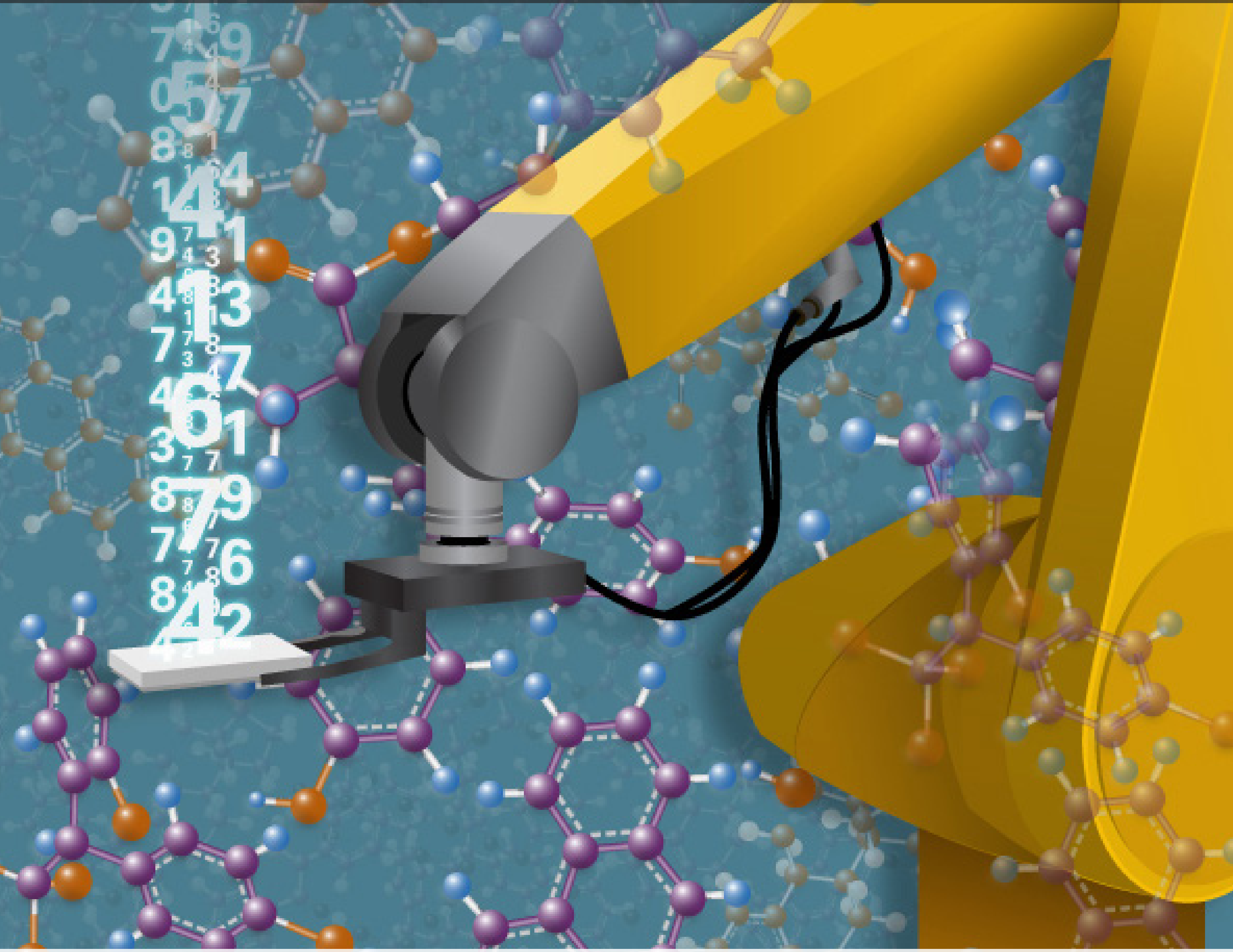
TOX21 CHALLENGE TO BUILD PREDICTIVE MODELS OF NUCLEAR RECEPTOR AND STRESS RESPONSE PATHWAYS AS MEDIATED BY EXPOSURE TO ENVIRONMENTAL TOXICANTS AND DRUGS

EDITED BY: Ruili Huang and Menghang Xia

PUBLISHED IN: Frontiers in Environmental Science



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-197-5
DOI 10.3389/978-2-88945-197-5

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

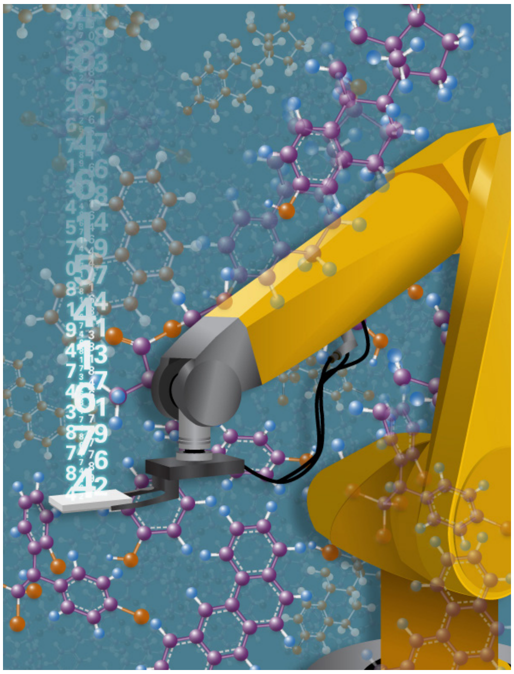
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - 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: **researchtopics@frontiersin.org**

TOX21 CHALLENGE TO BUILD PREDICTIVE MODELS OF NUCLEAR RECEPTOR AND STRESS RESPONSE PATHWAYS AS MEDIATED BY EXPOSURE TO ENVIRONMENTAL TOXICANTS AND DRUGS

Topic Editors:
Ruili Huang, National Center for Advancing Translational Sciences, National Institutes of Health, USA
Menghang Xia, National Center for Advancing Translational Sciences, National Institutes of Health, USA



The Tox21 robot screens the Tox21 10K library of environmental chemicals against a panel of in vitro assays in quantitative high throughput screening (qHTS) format generating millions of robust data points that can be applied to build computational models for toxicity prediction. This picture shows the robot arm holding a 1536-well assay plate for screening. Examples of chemical structures are displayed in the background.
Image by Palladian Partners, Inc

Tens of thousands of chemicals are released into the environment every day. High-throughput screening (HTS) has offered a more efficient and cost-effective alternative to traditional toxicity tests that can profile these chemicals for potential adverse effects with the aim to prioritize a manageable number for more in depth testing and to provide clues to mechanism of toxicity.

The Tox21 program, a collaboration between the National Institute of Environmental Health Sciences (NIEHS)/National Toxicology Program (NTP), the U.S. Environmental Protection Agency's (EPA) National Center for Computational Toxicology (NCCT), the National Institutes of Health (NIH) National Center for Advancing Translational Sciences (NCATS), and the U.S. Food and Drug Administration (FDA), has generated quantitative high-throughput screening (qHTS) data on a library of 10K compounds, including environmental chemicals and drugs, against a panel of nuclear receptor and stress response pathway assays during its production phase (phase II). The Tox21 Challenge, a worldwide modeling competition, was launched that asks a "crowd" of researchers to use these data to elucidate the extent to which the interference of biochemical and cellular pathways by compounds can be inferred from chemical structure data. In the Challenge participants were asked to model twelve assays related to nuclear receptor and stress response pathways using the data generated against the Tox21 10K compound library as the training set. The computational models built within this Challenge are expected to improve the community's ability to prioritize novel chemicals with respect to potential concern to human health. This research topic presents the resulting computational models with good predictive performance from this Challenge.

Citation: Huang, R., Xia, M., eds. (2017). Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Toxicants and Drugs. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-197-5

Table of Contents

Section 1: Editorial

Section 2: Tox21 Challenge overview

Section 3: Winning models from the Challenge

Section 4: Other modeling methods with good predictive performance

frontiers
in Environmental Science

EDITORIAL
published: 25 January 2017
doi: 10.3389/fenvs.2017.00003

Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs

Ruili Huang and Menghang Xia**

Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD, USA

Keywords: Tox21, HTS, nuclear receptor, stress response, predictive model, QSAR, *in vitro* assay

OPEN ACCESS

Edited by:
Yang Liu,
Emory University, USA

Reviewed by:
Naresh Singhal,
University of Auckland, New Zealand

****Correspondence:***
Ruili Huang
huangru@mail.nih.gov
Menghang Xia
mxia@mail.nih.gov

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *11 August 2016*
Accepted: *12 January 2017*
Published: *25 January 2017*

Citation:
Huang R and Xia M (2017) Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. Front. Environ. Sci. 5:3. doi: 10.3389/fenvs.2017.00003

Editorial on the Research Topic

Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs

Tens of thousands of chemicals are released into the environment every year. High-throughput screening (HTS) has offered a more efficient and cost-effective alternative to traditional toxicity tests to profile these chemicals for potential adverse effects with the aim to prioritize a manageable number for in depth testing and to provide clues to their mechanisms of toxicity. The Tox21 program (NRC, 2007; Collins et al., 2008; Kavlock et al., 2009; Tice et al., 2013), a collaboration among the National Institute of Environmental Health Sciences (NIEHS)/National Toxicology Program (NTP), the U.S. Environmental Protection Agency's (EPA) National Center for Computational Toxicology (NCCT), the National Institutes of Health (NIH) National Center for Advancing Translational Sciences (NCATS), and the U.S. Food and Drug Administration (FDA), has generated quantitative high-throughput screening (qHTS) data (>50 million data points) on a library of 10K compounds, including environmental chemicals and drugs, against a panel of nuclear receptor and stress response pathway assays during its production phase (phase II) (Huang et al., 2016). A worldwide modeling competition, the Tox21 Data Challenge, was launched that asked a crowd of researchers to use these data as the training set to elucidate the extent to which the interference of biochemical and cellular pathways by compounds can be inferred from chemical structure data. This E-book comprises articles describing computational models with good predictive performance that resulted from this challenge (Huang et al.).

Machine learning methods have been widely used by the computational modeling community for the prediction of biological activity induced by small molecules. The Tox21 Data Challenge provides the unique opportunity to compare the predictive abilities of different computational methods for biological activity, specifically, those related to toxicity. The Challenge participants employed a wide range of chemical descriptors and/or fingerprints for small molecule representation, and machine learning algorithms for modeling.

Models employed the deep learning algorithm showed the best predictive performance. Deep Learning, as a field of machine learning, has gained popularity in the recent years, having been widely applied in the fields such as signal and information processing, speech recognition, as well as physics and life sciences. Deep Learning has also been applied to predict the outcome of biological assays. Mayr and coauthors, were the first to apply Deep Learning to computational toxicity (Mayr et al.). Deep Learning enabled them to construct a hierarchy of chemical features that combines the best of the features to ensembles, resulting in models that outperformed most other computational methods. Other than the high performance but computationally intensive Deep Learning method, simple traditional machine learning methods also attained success in predicting a number of assay activities. Abdelaziz and coauthors developed consensus models with methods implemented within the OCHEM (http://www.ochem.eu) web-based platform using 10 different descriptor sets and the associative neural networks (ASNN) algorithm (Abdelaziz et al.). These consensus models achieved the best overall balanced accuracy across all assays and top performance in the ATAD5 and mitochondrial membrane potential disruption assays. Their stratified bagging contributed models, and the selection of consensus models, were optimized to achieve the best balanced accuracy. Barta employed the ensemble approach for model development, combining various fingerprinting tools with different machine learning techniques, and applied assorted feature selection methods (Barta). Barta found that multi-tree ensemble methods, such as Random Forests and Extra Trees, produce reliable predictions and are insensitive to dimensionality extremes. These models achieved the best performance in predicting compound activities against AR, aromatase, and p53. Random Forest was also the method of choice for Uesawa, who produced the best performing ER-LBD model, calculated multiple descriptors and applied Random Forest for descriptor selection and model generation (Uesawa).

Other articles described models that employed classic machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), k Nearest Neighbor (kNN), and Naïve Bayes (Drwal et al.), with different combinations of molecular descriptors, each with their own spin on the specific implementation of these methods for model construction. Capuzzi and coauthors found that their models built with Deep Neural Networks performed better than those developed with simple machine learning algorithms and that dataset balancing had a detrimental effect on prediction accuracy (Capuzzi et al.). Stefaniak evaluated combinations of various attribute selection methods and machine learning algorithms and determined that combining the Best First method for attribute selection with the Rotation Forest/ADTree classifier produced the best models (Stefaniak). Koutsoukas and coauthors utilized circular molecular fingerprints combined with Random Forest and SVM (Koutsoukas et al.). Ribay and coauthors also applied the biological response profile of chemicals from public data sources toward model construction, and found significant improvement in model performance compared to models built with chemical structure information alone (Ribay et al.).

Other than the Deep Learning techniques, there is no clear indication of which machine learning algorithm and/or molecular descriptor, or a specific combination of the two, has significant advantage over the others. Specific implementation of the methods and application of the optimal methods to the most fitting dataset seem to make the most difference. Common strategies employed by the best performing models show that consensus modeling and the diversity of descriptors tend to improve the predictive performance of models. The top performing models reached prediction accuracies close to the level of experimental errors (Huang et al.), demonstrating the feasibility of applying these models as screening tools for chemical prioritization.

The articles from this e-book present a groundbreaking direction for toxicological related testing and are intended to help improve the understanding of how chemicals could disrupt biological pathways and result in toxicity. Specifically, the computational models generated from this Challenge can be applied to predict the potential of those environmental chemicals with limited information to disrupt nuclear receptor and cellular stress response pathways. The computational models built within this Challenge are expected to improve the community's ability to prioritize novel chemicals with respect to potential concern to human health. The best performing models are currently being made publicly accessible to the scientific community (Abdelaziz et al.; Mayr et al.) to help facilitate chemical risk assessment.

AUTHOR CONTRIBUTIONS

RH and MX wrote the editorial.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Programs of the National Toxicology Program (Interagency agreement #Y2-ES-7020-01), National Institute of Environmental Health Sciences, the U.S. Environmental Protection Agency (Interagency Agreement #Y3-HG-7026-03), and the National Center for Advancing Translational Sciences, National Institutes of Health. We would like to take this opportunity to thank all the authors for their excellent contributions that made this series a success, and the reviewers for their insightful comments and remarks.

REFERENCES

Collins, F. S., Gray, G. M., and Bucher, J. R. (2008). Toxicology. Transforming environmental health protection. *Science* 319, 906–907. doi: 10.1126/science.1154619

Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., et al. (2016). Modelling the Tox21 10 K chemical profiles for *in vivo* toxicity prediction and mechanism characterization. *Nat. Commun.* 7:10425. doi: 10.1038/ncomms10425

Kavlock, R. J., Austin, C. P., and Tice, R. R. (2009). Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Anal. 29*, 485–487. discussion: 492–487. doi: 10.1111/j.1539-6924.2008.01168.x

NRC (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press.

