder signs and symptoms for a previously unseen set of anti-inflammatory agents (AUC = 80%) and heart failures for an unseen set of anti-neoplastic agents (AUC = 76%). Besides, independent testing on transcriptional data from the Drug Toxicity Signature Generation Center (DToxS) produces similar results in terms of accuracy and shows an average AUC of 72% for previously seen drugs and 60% for unseen respectively. Given the ubiquitous manifestation of multiple drug adverse effects in every human organ, the methodology is expected to be applicable to additional tissue-specific side effects beyond cardiotoxicity.

**Keywords: machine learning, cardiotoxic adverse effect, safety pharmacology, bioinformatics and computational biology, *in silico* analysis**

***Edited by:***
*Jianfeng Pei,*
*Peking University, China*

***Reviewed by:***
*Zhi-Liang Ji,*
*Xiamen University, China*
*Cao Dongsheng,*
*Central South University, China*
*Feng Zhu,*
*Zhejiang University, China*

****Correspondence:***
*Polina Mamoshina*
*polina.mamoshina@cs.ox.ac.uk*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *17 October 2019*
***Accepted:*** *21 April 2020*
***Published:*** *21 May 2020*

***Citation:***
*Mamoshina P, Bueno-Orovio A and Rodriguez B (2020) Dual Transcriptomic and Molecular Machine Learning Predicts all Major Clinical Forms of Drug Cardiotoxicity. Front. Pharmacol. 11:639. doi: 10.3389/fphar.2020.00639*

## INTRODUCTION

Drug cardiotoxicity significantly limits the application of numerous therapies, and also slows down the drug research and development process (Cook et al., 2014; Onakpoya et al., 2016). As the attrition rate due to cardiotoxicity remains high, the need and importance of novel approaches capable of efficient safety testing has been widely emphasized, but not solved (Cook et al., 2014; Waring et al., 2015). Human-based approaches exploiting *in silico* methods have been postulated as the most promising alternative to costly animal experiments (Lawrence et al., 2008; Vicente et al., 2018), which frequently exhibit limited translation ability to human (Mak et al., 2014; Rodriguez et al., 2016).

In this regard, great progress has been made for example to evaluate the ability of *in silico* models to assess and predict the clinical risk of drug-induced arrhythmias (Lancaster and Sobie, 2016; Passini et al., 2017; Dutta et al., 2017). However, less attention has been paid to the prediction of other forms of drug-induced cardiotoxicity, such as cardiomyopathies, heart failure, myocardial ischemia or myocarditis (Mladěnka et al., 2018). Novel approaches are therefore needed to account for the wider spectrum of possible cardiovascular drug side effects beyond those mainly linked to adverse electrophysiological interactions.

Machine learning methods are gaining recognition in biological data analysis (Mamoshina et al., 2016; Lin and Lane, 2017; Lo et al., 2018). However, less than a handful of studies have addressed drug cardiotoxicity prediction beyond drug-induced arrhythmias. Huang and colleagues used protein-protein interactions to predict general cardiotoxicity for 578 drugs, using a support vector machine method (Huang et al., 2011). Using transcriptional profiles and fingerprints of 251 drugs, Wang et al. focused on prediction of gastrointestinal, liver and kidney toxicities, and myocardial infraction, a single form of cardiotoxicity, using an extra trees algorithm for multi-label classification (Wang et al., 2016). Messinis et al. (2018) developed a transcriptomic-based predictor of drug-induced cardiomyopathy with 31 drugs. Importantly, although all these studies reported relatively good accuracies (0.68 (Huang et al., 2011), 0.80 (Wang et al., 2016) and 1 (Messinis et al., 2018), respectively) under different cross-validation strategies (random split of samples or leave-one-drug out), none of them conducted an independent validation on drugs previously unseen by the trained model. This is crucial to ensure the translatability of proposed approaches to real-world applications, and thus an important limitation of previous work. In addition, none of the previous algorithms was developed to predict all major forms of drug cardiotoxicity.

Our hypothesis is that molecular and structural properties of drugs combined with their associated transcriptional changes in gene expression represent a suitable strategy to characterize their cardiac safety. The goal of this work is therefore to tackle four main challenges in cardiotoxicity prediction, namely prediction of six cardiotoxicity types, addressing the class imbalance problem, robust validation with independent datasets, and combination of transcriptional data and molecular descriptors. Our aim is to develop an independently-validated supervised machine-learning-based approach for the simultaneous prediction of all major forms of drug cardiotoxicity in human, using a substantial dataset of transcriptional and molecular descriptors compiled from diverse publicly-available data repositories. Our proposed approach specifically accounts for independent validation, the challenges of severe safety class imbalance and the preservation of relationships between different drug-induced cardiotoxicities. Addressing class imbalance is crucial as available datasets are usually heavily unbalanced (i.e., unequal distribution between drug types and/or cardiotoxic classes) (Banerjee et al., 2018), especially in large datasets curated automatically. This limits the generalization ability of data-driven methods, and in particular of unsupervised ones. In this work, we overcome these challenges by the application of supervised classifiers, as they generally demonstrate higher predictive performance on unbalanced biological data (Miller et al., 2008). Importantly, we demonstrate prediction of all six main forms of cardiotoxicity related to drug action in human. Through this work, we therefore significantly increase the domain of applicability and translation capabilities of machine-learning for cardiotoxicity prediction in preclinical drug evaluation. The findings and methodologies are expected to be generalizable to other organ-specific side effects.

# MATERIALS AND METHODS

## Data Preparation

The first step was to curate a database of cardiotoxic and matching safe drugs (**Figure 1B** and **Table 1**), using diverse publicly-available knowledge and data repositories, including DrugBank (Wishart et al., 2008) (www.drugbank.com), Connectivity map Project (https://clue.io/cmap) (Subramanian et al., 2017), SIDER(Kuhn et al., 2016) (sideeffects.embl.de), MedDRA (https://bioportal.bioontology.org/ontologies/MEDDRA) and MESH (https://www.ncbi.nlm.nih.gov/mesh).

Names and IDs were retrieved from the MedDRA dictionary for all cardiac (MedDRA ID 10007541) and vascular (MedDRA ID 10047065) disorders. Interactions of chemicals in the form of STITCH compound identifiers and their MedDRA terms of side effects were downloaded from the publicly available SIDER database. We used PubChem Compound Identifiers (CIDs) to match this list to the information on drug targets, drug status ('approved', 'investigational', etc.), and drug SMILES notations obtained from the Drugbank database. For drug target information, only experimentally verified interactions (such as inhibition, activation and intercalation between drugs and proteins or other molecules, like DNA) as provided by Drugbank were considered for this work. Compounds linked to the MedDRA term 'cardiac disorders' were labeled as drugs with cardiotoxicity reports and were considered as 'positive cases' in further model contraction. Compounds with same targets and no record of cardiac disorders in the database were considered as safe and as 'negative cases' in further model contraction. Safe compounds were additionally filtered by their status, and only approved compounds currently on the market were used for analysis, resulting in 26 (out of 759) drugs being removed from the analysis. This was performed to prevent possible unreported toxicities in drugs considered safe but withdrawn from the market. The unsafe group includes both marketed and withdrawn drugs, including drugs withdrawn due to cardiac side effects. Information about therapeutic classes of drugs were collected from the MESH medical vocabulary.

In this work, we used the Connectivity map project as a source of gene expression cell responses to drugs, or drug transcriptional profiles. These were measured using an L1000 high-throughput profiling method. The L1000 fluorescent assay allows the detection and quantification of the expression of up to 978 landmark and 80 control transcripts simultaneously in each

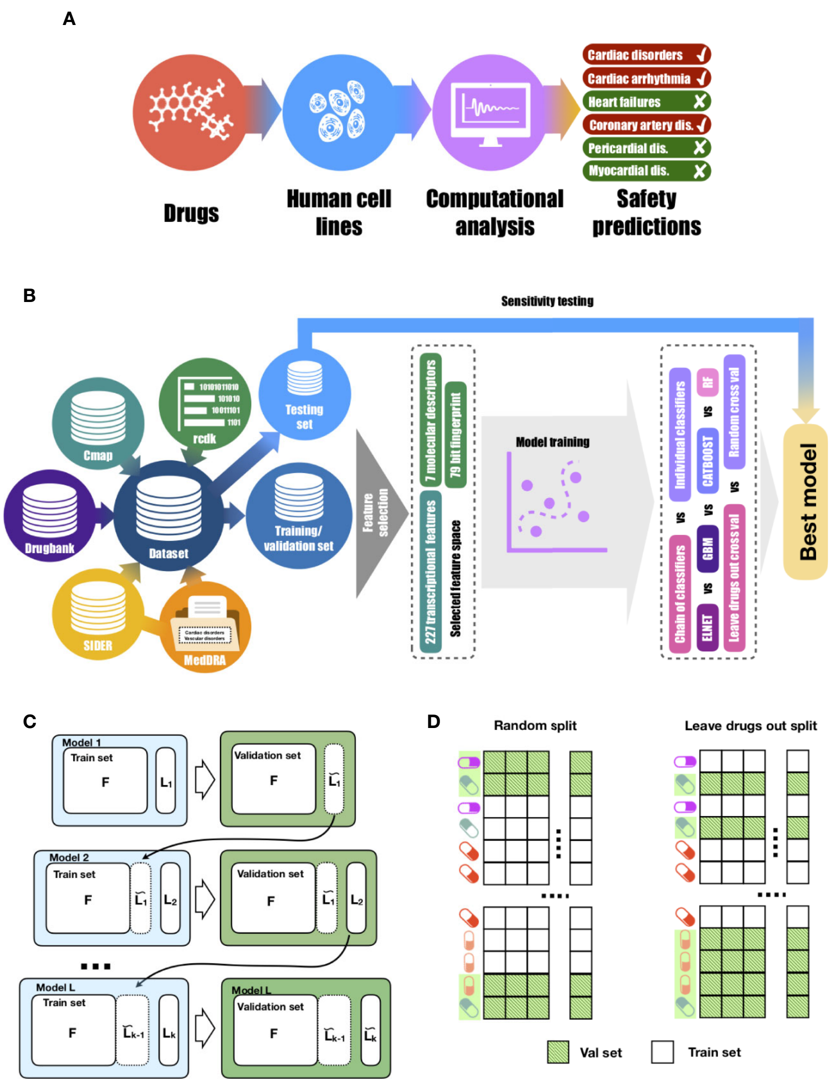


**FIGURE 1 |** Proposed approach for drug-induced cardiotoxicty prediction. **(A)** *In silico* prediction of multiple forms of drug-induced cardiac adverse reactions. **(B)** The design of the proposed machine learning model builds on a comprehensive database of drugs linked to their transcriptional profiles, molecular descriptors, fingerprints and safety information. Our database was collated using several publicly-available knowledge and data repositories, as detailed in **Table 1**. The rcdk package was used to calculate the set of molecular descriptors and fingerprints. After collation, the whole dataset was split into training and testing sets of unique sets of drugs in each. Using training data, the set of the most predictive features was selected. Those features were later used to train multi-label models (chain of classifiers with nested stacking vs sets of individual classifier). Training was performed with two cross-validation strategies: random vs leave-drug-out. **(C)** Proposed chain of classifiers with nested stacking model for multi-label prediction of drug cardiac safety. Each classifier *L* in the chain is trained on a set of *F* features and the *L-1* labels predicted by the previous *L-1* classifiers. **(D)** Considered cross-validation approaches. Random cross-validation results in validation and testing sets having profiles of drugs from training sets, which leads to inflated performance on validation and low generalisation abilities on unseen test data compared to leave-drug-out cross-validation.

**TABLE 1 |** Summary of databases and knowledge portals used in the study.

| Name | Type of data | Link | Reference |
|---|---|---|---|
| DrugBank | Drug structure, list of targets | www.drugbank.com | (Wishart et al., 2008) |
| MedDRA | Side effect hierarchy | https://bioportal.bioontology.org/ontologies/MEDDRA | |
| SIDER | Drug safety information | sideeffects.embl.de | (Kuhn et al., 2016) |
| MESH | Pharmacological classes information | https://www.ncbi.nlm.nih.gov/mesh | |
| Connectivity map Project | Transcriptional profiles | https://clue.io/cmap | (Subramanian et al., 2017) |
| DToxS | Transcriptional profiles | https://martip03.u.hpc.mssm.edu/index.php | |

well of 384-well plate, where each well can contain a separate drug profile. This massive scale expression data is available in multiple levels starting from raw fluorescent intensity values from each well to replicate collapsed scores for drugs. In this work, we used well-established 'core' human cell lines with drug transcriptional profiles available. We explored the provided transcriptional profiles (normalized across each scan plate) of the 977 'landmark genes' (Level 3a—NORM, as described in detail in https://clue.io/connectopedia/data_levels), for six cell lines (A549 and MCF7 for training and validation and PHH, SKB, SKM1, A673 for sensitivity testing), two incubation times (6 and 12 h), and multiple drug concentrations. To link the drug transcriptional profiles provided by the Connectivity map project to the information about their side effects, targets and status, we utilized the Chemical Translation Service (http://cts.fiehnlab.ucdavis.edu/) to match PubChem CIDs to their corresponding Broad IDs. In total, we collected a database of 1,131 drugs (fully analyzed for relationships across cardiotoxicity types), 357 of which had transcriptional profiles (used for prediction), with a total of 9,933 samples. Samples refer to independent transcriptional profiles of a drug at a given cell line, incubation time or concentration.

## Calculation of Molecular Descriptors and Fingerprints

We used the 'rcdk' package (Guha et al., ) to calculate seven molecular descriptors widely used in drug property prediction (Dong et al., 2015; Zhang et al., 2016): molecular weight (MW), partition coefficient (XLogP), atomic polarisabilities (apol), topological polar surface area (TopoPSA), polar surface area expressed as a ratio to molecular size (tpsaEfficiency), Ghose-Crippen LogKow (ALogP) and molar refractivity (AMR). We also calculated the commonly used 79-bit 'estate' fingerprint. These are widely used descriptors in drug property prediction and have been shown to characterize drug properties, including safety (Dong et al., 2015; Zhang et al., 2016).

## Selection of Transcriptional Features

In order to evaluate the predictive power of individual transcriptional profiles to cardiac safety prediction, following (Cai et al., 2018) two selection methods (correlation-based, wrapped-based) were considered. Correlation-based methods aim to identify transcriptional features (genes) highly correlated with each cardiotoxicity form. Wrapped-based methods use predictive models to score all combinations of feature subsets for each form of cardiotoxicity. As a correlation-based method, the 'select.cfs' function from the Biocomb R package (Novoselova et al., 2018) was used, while the Boruta algorithm implemented in the Boruta R package (Lagani et al., 2017) was used as a wrapper-based method. This way, for each cardiotoxicity form, we identify two subsets of genes.

Cohen's Kappa scores (Cohen, 1960), calculated by the 'Kappa.test' function from the fmsb R package, were used (i) to estimate the accuracy of classifiers, and (ii) to evaluate the similarities between vectors. The evaluation of similarities between vectors was applied to binary vectors of cardiotoxicity types, and between transcriptional features vectors selected using either correlation-based or wrapper-based methods, as described above. The Kappa scores are given by:

$$\boldsymbol{\kappa} = 1 - \frac{1 - \mathbf{p_o}}{1 - \mathbf{p_e}} \tag{1}$$

where $p_o$ is the relative observed agreement between two binary vectors, and $p_e$ is the expected agreement between predicted and actual values. Values smaller than 0 demonstrate poor agreement and values from 0.81 to 1 correspond to almost perfect agreement. To analyze whether the chosen genes were associated with the same or different biological functions, we also intersected the lists of determined genes with the Reactome database of pathways (Fabregat et al., 2018).

## Training, Validation and Testing Set Design

Models were trained on the expression values of relevant genes, seven molecular descriptors values and 79 fingerprint values, calculated as detailed above (340 features in total). We randomly split the entire drug dataset by protein targets (information obtained from the Drugbank) into unique training (291 drugs, 8,237 samples) and testing (66 drugs, 1,696 samples) sets. By such design, both datasets only overlap by a set of protein targets, but drugs on the testing phase are completely unseen during training, therefore facilitating preclinical translation to novel chemicals. This strategy was also enforced during model development, where for each cardiotoxicity label its respective training and validation sets were preserved completely non-overlapping by drugs with leave-drug-out cross-validation strategy (**Figure 1D**). We collapsed samples, so each drug profile referred to gene expression values for each individual drug with one cell line, incubation time and concentration. Models were trained on matrices size of 340 × 1,154 and tested on 340 × 746.

To benchmark the performance of models and select the best set of parameters, we used leave-drug-out cross-validation in contrast to random cross-validation (**Figure 1C**). This cross-validation strategy is crucial to accurately assess the performance of the model on unseen drugs, and therefore evaluate its translational potential into real-world practice. We performed the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002), implemented in the DMwR R package (Torgo, 2013), on the training set in cross-validation to avoid overfitting at any stage. To determine the generalisation ability of methods for novel drugs, we assessed the best-performing models on the selected testing set of 66 unique drugs.

## Chain of Classifiers With Nested Stacking

To predict cardiovascular safety we employed a chain classifier with nested stacking, which takes into account label dependencies (**Figure 1C**). The chain of classifiers with nested stacking (Senge et al., 2019) is a model that receives a feature vector and maps it to a set of labels. Each classifier in the chain is trained on a set of features and the set of labels predicted by the previous classifiers. We used cardiotoxicity types as labels and gene expression values along with molecular descriptors and fingerprints as features. We used the following order of cardiotoxicity types obtained from MedDRA (see 'Data Preparation'): 'Vascular disorders', 'Cardiac disorder signs and symptoms', 'Cardiac arrhythmias', 'Heart failure', 'Coronary artery diseases', 'Pericardial disorders' and 'Myocardial disorders'. This order was based on the number of drugs related to those side effects. Because the first model will not receive the information from other cardiotoxicity types, we introduced vascular disorders (which is also related to cardiac disorders) as the first disorder for prediction, in order to minimize the effect of the first position for 'Cardiac disorder sign and symptoms'. However, the accuracy of 'Vascular disorder' prediction was not used in the evaluation of the model performance. This way, the chain of classifiers takes a set of features (transcriptional, molecular descriptors and fingerprints) and is tasked to predict whether the drug has cardiotoxicity reports ('positive case') or not ('negative case'), for six cardiotoxicity types.

To determine which algorithm accounts best for the observed data, we adapted several supervised binary classification algorithms widely used in bioinformatics: elastic net logistic regression (glmnet R library by Friedman et al. (Friedman et al., 2010)), random forest (ranger R library by Marvin et al. (Wright and Ziegler, 2017)), gradient boosting (gbm R library by Ridgeway et al. (Greenwell et al., 2007 )) and categorical boosting (catboost R library by Prokhorenkova et al. (Prokhorenkova et al., 2018)). All models were optimized with Latin hypercube sampling of parameters (clhs R library by Roudier (Minasny and McBratney, 2006)) towards maximum Matthews correlation coefficient (MCC):

$$\mathbf{MCC} = \frac{\sum \mathbf{tp} \times \sum \mathbf{tn} - \sum \mathbf{fp} \times \sum \mathbf{fn}}{\sqrt{(\sum \mathbf{tp} + \sum \mathbf{fp})(\sum \mathbf{tp} + \sum \mathbf{fn})(\sum \mathbf{tn} + \sum \mathbf{fp})(\sum \mathbf{tn} + \sum \mathbf{fn})}} \qquad (2)$$

where *tp* (true positives) and *tn* (true negatives) are the number of unsafe and safe compounds predicted correctly, respectively, and *fp* (false positives) and *fn* (false negatives) are the number of safe and unsafe drugs predicted wrongly, respectively. An MCC of 0 indicates that the prediction is not better than a random prediction, an MCC of 1 indicates perfect prediction or total agreement, and an MCC of −1 indicates total disagreement.

The optimized parameters are supplied in **Supplementary Table 4**. We trained models with five-fold cross-validation selected to leave drugs out to compensate for overfitting, and to receive more robust performance metrics. Once trained, to predict the cardiac safety of any unseen chemicals, the model only receives as inputs their transcriptional features, molecular descriptors, and fingerprints.

## Model Comparison

In this study, our proposed chain of classifiers with nested stacking for multi-label classification of drug cardiotoxicity is compared against a set of independent binary classifiers by cardiotoxicity types (meaning the drug has at least one side effect). We adapted the same four classification algorithms (elastic net logistic regression, random forest, gradient boosting and categorical boosting) for this task. We adjusted the set of hyperparameters and validation and tested models as described in the previous section. The optimized parameters for each model are supplied in **Supplementary Table 4**.

## Model Evaluation

In addition to MCC, the following metrics were used to evaluate model performance for each cardiotoxicity forms:

$$Accuracy = \frac{\sum tp + \sum tn}{\sum tp + \sum tn + \sum fp + \sum fn}, \qquad (3)$$

where *tp* is a number of correctly predicted drugs with cardiotoxicity reports, *tn* is a number of correctly predicted drugs without cardiotoxicity reports, *fn* is a number of incorrectly predicted drugs with cardiotoxicity reports and *fp* is a number of incorrectly predicted without cardiotoxicity reports. Accuracy shows the ratio of correctly predicted drugs to a total number of drugs.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \qquad (4)$$

or F1 score, where

$$precision = \frac{\sum tp}{\sum tp + \sum fp}$$

and

$$recall = \frac{\sum tp}{\sum tp + \sum fn},$$

where *tp* is a number of correctly predicted drugs with cardiotoxicity reports, *tn* is a number of correctly predicted drugs without cardiotoxicity reports, *fn* is a number of incorrectly predicted drugs with cardiotoxicity reports and *fp* is a number of incorrectly predicted without cardiotoxicity reports. *Precision* equals the fraction of correctly predicted unsafe compounds in all compounds predicted as unsafe, whereas *recall* shows the sensitivity of a model and equals the fraction of correctly predicted unsafe compounds out of all real unsafe compounds.

## External Validation

As external validation data, we downloaded gene expression drug profiles from the Drug Toxicity Signature Generation Center (DToxS) website (https://martip03.u.hpc.mssm.edu/index.php). This website provides access to expression data of PromoCell cardiomyocytes (up to four lines) incubated with FDA approved drugs. In total, we obtained 1,338 samples, which were collapsed in the same manner as Connectivity map data, so each drug profile referred to gene expression values for one cell line, incubation time and concentration. As a result, models were tested on 654 profiles of 51 drugs, 18 of which were for the same drugs used for training and validation of models. We used 'rcdk' package to calculate the same fingerprints and set of molecular descriptors as for the drug dataset that training and testing.

$$\mathbf{AUC} = \int_{-\infty}^{\infty} \mathbf{TPR(T)}\left(-\mathbf{FPR}'\,\mathbf{(T)dT}\right) \qquad (5)$$

or area under the receiver operating characteristic (ROC) curve, where TPR is the true positive rate (identical to recall) and FPR is the false positive rate. AUC measures the diagnostic ability of a predictor, where an AUC of 0.5 indicates that the prediction is not better than a random prediction, and an AUC of 1 indicates perfect prediction. The pROC R library by Robin et al. (2011) was used to calculate AUC values for the classifiers.

# RESULTS

## Enriched Analysis and Prediction of Six Drug-Induced Cardiotoxicity Forms Using Transcriptional and Molecular Data

**Figure 1** describes the computational and dataset framework defined through this study. The six main drug-induced cardiotoxicity forms were identified from MedDRA as the focus for prediction: 'Cardiac disorders signs and symptoms', 'Cardiac arrhythmias', 'Heart failure', 'Coronary artery disease', 'Pericardial disorders', and 'Myocardium disorders' (**Figure 1A**).

Then, a large dataset of drugs was collected from diverse publicly-available data repositories (**Figure 1B**), including two sources of information: transcriptional profiles and derived molecular descriptors and fingerprints. This yielded information on 1,131 drugs, 357 of which had transcriptional profiles with a total of 9,933 samples. As a strategy for validation, these were split into unique training (291 drugs, 8,237 samples) and testing (66 drugs, 1,696 samples) sets. Training was blinded to drugs on the testing set, hence facilitating preclinical translation to novel unseen chemicals (**Figure 1C**).

Transcriptional and molecular descriptors for the drugs were used as inputs to the machine learning algorithms (**Figure 1D**). All machine-learning models were evaluated in performance on the blinded testing dataset, also considering two independent (random and leave-drug-out) cross-validation strategies (**Figure 1D**). Further details on study design are provided in *Methods*.

**Figure 2A** shows the number of drugs labeled as unsafe for the six groups of cardiotoxicity forms considered. Notably, 49% (46 out of 93) of antineoplastic drugs are reported to cause 'cardiac disorders and signs and symptoms', indicating a high prevalence across cardiotoxicity forms and drug classes (**Supplementary Table 1**). On the other hand, 24% of CNS, 21% of CV, and 27% of antineoplastic agents produced cardiac arrhythmias, and 23% of CV and 25% of antineoplastic agents induced coronary artery disease (**Supplementary Table 1**). The prevalence of heart failure, myocardial disorders and pericardial disorders was lower for all drug classes, although still significant in some cases (for example, 14% of antineoplastic agents producing heart failure; **Supplementary Table 1**). We also evaluated the level of association between cardiotoxicity forms in terms of Cohen's Kappa (**Figure 2B**). 'Cardiac arrhythmias' demonstrate a substantial association with both 'cardiac disorder signs and symptoms' and 'coronary artery

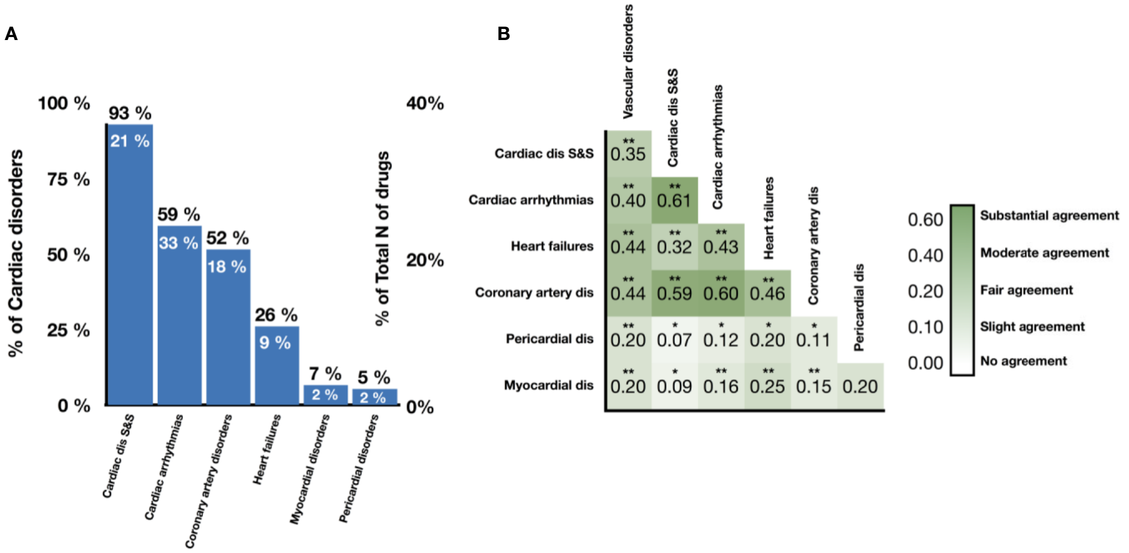


**FIGURE 2 |** Prevalence and association of different cardiotoxicity forms in the drug database. **(A)** Proportion of drugs labeled as unsafe for given cardiotoxicity forms out of all unsafe drugs (numbers in black) and all drugs (numbers in white). **(B)** Levels of association between cardiotoxicity forms, measured in terms of Cohen's Kappa. Symbols represent statistical significance level by Z-test (**p* <0.05, ***p* <0.01). 'Cardiac dis S&S' is the MedDRA term for cardiac disorder signs and symptoms, 'dis' is for disorders.

disorders', with Cohen's Kappa values of 0.61 and 0.60 respectively. Interestingly, 'heart failure' demonstrates a lower agreement with 'cardiac disorder signs and symptoms' compared to 'cardiac arrhythmias' and 'coronary artery disorders' but higher agreement with 'myocardial disorders'.

Based on the strong associations observed between drug-induced cardiotoxicity forms, we concluded that a prediction model should leverage dependance between side effects, and be able to re-use the information learned about the molecular basis of one side effect to better understand the molecular basis of others. This motivates our formulation of the cardiotoxicity prediction task as a multi-label classification problem, and our proposed machine-learning architecture as a chain of classifiers (**Figure 1D**), in order to preserve such relationships between drug-induced cardiotoxicity forms.

## Candidate Genes and Pathways Associated With Cardiotoxicity

We then evaluated whether the association between cardiotoxicity forms identified in **Figure 2** was also evident from transcriptional data, either in terms of genes or pathways. We hypothesized that transcriptional data would reveal underlying information on cardiotoxicity types when genetic pathways, rather than individual genes, are considered in the analysis. **Figures 3A, B** show the comparison of the gene vectors ranked as important by two feature selection methods for the different cardiotoxicity forms. Similarly, **Figures 3C, D** show this comparison in terms of the Reactome pathways to which those genes are related.

Analysis of the list of genes showed little to no intersection between them (**Supplementary Table 2**). Conversely, the consideration of pathways significantly improved the agreement (**Supplementary Table 3**). Gene and associated pathways for 'cardiac disorders', 'cardiac arrhythmias' and 'heart failure' identified by both methods display moderate to high agreement, whereas for pericardial disorders no significant agreement was found. Interestingly, cardiac arrhythmias and coronary artery diseases show high similarity in both selected vectors of genes and pathways, and of label vectors of drugs.

Interestingly, G protein-coupled receptor transduction was selected as important by both methods for the prediction of cardiac disorder signs and symptoms ('G alpha (q) signaling events' and 'G alpha (s) signaling events') and heart failure ('G-alpha (i) signaling events'). IGF1R and IGF1R-related signaling were also among the genes and Reactome terms selected by both methods for cardiac disorder signs and symptoms and for pericardial disorders. MAMLD1 gene, included in the Notch signaling pathways, was the only one ranked as important for predicting cardiac arrhythmias by both selection procedures. This further suggests association between different cardiotoxicity forms also at the feature level, which again motivates us to use a chain of classifiers to keep relations between cardiotoxicity forms during prediction.

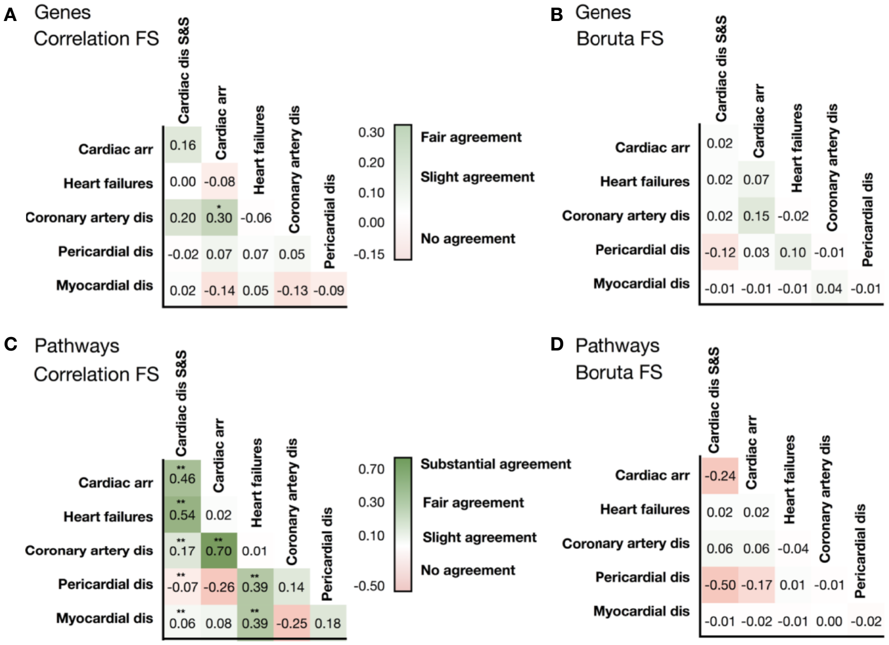


**FIGURE 3 |** Similarities in the list of selected genes identified by **(A)** the correlation feature selection, and **(B)** the Boruta wrapper-based algorithm. Similarities in the list of pathways associated with genes identified by correlation-based feature selection **(C)**, and the wrapper-based algorithm **(D)**. Levels of association between cardiotoxicity forms are measured in terms of Cohen's Kappa. Symbols represent statistical significance level by Z-test (*$p$ <0.05, **$p$ <0.01). 'Cardiac dis S&S' is the MedDRA term for cardiac disorder signs and symptoms, 'dis' is for disorders.

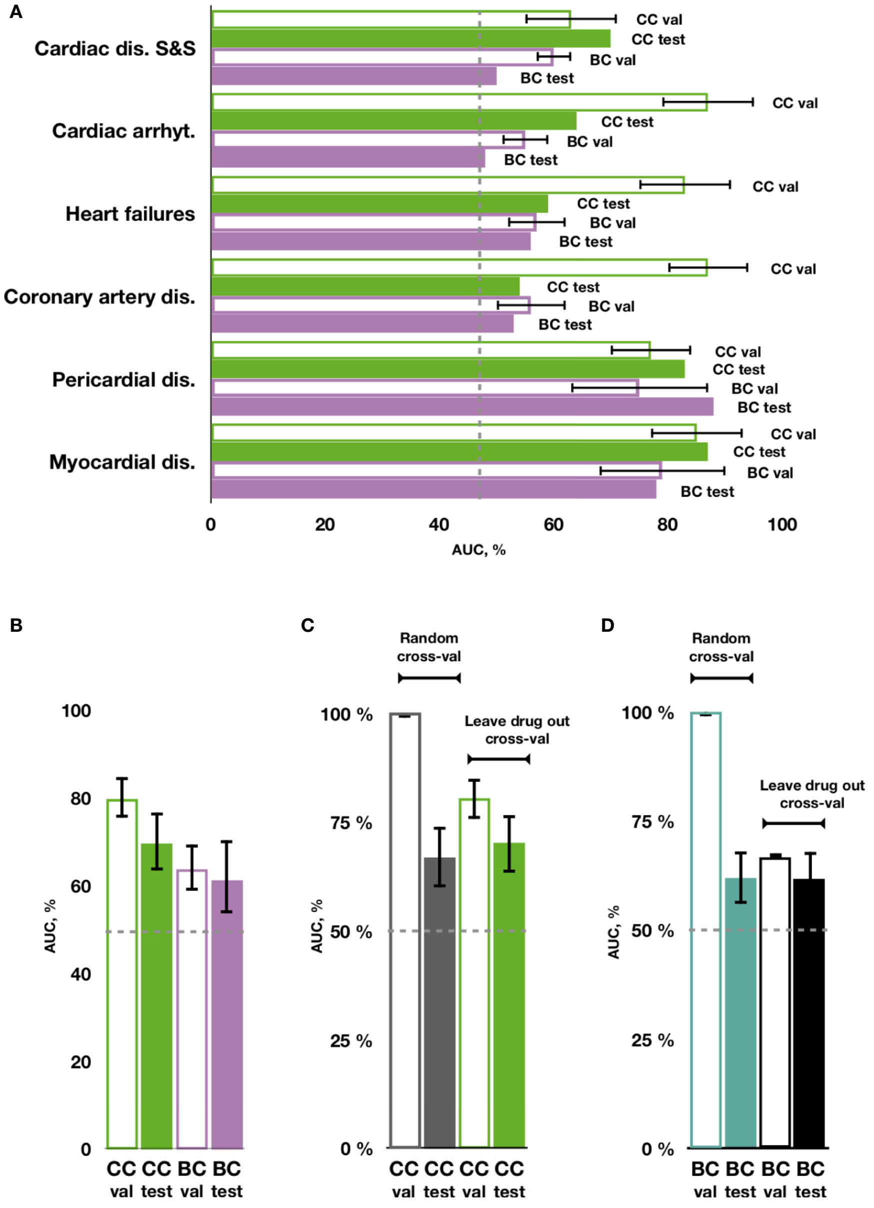


**FIGURE 4 |** Prediction of cardiotoxicity forms using the chain of classifiers with nested stacking (CC) versus a set of binary classifiers (BC). **(A)** AUC of best performing CC and BC in safety drug prediction on validation (val) and testing sets (test) for each independent cardiac disorder. **(B)** Average AUC across all cardiac disorders for CC versus BC in validation and test sets. **(C, D)** Comparison between random vs leave-drug-out cross-validation strategies. Average performance across all labels is shown for the best performing chain of classifiers (CC) model, trained with leave-drug-out vs random cross-validation strategies, and for a set of independent binary classifiers (BC). Random cross-validation significantly inflates the accuracy of the trained models compared to leave-drug-out validation. Cardiac dis S&S': cardiac disorder signs and symptoms, 'dis': disorders, 'arrhyt': arrhythmias.

## Machine Learning Prediction of Drug-Induced Cardiotoxicity Forms: The Importance of Leave-Drug-Out Cross-Validation

Using transcriptional and molecular features, all investigated forms of drug-induced cardiotoxicity were predicted with relatively good accuracy using the proposed chain of classifiers model with nested stacking trained with leave-drug-out cross-validation, and for all algorithms considered (elastic net logistic regression, gradient boosting, categorical boosting, and random forests). The best results were obtained for the chain of random forest classifiers, with an average AUC of 0.79 and an average MCC of 0.38 across all cardiotoxicity forms on validation, and 0.66 and 0.15 on testing (**Figure 4A**, **Table 2**, **Supplementary Figure 1** and **Supplementary Table 4**). The second best results were obtained with a chain of gradient boosting classifiers, with

**TABLE 2 |** The performance of multi-label classification models trained on transcriptional profiles and molecular descriptors and fingerprints of drugs on the validation and the testing set. The values are reported for Area under the receiver operating characteristic (ROC) curve (AUC; upper value), % and Matthews correlation (MCC; lower value).