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* 121, 756–765. doi: 10.1289/ehp.1205784

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.

ORIGINAL RESEARCH
published: 14 January 2016
doi: 10.3389/fenvs.2015.00085

Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs

Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A. Shahane, Anna Rossoshek and Anton Simeonov*

Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD, USA

OPEN ACCESS

Edited by:
Juergen Pilz,
Alpen-Adria Universitaet Klagenfurt,
Austria

Reviewed by:
Christina Mitsakou,
King's College London, UK
Ijaz Hussain,
Quaid-i-Azam University, Islamabad,
Pakistan

***Correspondence:**
Ruili Huang
huangru@mail.nih.gov

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Environmental Science

Received: *31 August 2015*
Accepted: *22 December 2015*
Published: *14 January 2016*

Citation:
Huang R, Xia M, Nguyen D-T, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A and Simeonov A (2016) Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. Front. Environ. Sci. 3:85. doi: 10.3389/fenvs.2015.00085

Tens of thousands of chemicals with poorly understood biological properties are released into the environment each day. High-throughput screening (HTS) is potentially a more efficient and cost-effective alternative to traditional toxicity tests. Using HTS, one can profile chemicals for potential adverse effects and prioritize a manageable number for more in-depth testing. Importantly, it can provide clues to mechanism of toxicity. The Tox21 program has generated >50 million quantitative high-throughput screening (qHTS) data points. A library of several thousands of compounds, including environmental chemicals and drugs, is screened against a panel of nuclear receptor (NR) and stress response (SR) pathway assays. The National Center for Advancing Translational Sciences (NCATS) has organized an international data challenge in order to "crowd-source" data and build predictive toxicity models. This Challenge asks a "crowd" of researchers to use these data to elucidate the extent to which the interference of biochemical and cellular pathways by compounds can be inferred from chemical structure data. The data generated against the Tox21 library served as the training set for this modeling Challenge. The competition attracted participants from 18 different countries to develop computational models aimed at better predicting chemical toxicity. The winning models from nearly 400 model submissions all achieved >80% accuracy. Several models exceeded 90% accuracy, which was measured by area under the receiver operating characteristic curve (AUC-ROC). Combining the winning models with the knowledge already gained from Tox21 screening data are expected to improve the community's ability to prioritize novel chemicals with respect to potential human health concern.

Keywords: Tox21, HTS, nuclear receptor, stress response, predictive model, QSAR, *in vitro* assay

INTRODUCTION

Humans are exposed to many different chemicals during the course of their lifetimes through various sources including food, household cleaning products, and medicines. In some cases, these chemicals can be toxic. In fact, more than 30% of promising pharmaceuticals have failed in human clinical trials because they were found to be toxic despite promising pre-clinical studies in animal models (Kola and Landis, 2004). Creating rapid and efficient methods for assessing chemical toxicity has the potential to improve how scientists evaluate environmental chemicals, develop new medicines, and even foster decisions made by regulatory agencies on whether or not these chemicals should be made available. More than 80,000 chemical compounds are registered for use in the U.S., and for 95% of them, there is no data on human exposure to inform society about their effects on health (Judson et al., 2009). The use of *in silico* approaches, such as quantitative structure-activity relationship (QSAR) models that infer biological activity from chemical structure similarity, is a viable alternative to fill in the gap where experimental data is lacking (Muster et al., 2008; Vedani and Smiesko, 2009). These models could be applied to all the chemicals of environmental concern and obtain an estimate on their toxicity potential in a matter of hours of computational time. Chemicals estimated to have a high potential for toxicity, which would be a much smaller number, could be prioritized for experimental evaluation and validation. In addition, these models could also identify structural features of a chemical that are responsible for its toxic activity, which could serve as structural alerts for toxicity (Sanderson and Earnshaw, 1991; Saiakhov and Klopman, 2008). Combining these computational models with existing experimental data will make chemical prioritization more time and cost efficient.

The U.S. Tox21 program (NRC, 2007; Collins et al., 2008; Kavlock et al., 2009; Tice et al., 2013), a collaboration between the National Institute of Environmental Health Sciences (NIEHS)/National Toxicology Program (NTP), the U.S. Environmental Protection Agency's (EPA) National Center for Computational Toxicology (NCCT), the National Institutes of Health (NIH) National Center for Advancing Translational Sciences (NCATS), and the U.S. Food and Drug Administration (FDA), is aimed at developing better toxicity assessment methods. The goal is to quickly and efficiently test whether certain chemicals have the potential to disrupt processes in the human body that may lead to adverse health effects. The Tox21 consortium leverages its partners' resources and expertise to predict more effectively how a collection of ∼10,000 compounds (referred to as Tox21 10K library) composed of environmental chemicals and approved drugs will affect human health and the environment. The Tox21 10K library has been tested in a quantitative high-throughput screening (qHTS) format against a panel of nuclear receptor (NR) (Huang et al., 2011, 2014; Hsu et al., 2014; Chen et al., 2015) and stress response (SR) pathway assays (Attene-Ramos et al., 2015; Nishihara et al., 2016), producing over 50 million data points to date (PubChem, 2013b). These data can serve as a knowledge-base to correlate chemical structures to their biological activities to develop QSAR models. To encourage the mining and usage of these data now publicly available, NCATS launched the Tox21 Data Challenge 2014[1], the goal of which was to "crowdsource" data analysis by independent researchers to reveal how well they can predict compounds' interference in cellular and biochemical pathways resulting in potential toxicity by using only chemical structure data. The Challenge's computational models could become part of the decision-making tools for government agencies in determining which environmental chemicals and drugs are of the greatest potential concern to human health.

TABLE 1 | Tox21 assays used in subchallenges.

Assay ID	Assay	PubChem AID
NR-AhR	Aryl hydrocarbon receptor	743122
NR-Aromatase	Aromatase	743139
NR-AR	Androgen receptor, full length	743040
NR-AR-LBD	Androgen receptor, LBD	743053
NR-ER	Estrogen receptor alpha, full length	743079
NR-ER-LBD	Estrogen receptor alpha, LBD	743077
NR-PPAR-gamma	Peroxisome proliferator-activated receptor gamma	743140
SR-ARE	Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element	743219
SR-ATAD5	ATAD5	720516
SR-HSE	Heat shock factor response element	743228
SR-MMP	Mitochondrial membrane potential	720637
SR-p53	p53	720552

Here, we describe the Challenge and provide an overall summary of the results. Data from 12 assays were selected based on data quality and public interests for this Challenge (**Table 1**). The Challenge was divided into subchallenges. In subchallenges 1–12, participants were asked to model compound activity for each one of the 12 assays. In subchallenges 13 and 14, participants were asked to model all NR pathway assays (NR Panel Challenge) and all SR pathway assays (SR Panel Challenge). In the final subchallenge, 15 (Grand Challenge), participants were asked to build models for all 12 assays. The Tox21 10K dataset was used for model training. Data generated on part of the LOPAC^{1280} (Library of Pharmacologically Active Compounds) collection was used for testing. For final model evaluation and scoring, a new set of compounds provided by the EPA, for which no experimental data were available at the time of the Challenge, was screened against the 12 assays. This new set of data together with the rest of the LOPAC data was used to evaluate the final model submissions. The Challenge was launched on July 16, 2014 and closed for scoring on November 14, 2014. Participants were encouraged to enter the competition as teams. One winning team with the best predictive model was selected for each subchallenge, and the winners were announced on January 26, 2015. One hundred and twenty five participants representing 18 different countries registered for the Challenge (**Figure 1**). Three hundred and seventy eight model submissions from 40 teams were received for final evaluation (**Figure 1**).

[1] https://tripod.nih.gov/tox21/challenge/

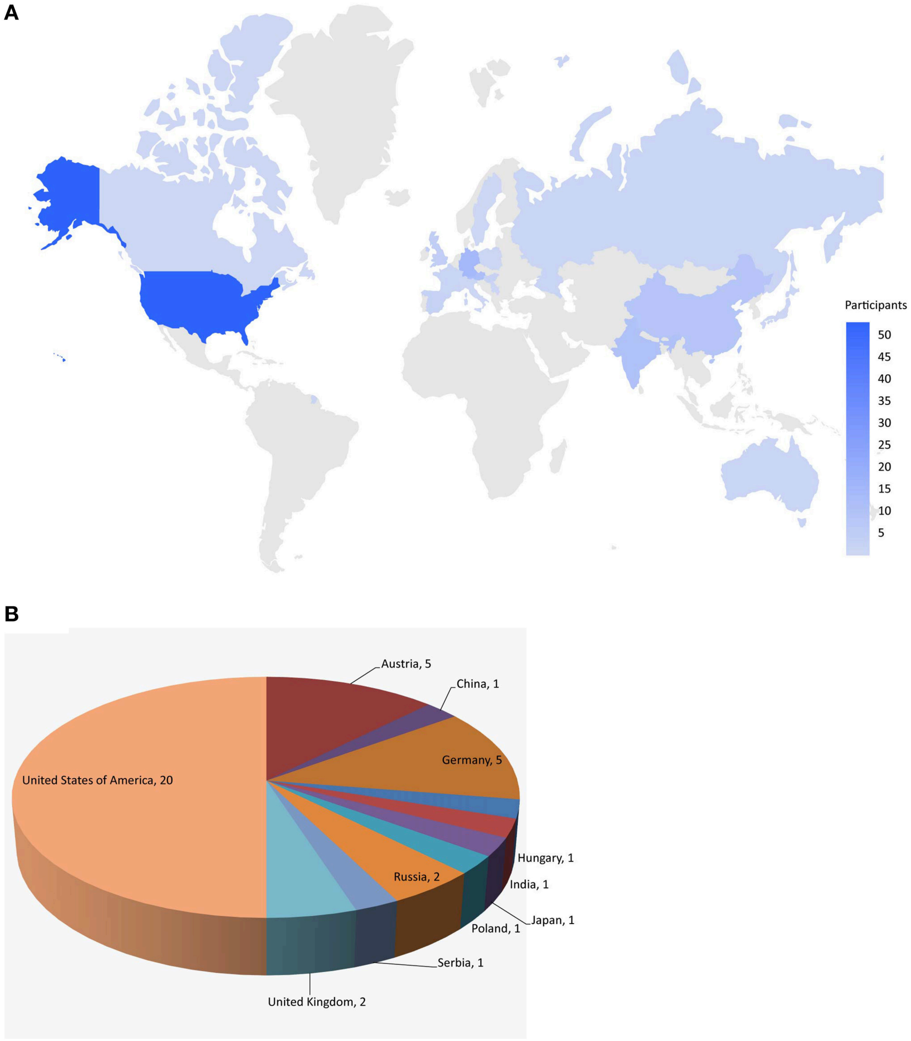


FIGURE 1 | Worldwide challenge participation. (A) Distribution of individuals registered for the challenge. **(B)** Distribution of teams that submitted models for final evaluation.

METHODS

The qHTS data generated on the Tox21 10K compound collection are publicly available (PubChem, 2013a,b). The 12 assays were selected based on data quality, active rate, and toxicological relevance for the Tox21 Challenge and their PubChem assay IDs (AIDs) are listed in **Table 1**. All of the compounds in the Tox21 10K collection went through analytical quality control (QC) to test for their purity and identity. The samples that failed QC were excluded from the training set for the Challenge. Based on the concentration response data, each compound in each assay was assigned one of three possible activity outcomes: active, inactive, and inconclusive (Huang et al., 2014; Attene-Ramos et al., 2015). The compounds that showed inconclusive activity in all 12 assays were filtered out, thus leaving 8043 samples for the training set. The LOPAC^{1280} collection (Sigma-Aldrich) contained 1280 compounds, 688 of which overlapped with the Tox21 10K compounds. The non-overlapping 592 LOPAC compounds were randomly split into two sets of equal size, with 296 compounds in each set. One set was provided to the Challenge participants for model testing and the other was held back for final evaluation. An additional set of 345 compounds, for which no experimental data was available at the time of the Challenge, was provided by the EPA as an extension to the Tox21 10K collection. The training, test, and final evaluation sets appeared to cover similar chemical structure spaces as shown by the 3D plots generated using principal components 1–3 generated from the 729-bit ChemoTyper[2] fingerprints (Supplementary Figure 1). The chemical structures of these compounds were provided to the Challenge participants to generate activity predictions. While in parallel, these compounds were also screened against the 12

[2] https://chemotyper.org/

assays to generate experimental data. The experimental screens were finished at the same time as the final model submission was closed. These newly generated assay data together with the 296 LOPAC compounds (641 compounds total) were used as the final evaluation set to score and rank the model submissions to determine the winners. All datasets were posted online[3] for registered participants to download, which are now open to the public.

Challenge participants were asked to provide an estimate of the probability of a chemical being active in an assay as well as an active/inactive call. The performance of the model was evaluated by the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC) using the activity estimates produced by the model. The ROC curve is a plot of sensitivity [TP/(TP+FN)] vs. (1-specificity [TN/(TN+FP)]) (Zweig and Campbell, 1993), where TP = true positive (number of active compounds also predicted as active), FP = false positive (number of inactive compounds predicted as active), TN = true negative (number of inactive compounds also predicted as inactive), and FN = false negative (number of active compounds predicted as inactive). A perfect model would have an AUC-ROC of 1 and an AUC-ROC of 0.5 indicates a random classifier. In cases where there was a tie between the AUC-ROC scores from two teams, the balanced accuracy (BA = (specificity + sensitivity)/2) calculated based on the active/inactive calls was used to determine the final ranking. Teams were expected to provide a prediction on the activity of every compound in the final evaluation set. Missing predictions were counted as false positive or false negative in the scoring process. Teams were asked, in addition, to provide a description of the prediction method they used, which should be embodied in a set of algorithms and a software system, for the Challenge organizers to directly use to verify the results. Challenge rules and scoring criteria were also posted online[4], where registered Challenge participants were able to upload their model predictions and method descriptions.

Consensus Modeling

A consensus model (Eduati et al., 2015) was built for each assay based on all the submitted models for that assay, such that the probability of a chemical being active in an assay is determined by combining predictions made by all individual models. Each individual model is also weighed by its predictive performance on the final evaluation set, as measured by the AUC-ROC score, such that better performing models would contribute more to the consensus prediction. Specifically, for the consensus model, the probability C of chemical i being active is calculated as follows:

$$C_i = \sum_{j=1}^{n} w_j \cdot P_j \tag{1}$$

where n is the total number of models that provided predictions for chemical i, P_j is the predicted probability of chemical i being active by model j, and w_j is the weight of model j, which is the AUC-ROC score on the final evaluation set obtained by model j. C_i is thus the consensus prediction of the activity of chemical i in an assay. The performances of the consensus models are evaluated by generating the AUC-ROC scores on the final evaluation set using these consensus probabilities as predictors.