| Cardiotoxicity form | Set | | Cardiac dis S&S | Cardiac arrhythmias | Heart failures | Coronary artery dis | Pericardial dis | Myocardial dis | Mean ± 0.5 SD |
|---|---|---|---|---|---|---|---|---|---|
| Leave drug out cross-validation strategy | | | | | | | | | |
| Chain of classifiers with nested stacking | | | | | | | | | |
| Dual feature set | | | | | | | | | |
| RF | Validation | AUC | 63 | 88 | 79 | 91 | 64 | 88 | 79 ± 6 |
| | | MCC | 0.21 | 0.50 | 0.29 | 0.68 | 0.13 | 0.46 | 0.38 ± 0.10 |
| | *Testing* | *AUC* | *70* | *64* | *58* | *54* | *83* | *87* | ***66 ± 9*** |
| | | *MCC* | *0.33* | *0.37* | *0.04* | *0.10* | *0.14* | *–0.08* | ***0.15 ± 0.9*** |
| ELNET | Validation | AUC | 62 | 67 | 65 | 55 | 71 | 65 | 64 ± 3 |
| | | MCC | 0.20 | 0.23 | 0.19 | –0.03 | 0.6 | 0.37 | 0.17 ± 0.07 |
| | *Testing* | *AUC* | *67* | *65* | *63* | *49* | *66* | *70* | *63 ± 4* |
| | | *MCC* | *0.28* | *0.28* | *–0.08* | *–0.08* | *–0.03* | *0.01* | *0.06 ± 0.09* |
| GBM | Validation | AUC | 67 | 87 | 74 | 90 | 60 | 86 | 77 ± 6 |
| | | MCC | 0.29 | 0.55 | 0.35 | 0.66 | 0.16 | 0.54 | 0.42 ± 0.09 |
| | *Testing* | *AUC* | *66* | *62* | *51* | *55* | *60* | *64* | *60 ± 3* |
| | | *MCC* | *0.22* | *0.26* | *–0.08* | *0.03* | *0.12* | *–0.04* | *0.08 ± 0.07* |
| CATBOOST | Validation | AUC | 64 | 84 | 64 | 88 | 65 | 53 | 70 ± 7 |
| | | MCC | 0.30 | 0.68 | 0.32 | 0.65 | 0.24 | 0.02 | 0.37 ± 0.13 |
| | *Testing* | *AUC* | *67* | *63* | *59* | *57* | *67* | *49* | *60 ± 3* |
| | | *MCC* | *0.27* | *0.18* | *–0.07* | *0.09* | *0.35* | *0.0* | *0.14 ± 0.08* |
| Transcriptional features only | | | | | | | | | |
| RF | Validation | AUC | 63 | 83 | 82 | 90 | 61 | 89 | 78 ± 6 |
| | | MCC | 0.24 | 0.47 | 0.4 | 0.66 | –0.02 | 0.62 | 0.4 ± 0.13 |
| | *Testing* | *AUC* | *69* | *65* | *58* | *64* | *71* | *67* | *66 ± 2* |
| | | *MCC* | *0.22* | *0.19* | *0* | *0.21* | *0.11* | *0.11* | *0.14 ± 0.04* |
| Descriptors and molecular fingerprints only | | | | | | | | | |
| RF | Validation | AUC | 66 | 83 | 82 | 86 | 48 | 81 | 74 ± 7 |
| | | MCC | 0.24 | 0.55 | 0.37 | 0.51 | –0.08 | 0.19 | 0.3 ± 0.12 |
| | *Testing* | *AUC* | 63 | 72 | 42 | 56 | 62 | 58 | 59 ± 5 |
| | | *MCC* | 0.21 | 0.22 | 0.01 | 0.16 | –0.08 | –0.07 | 0.08 ± 0.07 |
| A set of independent binary classifiers | | | | | | | | | |
| RF | Validation | AUC | 57 | 62 | 61 | 61 | 57 | 53 | 58 ± 2 |
| | | MCC | 0.14 | 0.12 | 0.06 | –0.03 | 0.0 | 0.17 | 0.08 ± 0.04 |
| | *Testing* | *AUC* | *64* | *63* | *59* | *59* | *45* | *58* | *58 ± 3* |
| | | *MCC* | *0.21* | *0.18* | *–0.02* | *0.12* | *–0.01* | *–0.01* | *0.08 ± 0.05* |
| ELNET | Validation | AUC | 60 | 58 | 61 | 55 | 72 | 64 | 62 ± 3 |
| | | MCC | 0.12 | 0.15 | 0.10 | –0.08 | 0.06 | 0.17 | 0.09 ± 0.05 |
| | *Testing* | *AUC* | *65* | *67* | *61* | *52* | *64* | *65* | *62 ± 3* |
| | | *MCC* | *0.19* | *0.25* | *–0.03* | *–0.08* | *–0.04* | *–0.05* | *0.04 ± 0.07* |
| GBM | Validation | AUC | 59 | 61 | 50 | 49 | 58 | 66 | 57 ± 3 |
| | | MCC | 0.11 | 0.22 | –0.02 | –0.03 | –0.14 | 0.37 | 0.08 ± 0.09 |
| | *Testing* | *AUC* | *67* | *61* | *54* | *57* | *50* | *63* | *59 ± 3* |
| | | *MCC* | *0.21* | *0.20* | *–0.02* | *0.04* | *–0.03* | *0.05* | *0.08 ± 0.05* |
| CATBOOST | Validation | AUC | 62 | 60 | 61 | 56 | 54 | 70 | 60 ± 3 |
| | | MCC | 0.23 | 0.26 | 0.20 | 0.28 | 0.07 | 0.37 | 0.24 ± 0.05 |
| | *Testing* | *AUC* | *72* | *66* | *66* | *60* | *55* | *61* | *63 ± 3* |
| | | *MCC* | *0.37* | *0.22* | *0.0* | *0.11* | *0.0* | *0.06* | *0.13 ± 0.07* |
| Random cross-validation strategy | | | | | | | | | |
| Chain of classifiers with nested stacking | | | | | | | | | |
| RF | Validation | AUC | 92 | 96 | 93 | 95 | 94 | 92 | **94 ± 1** |
| | | MCC | 0.70 | 0.77 | 0.53 | 0.72 | 0.44 | 0.26 | **0.57 ± 0.10** |
| | *Testing* | *AUC* | *68* | *62* | *52* | *50* | *73* | *79* | *64 ± 6* |
| | | *MCC* | *0.21* | *0.28* | *–0.09* | *–0.04* | *0.27* | *–0.07* | *0.09 ± 0.09* |
| A set of independent binary classifiers | | | | | | | | | |
| RF | Validation | AUC | 90 | 88 | 74 | 83 | 86 | 77 | **83 ± 3** |
| | | MCC | 0.67 | 0.61 | 0.25 | 0.57 | 0.56 | 0.16 | **0.47 ± 0.11** |
| | *Testing* | *AUC* | *66* | *63* | *59* | *56* | *45* | *58* | *58 ± 4* |
| | | *MCC* | *0.20* | *0.20* | *0.03* | *–0.09* | *–0.01* | *–0.01* | *0.05 ± 0.06* |

*Validation set and testing set performance are shown in the upper and the lower cells respectively and the best performance is shown in bold. RF is for random forest, ELNET is for elastic net logistic regression, GBM is for gradient boosting machines, CATBOOST is for categorical boosting. Cardiac dis S&S is for cardiac disorders signs and symptoms MedDRA term,dis is for disorders. Results for testing are shown in italics.*

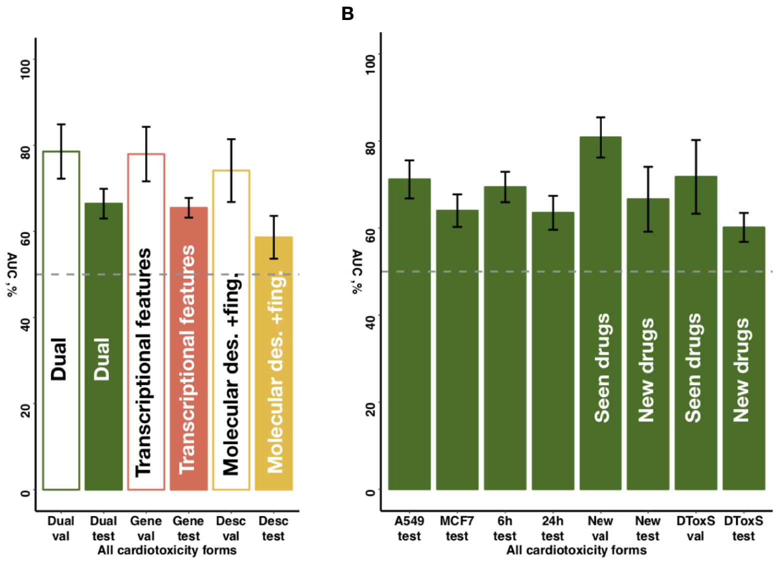


**FIGURE 5 |** Sensitivity of predictive accuracy to feature types, cell lines and incubation times. **(A)** Merging of molecular descriptors and fingerprints and transcriptional features results in increased average performance across all drug classes. **(B)** The best performing model demonstrates similar predictive accuracy across different cell lines (A549, MFC7) and incubation times (6 and 24 h) incubation times. For new cell lines (PHH, SKB, SKM1, A673), the model discriminates more accurately seen drugs with unseen cell lines ('New val') than unseen drugs ('New test'). For new data type (DToxS), the model also discriminates more accurately seen drugs ('DToxS val') than unseen drugs ('DToxS test').

average AUC of 0.71 and average MCC of 0.24 on validation, and 0.66 and 0.15 on test set. The chain of categorical boosting classifiers showed average AUC of 0.77 and average MCC of 0.51 on validation, with average AUC of 0.65 and average MCC of 0.13 for testing. Finally, elastic net demonstrated the most modest performance among the trained set of chain classifiers, achieving an average AUC of 0.66 and average MCC of 0.16 on validation, with average AUC of 0.60 and average MCC of 0.11 for testing. Following these results, we selected a chain of random forest classifiers as the best model and evaluated its performance in detail, including validation on new cell types, external independent dataset and across cardiotoxicity types and pharmacological classes of drugs.

Importantly, cardiotoxicity types predicted with the best performing chain of classifiers model kept similar relationships to the ones shown in the original data (**Figure 2** and **Supplementary Figure 2**). This was particularly clear for the predominant associations between cardiac disease signs and symptoms, cardiac arrhythmias and coronary artery disease.

With individual binary predictors, the best set of random forest classifiers only obtained an average AUC of 0.67 and an average MCC of 0.08 across cardiotoxicity types on validation, with an average AUC of 0.62 and average MCC of 0.16 on testing (**Figure 4A**, **Table 2**, **Supplementary Figure 1** and **Supplementary Table 4**). Therefore, while cardiotoxicy types were predicted with different accuracies, the inclusion of information about other cardiotoxicities improved prediction accuracy for all cardiotoxicity types (**Figure 4B**). This was also observed on five different partitions of the entire dataset (**Supplementary Figure 3**), where the chain of classifiers outperformed the sets of individual binary classifiers. In line with that, the exclusion of cardiotoxicity forms resulted in a decreased prediction accuracy of sequential labels (**Supplementary Table 4**).

On the contrary to leave-drug-out, in the case of random cross-validation, samples using the same drugs may be present in training and testing datasets, and thus predictors learn associations between individual drugs and their safety rather than general features related to cardiotoxicity forms. This may produce unrealistically high results, indeed overestimating the accuracy of prediction.

To investigate these aspects in further detail, we compared predictions with our proposed chain of random forest classifiers with nested stacking and a set of independent random forest classifiers, both trained with either leave-drug-out or random cross-validation strategies. When validated and optimized with random cross-validation, both models demonstrate almost perfect accuracy on validation (averages for all cardiotoxicity types: AUC = 1, MCC = 0.95 for chain of classifiers; AUC = 1, MCC = 0.97 for independent predictors; **Figures 4C, D**, **Table 2** and **Supplementary Table 4**). Although apparently outperforming the models trained with leave-drug-out cross-validation, models trained with random cross-validation were however less accurate when predicting cardiotoxicity types of previously unseen drugs (**Figures 4C, D**, **Table 2** and **Supplementary Table 4**). Quantitatively, the chain of classifiers trained and optimized with a random cross-validation strategy exhibited an AUC of 0.67 and MCC of 0.13, compared to an AUC of 0.70 and MCC of 0.16 for leave-drug-out cross-validation (average values for all cardiotoxicity types).

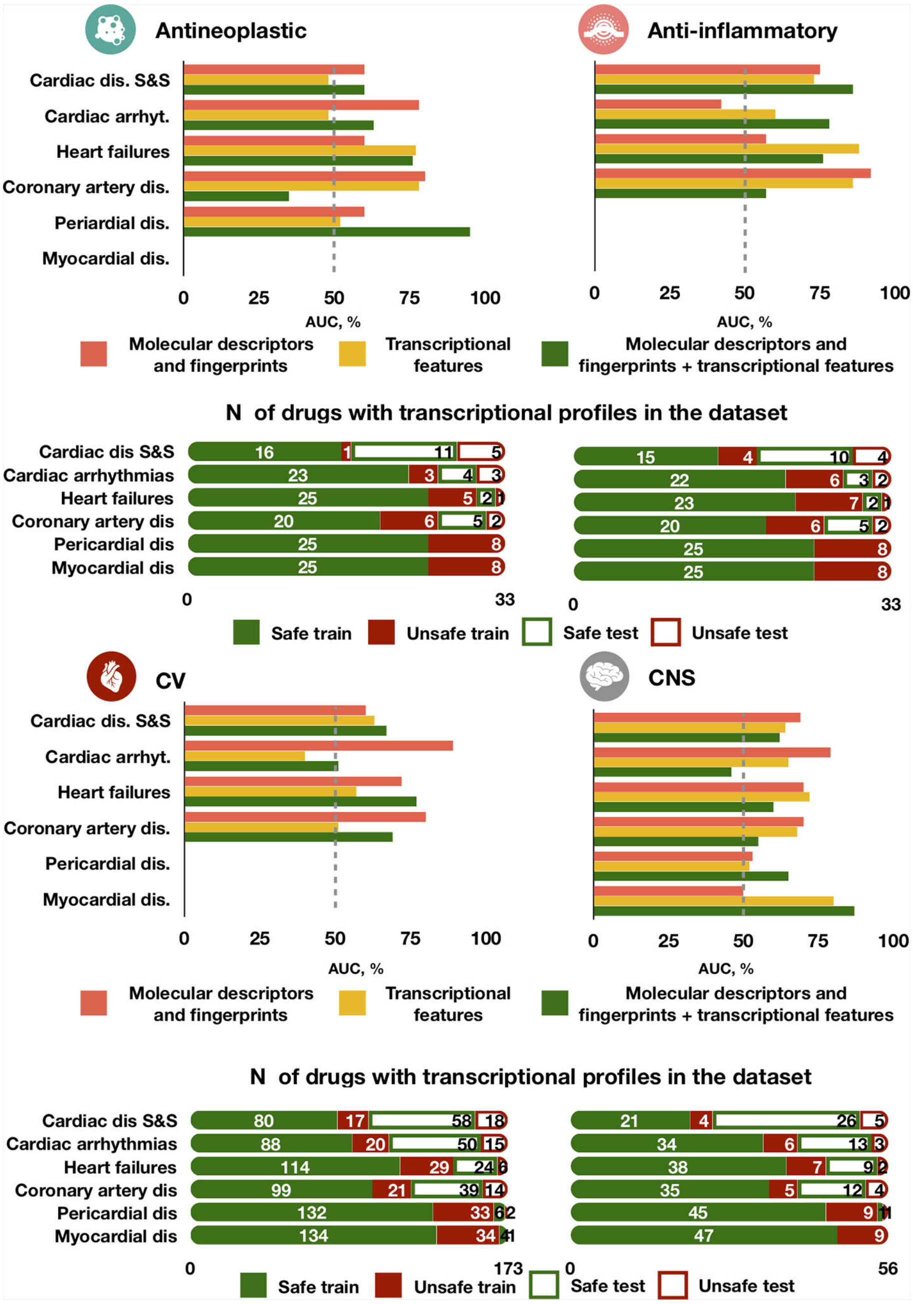


**FIGURE 6 |** Predictive accuracy across drug classes, cardiotoxicity forms and feature types. The best performing model trained on either molecular descriptors and fingerprints, transcriptional features or both demonstrates different performance in predicting types of agents. For each drug class, the top bar plot shows the AUC of the best predictor and the bottom bar plot displays the number of safe and unsafe drugs. CV for cardiovascular agents, CNS for Central nervous system agents.

Similarly, a set of individual classifiers with random cross-validation showed an AUC of 0.62 and MCC of 0.09, compared to an AUC of 0.62 and MCC of 0.15 in the case of leave-drug-out cross-validation.

To further explore the predictive value of individual feature sets, we compared for the entire dataset the proposed chain of random forest classifiers with nested stacking separately trained on each feature set. Models trained on both feature types outperformed models trained only on one feature, showing on validation an AUC of 0.80 for both feature types vs an AUC of 0.62 for molecular descriptors and fingerprints, or an AUC of 0.76 for transcriptional features only (**Figure 5A**, **Table 2**, **Supplementary Table 4**). Similarly, the dual transcriptomic and molecular classifier is more accurate on testing. The model trained only on molecular descriptors and fingerprints achieves 0.65 AUC, being 0.66 AUC when trained only for transcriptional features.

## External Validation

To further investigate the predictive power of the developed predictor and to assess its performance on different source of transcriptional data, we additionally analyzed the DToxS dataset (https://martip03.u.hpc.mssm.edu/index.php).

Notably, while DToxS provides gene expression data measured with a different technique (RNAseq) and with different cell lines (PromoCell cardiomyocytes), the proposed chain of random forest classifiers with nested stacking still achieved good accuracy when discriminating previously seen drugs (average AUC of 72%, **Figure 5B**). The accuracy partially drops when predicting cardiotoxicity forms for unseen drugs (average AUC of 60%, **Figure 5B**). Interestingly, while the model trained on molecular descriptors and fingerprints only shows superior accuracy in predicting seen drugs (average AUC of 90%, **Supplementary Table 4**), the dual transcriptomic and molecular classifier is still more accurate on testing for unseen drugs. For example, it is able to differentiate safe from drugs with reports of cardiac arrhythmias with an AUC of 70% (vs 58% for molecular descriptors and fingerprints, **Supplementary Table 4**). The transcriptional feature only model shows less accuracy when tested on new data, with an average AUC for all labels of 57% (**Supplementary Table 4**). However, this model is more accurate in predicting cardiac disorder signs and symptoms (AUC of 65% vs 56% and 52%, for dual and molecular descriptors and fingerprints models, respectively; **Supplementary Table 4**).

## Predictive Accuracy Across Drug Classes and Cardiotoxicity Forms

The best model (chain of random forest classifiers with nested stacking) demonstrated different predictive accuracy across cardiotoxicity forms and drug classes (**Figure 6** and **Supplementary Table 4**). For example, for antineoplastic drugs, the best model predicted 'pericardial disorders' more accurately (**Figure 6**; AUC = 0.95) than the average across all drugs, also achieving high accuracy in the prediction of 'heart failure' (AUC = 0.76). 'Cardiac disorder signs and symptoms', 'cardiac arrhythmias' and 'heart failure' were also more accurately predicted in the case of anti-inflammatory drugs compared to other cardiotoxicity types (**Figure 6**; AUCs of 0.86, 0.78 and 0.76, respectively).

High accuracy was also achieved in the prediction of 'myocardial disorders' by cardiovascular agents, as well as for 'heart failure' induced by central nervous system agents (both with AUC = 0.77, **Figure 6**). On the contrary, predictions for 'cardiac arrhythmias' were close or worse than random guessing for cardiovascular and central nervous system agents (AUCs of 0.51 and 0.46, respectively) and for cardiovascular agents (AUC = 0.51). Greater error in the prediction of 'coronary artery disorders' was found for antineoplastic agents (AUC = 0.35). Some of these cases of limited performance may be partially explained by a small frequency of side effects for specific drug types. For example, only 16 out of 56 total cardiovascular agents in our dataset are known to cause cardiac arrhythmias and coronary artery diseases (**Figure 6**). However, for antineoplastic drugs the frequency of coronary artery disease (7/33) is the same than for cardiac arrhythmias (7/33) and bigger than for heart failure (3/33), but the predictor performs better in predicting the latter two than coronary artery disease, indicating a possible dependency on feature types.

Interestingly, the model trained individually on molecular descriptors and fingerprints demonstrated higher accuracy than the transcriptomic or dual predictors in safety predictions of 'cardiac arrhythmias' for cardiovascular, antineoplastic and central nervous system agents, but not for anti-inflammatory agents (**Figure 6**). For central nervous system agents, in general, the molecular descriptor-based predictor is more accurate. Notably, for 'coronary artery disorders', a combination of the two feature types leads to a decrease in accuracy compared to individual feature set predictors.

For interpretability of these prediction differences, we analyzed the feature space in testing against the best model predictions when trained on either individual or combined feature types (**Supplementary Table 5** and **Supplementary Figures 4–8**). Different transcriptional features and molecular descriptors were associated with the accuracy of prediction. For instance, drug samples with higher expression values of Alpha-Synuclein (SNCA) or Heat Shock Protein 8 (HSPA8) genes were more often predicted incorrectly by all three predictors (**Supplementary Figures 4** and **5**). Drugs with a higher number of atomic bonds (nAtomBond) (**Supplementary Figure 6**) were classified less correctly when training on the combined feature set or only on molecular descriptors. The same was observed for drugs with higher values of polar surface area to molecular size ratio (tpsaEffiency) or topological polar surface area based on fragment contributions (TopoPSA) (**Supplementary Figures 7** and **8**).

# DISCUSSION

This study presents the first machine learning approach capable of predicting six forms of drug-induced cardiotoxicity from both gene expression and molecular descriptors data. Importantly, the algorithm (based on a chain of classifiers) is specifically developed to incorporate relationships between cardiotoxicity forms, identified in our data analysis, and to tackle class imbalance between cardiotoxic and safe drugs. We demonstrate high accuracy with the strictest validation strategy using drug datasets not used in training, and its importance compared to random cross-validation on samples. A further specific contribution of this study is the large comprehensive dataset of 1,131 drugs curated and collected from publicly available resources. This can provide a useful benchmark for future studies. Thus, we propose a novel and robust solution for preclinical drug safety testing that can potentially be expanded to other organs' toxicities.

To implement this solution, we first collected and analyzed a large dataset of 1,131 drugs from publicly available databases.

They include both safe and cardiotoxic drugs, 357 of which with cellular transcriptional profiles and a total of 9,933 samples available. Secondly, we proposed and implemented a chain of classifiers with nested stacking approach that classifies drugs by their risk, able to relatively accurately predict up to six forms of cardiotoxicity. Our method achieves a 0.80 average AUC across all cardiotoxicity types on a leave-drug-out cross-validation strategy on 291 drugs (8,237 samples). Further validation of the method on new and previously unseen 66 drugs (1,696 samples) with multiple mechanisms of action demonstrated that the proposed model holds high generalisation abilities compared to sets of individual classifiers. Models trained with leave-drug-out cross-validation were able to discriminate between safe and unsafe drugs with a 0.70 average AUC across all cardiotoxicity types. The model demonstrated higher accuracy for specific adverse drug effects and type of agents, and in particular for pericardial disease, cardiac disease and symptoms, heart failure and myocardial disease for antineoplastic, anti-inflammatory, cardiovascular and central nervous system agents, respectively. These results suggest the translational potential of the proposed approach towards applications in a pre-clinical context.

The combined dataset collected in this study demonstrated associations between forms of drug-induced cardiac complications, which are in agreement with the known literature on cardiac comorbidities. Clinical reports have evidenced a significant association between heart failure and other cardiovascular comorbidities, such as atrial fibrillation, ischemic heart disease and arrhythmias (Lawson et al., 2018; Kendir et al., 2018). Patients with a history of coronary heart disorders have been also shown to have a higher incidence of atrial fibrillation, one of the most common forms of cardiac arrhythmias (Naser et al., 2017). In our work, coronary artery disorders, which include ischemia and myocardial infarction, demonstrated a significant similarity to cardiac arrhythmias in terms of the drugs they are related to. We showed the same for heart failure and cardiac disorders, which displayed a moderate and substantial agreement to cardiac arrhythmias.

Notably, the list of genes identified as most important features (obtained by using two completely distinct feature selection methods while counting distinct genes for each label) demonstrates a significant amount of intersection at the level of associated pathways. For example, pathways related to Notch signaling were ranked as important for cardiac arrhythmia prediction by both methods. Previous evidence has shown the importance of Notch signaling in heart development and cardiac disease, including malignant congenital arrhythmias (D'Amato et al., 2016). IGF signaling and IGF1R were also selected consentaneously by both algorithms for cardiac disorder signs and symptoms and pericardial disorders, evidencing their key role in heart tissue functioning (Troncoso et al., 2014). In spite of being profiled using cancer cell lines, the selected features seem biologically relevant to cardiotoxicity, given their human origin. This emphasizes the potential benefits of using the combined approach for feature selection. At the same time, further detailed investigation of the feature importance list could help evaluate the proposed genes as possible therapeutic targets for cardiovascular therapies.

Our method takes advantage of the chain of classifiers approach. This approach significantly outperformed binary classification approaches that treat each label independently, with an improvement of 12.9% in terms of AUC (from AUC of 0.62 to 0.70). This highlights the importance of incorporating information about related adverse reactions in predicting drug safety. While demonstrating good generalisation abilities on unseen data, the model showed different performance across cardiotoxicity types depending on the type of agents. Cardiac arrhythmia-related safety was predicted more accurately for cancer and anti-inflammatory agents than for cardiovascular and central nervous system agents. This might be improved by the introduction of information more relevant to specific mechanisms of arrhythmogenesis.

The accumulated body of evidence suggests that gene expression signatures alone could also be used as a biomarker of cell response to drugs (Aliper et al., 2016; Xie et al., 2018). Our results, in line with previous studies (Wang et al., 2016), demonstrate that coupling of transcriptional profiles and molecular descriptors indeed improves the predictive power of algorithms. Thus, the combination of both feature types indeed increases the mean accuracy of a chain of classifiers by 8% (AUC of 0.65 to 0.70).

Previous research (Wang et al., 2016; Messinis et al., 2018) reported good accuracy in the prediction of multiple adverse reactions and myocardial infarction. However, such approaches were evaluated using drugs included in the training, rather than in an independent dataset as in our study, and neglected existing dependencies between various forms of cardiotoxicity, as demonstrated here. Indeed, our findings suggest that the retention of the information about cardiotoxicity types dependencies results in greater accuracy (**Figures 4A, B**). At the same time, we show that models trained with random cross-validation may display a significantly inflated performance on validation (**Figures 4C, D**), however becoming less accurate when predicting previously unseen drugs. In general, a leave-drug-out cross-validation strategy demonstrated a more robust performance compared to random cross-validation, the latter evidencing inflated accuracy metrics, which in turn may complicate model optimisation and overstate their expected generalization ability. Our proposed chain of classifiers model with nested stacking has indeed better generalization for multi-label predictions than previous models such as that by Wang and colleagues (Wang et al., 2016), and demonstrates a superior performance compared to models based on sets of individual classifiers.

Our model can predict both acute (cardiac arrhythmias) and chronic (coronary artery disorders and heart failures) effects of drugs, based on clinical human responses collected *via* SIDER. Chronic drug-induced cardiac changes are often irreversible and their effect is delayed. Therefore their prediction poses a challenge, as they require long-term animal experimentation and the drug effect on the animal cardiac system is highly variable and hard to translate into the human clinic (Lamberti et al., 2014). Another key advantage of our approach is the use of perturbation databases such as LINCS and Connectivity map. They constitute great resources for preclinical applications (Musa et al., 2018) and even have been used extensively to identify novel drug candidates that confirmed their effectiveness experimentally (Han et al., 2018). Providing a cheaper alternative to animal models, computational methods that integrate transcriptional responses of drugs and their molecular characteristics, such as proposed, can be used prior to animal experiments identifying drug cardiotoxicity early in the preclinical phase.

Using transcriptional signatures, molecular descriptors and fingerprints, the general methodology proposed in this study could be further applied to other tissue-specific side effects and organs. We presume that our approach can be also extended to other areas including drug target prediction, where information about the multi-label properties of drugs or multi-target properties plays a vital role (Ramsay et al., 2018).

A current limitation of the database used in this study is the absence of isomers (drugs with similar chemical structure, and hence similar molecular descriptors and fingerprints) with different safety profiles. While we expect such chemicals to have different transcriptional profiles, and therefore to be discriminated by their transcriptional features, further analysis is required to test this hypothesis. Our model demonstrated good generalisation properties under completely new cell lines (**Figure 5B**). However, cross-platform differences in the acquisition of transcriptome data pose additional challenges for future use, and validation on an external dataset of RNAseq expression samples showed a moderate loss of prediction accuracy (**Figure 5B**). In addition, a limiting factor in the number of drugs used in this study was the availability of drug transcriptional profiles (rather than information on drug-induced cardiotoxicities), and the release of additional datasets would be a valuable resource for future studies. This study is also constrained by the feature space explored, with only 971 of landmark genes analyzed, seven molecular descriptors, and one type of fingerprints used for model contraction. Inclusion of information about expression values of other genes or more comprehensive descriptors and fingerprints might increase the predictive power of models and bring more insights. Machine-learning algorithms, however, are known to be limited in their ability to provide an interpretation of learnt associations. Mechanistic models coupled with machine-learning-based approaches represent an alternative attractive approach, with the potential to shed light on aspects of the underlying cardiotoxicity mechanisms (Lancaster and Sobie, 2016). Whereas these aspects fall beyond the goal of our study in presenting our proposed approach, they deserve future consideration in order to refine its predictive power, so it does comparison against other multi-target machine-learning algorithms.

## DATA AVAILABILITY STATEMENT

The full curated dataset, together with a dedicated R package for cardiotoxicity prediction, are freely available upon request.

## AUTHOR CONTRIBUTIONS

PM planned study, performed analysis and prepared the manuscript. AB-O and BR planned the study and prepared the manuscript.

## FUNDING

The authors are financially supported by a BHF Intermediate Basic Science Research Fellowship (to AB-O, FS/17/22/32644), a Wellcome Trust Senior Research Fellowship in Basic Biomedical Sciences (to BR, 100246/Z/12/Z and 214290/Z/18/Z) and the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 675451 and 823712 (CompBioMed project). BR and AB-O also acknowledge additional support from an Impact for Infrastructure Award of the National Centre for the Replacement, Refinement & Reduction of Animals in Research (NC/P001076/1). PM is funded by Insilico Medicine Hong Kong Ltd. The funders did not have a role in the study design, data collection, and decision to publish.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2020.00639/full#supplementary-material

## REFERENCES

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol. Pharm.* 13, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Banerjee, P., Dehnbostel, F. O., and Preissner, R. (2018). Prediction Is a Balancing Act: Importance of Sampling Methods to Balance Sensitivity and Specificity of Predictive Models Based on Imbalanced Chemical Data Sets. *Front. Chem.* 6, 362. doi: 10.3389/fchem.2018.00362

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing* 300, 70–79. doi: 10.1016/j.neucom.2017.11.077

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., et al. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discovery* 13, 419–431. doi: 10.1038/nrd4309

D'Amato, G., Luxán, G., and de la Pompa, J. L. (2016). Notch signalling in ventricular chamber development and cardiomyopathy. *FEBS J.* 283, 4223–4237. doi: 10.1111/febs.13773

Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., et al. (2015). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* 7, 60. doi: 10.1186/s13321-015-0109-z

Dutta, S., Chang, K. C., Beattie, K. A., Sheng, J., Tran, P. N., Wu, W. W., et al. (2017). Optimization of an In silico Cardiac Cell Model for Proarrhythmia Risk Assessment. *Front. Physiol.* 8, 616. doi: 10.3389/fphys.2017.00616

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Kim, H. K., et al. (2018). Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw* 33, 1–22. doi: 10.18637/jss.v033.i01

Greenwell, B., Boehmke, B., Cunningham, J.GBM Developers (2007). Generalized Boosted Models: A guide to the gbm package.

Guha, R., Charlop-Powers, Z., and Schymanski, E. Package "rcdk." 2018.

Han, H.-W., Hahn, S., Jeong, H. Y., Jee, J.-H., Nam, M.-O., Kim, H. K., et al. (2018). LINCS L1000 dataset-based repositioning of CGP-60474 as a highly potent anti-endotoxemic agent. *Sci. Rep.* 8, 14969. doi: 10.1038/s41598-018-33039-0

Huang, L. C., Wu, X., and Chen, J. Y. (2011). Predicting adverse side effects of drugs. *BMC Genomics* 12, S11. doi: 10.1186/1471-2164-12-S5-S11

Kendir, C., van den Akker, M., Vos, R., and Metsemakers, J. (2018). Cardiovascular disease patients have increased risk for comorbidity: A cross-sectional study in the Netherlands. *Eur. J. Gen. Pract.* 24, 45–50. doi: 10.1080/13814788.2017.1398318

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44, D1075–D1079. doi: 10.1093/nar/gkv1075

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *J. Stat. Software* 80 (7), 1–25. doi: 10.18637/jss.v080.i07

Lamberti, M., Giovane, G., Garzillo, E. M., Avino, F., Feola, A., Porto, S., et al. (2014). Animal models in studies of cardiotoxicity side effects from antiblastic drugs in patients and occupational exposed workers. *BioMed. Res. Int.* 2014, 240642. doi: 10.1155/2014/240642

Lancaster, M. C., and Sobie, E. A. (2016). Improved Prediction of Drug-Induced Torsades de Pointes Through Simulations of Dynamics and Machine Learning Algorithms. *Clin. Pharmacol. Ther. U. S.* 100, 371–379. doi: 10.1002/cpt.367

Lawrence, C. L., Pollard, C. E., Hammond, T. G., and Valentin, J.-P. (2008). In vitro models of proarrhythmia. *Br. J. Pharmacol.* 154, 1516–1522. doi: 10.1038/bjp.2008.195

Lawson, C. A., Solis-Trapala, I., Dahlstrom, U., Mamas, M., Jaarsma, T., Kadam, U. T., et al. (2018). Comorbidity health pathways in heart failure patients: A sequences-of-regressions analysis using cross-sectional data from 10,575 patients in the Swedish Heart Failure Registry. *PloS Med.* 15, e1002540. doi: 10.1371/journal.pmed.1002540

Lin, E., and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomark Res. BioMed. Central* 5, 2. doi: 10.1186/s40364-017-0082-y

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* 23, 1538–1546. doi: 10.1016/j.drudis.2018.05.010

Mak, I. W., Evaniew, N., and Ghert, M. (2014). Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* 6, 114–118.

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of Deep Learning in Biomedicine. *Mol. Pharm.* 13, 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982

Messinis, D. E., Melas, I. N., Hur, J., Varshney, N., Alexopoulos, L. G., and Bai, J. P. F. (2018). Translational systems pharmacology-based predictive assessment of drug-induced cardiomyopathy. *CPT Pharmacometrics Syst. Pharmacol.* 7, 166–174. doi: 10.1002/psp4.12272

Miller, D. J., Wang, Y., and Kesidis, G. (2008). Emergent unsupervised clustering paradigms with potential application to bioinformatics. *Front. Biosci. U. S.* 13, 677–690. doi: 10.2741/2711

Minasny, B., and McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Comput. Geosci.* 32, 1378–1388. doi: 10.1016/j.cageo.2005.12.009

Mladěnka, P., Applová, L., Patočka, J., Costa, V. M., Remiao, F., Pourová, J., et al. (2018). Comprehensive review of cardiovascular toxicity of drugs and related agents. *Med. Res. Rev.* 38, 1332–1403. doi: 10.1002/med.21476

Musa, A., Ghoraie, L. S., Zhang, S.-D., Glazko, G., Yli-Harja, O., Dehmer, M., et al. (2018). A review of connectivity map and computational approaches in pharmacogenomics. *Brief Bioinform.* 19, 506–523. doi: 10.1093/bib/bbw112

Naser, N., Dilic, M., Durak, A., Kulic, M., Pepic, E., Smajic, E., et al. (2017). The Impact of Risk Factors and Comorbidities on The Incidence of Atrial Fibrillation. *Mater. Socio. Med. ScopeMed. Int. Med. J. Manage. Indexing Syst.* 29, 231. doi: 10.5455/msm.2017.29.231-236

Novoselova, N., Wang, J., Pessler, F., and Klawonn, F. (2018). Biocomb: Feature Selection and Classification with the Embedded Validation Procedures for Biomedical Data Analysis. https://rdrr.io/cran/Biocomb/

Onakpoya, I. J., Heneghan, C. J., and Aronson, J. K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med. England* 14, 10. doi: 10.1186/s12916-016-0553-2

Passini, E., Britton, O. J., Lu, H. R., Rohrbacher, J., Hermans, A. N., Gallacher, D. J., et al. (2017). Human in silico drug trials demonstrate higher accuracy than animal models in predicting clinical pro-arrhythmic cardiotoxicity. *Front. Physiol. Front. Media S.A.* 8, 668. doi: 10.3389/fphys.2017.00668

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features, in: *32nd Conference of Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada: Curran Associates Inc. p. 6639–6649.

Ramsay, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E., and Bolognesi, M. L. (2018). A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7, 3. doi: 10.1186/s40169-017-0181-2

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* 12, 77. doi: 10.1186/1471-2105-12-77

Rodriguez, B., Carusi, A., Abi-Gerges, N., Ariga, R., Britton, O., Bub, G., et al. (2016). Human-based approaches to pharmacology and cardiology: an interdisciplinary and intersectorial workshop. *Europace* 18, 1287–1298. doi: 10.1093/europace/euv320

Senge, R., del Coz, J. J., and Hüllermeier, E. (2019). Rectifying Classifier Chains for Multi-Label Classification. *arXiv*.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. *Cell Press;* 171, 1437–1452.e17. doi: 10.1016/j.cell.2017.10.049

Torgo, L. (2013) *DMwR [Internet]. [cited 2019 Oct 11]*. Available from: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Troncoso, R., Ibarra, C., Vicencio, J. M., Jaimovich, E., and Lavandero, S. (2014). New insights into IGF-1 signaling in the heart. *Trends Endocrinol. Metab.* 25, 128–137. doi: 10.1016/j.tem.2013.12.002

Vicente, J., Zusterzeel, R., Johannesen, L., Mason, J., Sager, P., Patel, V., et al. (2018). Mechanistic Model-Informed Proarrhythmic Risk Assessment of Drugs: Review of the "CiPA" Initiative and Design of a Prospective Clinical Validation Study. *Clin. Pharmacol. Ther.* 103, 54–66. doi: 10.1002/cpt.896

Wang, Z., Clark, N. R., and Ma'ayan, A. (2016). Drug Induced Adverse Events Prediction with the LINCS L1000 Data. *Bioinformatics* 32, 2338–2345. doi: 10.1093/bioinformatics/btw168

Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., et al. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discovery* 14, 475. doi: 10.1038/nrd4609

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi: 10.1093/nar/gkm958

Wright, M. N., and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Software* 77 (1), 1–17. doi: 10.18637/jss.v077.i01

Xie, L., He, S., Zhang, Z., Lin, K., Bo, X., Yang, S., et al. (2018). Domain-Adversarial Multi-Task Framework for Novel Therapeutic Property Prediction of Compounds. *Bioinformatics* 36, 2848–2855. doi: 10.1093/bioinformatics/btaa063

Zhang, C., Zhou, Y., Gu, S., Wu, Z., Wu, W., Liu, C., et al. (2016). In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicol. Res.* 5, 570–582. doi: 10.1039/C5TX00294J

**Conflict of Interest:** PM is associated with Insilico Medicine Hong Kong Ltd, engaged in drug discovery research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

PERSPECTIVE
published: 30 June 2020
doi: 10.3389/fphar.2020.00770

# Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery

*Izumi V. Hinkson, Benjamin Madej and Eric A. Stahlberg* on behalf of the ATOM Consortium*

*Frederick National Laboratory for Cancer Research, Frederick, MD, United States*

Conventional drug discovery is long and costly, and suffers from high attrition rates, often leaving patients with limited or expensive treatment options. Recognizing the overwhelming need to accelerate this process and increase success, the ATOM consortium was formed by government, industry, and academic partners in October 2017. ATOM applies a team science and open-source approach to foster a paradigm shift in drug discovery. ATOM is developing and validating a precompetitive, preclinical, small molecule drug discovery platform that simultaneously optimizes pharmacokinetics, toxicity, protein-ligand interactions, systems-level models, molecular design, and novel compound generation. To achieve this, the ATOM Modeling Pipeline (AMPL) has been developed to enable advanced and emerging machine learning (ML) approaches to build models from diverse historical drug discovery data. This modular pipeline has been designed to couple with a generative algorithm that optimizes multiple parameters necessary for drug discovery. ATOM's approach is to consider the full pharmacology and therapeutic window of the drug concurrently, through computationally-driven design, thereby reducing the number of molecules that are selected for experimental validation. Here, we discuss the role of collaborative efforts such as consortia and public-private partnerships in accelerating cross disciplinary innovation and the development of open-source tools for drug discovery.

**Keywords: artificial intelligence, machine learning, drug discovery and development, data science *in silico* modeling**

OPEN ACCESS

***Edited by:***
*Alex Zhavoronkov,*
*Biogerontology Research Foundation,*
*United Kingdom*

***Reviewed by:***
*Evgeny Putin,*
*Insilico Medicine, Hong Kong*
*Niu Huang,*
*National Institute of Biological Sciences (NIBS), China*
*Evelyne Bischof,*
*Shanghai University of Medicine & Health Sciences, China*

****Correspondence:***
*Eric A. Stahlberg*
*eric.stahlberg@nih.gov*

***Specialty section:***
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

***Received:*** *08 October 2019*
***Accepted:*** *11 May 2020*
***Published:*** *30 June 2020*

***Citation:***
*Hinkson IV, Madej B and Stahlberg EA (2020) Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery. Front. Pharmacol. 11:770. doi: 10.3389/fphar.2020.00770*

## INTRODUCTION

Preclinical drug discovery typically takes five and a half years and accounts for about one third of the cost of drug development (Paul et al., 2010). The process is largely empirical with a sequential, iterative approach to optimizing key drug discovery parameters—efficacy, pharmacokinetics (PK), safety, and developability. Millions of molecules are tested, thousands are produced, and most fail to progress in preclinical or clinical settings (Shannon Decker and Atkinson, 2007; Mohs and Greig, 2017). Furthermore, translation from R&D to the clinic is insufficient with a success rate of less than 10%, and safety liabilities and poor efficacy cited as the main causes of attrition (Miller et al., 2017; Lowe, 2019).