RESULTS AND DISCUSSION

Challenge Participation

The training dataset was made available to the Challenge participants at the time of the Challenge launch in July 2014. The test dataset was provided in early August 2014, when a Leaderboard was also created on the Challenge website for teams to submit their predictions on the test set. Teams were allowed to train and test their models using the Leaderboard until October 2014, at which point the Leaderboard was closed, the test dataset was released to the participants to test and improve models on their own, and the Challenge began to accept model submissions for final evaluation. Fifty-three teams participated in the Challenge by submitting a model either at the testing stage or for final evaluation. Final model submission was closed in November 2014 when the scoring started. Teams were allowed an additional month to submit their method descriptions. Final model performance scores and ranking were made available to all teams who submitted a model for final evaluation on the Challenge website in January 2015. The top ranking teams and their scores were posted on the Challenge website[5] and the winning teams (**Table 2**) were announced on the NCATS website, January 26, 2015[6]. For the final model evaluation, we received 378 model submissions from 40 teams (**Figure 1**), averaging 32 models per assay/subchallenge.

Model Performance

The performances of the submitted models measured by AUC-ROC and BA are shown in **Figure 2**. All winning models performed well with AUC-ROC scores ranging from 0.81 to 0.95 (1 is the perfect score) and BAs ranging from 0.68 to 0.90. The BA values were found generally lower than the AUC-ROC scores because the teams were asked to decide on their own the most appropriate cutoffs to make the active/inactive calls based on their training and testing results. This task tested the contestant's ability to select the right cutoff using the ROC. If the optimal cutoff was selected, the BA should have been very close to the AUC-ROC value.

Subchallenges SR-MMP and NR-AhR received the best performing models with the best AUC-ROC scores >0.9 and average AUC-ROC scores >0.8. The models received for the other subchallenges were comparable on average, with the NR-AR and NR-AR-LBD models achieving the lowest average performance scores (~0.7). A common confounding factor that affected model performance was data quality. We checked the reproducibility of the training and the final evaluation datasets against the model performances (**Figure 3**). All datasets used for this Challenge were found to be of high quality with >90% reproducibility. No correlation was found between data reproducibility and the average AUC-ROC score per subchallenge, as all datasets were highly reproducible and the

[3] https://tripod.nih.gov/tox21/challenge/data.jsp

[4] https://tripod.nih.gov/tox21/challenge/submissions.jsp

[5] https://tripod.nih.gov/tox21/challenge/leaderboard.jsp

[6] http://www.ncats.nih.gov/news-and-events/features/tox21-challenge.html

TABLE 2 | Tox21 challenge winners.

Team name	Challenge assay(s)	Team member(s)	Organization(s)
Bioinf@JKU	Grand Challenge (all 12 assays) Stress Response Panel NR-AhR SR-ARE	Günter Klambauer, Ph.D. Sepp Hochreiter, Ph.D. Andreas Mayr, M.Sc. Thomas Unterthiner, M.Sc.	Institute of Bioinformatics, Johannes Kepler University Linz, Austria
Bioinf@JKU-ensemble1	NR-ER SR-HSE	Günter Klambauer, Ph.D. Sepp Hochreiter, Ph.D. Andreas Mayr, M.Sc. Thomas Unterthiner, M.Sc. Herbert Zaunmair	Institute of Bioinformatics, Johannes Kepler University Linz, Austria
Bioinf@JKU-ensemble3	NR-AR-LBD	Günter Klambauer, Ph.D. Sepp Hochreiter, Ph.D. Ulrich Bodenhofer, Ph.D. Andreas Mayr, M.Sc. Thomas Unterthiner, M.Sc.	Institute of Bioinformatics, Johannes Kepler University Linz, Austria
Bioinf@JKU-ensemble4	Nuclear Receptor Signaling Panel NR-PAR-gamma	Günter Klambauer, Ph.D. Sepp Hochreiter, Ph.D. Birgit Hauer Andreas Mayr, M.Sc. Thomas Unterthiner, M.Sc.	Institute of Bioinformatics, Johannes Kepler University Linz, Austria
AMAZIZ	SR-ATAD5 SR-MMP	Ahmed M. Abdelaziz Sayed	Technical University of Munich
Dmlab	NR-AR Aromatase p53	Gergõ Barta, M.Sc.	Budapest University of Technology and Economics
Microsomes	NR-ER-LBD	Yoshihiro Uesawa, Ph.D.	Department of Clinical Pharmaceutics, Meiji Pharmaceutical University

best performing models were already reaching the level of assay precision.

Active rate or data balance is another common factor that affects model performance. Models built on less balanced data or assays with lower active rates (e.g., <5%) are generally of lower quality. There are different computational approaches to balance data and enhance model performance, but if the number of actives is too low, the information that can be retrieved from the active chemical structures that the model is trained on will be limited, nonetheless. Active rate was taken into consideration when selecting assays for the Challenge such that assays with extremely low active rates (e.g., <2%) were excluded. The active rates of the assays used in the Challenge were compared against the model performances as well (**Figure 3**), and a positive correlation was found between the two ($r = 0.63$, $p = 0.03$), i.e., models built for assays with higher active rates tend to perform better. For example, the assays with the best performing models, SR-MMP and NR-AhR, had >10% active rates, whereas the NR-AR assays that received the lowest average model performances had <5% actives.

Consensus Modeling—Wisdom of the Crowd

The goal of this Challenge was to rely on the wisdom of the crowd to identify high quality models that could aid chemical toxicity assessment, and previous challenges have shown that aggregation of predictions, which leverage the collective insight of all participants, can provide a more robust estimate than any individual model (Marbach et al., 2012; Eduati et al., 2015). We generated consensus models by aggregating the individual model predictions and tested the performance of the models on the final evaluation set. The consensus models performed on a par with the winning models (**Figure 4**). We tried a few different aggregation approaches. When we averaged all individual model predictions to produce the consensus prediction, the consensus model performed better than 86% of the individual models for each subchallenge, on average. We then weighed the predictions from each individual model by their AUC-ROC score, such that the better performing models would contribute more to the consensus prediction. This approach improved the performance of the consensus models by outperforming 87% of the individual models. To further reduce the impact of poor performing individual models, we only included the top performing models (AUC-ROC > 0.8) from each subchallenge. In this case, the consensus model performed better than 96% of the individual models for each subchallenge. For 6 out of the 12 subchallenges, the consensus model outperformed the winning model. Interestingly, even though weighing all individual predictions equally (including the worst individual models) resulted in less than optimal consensus models, the

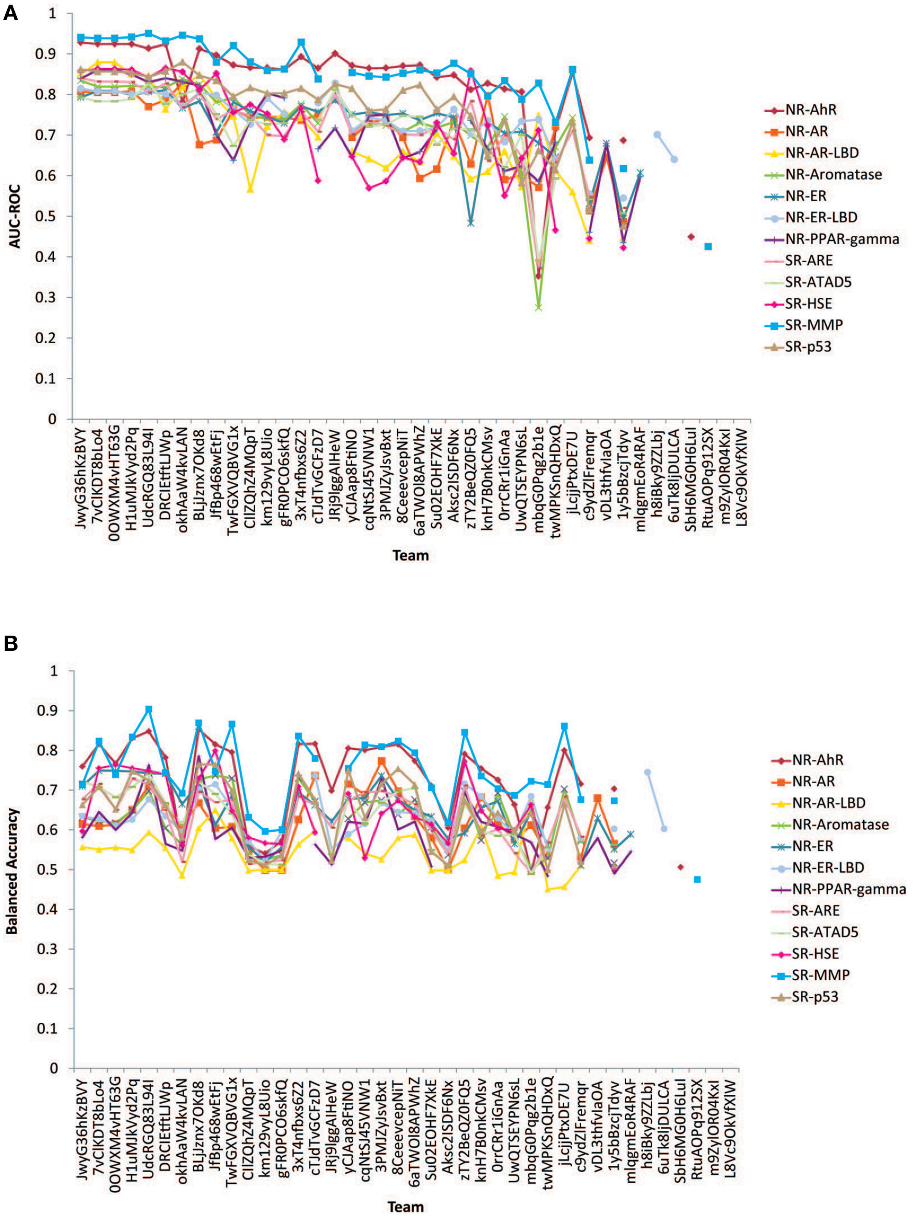


FIGURE 2 | Performances of models received for the 12 subchallenges for final evaluation. (A) Measured by AUC-ROC **(B)** Measured by balanced accuracy.

consensus models still outperformed the individual models 86% of the time, portraying the wisdom of the crowd.

In addition, we checked the compounds that were frequently predicted correctly or incorrectly by teams, and calculated a correct prediction rate for each compound in the final evaluation set. We then looked at the activity outcome distribution of each compound in the 12 assays used in the Challenge. In each assay, there were often a number of compounds for which no conclusive activity call could be made. Some compounds showed inconclusive activity in more assays than others. Based on this information, we also calculated an inconclusive rate for each compound in the final evaluation set. When the two parameters were compared, we found a strong negative correlation between the correct prediction rate and the inconclusive rate of compounds ($r = -0.75$, $p < 10^{-20}$). Inclusive outcomes were excluded when evaluating model performances, but the compounds that tend to produce inconclusive outcomes still appeared to be less predictable than compounds for which the activity was often clear. This observation suggests that there might be certain characteristics of the frequent inconclusive compounds that make them "unpredictable" and outliers/violators of the structure-dictates-activity rule. These compounds and their assay activities will be examined in more detail in a follow up study. Nevertheless,

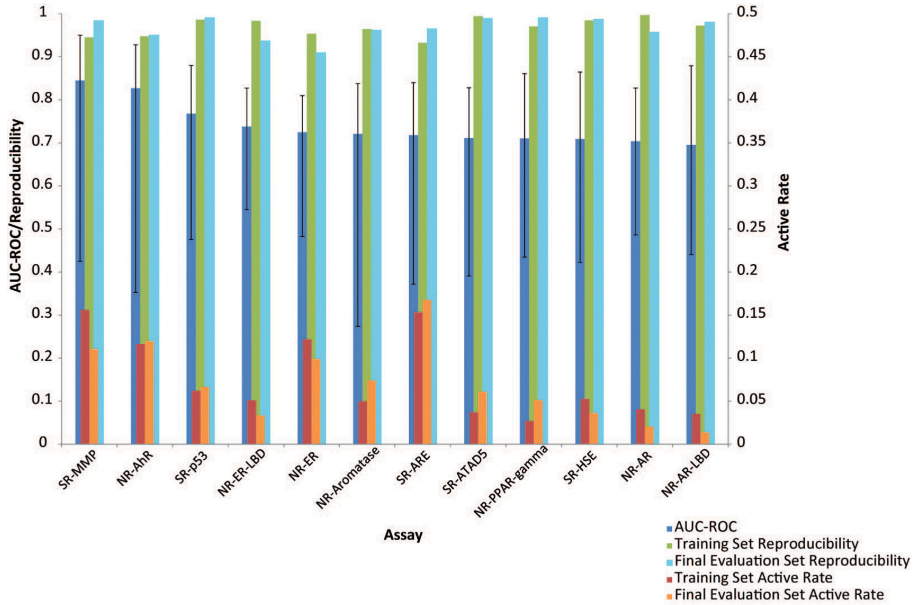


FIGURE 3 | Factors influencing model performance (displayed as mean AUC-ROC ± standard deviation). Assays that had higher active rates received better models. No correlation was found between data reproducibility and model performance because all assays were highly reproducible.

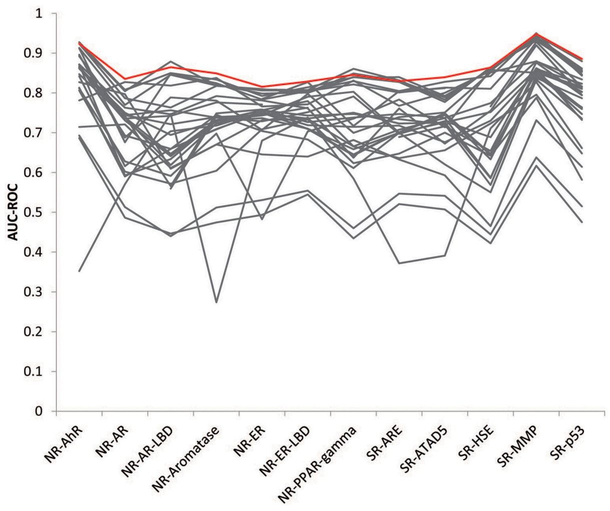


FIGURE 4 | Wisdom of the crowd. The best consensus models (colored in red) outperformed 96% of the individual models (colored in gray) on average and half of the winning models.

this type of information/insight could only be learned through a crowdsource exercise like this Challenge.

Methods used by Winning Teams

A wide range of chemical descriptors and/or fingerprints, and machine learning algorithms were employed by the winning teams, including both public tools and commercial or custom in-house software. The sources of chemical descriptors included MOE (Chemical Computing Group Inc., Montreal, Canada), ChemAxon (ChemAxon LLC., Cambridge, MA), Dragon (Talete SRL, Milan, Italy), PaDel (Yap, 2011), RDKit[7], PubChem fingerprint[8], GSFrag (Tetko et al., 2005), ISIDA fragments (Ruggiu et al., 2010), ESTATE indices (Hall and Kier, 1995), AlogPS (Tetko and Tanchuk, 2002), CDK (Steinbeck et al., 2003), inductive descriptors (Cherkasov, 2005), Adriana.Code (Molecular Networks GmbH, Erlangen, Germany), QNPR (Thormann et al., 2007), MERA, and MerSy (Bartashevich et al., 2002), to list a few. Examples of modeling algorithms included Random Forest (Breiman, 2001), deep neural networks (Schmidhuber, 2014), support vector machines (SVM) (Cortes and Vapnik, 1995), Elastic Nets (Zou and Hastie, 2005), Gradient Boosting Decision Trees (Friedman, 1999), Extra Trees (Geurts et al., 2006), associative neural networks (Tetko, 2008), and k-Nearest Neighbors (Altman, 1992). SVM appeared to be a popular algorithm choice among the winning teams. The winners commonly used multiple descriptor types and applied feature selection to select the most relevant descriptors, employed multiple modeling algorithms, and applied consensus models to make the final predictions. In addition to what the Challenge provided, the Grand Challenge winner also used outside data, such as data from literature and public databases including PubChem and ChEMBL (Gaulton et al., 2012).

CONCLUSIONS

The Tox21 Data Challenge produced high quality winning models, thus confirming the ability of computational approaches

[7] http://www.rdkit.org/
[8] https://pubchem.ncbi.nlm.nih.gov/

to provide meaningful predictions of toxicity responses in terms of pathway disruption upon environmental compound exposure using (only) chemical structure information. The combination of the individual models from all participating teams produced better performing consensus models, some of which even outperformed the winning models, showing the wisdom of the crowd. The high predictive performance of these models also serves as a validation of the quality of datasets produced from the Tox21 qHTS assays, which were the basis for this Challenge. The winning models will be made publicly available so that they can be applied to other chemical sets for which no experimental data are available and used to prioritize chemicals for more in-depth toxicity evaluation. All winning models, or better performing consensus models, can be applied in parallel to establish activity/toxicity profiles for these data poor chemicals. Compared to the other challenge participants, the winning teams often applied multiple descriptor types with feature selection, and multiple modeling algorithms to reach consensus predictions. As a follow up study, we will compare in detail the methods used by different teams to determine if there are specific techniques that enabled the winning models to outperform other models.

AUTHOR CONTRIBUTIONS

RH, MX, AR, and AS designed, developed and managed the project. SS, JZ, and SAS performed the experiments and collected data. RH designed the analyses and performed computational analyses of challenge outcomes. DN and TZ implemented the challenge website, participated in the scoring of predictions, collection of code, methods, and outcomes. RH and TZ scored the model submissions. RH managed the challenge and wrote the manuscript. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Programs of the National Toxicology Program (Interagency agreement #Y2-ES-7020-01), National Institute of Environmental Health Sciences, the U.S. Environmental Protection Agency (Interagency Agreement #Y3-HG-7026-03), and the National Center for Advancing Translational Sciences, National Institutes of Health. We would also like to thank Samuel Michael and Carleen Klumpp-Thomas for assisting with the screens, Paul Shinn, Misha Itkin, and Danielle VanLeer for compound management, William Leister for the Tox21 10K library quality control, and Deborah Ngan and Nicole Spears for proof reading the manuscript. The views expressed in this article are those of the authors and do not necessarily reflect the statements, opinions, views, conclusions, or policies of the National Center for Advancing Translational Sciences, National Institutes of Health, or the United States government. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2015.00085

REFERENCES

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46, 175–185.

Attene-Ramos, M. S., Huang, R., Michael, S., Witt, K. L., Richard, A., Tice, R. R., et al. (2015). Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ. Health Perspect.* 123, 49–56. doi: 10.1289/ehp.1408642

Bartashevich, E. V., Potemkin, V. A., Grishina, M. A., and Belik, A. V. (2002). A method for multiconformational modeling of the three-dimensional shape of a molecule. *J. Struct. Chem.* 43, 1033–1039. doi: 10.1023/A:1023611131068

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

Chen, S., Hsieh, J. H., Huang, R., Sakamuru, S., Hsin, L. Y., Xia, M., et al. (2015). Cell-based high-throughput screening for aromatase inhibitors in the Tox21 10K library. *Toxicol. Sci.* 147, 446–457. doi: 10.1093/toxsci/kfv141

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

Collins, F. S., Gray, G. M., and Bucher, J. R. (2008). Toxicology. Transforming environmental health protection. *Science* 319, 906–907. doi: 10.1126/science.1154619

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

Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.* 33, 933–940. doi: 10.1038/nbt.3299

Friedman, J. H. (1999). Greedy function approximation: a gradient boosting machine. Available online at: http://statweb.stanford.edu/~jhf/ftp/trebst.pdf

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–D1107. doi: 10.1093/nar/gkr777

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

Hall, L. H., and Kier, L. B. (1995). Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* 35, 1039–1045. doi: 10.1021/ci00028a014

Hsu, C. W., Zhao, J., Huang, R., Hsieh, J. H., Hamm, J., Chang, X., et al. (2014). Quantitative high-throughput profiling of environmental chemicals and drugs that modulate farnesoid X receptor. *Sci. Rep.* 4:6437. doi: 10.1038/srep06437

Huang, R., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., et al. (2014). Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep.* 4:5664. doi: 10.1038/srep05664

Huang, R., Xia, M., Cho, M. H., Sakamuru, S., Shinn, P., Houck, K. A., et al. (2011). Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. *Environ. Health Perspect.* 119, 1142–1148. doi: 10.1289/ehp.1002952

Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., et al. (2009). The toxicity data landscape for environmental chemicals. *Environ. Health Perspect.* 117, 685–695. doi: 10.1289/ehp.0800168

Kavlock, R. J., Austin, C. P., and Tice, R. R. (2009). Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Anal.* 29, 485–487. discussion: 492–487. doi: 10.1111/j.1539-6924.2008.01168.x

Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715. doi: 10.1038/nrd1470

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., and Pähler, A. (2008). Computational toxicology in drug development. *Drug Discov. Today* 13, 303–310. doi: 10.1016/j.drudis.2007.12.007

Nishihara, K., Huang, R., Zhao, J., Shahane, S. A., Witt, K. L., Smith-Roe, S. L., et al. (2016). Identification of genotoxic compounds using isogenic DNA repair deficient DT40 cell lines on a quantitative high throughput screening platform. *Mutagenesis* 31, 69–81. doi: 10.1093/mutage/gev055

NRC (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press.

PubChem (2013a). *Tox21 Phase II Compound Collection* [Online]. Available online at: http://www.ncbi.nlm.nih.gov/pcsubstance/?term=tox21 (Accessed Dec 4, 2013).

PubChem (2013b). *Tox21 Phase II Data* [Online]. Available online at: http://www.ncbi.nlm.nih.gov/pcassay?term=tox21 (Accessed Nov 16, 2013).

Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. (2010). ISIDA property-labelled fragment descriptors. *Mol. Inform.* 29, 855–868. doi: 10.1002/minf.201000099

Saiakhov, R. D., and Klopman, G. (2008). MultiCASE expert systems and the REACH initiative. *Toxicol. Mech. Methods* 18, 159–175. doi: 10.1080/15376510701857460

Sanderson, D. M., and Earnshaw, C. G. (1991). Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum. Exp. Toxicol.* 10, 261–273. doi: 10.1177/096032719101000405

Schmidhuber, J. (2014). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43, 493–500. doi: 10.1021/ci025584y

Tetko, I. V. (2008). Associative neural network. *Methods Mol. Biol.* 458, 185–202. doi: 10.1007/978-1-60327-101-1_10

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., et al. (2005). Virtual computational chemistry laboratory–design and description. *J. Comput. Aided Mol. Des.* 19, 453–463. doi: 10.1007/s10822-005-8694-y

Tetko, I. V., and Tanchuk, V. Y. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* 42, 1136–1145. doi: 10.1021/ci025515j

Thormann, M., Vidal, D., Almstetter, M., and Pons, M. (2007). Nomen est omen: quantitative prediction of molecular properties directly from IUPAC names. *Open Appl. Inform. J.* 1, 28–32. doi: 10.2174/1874136300701010028

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* 121, 756–765. doi: 10.1289/ehp.1205784

Vedani, A., and Smiesko, M. (2009). *In silico* toxicology in drug discovery - concepts based on three-dimensional models. *Altern. Lab. Anim.* 37, 477–496.

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

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39, 561–577.

Conflict of Interest Statement: The author declares 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 Environmental Science

ORIGINAL RESEARCH
published: 02 February 2016
doi: 10.3389/fenvs.2015.00080

DeepTox: Toxicity Prediction using Deep Learning

Andreas Mayr[1,2†], Günter Klambauer[1†], Thomas Unterthiner[1,2†] and Sepp Hochreiter[1]*

[1] Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria, [2] RISC Software GmbH, Johannes Kepler University Linz, Hagenberg, Austria

The Tox21 Data Challenge has been the largest effort of the scientific community to compare computational methods for toxicity prediction. This challenge comprised 12,000 environmental chemicals and drugs which were measured for 12 different toxic effects by specifically designed assays. We participated in this challenge to assess the performance of Deep Learning in computational toxicity prediction. Deep Learning has already revolutionized image processing, speech recognition, and language understanding but has not yet been applied to computational toxicity. Deep Learning is founded on novel algorithms and architectures for artificial neural networks together with the recent availability of very fast computers and massive datasets. It discovers multiple levels of distributed representations of the input, with higher levels representing more abstract concepts. We hypothesized that the construction of a hierarchy of chemical features gives Deep Learning the edge over other toxicity prediction methods. Furthermore, Deep Learning naturally enables multi-task learning, that is, learning of all toxic effects in one neural network and thereby learning of highly informative chemical features. In order to utilize Deep Learning for toxicity prediction, we have developed the DeepTox pipeline. First, DeepTox normalizes the chemical representations of the compounds. Then it computes a large number of chemical descriptors that are used as input to machine learning methods. In its next step, DeepTox trains models, evaluates them, and combines the best of them to ensembles. Finally, DeepTox predicts the toxicity of new compounds. In the Tox21 Data Challenge, DeepTox had the highest performance of all computational methods winning the grand challenge, the nuclear receptor panel, the stress response panel, and six single assays (teams "Bioinf@JKU"). We found that Deep Learning excelled in toxicity prediction and outperformed many other computational approaches like naive Bayes, support vector machines, and random forests.

Keywords: Deep Learning, deep networks, Tox21, machine learning, tox prediction, toxicophores, challenge winner, neural networks

Edited by:
*Ruili Huang,
National Institutes of Health National Center for Advancing Translational Sciences, USA*

Reviewed by:
*Michael Schmuker,
University of Sussex, UK
Johannes Mohr,
Technische Universität Berlin, Germany*

***Correspondence:**
*Sepp Hochreiter
hochreit@bioinf.jku.at*

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *31 August 2015*
Accepted: *04 December 2015*
Published: *02 February 2016*

Citation:
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

1. INTRODUCTION

Humans are exposed to an abundance of chemical compounds via the environment, nutrition, cosmetics, and drugs. To protect humans from potentially harmful effects, these chemicals must pass reliable tests for adverse effects and, in particular, for toxicity. A compound's effects on human health are assessed by a large number of time- and cost-intensive *in vivo* or *in vitro* experiments. In particular, numerous methods rely on animal tests, trading off additional safety against ethical concerns. The aim of the "Toxicity testing in the Twenty-first century" initiative is to develop more efficient and less time-consuming approaches to predicting how chemicals affect human health (Andersen and Krewski, 2009; Krewski et al., 2010). The most efficient approaches employ computational models that can screen large numbers of compounds in a short time and at low costs (Rusyn and Daston, 2010). However, computational models often suffer from insufficient accuracy and are not as reliable as biological experiments. In order for computational models to replace biological experiments, they must achieve comparable accuracy. Within the "Tox21 Data Challenge" (Tox21 challenge), the performance of computational methods for toxicity testing was assessed in order to judge their potential to reduce *in vitro* experiments and animal testing.

The Tox21 challenge organizers invited participants to build computational models to predict the toxicity of compounds for 12 toxic effects (see **Figure 1**). These toxic effects comprised stress response effects (SR), such as the heat shock response effect (SR-HSE), and nuclear receptor effects (NR), such as activation of the estrogen receptor (NR-ER). Both SR and NR effects are highly relevant to human health, since activation of nuclear receptors can disrupt endocrine system function (Chawla et al., 2001; Grün and Blumberg, 2007), and activation of stress response pathways can lead to liver injury or cancer (Bartkova et al., 2005; Labbe et al., 2008; Jaeschke et al., 2012). For constructing computational models, high-throughput screening assay measurements of these twelve toxic effects were provided. The training set consisted of the Tox21 10K compound library, which includes environmental chemicals and drugs (Huang et al., 2014). For a set of 647 new compounds, computational models had to predict the outcome of the high-throughput screening assays (see **Figure 1**). The assay measurements for these test compounds were withheld from the participants and used to evaluate the performance of the computational methods. The "area under ROC curve" (AUC) was used as a performance criterion that reflects how well a method can rank toxic compounds higher than non-toxic compounds.