Patients are waiting for the field of drug discovery to innovate new processes that will help improve the success rate of pharmaceutical development, lower drug costs, and get medicines to the clinic more quickly. With the average cost of developing a new molecular entity at over $2 billion, in large part due to the costs of failures, researchers are challenged to work outside the conventional slow, sequential, and costly drug development paradigm to better meet the urgent needs of patients (Kramer et al., 2007; Munos, 2009; Mullin, 2014; DiMasi et al., 2016). To increase the generation of successful new molecular entities, a number of groups have called for more innovation around the culture of and approach to drug discovery (Munos, 2006; Papadaki and Hirsch, 2013; Parekh et al., 2015). In particular, because so much of the cost of development stems from the cost of failures, approaches that improve our ability to distinguish early which molecules will ultimately succeed can have a disproportionate impact on improving the output of new medicines illustrate the potential for accelerating drug discovery through artificial intelligence (AI)-driven approaches (Ringel et al., 2013).

The demonstrations of ML for polypharmacological drug design, deep neural nets for predicting quantitative structure-activity relationships (QSAR), and generative molecular design through the use of variational autoencoders and generative adversarial networks (Besnard et al., 2012; Ma et al., 2015; Blaschke et al., 2018) hold great promise. To this end, significant interest has been raised in the application of approaches that combine AI, simulation, and experimentation to drug discovery (Vamathevan et al., 2019). Recognizing the compelling need for a paradigm shift in drug development, the ATOM consortium was established in October 2017[1]. ATOM's founders, the Frederick National Laboratory for Cancer Research (FNLCR, on behalf of the National Cancer Institute), Lawrence Livermore National Laboratory (LLNL, on behalf of the Department of Energy), GSK (GlaxoSmithKline), and the University of California, San Francisco (UCSF), have joined forces to leverage resources toward the common goal of benefiting patients. ATOM is applying an integrated approach to combine capabilities such as high-performance computing, human-relevant *in vitro* experimentation, data-driven and mechanistic modeling, and curation of pharmacological data toward the development of a novel preclinical drug discovery and development platform.

## Drug Discovery Consortia

As the complexity of biomedical research questions has increased, so too has the need to bring together expertise and resources from multiple disciplines and organizations (Cooke et al., 2015). Consequently, several articles by thought leaders have called for more collaboration in the drug development process (Altshuler et al., 2010; Dahlin et al., 2015; Alteri and Guizzaro, 2018; Takebe et al., 2018; Chaturvedula et al., 2019). Open innovation and open-source research strategies which emphasize the value of collaboration and use of both internal and external information, are creating the opportunity for the drug research and development industry to leverage know-how from across organizations (Munos, 2006; Hunter and Stephens, 2010; Owens, 2016). Cross-industry collaboration is particularly important in the application of computational approaches to drug discovery, where for instance, most companies have one or fewer drugs approved per year, far too small a sample size to support these approaches (Munos, 2009). The advantages of bringing together organizations into public-private partnerships (PPP) and consortia include not just scale, but also new-found agility and increased creativity alongside risk reduction and cost sharing (Papadaki and Hirsch, 2013; Slusher et al., 2013; Rosenberg, 2017; Kuchler, 2019). In fact, the US Food and Drug Administration (FDA), acknowledges the critical role of PPPs and consortia with respect to the innovation and modernization of medical product development (Maxfield et al., 2017).

One notable example of cross-sector collaboration is the Merk Molecular Activity Challenge[2] where the pharmaceutical company provided contestants with a training set of molecular descriptors and activities and a test set of descriptors only, and spurred the development of innovative ML methods for QSAR (Ma et al., 2015). In the last 2 years, new academic-industry consortia projects have emerged, focusing on applications of ML in drug discovery. The Machine Learning for Pharmaceutical Discovery and Synthesis Consortium, with membership from three Massachusetts Institute of Technology departments and several leading pharmaceutical companies, focuses on the application of ML to automate drug discovery and synthesis[3]. Summer 2019 saw the start of a new Innovative Medicines Initiative collaborative project led by Janssen, dubbed Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY)[4] (Kuchler, 2019). With a 3-year timeframe, the MELLODDY project focuses on employing federated ML to foster sharing data insights while preserving organizational intellectual property. Pharmaceutical industry participants will train models on their own proprietary data and share those models to increase the impact of AI and ML in the industry.

As an open consortium backed by major public entities, the Department of Energy, the National Cancer Institute, and the University of California Office of the President, as well as pharmaceutical leader GSK, the Accelerating Therapeutics for Opportunities in Medicine consortium (ATOM) is committed to creating new tools for drug discovery that can be shared broadly and benefit the public good. Computational approaches to drug design hold the potential to drastically improve the field's ability to generate novel drugs for patients in need. Harnessing advances in computational power and AI, ATOM is building a new, comprehensive, integrated platform for efficient molecular property prediction, optimization, and design. Drawing from team science, open innovation, and open-source concepts, the ATOM platform combines ML, simulation, and experimentation to generate novel drug candidates more rapidly than traditional approaches. ATOM's current scope focuses within the area of

[1] atomscience.org

[2] www.kaggle.com/c/MerckActivity/data

[3] news.mit.edu/2018/applying-machine-learning-to-challenges-in-pharmaceutical-industry-0517

[4] www.imi.europa.eu/projects-results/project-factsheets/melloddy

preclinical drug discovery, but its outcomes aim to benefit not only the member organizations and their immediate stakeholders, but the biomedical community at large including academicians, start-ups, private industry, clinicians, and patients.

## AI-Driven Drug Discovery

Drug discovery is relying increasingly on computational and AI-driven methods. Collaborative efforts that combine scientific know-how and computational power are being stood up to incubate innovative methods while sharing risk and accelerating progress. In the past decade significant advances have been made to accelerate the drug discovery process such as the development of computational and AI-based methods for virtual screening and *in silico* drug design. Moving beyond structure-based approaches and virtual screens, several seminal publications have demonstrated the use of generative adversarial networks and variational autoencoders for *de novo* drug design (Kadurin et al., 2017; Olivecrona et al., 2017; Gómez-Bombarelli et al., 2018; Merk et al., 2018; Polykovskiy et al., 2018; Putin et al., 2018; Segler et al., 2018; Ståhl et al., 2019; Hong et al., 2020). For example, a recently published deep generative model demonstrated the design of small-molecule drug candidates for discoidin domain receptor 1 prioritizing synthetic feasibility, efficacy, and uniqueness with respect to known small molecules, showcasing the ability to rapidly discover drugs at low cost (Zhavoronkov et al., 2019).

### Collaborative AI-Driven Drug Discovery at ATOM

The promise of AI-driven drug design carries with it, several challenges—the need for appropriate datasets, ability to generate and test evolving biological hypotheses, multi-parameter optimization, reduction in design-make-test-analyze cycle times, and adaptability of research culture (Schneider et al., 2020). ATOM is tackling these challenges through the collaborative development of a preclinical, open-source, small-molecule drug discovery platform (Chaturvedula et al., 2019). The initial stages have focused on building computational infrastructure, curating preclinical data from both GSK and public sources, and creating and testing data-driven modeling capabilities.

ATOM has developed a data-driven modeling pipeline capable of rapidly building and optimizing ML models for bioassay activity and molecular property predictions. This modeling pipeline is important for developing predictive models for public and private pharmaceutical assay datasets. While ML-based techniques to predict drug properties from structures are regularly used in the field of computational drug design, there remains a need for an automated modular pipeline for common modeling tasks. Some key features for such a software package are to enable reproducibility, incorporate new models, support a variety of chemical representations, allow for hyperparameter optimization, and validate predictive performance (Dahl et al., 2014; Gilmer et al., 2017; Feinberg et al., 2018; Yang et al., 2019).

Existing commercial pipeline tools such as BIOVIA Pipeline Pilot are limited in their customizability and can be cost prohibitive to small academic research groups and start-up companies[5]. On the other end of the spectrum, open-source pipeline tools such as KNIME are useful as GUI-based platforms for data processing, model fitting, and analysis, (Berthold, 2008) but have yet to demonstrate the suitability for large scale model generation.

### The ATOM Modeling Pipeline (AMPL)

AMPL[6], or the ATOM Modeling Pipeline, extends the popular DeepChem[7] library and supports ML and molecular featurization tools (Minnich et al., 2020). AMPL is implemented as a Python library that integrates with existing data science ecosystems and utilities. AMPL automates and optimizes many common ML model fitting tasks that are performed for pharmaceutical datasets including model fitting, validation, and prediction. AMPL allows researchers to reproducibly train and test models, incorporate new models, and provide utilities for automated dataset characterization, model validation, and uncertainty quantification. AMPL is designed to be a versatile library that can interface with many services and tools.

AMPL allows users to build *in silico* models based on molecular properties to aid in drug discovery. With an initial focus on safety and pharmacokinetic modeling, AMPL has been extensively tested on activity and property assay datasets. In preparation for the initial release of the pipeline, 11,552 regression and classification models were built to evaluate data splitting algorithms, model types, and feature types (Minnich et al., 2020). AMPL supports a wide variety of dataset splitting algorithms for validation and testing, including random splits, Butina clustering, scaffold splits, and temporal splits. AMPL uses models from scikit-learn and DeepChem including random forest, XGBoost, fully connected neural network, and graph convolution neural network models. Small molecules were represented as SMILES strings using the RDKit cheminformatics library and the molecule validation and standardization tool, MolVS. AMPL's data curation module was applied to datasets to filter out compound assay values with wide variability, and to characterize the datasets with Tanimoto distances between chemical fingerprints or Euclidean distances between descriptor feature vectors. Several featurization approaches were compared including Extended Connectivity Fingerprints (ECFP), DeepChem graph convolution latent vectors, Mordred chemical descriptors, and Molecular Operating Environment (MOE) descriptors. Due to the modular nature of AMPL's implementation, extensions to the pipeline are available for additional splitting algorithms, model types, and feature types.

Hyperparameter optimization is an important task for cheminformatics ML model fitting that may improve model predictive performance. AMPL supports basic hyperparameter optimization functions including searches using basic linear grids, logistic grids, random searches, and user-specified searches. Model fitting for safety and pharmacokinetic parameters used AMPL's hyperparameter optimization module to explore model parameter combinations. Generally, hyperparameter optimization improved predictive performance on properties of external test sets except for certain cases with limited data or ECFP featurization.

AMPL automatically calculates standard model performance metrics for regression and classification models. The regression

[5] www.3dsbiovia.com/products/ collaborative-science/ biovia-pipeline-pilot/

[6] github.com/ATOMconsortium/AMPL

[7] github.com/deepchem/deepchem

performance statistics include $R^2$, mean absolute error, and mean square error to evaluate the level of agreement between the model predicted values and actual experimental ground truth values. AMPL also includes classification performance metrics such as precision and recall, area under the precision-recall curve (PRC-AUC), negative predictive value, cross entropy, and accuracy metrics. As previously described, model prediction uncertainty was calculated for several of PK datasets for comparison with model prediction error (Minnich et al., 2020). AMPL enables this type of uncertainty quantification analysis toward better understanding model predictions, uncertainty, and error.

AMPL is open-source, modular, and flexible, allowing for additions or extensions as needed. This makes data-driven modeling using modern ML libraries accessible to the wider scientific community including academic or government laboratories and small companies. AMPL is now available for download on Github[8]. The website includes detailed library documentation as well as example Jupyter notebooks to learn to use the pipeline.

### AMPL Validation

Bioassay data, specifically the half-maximal effective drug concentration (EC50), and the half-maximal inhibitory drug concentration (IC50), of known hepatic, central nervous system, cardiovascular, and cellular toxicity safety liabilities were used to benchmark safety models. Models were fit for assays such as BSEP, β2 adrenoceptor, muscarinic acetylcholine receptor, dopamine D2, voltage-gated potassium channels, and phospholipidosis induction. For each assay type, model hyperparameters were optimized resulting in 2,130 classification models with thresholds appropriate set for each assay. As described by Minnich et al, the predictive performance of the classification models was evaluated using common validation statistics including receiver operating characteristic area under the curve (ROC AUCs) built on safety datasets. Predictive performance varied based on assay type, dataset size, dataset split type, feature type, and model type, but overall produced many useful models for pharmaceutical safety properties (Minnich et al., 2020).

A diverse set of pharmacokinetic data including blood-to-plasma ratio, plasma protein binding, *in vivo* clearance, volume of distribution, hepatocyte clearance, and microsomal clearance, logD was used to fit predictive models with AMPL (Minnich et al., 2020). Nine thousand four hundred twenty-two regression models were fit for all the assay types and corresponding model parameters were evaluated for improvements to predictive performance as described by Minnich et al. General trends between different training and test splits, feature types, and model types were examined. When using neural network models with calculated descriptors for many of these PK datasets, model predictions with MOE descriptors were slightly better than predictions with open-source Mordred descriptors. Several PK datasets with larger numbers of measurements (10,000 or more) benefitted from DeepChem's graph convolutional neural network models with better predictions compared to experiment than ECFP or calculated descriptors. For smaller PK datasets, random forest models with MOE descriptors had slightly better performance than other feature and model combinations (Minnich et al., 2020).

AMPL is designed to automatically and rapidly build and evaluate cheminformatics models. Automation of deep learning model training, parallelized hyperparameter search, performance benchmarking, and data and model storage are essential for reproducible ML predictions in drug discovery. Given the wide range of activity and property assay types, the validation performed by Minnich et al. demonstrate there is no single best model fitting approach for every dataset. This underscores the need to rapidly search and fit predictive models for new datasets enabled by the AMPL software suite.

Two examples of model fitting on publicly accessible datasets are available with the AMPL repository. Each example describes a general method of curating datasets, fitting a ML model, and using the created model for new predictions. Example code is included to download the datasets from their original source, perform basic curation on the datasets, train a model on the curated datasets, and then load the fitted model for prediction on a withheld test set. In the first example, AMPL mimicked a DeepChem example model by fitting a model to a public aqueous solubility dataset using DeepChem's graph convolutional neural network model (Delaney, 2004). In a second example, AMPL was used to fit a predictive neural network model using Mordred descriptors for human liver microsomal clearance from a public PK dataset (Wenzel et al., 2019). The entire process of data curation to analysis and visualization for these sample datasets is automated and reproducible with the AMPL library and tools.

AMPL models can be applied toward related compounds to rapidly predict bioassay activity or safety and pharmacokinetic properties. In the context of ATOM, AMPL is a key component in the overall mission to accelerate the drug discovery process.

## CONCLUSIONS

Given heavy reliance on expensive and lengthy experimentation, the field of drug discovery is increasingly integrating both computational and AI-driven methods for virtual screening and *in silico* drug design. Further, the application of deep neural network architectures in generative design in conjunction with data-driven and mechanistic modeling for functional property prediction and an *in silico* framework for rapid lead optimization will drastically change how drug discovery is done.

Collaborative efforts have been employed in recent efforts to develop new capabilities where risks and required investment have been high. ATOM provides an avenue for collaborative AI-driven drug discovery that results in an open-source framework that broadens availability and an opportunity to raise the level of collaborative drug discovery efforts.

The AMPL serves as the initial step toward the development of an open-source preclinical drug design platform that will accelerate the process of getting more effective therapies to patients. Future efforts involve extending the modeling capability of AMPL toward the development of an open-source pre-clinical drug discovery platform (**Figure 1**).

[8] github.com/ATOMconsortium/AMPL

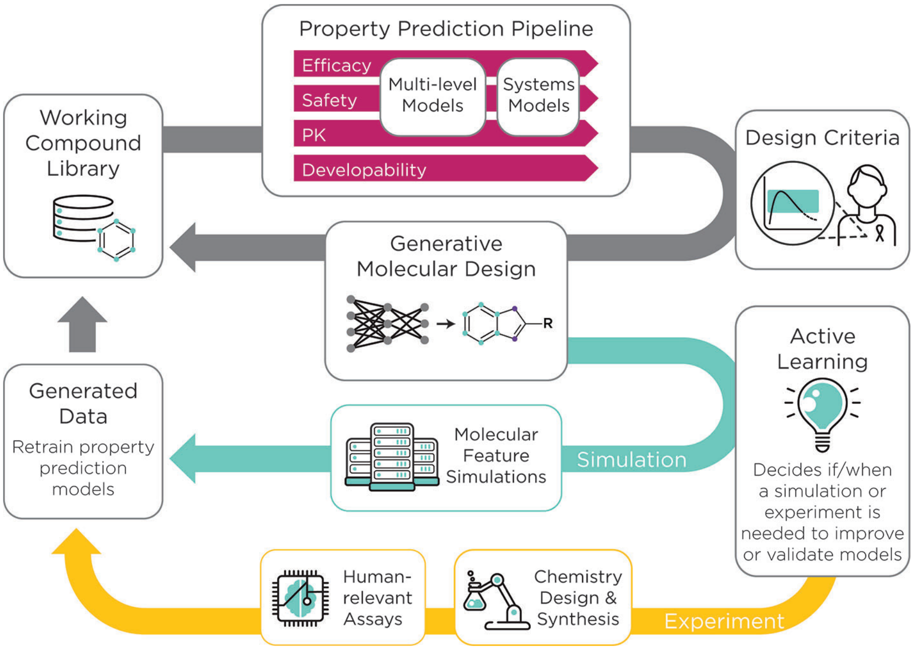

Property Prediction Pipeline
Efficacy
Safety
PK
Developability
Multi-level Models
Systems Models
Working Compound Library
Design Criteria
Generative Molecular Design
R
Active Learning
Decides if/when a simulation or experiment is needed to improve or validate models
Generated Data
Retrain property prediction models
Molecular Feature Simulations
Simulation
Human-relevant Assays
Chemistry Design & Synthesis
Experiment

**FIGURE 1** | The ATOM preclinical drug discovery workflow. ATOM is developing an active learning drug discovery framework that uses a compound library as input to a property prediction pipeline. The pipeline begins with historic data collected on a working compound library to train machine learning-based models for property prediction. Next, multi-level and systems-level models of efficacy, safety, and pharmacokinetics as well as developability are integrated to generate a set of drug design criteria. These parameters are simultaneously optimized for the generation of novel molecules by the generative molecular design framework. The multi-parameter optimization loop, in grey, can be run for numerous cycles. An active learning approach is used to decide whether a molecular simulation or experiment is needed to improve or validate the models. Data that result from these simulations and experiments are then used to re-train the property prediction models. The result of this workflow is a set of optimized drug candidates.

## Future Efforts

At ATOM, efforts are underway to integrate current and emerging computational capabilities with active learning in an AI-driven platform. ATOM is creating a generative molecular design framework that integrates predictive models from AMPL and initiates cycles of generative molecular design and multiparameter optimization. The goal of ATOM's generative molecular design framework is to propose novel small-molecule drug candidates with optimized properties based on design criteria such as potency, selectivity, cardiotoxicity, hepatoxicity, solubility, clearance, and synthetic accessibility[9]. New experimental and molecular simulation data will be selectively acquired to support the ML-based approach and will be integrated into the computational pipeline to kick start additional cycles of the molecular design and optimization. The integration of active learning will streamline time-consuming and costly experimentation and will guide the design of novel drug candidates (**Figure 1**). Collectively, these efforts usher in a paradigm shift in drug discovery that emphasizes collaboration, innovation, and the development of open-source tools.

## AUTHOR CONTRIBUTIONS

IH: manuscript writing and figure design. BM: manuscript writing. ES and ATOM consortium: manuscript and figure revision, approval of final manuscript.