The participants in the Tox21 challenge used a broad range of computational methods for toxicity prediction, most of which were from the field of machine learning. These methods represent the chemical compound by chemical descriptors, the features, which are fed into a predictor. Methods for predicting biological effects are usually categorized into similarity-based approaches and feature-based approaches. Similarity-based methods compute a matrix of pairwise similarities between compounds which is subsequently used by the prediction algorithms. These methods, which are based on the idea that similar compounds should have a similar biological effect include nearest neighbor algorithms (e.g., Kauffman and Jurs, 2001; Ajmani et al., 2006; Cao et al., 2012) and support vector machines (SVMs, e.g., Mahé et al., 2005; Niu et al., 2007; Darnag et al., 2010). SVMs rely on a kernel matrix which represents the pairwise similarities of objects. In contrast to similarity based methods, feature based methods either select input features (chemical descriptors) or weight them by a score or a model parameter. Feature-based approaches include (generalized) linear models (e.g., Luco and Ferretti, 1997; Sagardia et al., 2013), random forests, (e.g., Svetnik et al., 2003; Polishchuk et al., 2009), and scoring schemes based on naive Bayes (Bender et al., 2004; Xia et al., 2004). Choosing informative features for the task at hand is key in feature-

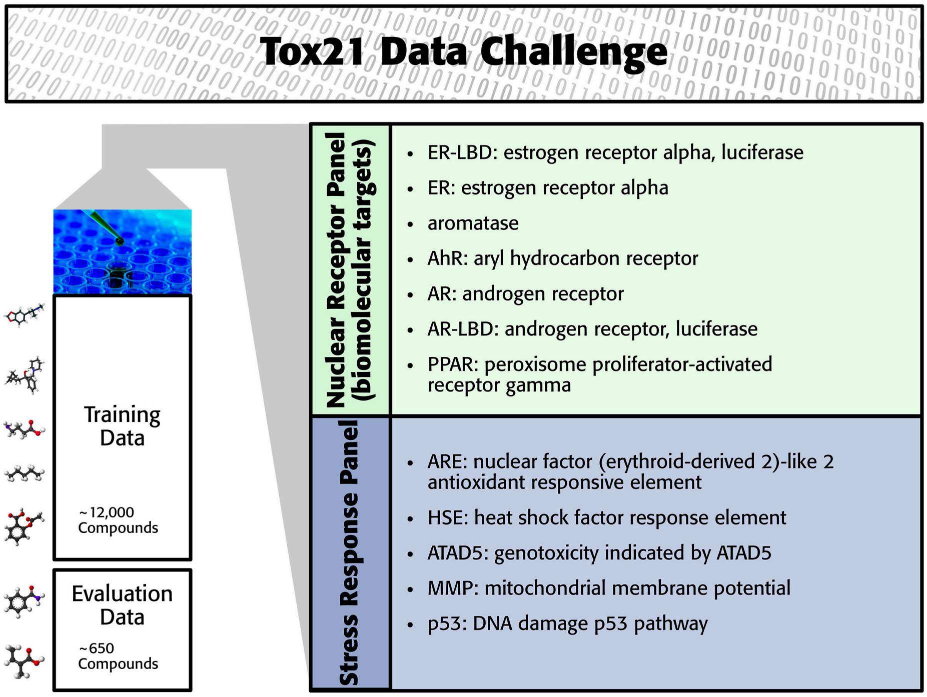

Tox21 Data Challenge
Training Data
~12,000 Compounds
Evaluation Data
~650 Compounds
Nuclear Receptor Panel (biomolecular targets)
• ER-LBD: estrogen receptor alpha, luciferase
• ER: estrogen receptor alpha
• aromatase
• AhR: aryl hydrocarbon receptor
• AR: androgen receptor
• AR-LBD: androgen receptor, luciferase
• PPAR: peroxisome proliferator-activated receptor gamma
Stress Response Panel
• ARE: nuclear factor (erythroid-derived 2)-like 2 antioxidant responsive element
• HSE: heat shock factor response element
• ATAD5: genotoxicity indicated by ATAD5
• MMP: mitochondrial membrane potential
• p53: DNA damage p53 pathway

FIGURE 1 | Overview of the Tox21 challenge dataset.

based methods and requires deep insights into chemical and biological properties and processes (Verbist et al., 2015), such as interactions between molecules (e.g., ligand-target), reactions and enzymes involved, and metabolic modifications of the molecules. Similarity-based approaches, in contrast, require a proper similarity measure between two compounds. The measure may use a feature-based, a 2D graph-based, or a 3D representation of the compound. Graph-based compound and molecule representations led to the invention of graph and molecule kernels (Kashima et al., 2003, 2004; Ralaivola et al., 2005; Mahé et al., 2006; Mohr et al., 2008; Vishwanathan et al., 2010; Klambauer et al., 2015). These methods are not able to automatically create task-specific or new chemical features. Deep Learning, however, excels in constructing new, task-specific features that result in data representations which enable Deep Learning methods to outperform previous approaches, as has been demonstrated in various speech and vision tasks.

Deep Learning (LeCun et al., 2015; Schmidhuber, 2015) has emerged as a highly successful field of machine learning. It has already impacted a wide range of signal and information processing fields, redefining the state of the art in vision (Cireşan et al., 2012a; Krizhevsky et al., 2012), speech recognition (Dahl et al., 2012; Deng et al., 2013; Graves et al., 2013), text understanding and natural language processing (Socher and Manning, 2013; Sutskever et al., 2014), physics (Baldi et al., 2014), and life sciences (Cireşan et al., 2013). MIT Technology Review selected it as one of the 10 technological breakthroughs of 2013. Deep Learning has already been applied to predict the outcome of biological assays (Dahl et al., 2014; Unterthiner et al., 2014, 2015; Ma et al., 2015), which made it our prime candidate for toxicity prediction.

Deep Learning is based on artificial neural networks with many layers consisting of a high number of neurons, called deep neural networks (DNNs). A formal description of DNNs is given in Section 2.2.1. In each layer Deep Learning constructs features in neurons that are connected to neurons of the previous layer. Thus, the input data is represented by features in each layer, where features in higher layers code more abstract input concepts (LeCun et al., 2015). In image processing, the first DNN layer detects features such as simple blobs and edges in raw pixel data (Lee et al., 2009; see **Figure 2**). In the next layers these features are combined to parts of objects, such as noses, eyes and mouths for face recognition. In the top layers the objects are assembled from features representing their parts such as faces.

The ability to construct abstract features makes Deep Learning well suited to toxicity prediction. The representation of compounds by chemical descriptors is similar to the representation of images by DNNs. In both cases the representation is hierarchical and many features within a layer are correlated. This suggests that Deep Learning is able to construct abstract chemical descriptors automatically. The constructed features can indicate functional groups or toxicophores (Kazius et al., 2005) as visualized in **Figure 3**.

The construction of indicative abstract features by Deep Learning can be improved by *Multi-task learning*. Multi-task learning incorporates multiple tasks into the learning process (Caruana, 1997). In the case of DNNs, different related tasks share features, which therefore capture more general chemical characteristics. In particular, multi-task learning is beneficial for a task with a small or imbalanced training set, which is common in computational toxicity. In this case, due to insufficient information in the training data, useful features cannot be constructed. However, multi-task learning allows this task to

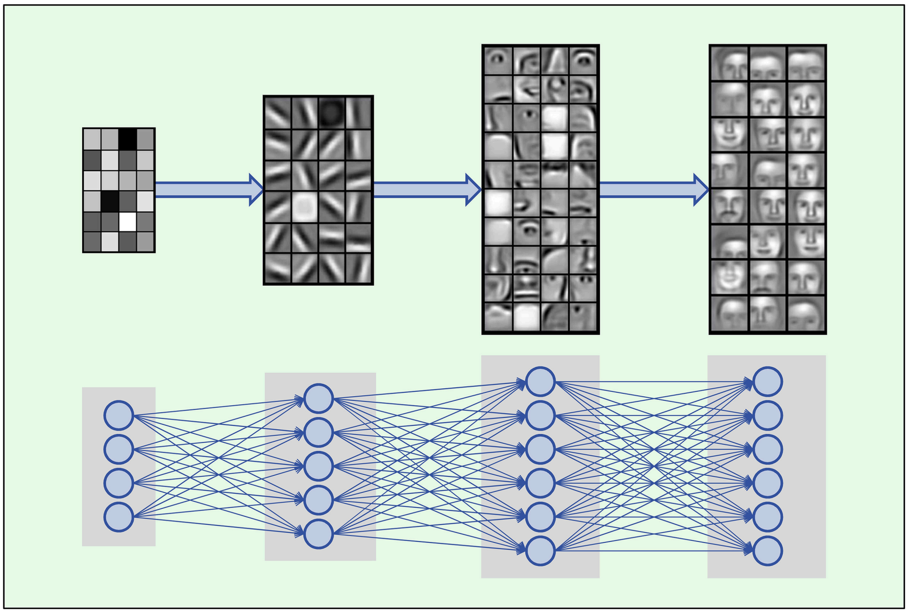


FIGURE 2 | Hierarchical composition of complex features. DNNs build a feature from simpler parts. A natural hierarchy of features arises. Input neurons represent raw pixel values which are combined to edges and blobs in the lower layers. In the middle layers contours of noses, eyes, mouths, eyebrows and parts thereof are built, which are finally combined to abstract features such as faces. Images adopted from Lee et al. (2011) with permission from the authors.

borrow features from related tasks and, thereby, considerably increases the performance.

Deep Learning thrives on large amounts of training data in order to construct indicative features (Krizhevsky et al., 2012) and, thereby, well-performing models. Recently, the availability of high-throughput toxicity assays provides sufficient data to use Deep Learning for toxicity prediction (Andersen and Krewski, 2009; Krewski et al., 2010; Shukla et al., 2010). In summary, Deep Learning is likely to perform well with the following prerequisites:

Large dataset: "Big data"	Several thousand data points must be available to allow the Deep Learning method to learn hierarchical representations of the data.
Many related input features	Multiple similar, i.e., correlated, inputs must be available. This allows very robust hidden representations.
Multi-task setting	Each data point has multiple possible output classes. The hidden representations can be shared across tasks, enhancing performance.

These three conditions are fulfilled for the Tox21 dataset: (1) High throughput toxicity assays have provided vast amounts of data. (2) Chemical compound descriptors are correlated. (3) A Multi-task setting is natural as different assays measure different but related toxic effects for the same compound (see **Figure 4**). To conclude, Deep Learning seems promising for computational toxicology because of its ability to construct abstract chemical features.

2. MATERIALS AND METHODS

For the Tox21 challenge, we used Deep Learning as key technology, for which we developed a prediction pipeline (DeepTox) that enables the use of Deep Learning for toxicity prediction. The DeepTox pipeline was developed for datasets with characteristics similar to those of the Tox21 challenge dataset and enables the use of Deep Learning for toxicity prediction. We first introduce the challenge dataset in Section 2.1. In Section 2.2 we then present, how we utilized Deep Learning for Toxicity prediction, while in Section 2.3 the DeepTox pipeline is explained.

2.1. Tox21 Challenge Data

In the Tox21 challenge, a dataset with 12,707 chemical compounds was given. This dataset consisted of a training dataset of 11,764, a leaderboard set of 296, and a test set of 647 compounds. For the training dataset, the chemical structures and assay measurements for 12 different toxic effects were fully available to the participants right from the beginning of the challenge, as were the chemical structures of the leaderboard set. However, the leaderboard set assay measurements were withheld by the challenge organizers during the first phase of the competition and used for evaluation in this phase, but were released afterwards, such that participants could improve their models with the leaderboard data for the final evaluation.

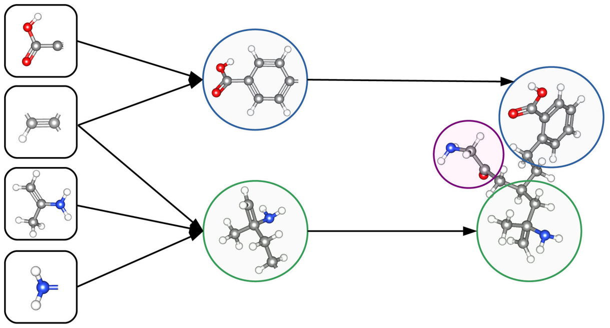


FIGURE 3 | Representation of a toxicophore by hierarchically related features. Simple features share chemical properties coded as reactive centers. Combining reactive centers leads to toxicophores that represent specific toxicological effects.

Table 1 lists the number of active and inactive compounds in the training and the leaderboard sets of each assay. The final evaluation was done on a test set of 647 compounds, where only the chemical structures were made available. The assay measurements were only known to the organizers and had to be predicted by the participants. In summary, we had a training set consisting of 11,764 compounds, a leaderboard set consisting of 296 compounds, both available together with their corresponding assay measurements, and a test set consisting of 647 compounds to be predicted by the challenge participants (see **Figure 1**). The chemical compounds were given in SDF format, which contains the chemical structures as undirected, labeled graphs whose nodes and edges represent atoms and bonds, respectively. The outcomes of the measurements were categorized (i.e., that is labeled) as "active," "inactive," or "inconclusive/not tested." Not all compounds were measured on all assays (see **Figure 4A**).

2.2. Deep Learning for Toxicity Prediction

Deep Learning is a highly successful machine learning technique that has already revolutionized many scientific areas. Deep Learning comprises an abundance of architectures such as deep neural networks (DNNs) or convolutional neural networks. We propose a DNNs for toxicity prediction and present the method's details and algorithmic adjustments in the following. First we introduce neural networks, and in particular DNNs, in Section 2.2.1. In Section 2.2.2, we then discuss key techniques that led to the success of DNNs compared to shallow and small neural networks. The objective that was minimized for the DNNs for toxicity prediction and the corresponding optimization algorithms are discussed in Section 2.2.3. We explain DNN hyperparameters and the DNN architectures used in Section 2.2.4. In Section 2.2.5, we describe the hardware that was employed to optimize the objectives of the DeepTox DNNs.

2.2.1. Deep Neural Networks

A neural network, and a DNN in particular, can be considered as a function that maps an input vector to an output vector. The mapping is parameterized by weights that are optimized in a learning process. In contrast to shallow networks, which have only one hidden layer and only few hidden neurons per layer,

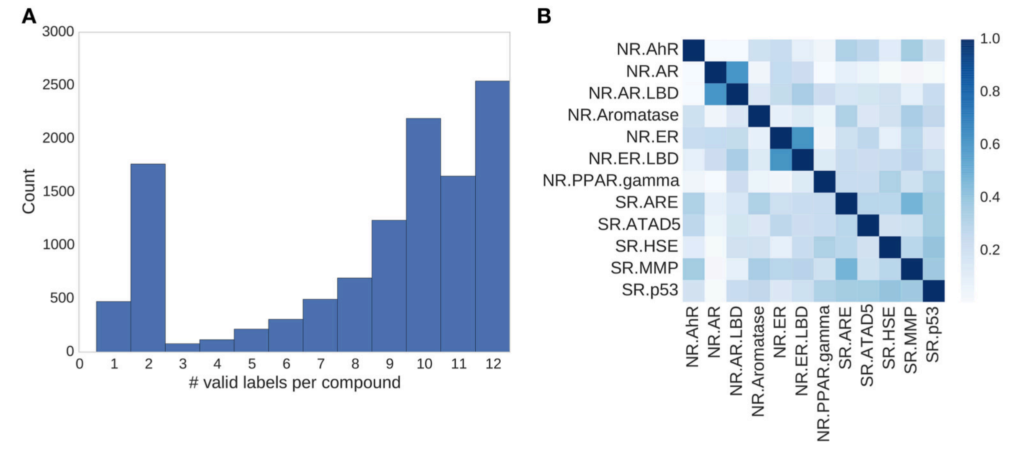


FIGURE 4 | Assay correlation. (A) Histogram showing the number of unambiguous assay label assignments per compound. Only ≈500 compounds had a label for just one assay, more than half (54%) of the compounds had labels for 10 or more tasks. **(B)** Absolute correlation coefficient between the different assays of the Tox21 challenge.

TABLE 1 | Number of active and inactive compounds in the training (Train) and the leaderboard (Leader) sets of each assay.

Set	Class	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
Train	Inactive	7219	8982	8296	6069	6866	8753	6760	8307	7722	6178	8097	7962
Train	Active	950	380	303	1098	360	338	937	446	428	1142	537	222
Leader	Inactive	241	289	249	186	196	247	238	277	257	200	241	252
Leader	Active	31	3	4	48	18	25	27	10	10	38	28	15

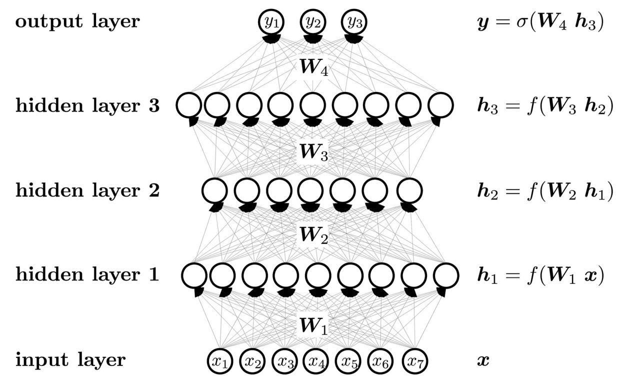

output layer
y_1 y_2 y_3
$\boldsymbol{y} = \sigma(\boldsymbol{W}_4\, \boldsymbol{h}_3)$
$\boldsymbol{W}_4$
hidden layer 3
$\boldsymbol{h}_3 = f(\boldsymbol{W}_3\, \boldsymbol{h}_2)$
$\boldsymbol{W}_3$
hidden layer 2
$\boldsymbol{h}_2 = f(\boldsymbol{W}_2\, \boldsymbol{h}_1)$
$\boldsymbol{W}_2$
hidden layer 1
$\boldsymbol{h}_1 = f(\boldsymbol{W}_1\, \boldsymbol{x})$
$\boldsymbol{W}_1$
input layer
x_1 x_2 x_3 x_4 x_5 x_6 x_7
$\boldsymbol{x}$

FIGURE 5 | Schematic representation of a DNN.

DNNs comprise many hidden layers with a great number of neurons. A DNN may have thousands of neurons in each layer (Cireşan et al., 2012b), which is in contrast to traditional artificial neural networks, that employ only a small number of neurons. The goal is no longer to just learn the main pieces of information, but rather to capture all possible facets of the input.

A neuron can be considered as an abstract feature with a certain activation value that represents the presence of this feature. A neuron is constructed from neurons of the previous layer, that is, the activation of a neuron is computed from the activation of neurons one layer below. The first layer is the "input layer," in which neuron activations are set to the value of the input vector. The last layer is the "output layer," where the activations represent the output vector. The intermediate layers are the "hidden layers," which give intermediate representations of the input vector.

Figure 5 visualizes the neural network mapping of an input vector to an output vector. A compound is described by the vector of its input features $\mathbf{x}$. The neural network NN maps the input vector $\mathbf{x}$ to the output vector $\mathbf{y}$. The activation value h_j^l of a neuron j in a layer l of the neural network is computed as the weighted sum over the values h_i^{l-1} of all neurons i in layer $(l-1)$, followed by the application of an activation function f. The weight w_{ji}^l scales the activation h_i^{l-1} of neuron i in layer $(l-1)$ before it is summed to compute the activation of neuron j in layer l. If the neural network has m layers, then the formulas are

$$\begin{aligned} \mathbf{y} &= \mathrm{NN}(\mathbf{x})\,, \\ \mathbf{h}^0 &= \mathbf{x}\,, \\ h_j^l &= f\Big(\sum_i w_{ji}^l \, h_i^{l-1} \Big)\,, \\ \mathbf{y} &= \mathbf{h}^m\,. \end{aligned}$$

In matrix notation, the activation of neurons is

$$\mathbf{h}^l = f\left(\mathbf{W}^l \, \mathbf{h}^{l-1}\right).$$

The output layer often has a special activation function, which is denoted by σ instead of f in **Figure 5**. Each neuron has a bias weight (i.e., a constant offset), that is added to the weighted sum for computing the activation of a neuron. To keep the notation uncluttered, these bias weights are not written explicitly, although they are model parameters like other weights.

2.2.2. Key Techniques for Deep Neural Networks

Recent algorithmic improvements in training DNNs enabled the success of Deep Learning: (1) "rectified linear units" (ReLUs) enforce sparse representations and counteract the vanishing gradient, (2) "dropout" for regularization, and (3) a cross-entropy objective combined with softmax or sigmoid activation.

One of the most successful inventions in the context of DNNs are rectified linear units (ReLUs) as activation functions (Nair and Hinton, 2010; Glorot et al., 2011). A ReLU f is the identity for positive values and zero otherwise. This activation function is called the "ramp function":

$$f(x) = \max(0, x).$$

Using ReLUs in DNNs leads to sparse input representations, which are robust against noise and advantageous for classifiers because classification is more likely to be easier in higher-dimensional spaces (Ranzato et al., 2008). Probably the most important advantage of ReLUs is that they are a remedy for the vanishing gradient (Hochreiter, 1991; Hochreiter et al., 2000), from which networks with sigmoid activation functions and many layers suffer. "Vanishing" means in this context that the length of a gradient decreases exponentially when propagated through the layers, ultimately becoming too small for learning in the lower(/est) layers. Another enabling technique is "dropout," which is one of the new regularization schemes that arose with the advent of DNNs in order to prevent overfitting—a serious problem for DNNs, as the number of hidden neurons is large and the complexity of the model class is very high. Dropout avoids co-adaption of units by randomly dropping units during training, that is, setting their activations and derivatives to zero (Hinton et al., 2012; Srivastava et al., 2014). The third technique that paved the way for the success of DNNs is the application of error functions such as cross-entropy and logistic-loss as objectives to be minimized. These error functions are combined with softmax or sigmoid activation functions in the output neurons.

2.2.3. DNN Learning, Objective and Optimization

The goal of neural network learning is to adjust the network weights such that the input-output mapping has a high predictive power on future data. We want to explain the training data, that is, to approximate the input-output mapping on the training data. Our goal is therefore to minimize the error between predicted and known outputs on that data. The training data consists of the output vector $\mathbf{t}$ for input vector $\mathbf{x}$, where the input vector is represented using d chemical features, and the length of the output vector is n, the number of tasks. Let us consider a *classification task*. For classification, the output component t_k for task k is binary, that is, $t_k \in \{0, 1\}$. In the case of toxicity prediction, the tasks represent different toxic effects, where zero indicates the absence and one the presence of a toxic effect. The neural network predicts the outputs y_k. In the output layer of the neural network a sigmoid activation function is used. Therefore, the neural network predicts outputs y_k, that are between 0 and 1, and the training data are perfectly explained if for all training examples all outputs k are predicted correctly, i.e., $y_k = t_k$. To penalize non-matching output-target pairs, an error function or objective is defined. Minimizing this error function means better aligning network outputs and targets. Typically, the cross-entropy is used as an error function for multi-class classification. In our case, we deal with *multi-task classification*, where multiple outputs can be one (multiple different toxic effects for one compound) or none can be one (no toxic effect at all). For the multi-task setting we use a logistic error function $-t_k \log(y_k) - (1 - t_k)\log(1 - y_k)$ for each output component k. If $t_k = y_k$, then only terms $(1 \log 1)$ or $(0 \log 0)$ appear, and the logistic error function is zero (note that $(0 \log 0)$ is defined to be zero). Otherwise, the logistic error function gives a positive value. The overall error function is the sum of these logistic error functions across all output components:

$$-\sum_{k=1}^{n} t_k \, \log(y_k) + (1 - t_k) \, \log(1 - y_k).$$

To cope with missing labels, we introduce a binary vector $\mathbf{m}$ for each sample, where m_k is one if the sample has a label for task k and zero otherwise. This leads to a slight modification to the above objective:

$$-\sum_{k=1}^{n} m_k \left(t_k \, \log(y_k) + (1 - t_k) \, \log(1 - y_k)\right).$$

Learning minimizes this objective with respect to the weights, as the outputs y_k are parametrized by the weights. The optimization problem is usually solved by gradient descent, which aims to minimize an objective function by iteratively adapting the parameters of the optimization problem in the direction of the steepest descent (the negative gradient) until a stationary point is found. A critical parameter is the step size or learning rate, i.e., how strongly the parameters are changed in the update direction. If a small step size is chosen, the parameters converge slowly to the local optimum. If the step size is too high, the parameters oscillate.

For neural networks, gradient descent can be applied with high computational efficiency by using the backpropagation algorithm (Werbos, 1974; Rumelhart et al., 1986). A computational simplification to computing a gradient over all training samples is *stochastic gradient descent* (Bottou, 2010). Stochastic gradient descent computes a gradient for an equally-sized set of randomly chosen training samples, *a mini-batch*, and updates the parameters according to this mini-batch gradient (Ngiam et al., 2011). The advantage of stochastic gradient descent is that the parameter updates are faster. The main disadvantage of stochastic gradient descent is that the parameter updates are more imprecise. For large datasets the increase in speed clearly outweighs the imprecision.

2.2.4. Hyperparameter Settings and DNN Network Architectures

The DeepTox pipeline assesses a variety of DNN architectures and hyperparameters. The networks consist of multiple layers of ReLUs, followed by a final layer of sigmoid output units, one for each task. One output unit is used for single-task learning. In the Tox21 challenge, the numbers of hidden units per layer were 1024, 2048, 4096, 8192, or 16,384. DNNs with up to four hidden layers were tested. Very sparse input features that were present in fewer than 5 compounds were filtered out, as these features would have increased the computational burden, but would have included too little information for learning. DeepTox uses stochastic gradient descent learning to train the DNNs (see Section 2.2.3), employing mini-batches of 512 samples. To regularize learning, both dropout (Srivastava et al., 2014) and L2 weight decay were implemented for the DNNs in the DeepTox pipeline. They work in concert to avoid overfitting (Krizhevsky et al., 2012; Dahl et al., 2014). Additionally, DeepTox uses early stopping, where the learning time is determined by cross-validation.

Table 2 shows a list of hyperparameters and architecture design parameters that were used for the DNNs, together with their search ranges. The best hyperparameters were determined by cross-validation using the AUC score as quality criterion. Even though multi-task networks were employed, the hyperparameters were optimized individually for each task. The evaluation of the models by cross-validation as implemented in the DeepTox pipeline is described in Section 2.3.4.

2.2.5. GPU Implementation

Graphics Processor Units (GPUs) have become essential tools for Deep Learning, because the many layers and units of a DNN give rise to a massive computational load, especially regarding CPU performance. Only through the recent advent of fast accelerated hardware such as GPUs has training a DNN model become feasible (Schmidhuber, 2015). As described in Section 2.2.1, the main equations of a neural net can be written in terms of matrix/vector operations, which are prime candidates for execution on massively parallel hardware architectures. Using state-of-the-art GPU hardware speeds up the training process by several orders of magnitude compared to using an optimized multi-core CPU implementation (Raina et al., 2009). Hence, we implemented the DNNs using the CUDA parallel computing platform and employed NVIDIA Tesla K40 GPUs to achieve speed-ups of 20–100x compared to CPU implementations (see Supplementary Section 5 for an overview on the computational resources that were used).

TABLE 2 | Hyperparameters considered for the neural networks.

Hyperparameter	Values considered
Scaling of predefined features	{standard-deviation, tanh, sqrt}
Number of Hidden Units	{1024, 2048, 4096, 8192, 16,384}
Number of Layers	{1, 2, 3, 4}
Backpropagation Learning Rate	{0.01, 0.05, 0.1}
Dropout usage/rate	{no, yes (50% Hidden Dropout, 20% Input Dropout)}
L2 Weight Decay	{0, 10^{-6}, 10^{-5}, 10^{-4}}

2.3. The DeepTox Pipeline

As mentioned above, we developed a pipeline, which enables the usage of DNNs for toxicity prediction. The pipeline receives raw training data and supplies predictions for new data. In detail "DeepTox" consists of: (1) cleaning and quality control of the data containing the chemical description of the compounds (Section 2.3.1), (2) creating chemical descriptors as input features for the models (Section 2.3.2), (3) model selection including feature selection if required by the model class (Section 2.3.3), (4) evaluating the quality of models in order to choose the best ones (Section 2.3.4), and (5) combining models to ensemble predictors (Section 2.3.5). The individual steps of the pipeline are visualized as boxes in **Figure 6**.

2.3.1. Data Cleaning and Quality Control

In the first step, DeepTox improves the quality of the training data. We had observed that the chemical substances in question are often mixtures of distinct chemical structures that are not connected by covalent bonds. Therefore, we introduced

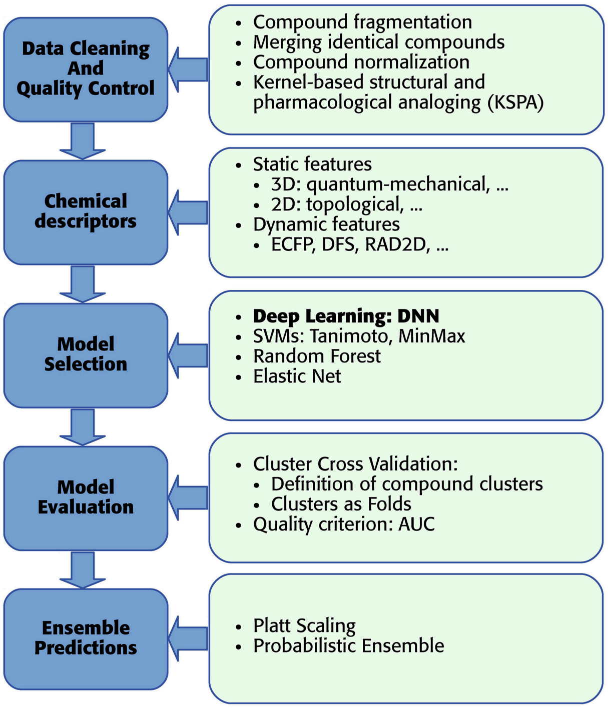


FIGURE 6 | DeepTox pipeline for toxicity prediction.

a fragmentation step to the DeepTox pipeline. In this step, these distinct structures are split into individual "compound fragments." Examples of frequently recurring compound fragments are Na^+ and Cl^- ions. Upon fragmentation, identical compound fragments can appear multiple times, which are merged by DeepTox. In this merging step, DeepTox semi-automatically labels merged compound fragments, removing contradictory and keeping agreeing measurements. Compound fragments that appear in multiple mixtures can have varying toxicity measurements since Tox21 testing was based on mixtures. If all measurements agree, the fragments are automatically labelled. For disagreeing measurements, an operator has to disentangle the contradictory measurements by assigning activities to compounds in the mixture. If this is impossible, the label is marked to be unknown. All fragments are then normalized by making "H"-atoms explicit and representing aromatic bonds/tautomers consistently, by calculating a canonical formula (Thalheim et al., 2010) using the software Chemaxon. After merging and normalization, the size of the dataset might be reduced. In the case of the Tox21 challenge dataset, 12,707 compounds were reduced to 8694 distinct fragments. To counteract the reduction in the training set size, an optional augmentation step was introduced to DeepTox: kernel-based structural and pharmacological analoging (KSPA), which has been very successful in toxicogenetics (Eduati et al., 2015). The central idea of KSPA is that public databases already contain toxicity assays that are similar to the assay under investigation. KSPA identifies these similar assays by high correlation values and adds their compounds and measurements to the given dataset. Thus, the dataset is enriched with both similar structures and similar assays from public data (see Supplementary Section 2). This typically leads to a performance improvement of Deep Learning methods due to increased datasets. Overall, the data cleaning and quality control procedure improves the predictive performance of the DNNs.