## FUNDING

This project has been funded in whole or in part with funds from University of California, San Francisco, funds from GlaxoSmithKline, federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E, and federal funds from U.S. Department of Energy, National Nuclear Security Administration under contract DE-AC52-07NA27344. The federal funds include the Cancer Moonshot funds that Congress passed the 21st Century Cures Act in December 2016, authorizing $1.8 billion in funding for the Cancer Moonshot over 7 years. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services nor the Department of Energy; nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical

[9] atomscience.org/abstracts-and-presentations/2019/9/25/generative-lead-optimization-of-de-novo-molecules-case-study-in-discovery-of-potent-selective-aurora-kinase-inhibitors-with-favorable-secondary-pharmacology

Research, Inc., for the National Cancer Institute. Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the Department of Energy, National Nuclear Security Administration.

## ACKNOWLEDGMENTS

We express thanks to ATOM team members John Baldoni, Frank Blanchard, Stacie Calad-Thomson, Ethan Dmitrovsky, Leonard Freedman, Mary Ellen Hackett, Deirdre Olynick, Michael Ringel, Tom Rush, and Jeremy Thomas for their advice on the manuscript. We thank the ATOM team for their research, technical, and partnership contributions to the projects described within. We also thank Frederick National Laboratory for Cancer Research, GlaxoSmithKline, Lawrence Livermore National Laboratory, University of California San Francisco, the National Cancer Institute, the Department of Energy, and the University of California Office of the President for guidance and consortium support.

## REFERENCES

Alteri, E., and Guizzaro, L. (2018). Be open about drug failures to speed up research. *Nature* 563 (7731), 317–319. doi: 10.1038/d41586-018-07352-7

Altshuler, J. S., Balogh, E., Barker, A. D., Eck, S. L., Friend, S. H., Ginsburg, G. S., et al. (2010). Opening up to precompetitive collaboration. *Sci. Transl. Med.* 2 (52), 52cm26. doi: 10.1126/scitranslmed.3001515

Berthold, M. R. E. A. (2008). "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*. Eds. C. B. H. Preisach, L. Schmidt-Thieme and R. Decker (Berlin, Heidelberg: Springer).

Besnard, J., Ruda, G. F., Setola, V., Abecassis, K., Rodriguiz, R. M., Huang, X.-P., et al. (2012). Automated design of ligands to polypharmacological profiles. *Nature* 492 (7428), 215–220. doi: 10.1038/nature11691

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* 37 (1-2), 1700123. doi: 10.1002/minf.201700123

Chaturvedula, A., Calad-Thomson, S., Liu, C., Sale, M., Gattu, N., and Goyal, N. (2019). Artificial Intelligence and Pharmacometrics: Time to Embrace, Capitalize, and Advance? *CPT Pharmacometr. Syst. Pharmacol.* 8 (7), 440–443. doi: 10.1002/psp4.12418

Cooke, N. J., Hilton, M. L.National Research Council (U.S.) and Committee on the Science of Team Science (2015). *"Enhancing the effectiveness of team science"* (Washington, D.C: The National Academies Press).

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task Neural Networks for QSAR Predictions. *ArXiv E-prints*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2014arXiv1406.1231D [Accessed June 01, 2014].

Dahlin, J. L., Inglese, J., and Walters, M. A. (2015). Mitigating risk in academic preclinical drug discovery. *Nat. Rev. Drug Discovery* 14 (4), 279–294. doi: 10.1038/nrd4578

Delaney, J. S. (2004). ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* 44 (3), 1000–1005. doi: 10.1021/ci034243x

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* 47, 20–33. doi: 10.1016/j.jhealeco.2016.01.012

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* 4 (11), 1520–1530. doi: 10.1021/acscentsci.8b00507

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). *Neural message passing for Quantum chemistry* (Sydney, NSW, Australia: JMLR.org).

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* 4 (2), 268–276. doi: 10.1021/acscentsci.7b00572

Hong, S. H., Ryu, S., Lim, J., and Kim, W. Y. (2020). Molecular Generative Model Based on an Adversarially Regularized Autoencoder. *J. Chem. Inf. Model* 60 (1), 29–36. doi: 10.1021/acs.jcim.9b00694

Hunter, J., and Stephens, S. (2010). Is open innovation the way forward for big pharma? *Nat. Rev. Drug Discovery* 9 (2), 87–88. doi: 10.1038/nrd3099

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8 (7), 10883–10890. doi: 10.18632/oncotarget.14073

Kramer, J. A., Sagartz, J. E., and Morris, D. L. (2007). The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat. Rev. Drug Discovery* 6 (8), 636–649. doi: 10.1038/nrd2378

Kuchler, H. (2019). Pharma groups combine to promote drug discovery with AI. *Financial Times June* 4, 2019.

Lowe, D. (2019). The Latest on Drug Failure and Approval Rates. *In The Pipeline* [Online]. Available from: https://blogs.sciencemag.org/pipeline/archives/2019/05/09/the-latest-on-drug-failure-and-approval-rates 2019].

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* 55 (2), 263–274. doi: 10.1021/ci500747n

Maxfield, K. E., Buckman-Garner, S., and Parekh, A. (2017). The Role of Public-Private Partnerships in Catalyzing the Critical Path. *Clin. Transl. Sci.* 10 (6), 431–442. doi: 10.1111/cts.12488

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* 37 (1-2), 1700153. doi: 10.1002/minf.201700153

Miller, S. M., Moos, W. H., Munk, B. H., and Munk, S. A. (2017). "10 - Drug discovery: Chaos can be your friend or your enemy," in *Managing the Drug Discovery Process*. Eds. W. H. Moos, S. M. Miller, B. H. Munk and S. A. Munk (Woodhead Publishing), 183–279. doi: 10.1016/B978-0-08-100625-2.00010-6

Minnich, A. J., McLoughlin, K., Tse, M., Deng, J., Weber, A., Murad, N., et al. (2020). AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* 60 (4), 1955–1968. doi: 10.1021/acs.jcim.9b01053

Mohs, R. C., and Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimers Dement (N. Y.)* 3 (4), 651–657. doi: 10.1016/j.trci.2017.10.005

Mullin, R. (2014). Tufts Study Finds Big Rise In Cost Of Drug Development. *Chem. Eng. News*. Available: https://cen.acs.org/articles/92/web/2014/11/Tufts-Study-Finds-Big-Rise.html [Accessed 2019].

Munos, B. (2006). Can open-source R&D reinvigorate drug research? *Nat. Rev. Drug Discovery* 5 (9), 723–729. doi: 10.1038/nrd2131

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery* 8 (12), 959–968. doi: 10.1038/nrd2961

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminformat.* 9 (1), 48. doi: 10.1186/s13321-017-0235-x

Owens, B. (2016). Data sharing: Access all areas. *Nature* 533, S71. doi: 10.1038/533S71a

Papadaki, M., and Hirsch, G. (2013). Curing consortium fatigue. *Sci. Transl. Med.* 5 (200), 200fs235. doi: 10.1126/scitranslmed.3006903

Parekh, A., Buckman-Garner, S., McCune, S., R. O. N. , Geanacopoulos, M., Amur, S., et al. (2015). Catalyzing the Critical Path Initiative: FDA's progress in drug development activities. *Clin. Pharmacol. Ther.* 97 (3), 221–233. doi: 10.1002/cpt.42

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* 9 (3), 203–214. doi: 10.1038/nrd3078

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., et al. (2018). Entangled Conditional Adversarial

Autoencoder for de Novo Drug Discovery. *Mol. Pharmaceut.* 15 (10), 4398–4405. doi: 10.1021/acs.molpharmaceut.8b00839

Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sanchez-Lengeling, B., Aspuru-Guzik, A., et al. (2018). Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* 58 (6), 1194–1204. doi: 10.1021/acs.jcim.7b00690

Ringel, M., Tollman, P., Hersch, G., and Schulze, U. (2013). Does size matter in R&D productivity? If not, what does? *Nat. Rev. Drug Discovery* 12, 901. doi: 10.1038/nrd4164

Rosenberg, A. (2017). *"UC launches drug discovery consortium"*. (Los Angeles, CA: University of California Newsroom).

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., et al. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery* 19 (5), 353–364. doi: 10.1038/s41573-019-0050-3

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* 4 (1), 120–131. doi: 10.1021/acscentsci.7b00512

Shannon Decker, E. A. S., and Atkinson, D. R. A. J. (2007). "Chapter 28 - Drug Discovery", in *Principles of Clinical Pharmacology*, *2nd ed.* Eds. C. E. Daniels, R. L. Dedrick and S. P. Markey (Academic Press), 439–447. doi: 10.1016/B978-012369417-1/50068-7

Slusher, B. S., Conn, P. J., Frye, S., Glicksman, M., and Arkin, M. (2013). Bringing together the academic drug discovery community. *Nat. Rev. Drug Discovery* 12 (11), 811–812. doi: 10.1038/nrd4155

Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., and Boström, J. (2019). Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. *J. Chem. Inf. Model.* 59 (7), 3166–3176. doi: 10.1021/acs.jcim.9b00325

Takebe, T., Imai, R., and Ono, S. (2018). The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. *Clin. Transl. Sci.* 11 (6), 597–606. doi: 10.1111/cts.12577

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* 18 (6), 463–477. doi: 10.1038/s41573-019-0024-5

Wenzel, J., Matter, H., and Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* 59 (3), 1253–1268. doi: 10.1021/acs.jcim.8b00785

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* 59 (8), 3370–3388. doi: 10.1021/acs.jcim.9b00237

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37 (9), 1038–1040. doi: 10.1038/s41587-019-0224-x

**Conflict of Interest:** The authors declare that this study received funding from GSK. The funder had the following involvement with the study: GSK is a founding member of ATOM and has contributed to R&D and consortium activities at ATOM.

TECHNOLOGY AND CODE
published: 18 December 2020
doi: 10.3389/fphar.2020.565644

# Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models

*Daniil Polykovskiy*[1]*, *Alexander Zhebrak*[1], *Benjamin Sanchez-Lengeling*[2], *Sergey Golovanov*[3], *Oktai Tatanov*[3], *Stanislav Belyaev*[3], *Rauf Kurbanov*[3], *Aleksey Artamonov*[3], *Vladimir Aladinskiy*[1], *Mark Veselov*[1], *Artur Kadurin*[1], *Simon Johansson*[4], *Hongming Chen*[4], *Sergey Nikolenko*[1,3,5]*, *Alán Aspuru-Guzik*[6,7,8,9]* *and Alex Zhavoronkov*[1]*

[1]*Insilico Medicine Hong Kong Ltd., Pak Shek Kok, Hong Kong,* [2]*Chemistry and Chemical Biology Department, Harvard University, Cambridge, MA, United States,* [3]*Neuromation OU, Tallinn, Estonia,* [4]*Molecular AI, DiscoverySciences, R&D, AstraZeneca, Gothenburg, Sweden,* [5]*Computer Science Department, National Research University Higher School of Economics, St. Petersburg, Russia,* [6]*Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, ON, Canada,* [7]*Department of Computer Science, University of Toronto, Toronto, ON, Canada,* [8]*CIFAR AI Chair, Vector Institute for Artificial Intelligence, Toronto, ON, Canada,* [9]*Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada*

OPEN ACCESS

**Edited by:**
*Jianxun Ding,*
*Chinese Academy of Sciences, China*

**Reviewed by:**
*Nazareno Paolocci,*
*Johns Hopkins University,*
*United States*
*Felix Zhou,*
*University of Oxford, United Kingdom*

***Correspondence:**
*Daniil Polykovskiy*
*daniil@insilico.com*
*Alex Zhavoronkov*
*alex@insilico.com*
*Alán Aspuru-Guzik*
*alan@aspuru.com*
*Sergey Nikolenko*
*snikolenko@gmail.com*

**Specialty section:**
*This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology*

**Received:** *25 May 2020*
**Accepted:** *26 October 2020*
**Published:** *18 December 2020*

**Citation:**
*Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M, Kadurin A, Johansson S, Chen H, Nikolenko S, Aspuru-Guzik A and Zhavoronkov A (2020) Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. Front. Pharmacol. 11:565644. doi: 10.3389/fphar.2020.565644*

Generative models are becoming a tool of choice for exploring the molecular space. These models learn on a large training dataset and produce novel molecular structures with similar properties. Generated structures can be utilized for virtual screening or training semi-supervized predictive models in the downstream tasks. While there are plenty of generative models, it is unclear how to compare and rank them. In this work, we introduce a benchmarking platform called Molecular Sets (MOSES) to standardize training and comparison of molecular generative models. MOSES provides training and testing datasets, and a set of metrics to evaluate the quality and diversity of generated structures. We have implemented and compared several molecular generation models and suggest to use our results as reference points for further advancements in generative chemistry research. The platform and source code are available at https://github.com/molecularsets/moses.

**Keywords: generative models, drug discovery, deep learning, benchmark, distribution learning**

## INTRODUCTION

The discovery of new molecules for drugs and materials can bring enormous societal and technological progress, potentially curing rare diseases and providing a pathway for personalized precision medicine (Lee et al., 2018). However, complete exploration of the huge space of potential chemicals is computationally intractable; it has been estimated that the number of pharmacologically-sensible molecules is in the order of $10^{23}$ to $10^{80}$ compounds (Kirkpatrick and Ellis, 2004; Reymond, 2015). Often, this search is constrained based on already discovered structures and desired qualities such as solubility or toxicity. There have been many approaches to exploring the chemical space *in silico* and *in vitro*, including high throughput screening, combinatorial libraries, and evolutionary algorithms (Hu et al., 2009; Curtarolo et al., 2013; Pyzer-Knapp et al., 2015; Le and Winkler, 2016). Recent works demonstrated that machine learning methods can produce new small molecules (Merk et al., 2018a; Merk et al., 2018b; Polykovskiy et al., 2018b; Zhavoronkov et al., 2019a) and peptides (Grisoni et al., 2018) showing biological activity.

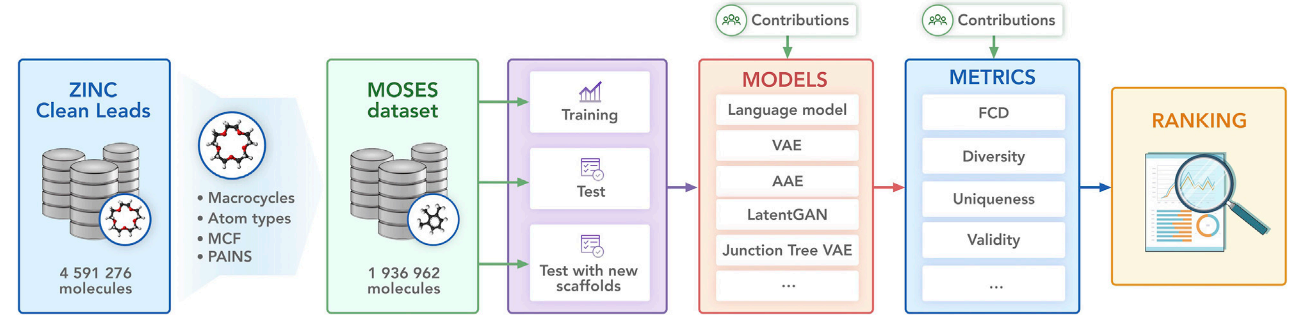


**FIGURE 1 |** Molecular Sets (MOSES) pipeline. The open-source library provides a dataset, baseline models, and evaluation metrics.

Over the last few years, advances in machine learning, and especially in deep learning, have driven the design of new computational systems for modeling increasingly complex phenomena. One approach that has been proven fruitful for modeling molecular data is deep generative models. Deep generative models have found applications in a wide range of settings, from generating synthetic images (Karras et al., 2018) and natural language texts (Yu et al., 2017), to the applications in biomedicine, including the design of DNA sequences (Killoran et al., 2017), and aging research (Zhavoronkov et al., 2019b). One important field of application for deep generative models lies in the inverse design of drug compounds (Sanchez-Lengeling and Aspuru-Guzik, 2018) for a given functionality (solubility, ease of synthesis, toxicity). Deep learning also found other applications in biomedicine (Mamoshina et al., 2016; Ching et al., 2018), including target identification (Mamoshina et al., 2018), antibacterial drug discovery (Ivanenkov et al., 2019), and drug repurposing (Aliper et al., 2016; Vanhaelen et al., 2017).

Part of the success of deep learning in different fields has been driven by ever-growing availability of large datasets and standard benchmark sets. These sets serve as a common measuring stick for newly developed models and optimization strategies (LeCun et al., 1998; Deng et al., 2009). In the context of organic molecules, MoleculeNet (Wu et al., 2018) was introduced as a standardized benchmark suite for regression and classification tasks. Brown et al. (2019) proposed to evaluate generative models on goal-oriented and distribution learning tasks with a focus on the former. We focus on standardizing metrics and data for the distribution learning problem that we introduce below.

In this work, we provide a benchmark suite—Molecular Sets (MOSES)—for molecular generation: a standardized dataset, data preprocessing utilities, evaluation metrics, and molecular generation models. We hope that our platform will serve as a clear and unified testbed for current and future generative models. We illustrate the main components of MOSES in **Figure 1**.

## Distribution Learning

In MOSES, we study distribution learning models. Formally, given a set of training samples $X_{\mathrm{tr}} = \{x_1^{\mathrm{tr}}, \ldots, x_N^{\mathrm{tr}}\}$ from an unknown distribution $p(x)$, distribution learning models approximate $p(x)$ with some distribution $q(x)$.

Distribution learning models are mainly used for building virtual libraries (van Hilten et al., 2019) for computer-assisted drug discovery. While imposing simple rule-based restrictions on a virtual library (such as maximum or minimum weight) is straightforward, it is unclear how to apply implicit or soft restrictions on the library. For example, a medicinal chemist might expect certain substructures to be more prevalent in generated structures. Relying on a set of manually or automatically selected compounds, distribution learning models produce a larger dataset, preserving implicit rules from the dataset. Another application of distribution learning models is extending the training set for downstream semi-supervized predictive tasks: one can add new unlabeled data by sampling compounds from a generative model.

The quality of a distribution learning model is a deviation measure between $p(x)$ and $q(x)$. The model can define a probability mass function $q(x)$ implicitly or explicitly. Explicit models such as Hidden Markov Models, n-gram language models, or normalizing flows (Dinh et al., 2017; Shi et al., 2019) can analytically compute $q(x)$ and sample from it. Implicit models, such as variational autoencoders, adversarial autoencoders, or generative adversarial networks (Kadurin et al., 2016; De Cao and Kipf, 2018; Gómez-Bombarelli et al., 2018) can sample from $q(x)$, but can not compute the exact values of the probability mass function. To compare both kinds of models, evaluation metrics considered in this paper depend only on samples from $q(x)$.

## Molecular Representations

In this section, we discuss different approaches to representing a molecule in a machine learning-friendly way (**Figure 2**): string and graph representations.

**String representations.** Representing a molecular structure as a string have been quickly adopted (Jaques et al., 2016; Guimaraes et al., 2017; Kadurin et al., 2017; Olivecrona et al., 2017; Yang et al., 2017; Kang and Cho, 2018; Popova et al., 2018; Putin et al., 2018; Segler et al., 2018) for generative models due to the abundance of sequence modeling tools such as recurrent neural networks, attention mechanisms, and dilated convolutions. Simplified molecular input line entry system (SMILES) (Weininger, 1988) is the most widely used string representation for generative machine learning models.

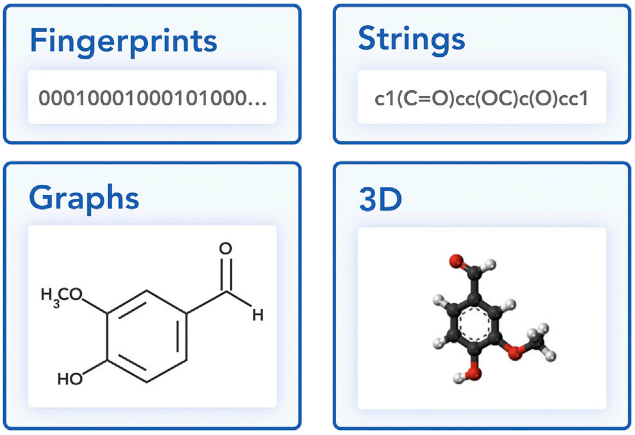


**FIGURE 2** | Different views on a vanillin molecule.

SMILES algorithm traverses a spanning tree of a molecular graph in depth-first order and stores atom and edge tokens. SMILES also uses special tokens for branching and edges not covered with a spanning tree. Note that since a molecule can have multiple spanning trees, different SMILES strings can represent a single molecule. While there is a canonicalization procedure to uniquely construct a SMILES string from a molecule (Weininger et al., 1989), ambiguity of SMILES can also serve as augmentation and improve generative models (Arús-Pous et al., 2019).

DeepSMILES (O'Boyle and Dalke, 2018) was introduced as an extension of SMILES that seeks to reduce invalid sequences by altering syntax for branches and ring closures. Some methods try to incorporate SMILES syntax into a network architecture to increase the fraction of valid molecules (Kusner et al., 2017; Dai et al., 2018). SELFIES (Krenn et al., 2019) defines a new syntax based on a Chomsky type-2 grammar augmented with self-referencing functions. International Chemical Identifier (InChI) (Stein et al., 2003) is a more verbose string representation which explicitly specifies a chemical formula, atoms' charges, hydrogens, and isotopes. However, Gómez-Bombarelli et al. (2018) reported that InChI-based models perform substantially worse than SMILES-based models in generative modeling—presumably due to a more complex syntax.

**Molecular graphs.** Graph representations have long been used in chemoinformatics for storing and processing molecular data. In a molecular graph, each node corresponds to an atom and each edge corresponds to a bond. Such graph can specify hydrogens either explicitly or implicitly. In the latter case, the number of hydrogens can be deduced from atoms' valencies.

Classical machine learning methods mostly utilize molecular descriptors extracted from such graphs. Deep learning models, however, can learn from graphs directly with models such as Graph Convolutional Networks (Duvenaud et al., 2015), Weave Networks (Wu et al., 2018), and Message Passing Networks (Gilmer et al., 2017). Molecular graph can also be represented as adjacency matrix and node feature matrix; this approach has been successfully employed in the MolGAN model (De Cao and Kipf, 2018) for the QM9 dataset (Ramakrishnan et al., 2014). Other approaches such as Junction Tree VAE (Jin et al., 2018) process molecules in terms of their subgraphs.

## Metrics

In this section, we propose a set of metrics to assess the quality of generative models. The proposed metrics detect common issues in generative models such as overfitting, imbalance of frequent structures or mode collapse. Each metric depends on a generated set $G$ and a test (reference) set $R$. We compute all metrics (except for validity) only for valid molecules from the generated set. We suggest generating 30, 000 molecules and obtaining $G$ as valid molecules from this set.

**Fraction of valid (Valid) and unique (Unique@k)** molecules report validity and uniqueness of the generated SMILES strings. We define validity using RDKit's molecular structure parser that checks atoms' valency and consistency of bonds in aromatic rings. In the experiments, we compute Unique@$K$ and for the first $K = 1,000$ and $K = 10,000$ valid molecules in the generated set. If the number of valid molecules is less than $K$, we compute uniqueness on all valid molecules. Validity measures how well the model captures explicit chemical constraints such as proper valence. Uniqueness checks that the model does not collapse to producing only a few typical molecules.

**Novelty** is the fraction of the generated molecules that are not present in the training set. Low novelty indicates overfitting.

**Filters** is the fraction of generated molecules that pass filters applied during dataset construction (see **Section 5**). While the generated molecules are often chemically valid, they may contain unwanted fragments: when constructing the training dataset, we removed molecules with such fragments and expect the models to avoid producing them.

**Fragment similarity (Frag)** compares distributions of BRICS fragments (Degen et al., 2008) in generated and reference sets. Denoting $c_f(A)$ a number of times a substructure $f$ appears in molecules from set $A$, and a set of fragments that appear in either $G$ or $R$ as $F$, the metric is defined as a cosine similarity:

$$\mathrm{Frag}(G,R) = \frac{\sum\limits_{f\in F}\left[c_f(G)\cdot c_f(R)\right]}{\sqrt{\sum\limits_{f\in F} c_f^2(G)}\sqrt{\sum\limits_{f\in F} c_f^2(R)}}. \qquad (1)$$

If molecules in both sets have similar fragments, Frag metric is large. If some fragments are over- or underrepresented (or never appear) in the generated set, the metric will be lower. Limits of this metric are [0,1].

**Scaffold similarity (Scaff)** is similar to fragment similarity metric, but instead of fragments we compare frequencies of Bemis–Murcko scaffolds (Bemis and Murcko, 1996). Bemis–Murcko scaffold contains all molecule's ring structures and linker fragments connecting rings. We use RDKit implementation of this algorithm which additionally considers carbonyl groups attached to rings as part of a scaffold. Denoting $c_s(A)$ a number of times a scaffold $s$ appears in molecules from set $A$, and a set of fragments that appear in either $G$ or $R$ as $S$, the metric is defined as a cosine similarity:

$$\mathrm{Frag}(G,R) = \frac{\sum_{s\in S}[c_s(G)\cdot c_s(R)]}{\sqrt{\sum_{s\in S}c_s^2(G)}\sqrt{\sum_{s\in S}c_s^2(R)}}. \tag{2}$$

The purpose of this metric is to show how similar are the scaffolds present in generated and reference datasets. For example, if the model rarely produces a certain chemotype from a reference set, the metric will be low. Limits of this metric are [0,1].

Note that both fragment and scaffold similarities compare molecules at a substructure level. Hence, it is possible to have a similarity one even when $G$ and $R$ contain different molecules.

**Similarity to a nearest neighbor (SNN)** is an average Tanimoto similarity $T(m_G, m_R)$ (also known as the Jaccard index) between fingerprints of a molecule $m_G$ from the generated set $G$ and its nearest neighbor molecule $m_R$ in the reference dataset $R$:

$$\mathrm{SNN}(G,R) = \frac{1}{|G|}\sum_{m_G\in G}\max_{m_R\in R}T(m_G,m_R), \tag{3}$$

In this work, we used standard Morgan (extended connectivity) fingerprints (Rogers and Hahn, 2010) with radius 2 and 1024 bits computed using RDKit library (Landrum, 2006). The resulting similarity metric can be interpreted as precision: if generated molecules are far from the manifold of the reference set, similarity to the nearest neighbor will be low. Limits of this metric are [0,1].

**Internal diversity (IntDiv$_p$)** (Benhenda, 2017) assesses the chemical diversity within the generated set of molecules $G$.

$$\mathrm{IntDiv}_p(G) = 1 - \sqrt[p]{\frac{1}{|G|^2}\sum_{m_1,m_2\in G}T(m_1,m_2)^p}. \tag{4}$$

This metric detects a common failure case of generative models—mode collapse. With mode collapse, the model produces a limited variety of samples, ignoring some areas of the chemical space. A higher value of this metric corresponds to higher diversity in the generated set. In the experiments, we report IntDiv$_1$ (G) and IntDiv$_2$ (G). Limits of this metric are [0,1].

**Fréchet ChemNet Distance (FCD)** (Preuer et al., 2018) is calculated using activations of the penultimate layer of a deep neural network ChemNet trained to predict biological activities of drugs. We compute activations for canonical SMILES representations of molecules. These activations capture both chemical and biological properties of the compounds. For two sets of molecules $G$ and $R$, FCD is defined as

$$\mathrm{FCD}(G,R) = \left\|\mu_G - \mu_R\right\|^2 + \mathrm{Tr}\left[\Sigma_G + \Sigma_R - 2(\Sigma_G\Sigma_R)^{1/2}\right] \tag{5}$$

where $\mu_G$, $\mu_R$ are mean vectors and $\Sigma_G$, $\Sigma_R$ are full covariance matrices of activations for molecules from sets $G$ and $R$ respectively. FCD correlates with other metrics. For example, if the generated structures are not diverse enough (low IntDiv$_p$) or the model produces too many duplicates (low uniqueness), FCD will decrease, since the variance is smaller. We suggest using FCD for hyperparameter tuning and final model selection. Values of this metric are non-negative, lower is better.

**Properties distribution** is a useful tool for visually assessing the generated structures. To quantitatively compare the distributions in the generated and test sets, we compute a 1D Wasserstein-1 distance between property distributions of generated and test sets. We also visualize a kernel density estimation of these distributions in the Experiments section. We use the following four properties:

- Molecular weight (MW): the sum of atomic weights in a molecule. By plotting histograms of molecular weight for the generated and test sets, one can judge if a generated set is biased toward lighter or heavier molecules.
- LogP: the octanol-water partition coefficient, a ratio of a chemical's concentration in the octanol phase to its concentration in the aqueous phase of a two-phase octanol/water system; computed with RDKit's Crippen (Wildman and Crippen, 1999) estimation.
- Synthetic Accessibility Score (SA): a heuristic estimate of how hard (10) or how easy (1) it is to synthesize a given molecule. SA score is based on a combination of the molecule's fragments contributions (Ertl and Schuffenhauer, 2009). Note that SA score does not adequately assess up-to-date chemical structures, but it is useful for assessing distribution learning models.
- Quantitative Estimation of Drug-likeness (QED): a [0,1] value estimating how likely a molecule is a viable candidate for a drug. QED is meant to capture the abstract notion of esthetics in medicinal chemistry (Bickerton et al., 2012). Similar to SA, descriptor limits in QED have been changing during the last decade and current limits may not cover latest drugs (Shultz, 2018).

## DATASET

The proposed dataset used for training and testing is based on the ZINC Clean Leads (Sterling and Irwin, 2015) collection which contains 4, 591, 276 molecules with molecular weight in the range from 250 to 350 Da, a number of rotatable bonds not greater than 7, and XlogP (Wang et al., 1997) not greater then 3.5. Clean-leads dataset consists of structures suitable for identifying hit compounds and they are small enough to allow for further ADMET optimization of generated molecules (Teague et al., 1999). We removed molecules containing charged atoms, atoms besides C, N, S, O, F, Cl, Br, H, or cycles larger than eight atoms. The molecules were filtered via custom medicinal chemistry filters (MCFs) and PAINS filters (Baell and Holloway, 2010). We describe MCFs and discuss PAINS in Supplementary Information 1. We removed charged molecules to avoid ambiguity with tautomers and pH conditions. Note that in the initial set of molecules, functional groups were present in both ionized and unionized forms.

The final dataset contains molecules, with internal diversity IntDiv$_1$ = 0.857; it contains 448, 854 unique Bemis-Murcko (Bemis and Murcko, 1996) scaffolds and 58, 315 unique BRICS

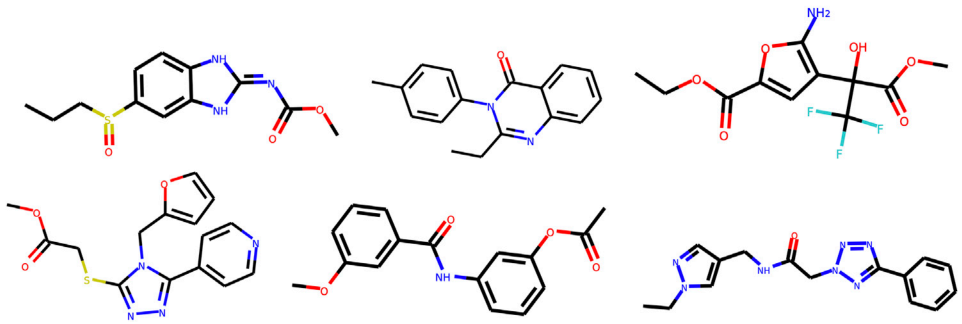


**FIGURE 3 |** Examples of molecules from MOSES dataset.

(Degen et al., 2008) fragments. We show example molecules in **Figure 3** and a representative diverse subset in Supplementary Information 2. We provide recommended split into three non-intersecting parts: train (1, 584, 664 molecules), test (176, 075 molecules) and scaffold test (176, 226 molecules). The scaffold test set has all molecules containing a Bemis-Murcko scaffold from a random subset of scaffolds. Hence, scaffolds from the scaffold test set differ from scaffolds in both train and test sets. We use scaffold test split to assess whether a model can produce novel scaffolds absent in the training set. The test set is a random subset of the remaining molecules in the dataset.

## BASELINES

We implemented several models that cover different approaches to molecular generation, such as character-level recurrent neural networks (CharRNN) (Preuer et al., 2018; Segler et al., 2018), Variational Autoencoders (VAE) (Kadurin et al., 2016; Blaschke et al., 2018; Gómez-Bombarelli et al., 2018), Adversarial Autoencoders (AAE) (Kadurin et al., 2016; Polykovskiy et al., 2018b), Junction Tree Variational Autoencoders (JTN-VAE) (Jin et al., 2018), LatentGAN (Prykhodko et al., 2019), and non-neural baselines.

Model comparison can be challenging since different training parameters (number of epochs, batch size, learning rate, initial state, optimizer) and architecture hyperparameters (hidden layer dimension, number of layers, etc.) can significantly alter their performance. For each model, we attempted to preserve its original architecture as published and tuned the hyperparameters to improve the performance. We used random search over multiple architectures for every model and selected the architecture that produced the best value of FCD. Models are implemented in *Python* 3 utilizing PyTorch (Paszke et al., 2017) framework. Please refer to the Supplementary Information three for the training details and hyperparameters.

**Character-level recurrent neural network (CharRNN)** (Segler et al., 2018) models a distribution over the next token given previously generated ones. We train this model by maximizing log-likelihood of the training data represented as SMILES strings.

**Variational autoencoder (VAE)** (Kingma and Welling, 2013) consists of two neural networks—an encoder and a decoder—that infer a mapping from high-dimensional data representation onto a lower-dimensional space and back. The lower-dimensional space is called the latent space, which is often a continuous vector space with normal prior distribution. VAE parameters are optimized to encode and decode data by minimizing reconstruction loss and regularization term in a form of Kullback-Leibler divergence. VAE-based architecture for the molecular generation was studied in multiple previous works (Kadurin et al. 2016; Blaschke et al. 2018; Gómez-Bombarelli et al. 2018). We combine aspects from these implementations and use SMILES as input and output representations.

**Adversarial Autoencoder (AAE)** (Makhzani et al., 2016) replaces the Kullback-Leibler divergence from VAE with an adversarial objective. An auxiliary discriminator network is trained to distinguish samples from a prior distribution and model's latent codes. The encoder then adapts its latent codes to minimize discriminator's predictive accuracy. The training process oscillates between training the encoder-decoder pair and the discriminator. Unlike Kullback-Leibler divergence that has a closed-form analytical solution only for a handful of distributions, a discriminator can be used for any prior distribution. AAE-based models for molecular design were studied in (Kadurin et al., 2016; Kadurin et al., 2017; Polykovskiy et al., 2018b). Similar to VAE, we use SMILES as input and output representations.

**TABLE 1 |** Performance metrics for baseline models: fraction of valid molecules, fraction of unique molecules from and molecules.

| Model | Valid (↑) | Unique@1k (↑) | Unique@10k (↑) |
|---|---|---|---|
| *Train* | *1.0* | *1.0* | *1.0* |
| HMM | 0.076 ± 0.0322 | 0.623 ± 0.1224 | 0.5671 ± 0.1424 |
| NGram | 0.2376 ± 0.0025 | 0.974 ± 0.0108 | 0.9217 ± 0.0019 |
| Combinatorial | **1.0 ± 0.0** | 0.9983 ± 0.0015 | 0.9909 ± 0.0009 |
| CharRNN | 0.975 ± 0.026 | **1.0 ± 0.0** | 0.999 ± 0.0 |
| VAE | 0.977 ± 0.001 | **1.0 ± 0.0** | 0.998 ± 0.001 |
| AAE | 0.937 ± 0.034 | **1.0 ± 0.0** | 0.997 ± 0.002 |
| JTN-VAE | **1.0 ± 0.0** | **1.0 ± 0.0** | **0.9996 ±0.0003** |
| LatentGAN | 0.897 ± 0.002 | **1.0 ± 0.0** | 0.997 ± 0.005 |

*Reported (mean ± SD) over three independent model initializations.*

**TABLE 2 |** Performance metrics for baseline models: fraction of molecules passing filters (MCF, PAINS, ring sizes, charge, atom types), novelty, and internal diversity.

| Model | Filters (↑) | Novelty (↑) | IntDiv$_1$ | IntDiv$_2$ |
|---|---|---|---|---|
| *Train* | *1.0* | *0.0* | *0.857* | *0.851* |
| HMM | 0.9024 ± 0.0489 | **0.9994 ± 0.001** | 0.8466 ± 0.0403 | 0.8104 ± 0.0507 |
| NGram | 0.9582 ± 0.001 | 0.9694 ± 0.001 | **0.8738 ± 0.0002** | 0.8644 ± 0.0002 |
| Combinatorial | 0.9557 ± 0.0018 | 0.9878 ± 0.0008 | 0.8732 ± 0.0002 | **0.8666 ± 0.0002** |
| CharRNN | 0.994 ± 0.003 | 0.842 ± 0.051 | 0.856 ± 0.0 | 0.85 ± 0.0 |
| VAE | **0.997 ± 0.0** | 0.695 ± 0.007 | 0.856 ± 0.0 | 0.85 ± 0.0 |
| AAE | 0.996 ± 0.001 | 0.793 ± 0.028 | 0.856 ± 0.003 | 0.85 ± 0.003 |
| JTN-VAE | 0.976 ± 0.0016 | 0.9143 ± 0.0058 | 0.8551 ± 0.0034 | 0.8493 ± 0.0035 |
| LatentGAN | 0.973 ± 0.001 | 0.949 ± 0.001 | 0.857 ± 0.0 | 0.85 ± 0.0 |

*Reported (mean ± SD) over three independent model initializations.*

**TABLE 3 |** Performance metrics for baseline models: Fréchet ChemNet Distance (FCD) and Similarity to a nearest neighbor (SNN).

| Model | FCD (↓) | | SNN (↑) | |
|---|---|---|---|---|
| | Test | TestSF | Test | TestSF |
| *Train* | *0.008* | *0.476* | *0.642* | *0.586* |
| HMM | 24.4661 ± 2.5251 | 25.4312 ± 2.5599 | 0.3876 ± 0.0107 | 0.3795 ± 0.0107 |
| NGram | 5.5069 ± 0.1027 | 6.2306 ± 0.0966 | 0.5209 ± 0.001 | 0.4997 ± 0.0005 |
| Combinatorial | 4.2375 ± 0.037 | 4.5113 ± 0.0274 | 0.4514 ± 0.0003 | 0.4388 ± 0.0002 |
| CharRNN | **0.073 ± 0.025** | **0.52 ± 0.038** | 0.601 ± 0.021 | 0.565 ± 0.014 |
| VAE | 0.099 ± 0.013 | 0.567 ± 0.034 | **0.626 ± 0.0** | **0.578 ± 0.001** |
| AAE | 0.556 ± 0.203 | 1.057 ± 0.237 | 0.608 ± 0.004 | 0.568 ± 0.005 |
| JTN-VAE | 0.3954 ± 0.0234 | 0.9382 ± 0.0531 | 0.5477 ± 0.0076 | 0.5194 ± 0.007 |
| LatentGAN | 0.296 ± 0.021 | 0.824 ± 0.030 | 0.538 ± 0.001 | 0.514 ± 0.009 |

*Reported (mean ± SD) over three independent model initializations. Results for random test set (Test) and scaffold split test set (TestSF).*

**TABLE 4 |** Fragment similarity (Frag), Scaffold similarity (Scaff).

| Model | Frag (↑) | | Scaf (↑) | |
|---|---|---|---|---|
| | Test | TestSF | Test | TestSF |
| *Train* | *1.0* | *0.999* | *0.991* | *0.0* |
| HMM | 0.5754 ± 0.1224 | 0.5681 ± 0.1218 | 0.2065 ± 0.0481 | 0.049 ± 0.018 |
| NGram | 0.9846 ± 0.0012 | 0.9815 ± 0.0012 | 0.5302 ± 0.0163 | 0.0977 ± 0.0142 |
| Combinatorial | 0.9912 ± 0.0004 | 0.9904 ± 0.0003 | 0.4445 ± 0.0056 | 0.0865 ± 0.0027 |
| CharRNN | **1.0 ± 0.0** | **0.998 ± 0.0** | 0.924 ± 0.006 | **0.11 ± 0.008** |
| VAE | 0.999 ± 0.0 | **0.998 ± 0.0** | **0.939 ± 0.002** | 0.059 ± 0.01 |
| AAE | 0.991 ± 0.005 | 0.99 ± 0.004 | 0.902 ± 0.037 | 0.079 ± 0.009 |
| JTN-VAE | 0.9965 ± 0.0003 | 0.9947 ± 0.0002 | 0.8964 ± 0.0039 | 0.1009 ± 0.0105 |
| LatentGAN | 0.999 ± 0.003 | **0.998 ± 0.003** | 0.886 ± 0.015 | 0.1 ± 0.006 |

*Reported (mean ± SD) over three independent model initializations. Results for random test set (Test) and scaffold split test set (TestSF).*

**Junction Tree VAE (JTN-VAE)** (Jin et al., 2018) generates molecules in two phases by exploiting valid subgraphs as components. In the first phase, it generates a tree-structured object (a junction tree) whose role is to represent the scaffold of subgraph components and their coarse relative arrangements. The components are valid chemical substructures automatically extracted from the training set. In the second phase, the subgraphs (nodes of the tree) are assembled together into a coherent molecular graph.