2.3.2. Chemical Descriptors

For Deep Learning, a large number of correlated features is favorable to achieve high performance (see Sections 1 and Krizhevsky et al., 2012). Hence, DeepTox calculates as many types of features as possible, which can be grouped into two basic categories: static and dynamic features. Static features are typically identified by experts as promising properties for predicting biological activity or toxicity. Examples are atom counts, surface areas, and the presence or absence of a predefined substructure in a compound. Since static features are defined a priori, the number of static features that represent a molecule is fixed. For the static features, DeepTox calculates a number of numerical features based on the topological and physical properties of each compound using off-the-shelf software (Cao et al., 2013). These static features include weight, Van der Waals volume, and partial charge information. DeepTox also calculates the presence and absence of 2500 predefined toxicophore features, i.e., patterns of substructures previously reported as toxicophores in the literature (e.g., Kazius et al., 2005), and standard binary and count features such as MACCS and PCFP. Dynamic features are extracted on the fly from the chemical structure of a compound in a prespecified way (e.g., ECFP fingerprint features, Rogers and Hahn, 2010) The DeepTox pipeline uses JCompoundMapper (Hinselmann et al., 2011) to create dynamic features. Dynamic features are often highly specific and therefore sparse. Even if a huge (possibly infinite) number of different dynamic features exists, handling the dataset would remain feasible, as absent features are not reported. Normally, either the presence of a feature (binary) or the count of a feature (discrete) is reported for each compound. While many of these sparse features may be uninformative, some dynamic features may be specific to toxic effects.

The DeepTox pipeline uses a large number of different types of static or dynamic features (see Supplementary Section 1). Different types of input features have substantially different scales and distributions which poses a problem for DNNs. To make all of them available in the same range, DeepTox both standardizes real-valued and count features and applies the tanh nonlinearity. If the software libraries fail to compute a particular feature, median-imputation is performed to substitute the missing value before standardization. The Tox21 dataset in particular comprised several thousands of static features and hundreds of millions of dynamic features that were sparsely coded.

2.3.3. DeepTox Model Selection and Complementary Models

Model Selection is the key step in the DeepTox pipeline. Its goal is to find a model that describes the training data (i.e., assay measurements of compounds) well and can be used to predict assay outcomes of unmeasured compounds.

The main workhorses in the model building part of the DeepTox pipeline are Deep Neural Networks (DNNs), which are described above. Here, we present complementary learning techniques that are included in the DeepTox model building part. These techniques include SVMs, random forests (RF), and elastic nets. These methods are used for cross-checking, supplementing the Deep Learning models, and for ensemble learning to complement DNNs. DeepTox considers both similarity-based method, such as SVMs, and feature-based methods, such as random random forests and elastic nets.

2.3.3.1. Support vector machines

SVMs are large-margin classifiers that are based on the concept of structural risk minimization. They are widely used in chemoinformatics (Mohr et al., 2010; Rosenbaum et al., 2011). SVMs are similarity-based machine learning methods and therefore depend on a kernel function that determines the similarity of two compounds.

The choice of similarity measure is crucial to the performance of SVMs. DeepTox uses a linear kernel as a similarity measure between two compounds **x** and **z**, and variations of the Tanimoto kernel:

- $K_{\text{linear}}(\mathbf{x}, \mathbf{z}) = \sum_{p \in \mathcal{P}} N(p, \mathbf{x}) \cdot N(p, \mathbf{z})$,
- $K_{\text{Minmax}}(\mathbf{x}, \mathbf{z}) = \frac{\sum_{p \in \mathcal{P}} \min\ N(p,\mathbf{x}),N(p,\mathbf{z})}{\sum_{p \in \mathcal{P}} \max\ N(p,\mathbf{x}),N(p,\mathbf{z})}$,
- $K_{\text{Minmax\_new}}(\mathbf{x}, \mathbf{z}) = \frac{\sum_{p \in \mathcal{P}\ N(p,\mathbf{x})+N(p,\mathbf{z})>0} \frac{\min(N(p,\mathbf{x}),N(p,\mathbf{z}))}{\max(N(p,\mathbf{x}),N(p,\mathbf{z}))}}{\sum_{p \in \mathcal{P}\ N(p,\mathbf{x})+N(p,\mathbf{z})>0} 1}$,

where $N(p, \mathbf{x})$ quantifies feature p for compound $\mathbf{x}$, and $\mathcal{P}$ features are considered for a set of compounds. For binary input features, $N(p, \mathbf{x})$ indicates whether a substructure p occurs in the molecule $\mathbf{x}$. For integer-valued input features, $N(p, \mathbf{x})$ is the standardized occurrence count of p in $\mathbf{x}$. For real-valued input features, $N(p, \mathbf{x})$ is the standardized value of a feature p for molecule $\mathbf{x}$.

Our novel MinMax kernel $K_{\mathrm{Minmax\_new}}(\mathbf{x}, \mathbf{z})$ allows continuous features (e.g., partial charges) to be combined with with discrete (e.g., atom counts) and binary (e.g., substructure indicators) features. Since only positive values are allowed, DeepTox splits continuous and count features into positive and negative parts after centering them by the mean or the median.

The hyperparameters for learning SVM models are the SVM regularization parameter, a shrinkage/growth parameter for the kernel similarity, and weights of kernel matrices. Hyperparameters were selected as for DNNs.

2.3.3.2. Random forests

Random forest (Breiman, 2001) approaches construct decision trees for classification, and average over many decision trees for the final classification. Each individual tree uses only a subset of samples and a subset of features, both chosen randomly. In order to construct decision trees, features that optimally separate the classes must be chosen at each node of the tree. Optimal features can be selected based on the information gain criterion or the Gini coefficient. The hyperparameters for random forests are the number of trees, the number of features considered in each step, the number of samples, the feature choice, and the feature type. Random forests require a preprocessing step that reduces the number of features. The t-test and Fisher's exact test were used for real-valued and binary features, respectively.

2.3.3.3. Elastic net

Elastic nets (Friedman et al., 2010; Simon et al., 2011) learn linear regression functions. They basically compute least-square solutions. However, in contrast to ordinary least squares the objective includes a penalty term—a weighted combination between the pure L1 and the pure L2 norm on the coefficients of the linear function. The L1 and L2 regularization leads to sparse solutions via the L1 term and to solutions without large coefficients via the L2 term. The L1 term selects features, and the L2 term prevents model overfitting due to over-reliance on single features. In the Tox21 challenge DeepTox used only static features for elastic net. Since elastic nets built this way typically showed poorer performance than Deep Learning, SVMs and random forests, they were rarely included in the ensembles of the Tox21 challenge.

2.3.4. Model Evaluation

DeepTox determines the performance of our methods by *cluster cross-validation*. In contrast to standard cross-validation, in which the compounds are distributed randomly across cross-validation folds, clusters of compounds are distributed. Concretely, we used Tanimoto similarity based on ECFP4 fingerprints and single linkage clustering to identify compound clusters. A similarity threshold of 0.7 gave us many small clusters that we then distributed randomly across the folds. DeepTox considers two aspects for defining the cross-validation folds: the ratio of actives to inactives and the similarity of compounds.

The ratio of actives to inactives in the cross-validation folds should be close to the ratio expected in future data. In the Tox21 challenge training dataset, a certain number of compounds were measured in only a few assays, whereas we expected the compounds in the final test set to be measured in all twelve assays. Therefore, in the cross-validation folds, only compounds with labels from at least eight of the twelve assays were included. Thus, we ensured that the ratios of actives to inactives in the cross-validation folds were similar to that in the final test data.

The compounds in different cross-validation folds should not be overly similar. A compound in the test fold that is similar to a compound in the training folds could easily be classified correctly by all methods simply based on the overall similarity. In this case, information about the performance of the methods is lost. To avoid that excessively similar compounds are in the test and in the training fold during model evaluation, DeepTox performs cluster cross-validation, which guarantees a minimum distance between compounds of all folds (even across all clusters) if single-linkage clustering is performed. In the challenge, the clusters that resulted from single-linkage clustering of the compounds were distributed among five cross-validation folds. The similarity measure for clustering was the chemical similarity given by ECFP4 fingerprints. In cluster cross-validation, cross-validation folds contain structurally similar compounds that often share the same scaffold or large substructures.

For the Tox21 challenge, the compounds of the leaderboard set were considered to be an additional cross-validation fold. Aside from computing a mean performance over the cross-validation folds, DeepTox also considered the performance on the leaderboard fold as an additional criterion for performance comparisons.

2.3.5. Ensembles of Models

DeepTox constructs ensembles that contain DNNs and complementary models. For the ensembles, the DeepTox pipeline gives high priority to DNNs, as they tend to perform better than other methods. The pipeline selects ensemble members based on their cross-validation performance and, for the Tox21 challenge dataset, their performance on the leaderboard set. DeepTox uses a variety of criteria to choose the methods that form the ensembles, which led to the different final predictions in the challenge. These criteria were the cross-validation performances and the performance on the leader board set, as well as independence of the methods. The performance criteria ensure that very high-performing models form the ensembles, while the independence criterion ensures that ensembles consist of models built by different methods, or that ensembles are built from different sets of features.

A problem that arises when building ensembles is that values predicted by different models are on different scales. To make the predictions comparable, DeepTox employs Platt scaling (Platt, 1999) to transform them into probabilistic predictions. Platt scaling uses a separate cross-validation run to supply probabilities. Note that probabilities predicted by models such as logistic regression are not trustworthy as they can overfit to the training set. Therefore, a separate run with predictions on unseen data must be performed to calibrate the predictions of a model in such a way that they are trustworthy probabilities. Since the arithmetic mean is not a reasonable choice for combining the predictions of different models, DeepTox uses a probabilistic approach with similar assumptions as naive Bayes (see Supplementary Section 3) to fully exploit the probabilistic predictions in our ensembles.

3. RESULTS

3.1. Benefit of Multi-Task Learning

We were able to apply multi-task learning in the Tox21 challenge because most of the compounds were labeled for several tasks (see Section 1). Multi-task learning has been shown to enhance the performance of DNNs when predicting biological activities at the protein level (Dahl et al., 2014). Since the twelve different tasks of the Tox21 challenge data were highly correlated, we implemented multi-task learning in the DeepTox pipeline.

To investigate whether multi-task learning improves the performance, we compared single-task and multi-task neural networks on the Tox21 leaderboard set. Furthermore, we computed an SVM baseline (linear kernel). **Table 3** lists the resulting AUC values and indicates the best result for each task in italic font. The results for DNNs are the means over 5 networks with different random initializations. Both multi-task and single-task networks failed on an assay with a very unbalanced class distribution. For this assay, the data contained only 3 positive examples in the leaderboard set. *For 10 out of 12 assays, multi-task networks outperformed single-task networks.*

TABLE 3 | Comparison: multi-task (MT) with single-task (ST) learning and SVM baseline evaluated on the leaderboard-set.

Task	AUC MT	AUC ST	AUC SVM
NR.AhR	0.8409	*0.8487*	0.8289
NR.AR	0.3459	*0.3755*	0.3344
NR.AR.LBD	*0.9289*	0.8799	0.8771
NR.Aromatase	*0.7921*	0.7523	0.7710
NR.ER	*0.6949*	0.6659	0.6962
NR.ER.LBD	*0.7272*	0.6532	0.6895
NR.PPAR.gamma	*0.7102*	0.6367	0.6653
SR.ARE	0.8017	0.7927	*0.8201*
SR.ATAD5	*0.7958*	0.7972	0.7310
SR.HSE	*0.8101*	0.7354	0.6697
SR.MMP	*0.8489*	0.8485	0.8256
SR.p53	*0.7487*	0.6955	0.6662

3.2. Learning of Toxicophore Representations

As mentioned in Section 1, neurons in different hidden layers of the network may encode toxicophore features. To check whether Deep Learning does indeed construct toxicophores, we performed separate experiments. In the challenge models, toxicophores (see Section 2.3.2) were used as input features. We removed these features to withhold all toxicophore-related substructures from the network input, and were thus able to check whether toxicophores were constructed automatically by DNNs.