**Latent Vector Based Generative Adversarial Network (LatentGAN)** (Prykhodko et al., 2019) combines an autoencoder and a generative adversarial network. LatentGAN pretrains an autoencoder to map SMILES structures onto latent vectors. A generative adversarial network is then trained to produce latent vectors for the pre-trained decoder.

**Non-neural baselines** implemented in MOSES are n-gram generative model, Hidden Markov Model (HMM), and a combinatorial generator. N-gram model collects statistics of n-grams frequencies in the training set and uses such distribution to sequentially sample new strings. Hidden Markov models utilize Baum-Welch algorithm to learn a probabilistic distribution over the SMILES strings. The model consists of several states $(s_1,\ldots,s_K)$, transition probabilities between states $p(s_{i+1} \mid s_i)$, and token emission

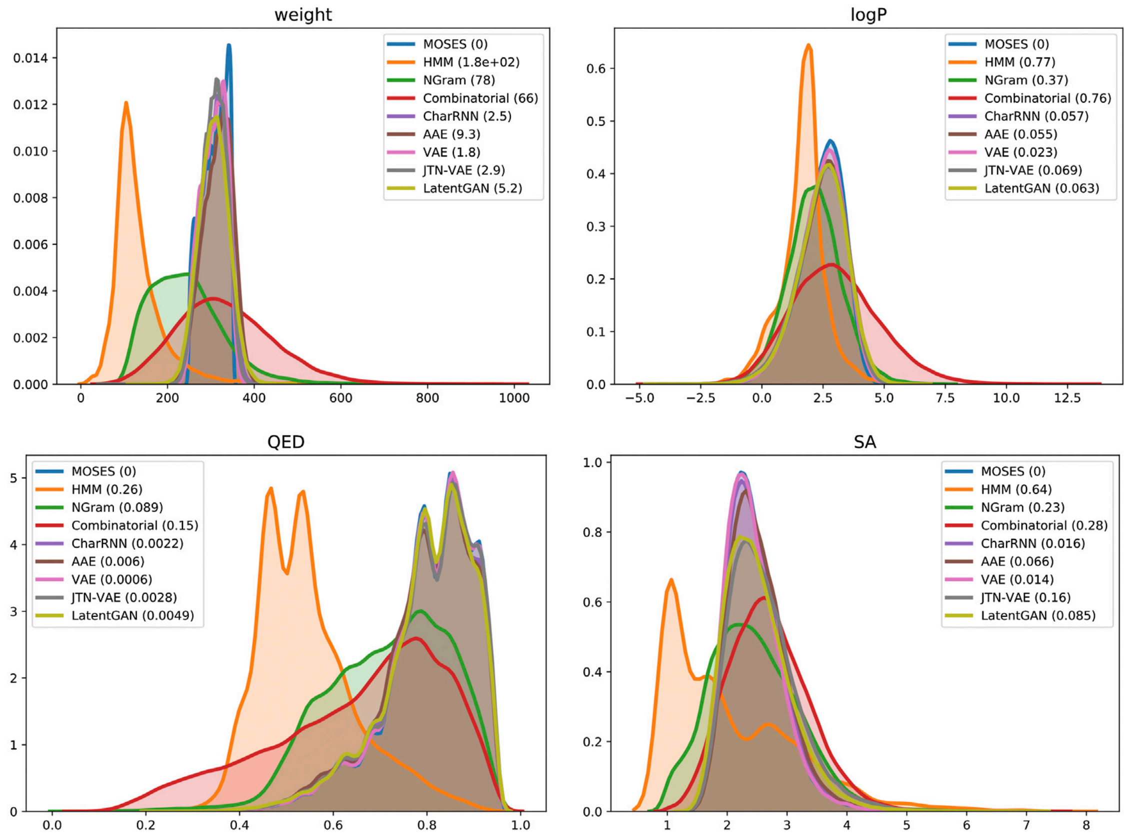


**FIGURE 4 |** Distribution of chemical properties for MOSES dataset and sets of generated molecules. In brackets—Wasserstein-1 distance to MOSES test set. Parameters: molecular weight, octanol-water partition coefficient (logP), quantitative estimation of drug-likeness (QED) and synthetic accessibility score (SA).

probabilities $p(x_i|s_i)$. Beginning from a "start" state, at each iteration the model samples a next token and state from emission and transition probabilities correspondingly. A combinatorial generator splits molecular graphs of the training data into BRICS fragments and generates new molecules by randomly connecting random substructures. We sample fragments according to their frequencies in the training set to model the distribution better.

## PLATFORM

The dataset, metrics and baseline models are provided in a GitHub repository https://github.com/molecularsets/moses and as a PyPI package molsets. To contribute a new model, one should train a model on MOSES train set, generate 30, 000 samples and compute metrics using the provided utilities. We recommend running the experiment at least three times with different random seeds to estimate sensitivity of the model to random parameter initialization. We store molecular structures in SMILES format; molecular graphs can be reconstructed using RDKit (Landrum, 2006).

## RESULTS

We trained the baseline models on MOSES train set and provide results in this section. In **Table 1** we compare models with respect to the validity and uniqueness metrics. Hidden Markov Model and NGram models fail to produce valid molecules since they have a limited context. Combinatorial generator and JTN-VAE have built-in validity constraints, so their validity is 100%.

**Table 2** reports additional properties of the generated set: fraction of molecules passing filters, fraction of molecules not present in the training set, and internal diversity. All modules successfully avoid forbidden structures (MCF and PAINS) even though such restrictions were only defined implicitly—using a training dataset. Combinatorial generator has higher diversity than the training dataset, which might be favorable for discovering new chemical structures. Autoencoder-based models show low novelty, indicating that these models overfit to the training set.

**Table 3** reports Fréchet ChemNet Distance (FCD) and similarity to a nearest neighbor (SNN). All neural network-based models show low FCD, indicating that the models successfully captured the statistics of the dataset. Surprisingly, a simple language model, character level RNN, shows the best results in terms of the FCD measure. Variational autoencoder (VAE) showed the best results in terms of SNN, but combined with low novelty we suppose that the model overfitted on the training set.

In **Table 4** we report similarities of substructure distributions—fragments and scaffolds. Scaffold similarity from the training set to the scaffold test set (TestSF) is zero by design. Note that CharRNN successfully discovered many novel scaffolds (11%), suggesting that the model generalizes well.

Finally, we compared distributions of four molecular properties in generated and test sets (**Figure 4**): molecular weight (MW), octanol-water partition coefficient (logP), quantitative estimation of drug-likeness (QED), and synthetic accessibility score (SA). Deep generative models closely match the data distribution; hidden Markov Model is biased toward lighter molecules, which is consistent with low validity: larger molecules impose more validity constraints. A combinatorial generator has higher variance in molecular weight, producing larger and smaller molecules than those present in the training set.

## DISCUSSION

From a wide range of presented models, CharRNN currently performs the best in terms of the key metrics. Specifically, it produces the best FCD, Fragment, and Scaffold scores, indicating that the model not only captured the training distribution well, but also did not overfit on the training set.

The presented set of metrics assesses models' performance from different perspectives; therefore, for each specific downstream task, one could consider the most relevant metric. For example, evaluation based on Scaf/TestSF score could be relevant when model's objective is to discover novel scaffolds. For a general evaluation, we suggest using FCD/Test metric that captures multiple aspects of other metrics in a single number. However, it does not give insights into specific issues that cause high FCD/Test values, hence more interpretable metrics presented in this paper are necessary to investigate the model's performance thoroughly.

## CONCLUSION

With MOSES, we have designed a molecular generation benchmark platform that provides a dataset with molecular structures, an implementation of baseline models, and metrics for their evaluation. While standardized comparative studies and test sets are essential for the progress of machine learning applications, the current field of *de novo* drug design lacks evaluation protocols for generative machine learning models. Being on the intersection of mathematics, computer science, and chemistry, these applications are often too challenging to explore for research scientists starting in the field. Hence, it is necessary to develop a transparent approach to implementing new models and assessing their performance. We presented a benchmark suite with unified and extendable programming interfaces for generative models and evaluation metrics.

This platform should allow for a fair and comprehensive comparison of new generative models. For future work on this project, we will keep extending the MOSES repository with new baseline models and new evaluation metrics. We hope this work will attract researchers interested in tackling drug discovery challenges.

## DATA AVAILABILITY STATEMENT

The data and code of the MOSES platform is available at https://github.com/molecularsets/moses.

## AUTHOR CONTRIBUTIONS

DP, AZhe, SG, OT, SB, RK, AA, AK, SJ, and HC designed and conducted the experiments; DP and AZhe, BS-L, VA, MV, SJ, HC, SN, AA-G, AZha wrote the manuscript.

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://arxiv.org/abs/1811.12823 (Polykovskiy et al., 2018a).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2020.565644/full#supplementary-material.

## REFERENCES

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530. doi:10.1021/acs.molpharmaceut.6b00248

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., et al. (2019). Randomized smiles strings improve the quality of molecular generative models. *J. Cheminf.* 11, 1–13. doi:10.1186/s13321-019-0393-0

Baell, J. B. and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi:10.1021/jm901137j

Bemis, G. W. and Murcko, M. A. (1996). The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi:10.1021/jm9602928

Benhenda, M. (2017). ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? Available from: https://arxiv.org/abs/1708.08227.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98. doi:10.1038/nchem.1243

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of generative autoencoder in de novo molecular design. *Mol. Inform.* 37, 1700123. doi:10.1002/minf.201700123

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C. (2019). Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* 59, 1096–1108. doi:10.1021/acs.jcim.8b00839

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15, 20170387. doi:10.1098/rsif.2017.0387

Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. (2018). "Syntax-directed variational autoencoder for structured data," in International conference on learning representations.

De Cao, N. and Kipf, T. (2018). "MolGAN: an implicit generative model for small molecular graphs," in ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008). On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* 3, 1503–1507. doi:10.1002/cmdc.200800178

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). *CVPR09.*ImageNet: a large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, June 20–25, 2009. IEEE.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. Available at: https://library.seg.org/doi/10.1190/segam2017-17559486.1

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems 28*. Editors C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (New York, NY: Curran Associates, Inc.), 2224–2232.

Ertl, P. and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* 1, 8. doi:10.1186/1758-2946-1-8

Ferrero, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S., and Levy, O. (2013). The high-throughput highway to computational materials design. *Nat. Mater.* 12, 191–201. doi:10.1038/nmat3568

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry," in Proceedings of the 34th international conference on machine learning. JMLR, 1263–1272

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a Data-Driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi:10.1021/acscentsci.7b00572

Grisoni, F., Neuhaus, C. S., Gabernet, G., Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Designing anticancer peptides by constructive machine learning. *ChemMedChem* 13, 1300–1302. doi:10.1002/cmdc.201800204

Guimaraes, G. L., Sanchez-Lengeling, B., Farias, P. L. C., and Aspuru-Guzik, A. (2017). Objective-Reinforced generative adversarial networks (ORGAN) for sequence generation models. Available at: https://arxiv.org/abs/1705.10843.

Hu, X., Beratan, D. N., and Yang, W. (2009). Emergent strategies for inverse molecular design. *Sci. China Ser. B-Chem.* 52, 1769–1776. doi:10.1007/s11426-009-0260-3

Ivanenkov, Y. A., Zhavoronkov, A., Yamidanov, R. S., Osterman, I. A., Sergiev, P. V., Aladinskiy, V. A., et al. (2019). Identification of novel antibacterials using machine learning techniques. *Front. Pharmacol.* 10, 913. doi:10.3389/fphar.2019.00913

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. (2016). Sequence tutor: conservative fine-tuning of sequence generation models with KL-control. Available at: https://arxiv.org/abs/1611.02796.

Jin, W., Barzilay, R., and Jaakkola, T. (2018). "Junction tree variational autoencoder for molecular graph generation," in Proceedings of the 35th international conference on machine learning. Editors J. Dy and A. Krause (Stockholmsmässan, Stockholm Sweden: PMLR), 2323–2332.

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2016). The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8, 10883–10890. doi:10.18632/oncotarget.14073

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* 14, 3098–3104. doi:10.1021/acs.molpharmaceut.7b00346

Kang, S. and Cho, K. (2018). Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* 59, 43–52. doi:10.1021/acs.jcim.8b00263

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation," in International conference on learning representations. ICLR. 1–26.

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. (2017). Generating and designing DNA with deep generative models. Available from: https://arxiv.org/abs/1712.06148.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding variational bayes," in International conference on learning representations.

Kirkpatrick, P. and Ellis, C. (2004). Chemical space. *Nature* 432, 823. doi:10.1038/432823a

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2019). Selfies: a robust representation of semantically constrained graphs with an example application in chemistry. Available at: https://grlearning.github.io/papers/59.pdf.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). "Grammar variational autoencoder,"in *Proceedings of the 34th international conference on machine learning*. Editors D. Precup and Y. W. Teh (Sydney, Australia: Proceedings of Machine Learning Research), Vol. 70. 1945–1954.

Labat, R., Fu, Y., and Lai, L. (1997). A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* 37, 615–621. doi:10.1021/ci960169p

Landrum, G. (2006). RDKit: open-source cheminformatics. Available at: http://www.rdkit.org/.

Le, T. C. and Winkler, D. A. (2016). Discovery and optimization of materials using evolutionary approaches. *Chem. Rev.* 116, 6107–6132. doi:10.1021/acs.chemrev.5b00691

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi:10.1109/5.726791

Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., et al. (2018). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* 9, 42. doi:10.1038/s41467-017-02465-5

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2016). "Adversarial autoencoders," in International conference on learning representations.

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Mol. Pharm.* 13, 1445–1454. doi:10.1021/acs.molpharmaceut.5b00982

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., et al. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* 9, 242. doi:10.3389/fgene.2018.00242

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018a). De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* 37, 1700153. doi:10.1002/minf.201700153

Merk, D., Grisoni, F., Friedrich, L., and Schneider, G. (2018b). Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid x receptor modulators. *Commun. Chem.* 1, 68. doi:10.1038/s42004-018-0068-1

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* 9, 48. doi:10.1186/s13321-017-0235-x

O'Boyle, N. and Dalke, A. (2018). DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. *ChemRxiv*. doi:10.26434/chemrxiv.7097960

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in pytorch," in NIPS workshop.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al. (2018a). Molecular sets (moses): a benchmarking platform for molecular generation models. Available from: https://arxiv.org/abs/1811.12823.

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., et al. (2018b). Entangled conditional adversarial autoencoder for de novo drug discovery. *Mol. Pharm.* 15, 4398–4405. doi:10.1021/acs.molpharmaceut.8b00839

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4, eaap7885. doi:10.1126/sciadv.aap7885

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. (2018). Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* 58, 1736–1741. doi:10.1021/acs.jcim.8b00234

Prykhodko, O., Johansson, S. V., Kotsias, P.-C., Arús-Pous, J., Bjerrum, E. J., Engkvist, O., et al. (2019). A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminf.* 11, 74. doi:10.1186/s13321-019-0397-9

Putin, E., Asadulaev, A., Vanhaelen, Q., Ivanenkov, Y., Aladinskaya, A. V., Aliper, A., et al. (2018). Adversarial threshold neural computer for molecular de novo design. *Mol. Pharm.* 15, 4386–4397. doi:10.1021/acs.molpharmaceut.7b01137

Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., and Aspuru-Guzik, A. (2015). What is High-Throughput virtual screening? a perspective from organic materials discovery. *Annu. Rev. Mater. Res.* 45, 195–216. doi:10.1146/annurev-matsci-070214-020823

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 1, 140022. doi:10.1038/sdata.2014.22

Reymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48, 722–730. doi:10.1021/ar500432k

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi:10.1021/ci100050t

Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365. doi:10.1126/science.aat2663

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi:10.1021/acscentsci.7b00512

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. (2019). "Graphaf: a flow-based autoregressive model for molecular graph generation," in International conference on learning representations.

Shultz, M. D. (2018). Two decades under the influence of the rule of five and the changing properties of approved oral drugs. *J. Med. Chem.* 62, 1701–1714. doi:10.1021/acs.jmedchem.8b00686

Stein, S. E., Heller, S. R., and Tchekhovskoi, D. V. (2003). "An open standard for chemical structure representation: the iupac chemical identifier." in International chemical information conference.

Sterling, T. and Irwin, J. J. (2015). Zinc 15 - ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi:10.1021/acs.jcim.5b00559

Teague, S. J., Davis, A. M., Leeson, P. D., and Oprea, T. (1999). The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed.* 38, 3743–3748. doi:10.1002/(SICI)1521-3773(19991216)38:24%3C3743::AID-ANIE3743%3E3.0.CO;2-U

van Hilten, N., Chevillard, F., and Kolb, P. (2019). Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.* 59, 644–651. doi:10.1021/acs.jcim.8b00737

Vanhaelen, Q., Mamoshina, P., Aliper, A. M., Artemov, A., Lezhnina, K., Ozerov, I., et al. (2017). Design of efficient computational workflows for in silico drug repurposing. *Drug Discov. Today* 22, 210–222. doi:10.1016/j.drudis.2016.09.019

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36. doi:10.1021/ci00057a005

Weininger, D., Weininger, A., and Weininger, J. L. (1989). Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Model.* 29, 97–101. doi:10.1021/ci00062a008

Wildman, S. A. and Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* 39, 868–873. doi:10.1021/ci990307l

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a

Yang, X., Zhang, J., Yoshizoe, K., Terayama, K., and Tsuda, K. (2017). ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* 18, 972–976. doi:10.1080/14686996.2017.1401424

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). "Seqgan: sequence generative adversarial nets with policy gradient," in Thirty-first AAAI conference on artificial intelligence.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019a). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.*, 37, 1038–1040. doi:10.1038/s41587-019-0224-x

Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., and Aliper, A. (2019b). Artificial intelligence for aging and longevity research: recent advances and perspectives. *Ageing Res. Rev.* 49, 49–66. doi:10.1016/j.arr.2018.11.003

**Conflict of Interest:** DP, AZhe, VA, MV, and AZha work for Insilico Medicine, a commercial artificial intelligence company. SG, OT, SB, RK, AA, and SN work for Neuromation OU, a company engaged in AI development through synthetic data and generative models. SJ and HC work for a pharmaceutical company AstraZeneca. AA-G is a cofounder and board member of, and consultant for, Kebotix, an artificial intelligence-driven molecular discovery company and a member of the science advisory board of Insilico Medicine.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read
for greatest visibility
and readership

**FAST PUBLICATION**
Around 90 days
from submission
to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative,
and constructive
peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers
acknowledged by name
on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data
and methods to enhance
research reproducibility

**DIGITAL PUBLISHING**
Articles designed
for optimal readership
across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics
track visibility across
digital media

**EXTENSIVE PROMOTION**
Marketing
and promotion
of impactful research

**LOOP RESEARCH NETWORK**
Our network
increases your
article's readership