We trained a multi-task deep network on the Tox21 data using exclusively ECFP4 fingerprint features, which had similar performance as a DNN trained on the full descriptor set (see Supplementary Section 4, Supplementary Table 1). ECFP fingerprint features encode substructures around each atom in a compound up to a certain radius. Each ECFP fingerprint feature counts how many times a specific substructure appears in a compound. After training, we looked for possible associations

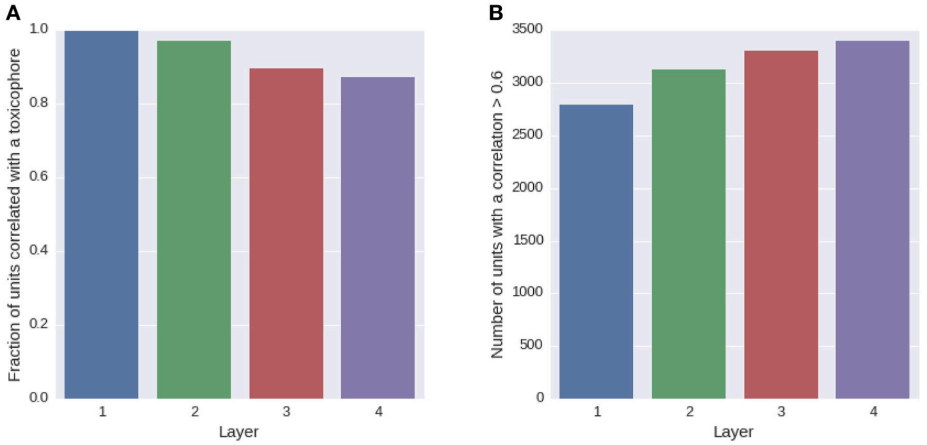


FIGURE 7 | Quantity of neurons with significant associations to toxicophores. (A) The histogram shows the fraction of neurons in a layer that yield significant correlations to a toxicophore. With an increasing level of the layer, the number of neurons with significant correlation decreases . **(B)** The histogram shows the number of neurons in a layer that exceed a correlation threshold of 0.6 to their best correlated toxicophore. Contrary to **(A)** the number of neurons increases with the network layer. Note that each layer consisted of the same number of neurons.

between all neurons of the networks and 1429 toxicophores, that were available as described in Section 2.3.2. We checked the associations using a *U*-test, in which a neuron was characterized by its activation over the compounds of the training set and a toxicophore was characterized by its presence/absence in the training set compounds. The alternative hypothesis for the test was that compounds containing the toxicophore substructure have different activations than compounds that do not contain the toxicophore substructure. Bonferroni multiple testing correction was applied afterwards, that is the *p*-values from the *U*-test were multiplied by the number of hypothesis, concretely the number of toxicophores (1429) times the number of neurons of the network (16,384). After this correction, 99% of neurons in the first hidden layer had a significant association with at least one toxicophore feature using a significance threshold of 0.05. The number of neurons with significant associations decreases with increasing level of the layer. In the second layer, there are 97% neurons with a significant association and 90 and 87% in the third and fourth layer, respectively (see **Figure 7A**). Next we investigated the correlation of known toxicophores to neurons in different layers to quantify their matching. To this end, we used the rank-biserial correlation which is compatible to the previously used *U*-test. To limit false detections, we constrained the analysis to estimates with a variance <0.01. We observed that higher layers have a higher number of neurons with rank-biserial correlation above 0.6 (see **Figure 7B**). This means features in higher layers match toxicophores more precisely.

The decrease in the number of neurons with significant associations with toxicophores through the layers and the simultaneous increase of neurons with high correlation can be explained by the typical characteristics of a DNN: In lower layers, features code for small substructures of toxicophores, while in higher layers they code for larger substructures or whole toxicophores. Features in lower layers are typically part of several higher layer features, and therefore correlate with more toxicophores than higher level features, which explains the decrease of neurons with significant associations to toxicophores. Features in higher layers are more specific and are therefore correlated more highly with toxicophores, which explains the

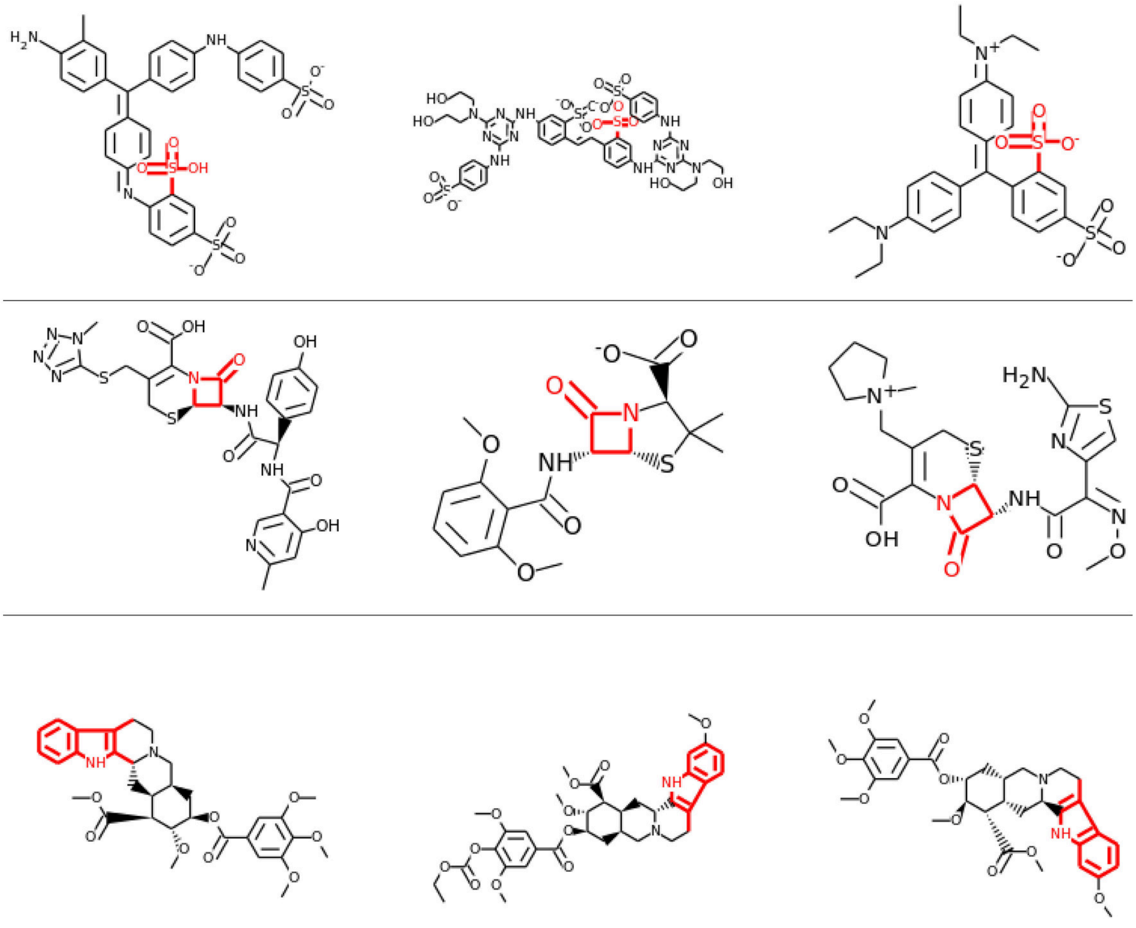


FIGURE 8 | Feature Construction by Deep Learning. Neurons that have learned to detect the presence of toxicophores. Each row shows a particular hidden unit in a learned network that correlates highly with a particular known toxicophore feature. The row shows the three chemical compounds that had the highest activation for that neuron. Indicated in red is the toxicophore structure from the literature that the neuron correlates with. The first row and the second row are from the first hidden layer, the third row is from a higher-level layer.

increase of neurons with high correlation values. Our findings underline that deep networks can indeed learn to build complex toxicophore features with high predictive power for toxicity.

Visual inspection of the results also confirmed that lower layers tended to learn smaller features, often focusing on single functional groups, such as sulfonic acid groups (see row 1 and 2 of **Figure 8**), while in higher layers the correlations tended to be with larger toxicophore clusters (row 3 of **Figure 8**). Most importantly, these learned toxicophore structures demonstrated that *Deep Learning can support finding new chemical knowledge that is encoded in its hidden units.*

3.3. Comparison of DNN and Complementary Methods

We selected the best-performing models from each method in the DeepTox pipeline based on an evaluation of the DeepTox cross-validation sets and evaluated them on the final test set. The methods we compared were DNNs, SVMs (Tanimoto kernel), random forests (RF), and elastic net (ElNet). **Table 4** shows the AUC values for each method and each dataset. We also provided the mean AUC over the NR and SR panel, and the mean AUC over all datasets. *The results confirm the superiority of Deep Learning over complementary methods for toxicity prediction by outperforming other approaches in 10 out of 15 cases.*

TABLE 4 | AUC Results for different learning methods as part of DeepTox evaluated on the final test set.

	AVG	NR	SR	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
DNN	0.837	0.827	0.851	0.923	0.778	0.825	0.829	0.804	0.775	0.791	0.811	0.863	0.930	0.860	0.856
SVM	0.832	0.819	0.849	0.919	0.822	0.748	0.818	0.819	0.781	0.799	0.798	0.848	0.946	0.854	0.827
RF	0.820	0.805	0.840	0.917	0.776	0.812	0.810	0.806	0.786	0.770	0.746	0.826	0.945	0.835	0.805
ElNet	0.803	0.787	0.826	0.897	0.788	0.692	0.778	0.763	0.768	0.765	0.805	0.844	0.924	0.818	0.799

TABLE 5 | The leading teams' AUC Results on the final test set in the Tox21 challenge.

	AVG	NR	SR	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
our method	**0.846**	**0.826**	**0.858**	**0.928**	0.807	**0.879**	**0.840**	0.834	0.793	**0.810**	0.814	**0.865**	0.942	0.862	**0.861**
AMAZIZ	0.838	0.816	0.854	0.913	0.770	0.846	0.805	0.819	**0.828**	0.806	0.806	0.842	**0.950**	0.843	0.830
dmlab	0.824	0.811	0.850	0.781	**0.828**	0.819	0.768	**0.838**	0.800	0.766	0.772	0.855	0.946	**0.880**	0.831
T	0.823	0.798	0.842	0.913	0.676	0.848	0.801	0.825	0.814	0.784	0.805	0.811	0.937	0.847	0.822
microsomes	0.810	0.785	0.814	0.901	–	–	0.804	–	0.812	0.785	**0.827**	–	–	0.826	0.717
filipsPL	0.798	0.765	0.817	0.893	0.736	0.743	0.758	0.776	–	0.771	–	0.766	0.928	0.815	–
Charite	0.785	0.750	0.811	0.896	0.688	0.789	0.739	0.781	0.751	0.707	0.798	0.852	0.880	0.834	0.700
RCC	0.772	0.751	0.781	0.872	0.763	0.747	0.761	0.792	0.673	0.781	0.762	0.755	0.920	0.795	0.637
frozenarm	0.771	0.759	0.768	0.865	0.744	0.722	0.700	0.740	0.726	0.745	0.790	0.752	0.859	0.803	0.803
ToxFit	0.763	0.753	0.756	0.862	0.744	0.757	0.697	0.738	0.729	0.729	0.752	0.689	0.862	0.803	0.791
CGL	0.759	0.720	0.791	0.866	0.742	0.566	0.747	0.749	0.737	0.759	0.727	0.775	0.880	0.817	0.738
SuperTox	0.743	0.682	0.768	0.854	–	0.560	0.711	0.742	–	–	–	–	0.862	0.732	–
kibutz	0.741	0.731	0.731	0.865	0.750	0.694	0.708	0.729	0.737	0.757	0.779	0.587	0.838	0.787	0.666
MML	0.734	0.700	0.753	0.871	0.693	0.660	0.701	0.709	0.749	0.750	0.710	0.647	0.854	0.815	0.645
NCI	0.717	0.651	0.791	0.812	0.628	0.592	0.783	0.698	0.714	0.483	0.703	0.858	0.851	0.747	0.736
VIF	0.708	0.702	0.692	0.827	0.797	0.610	0.636	0.671	0.656	0.732	0.735	0.723	0.796	0.648	0.666
Toxic Avg	0.644	0.659	0.607	0.715	0.721	0.611	0.633	0.671	0.593	0.646	0.640	0.465	0.732	0.614	0.682
Swamidass	0.576	0.596	0.593	0.353	0.571	0.748	0.372	0.274	0.391	0.680	0.738	0.711	0.828	0.661	0.585

3.4. Tox21 Data Challenge Results

The DeepTox pipeline, which is dominated by DNNs, consistently showed very high performance compared to all competing methods. It won a total of 9 of the 15 challenges and did not rank lower than fifth place in any of the subchallenges In particular, it achieved the best average AUC in both the SR and the NR panel, and additionally the best average AUC across the whole set of sub-challenges. It was thus declared winner of the Nuclear Receptor and the Stress Response panel, as well as the overall Tox21 Grand Challenge.

The leading teams' results (team names abbreviated) from all 12 subchallenges and the average results over the 12 subchallenges and the subchallenges that were part of the "Nuclear Receptor" and the "Stress Response" panel, respectively, are given in **Table 5**. The best results are indicated in bold with gray background, the second-best results with light gray background.

The Tox21 challenge result can be summarized as follows: *The Deep-Learning-based DeepTox pipeline clearly outperformed all competitors.*

4. DISCUSSION

In this paper, we have introduced the DeepTox pipeline for toxicity prediction based on Deep Learning.

Deep Learning is known to learn abstract representations of the input data with higher levels of abstractions in higher layers (LeCun et al., 2015). This concept has been relatively straightforward to demonstrate in image recognition, where simple objects, such as edges and simple blobs, in lower layers are combined to abstract objects in higher layers (Lee et al., 2009). In toxicology, however, it was not known how the data representations from Deep Learning could be interpreted. We could show that many hidden neurons represent previously known toxicophores (Kazius et al., 2005)—proven concepts which have formerly been handcrafted over decades by experts in the field. Naturally, we conclude that these representations also include novel, previously undiscovered toxicophores that are latent in the data. Using these representations, our pipeline outperformed methods that were specifically tailored to toxicological applications.

Successful deep learning is facilitated by Big Data and the use of graphical processing units (GPUs). In this case, Big Data is a blessing rather than a curse. However, Big Data implies a large computational demand. GPUs alleviate the problem of large computation times, typically by using CUDA kernels on Nvidia cards (Raina et al., 2009; Unterthiner et al., 2014, 2015; Clevert et al., 2015). Concretely, training a single DNN on the Tox21 dataset takes about 10 min on an Nvidia Tesla K40 with our optimized implementation. However, we had to train thousands of networks in order to investigate different hyperparameter settings via our cross-validation procedure, which is crucial for the performance of DNNs. The hyperparameter search was parallelized across multiple GPUs. Concluding, we consider the use of GPUs a necessity and recommend the use of multiple GPU units.

Similar to the successes in other fields (Dahl et al., 2012; Krizhevsky et al., 2012; Deng et al., 2013; Graves et al., 2013; Socher and Manning, 2013; Baldi et al., 2014; Sutskever et al., 2014), Deep Learning has increased the predictive performance of computational methods in toxicology. As confirmed by the NIH[1], the high quality of the models in the Tox21 challenge makes them suitable for deployment in leading-edge toxicological research. We believe that Deep Learning is highly suited to predicting toxicity and is capable of significantly influencing this field in the future.

FUNDING

The research was funded by ChemBioBridge and MrSymBioMath.

ACKNOWLEDGMENTS

We thank ChemAxon[2] for providing an academic license: Chem Base was used for structure searching and chemical database access and management and Standardizer was used for structure canonicalization and transformation, JChem 14.9.1.0, 2014. Further we thank Honglak Lee for the permission to use his images and the NVIDIA corporation for the GPU donations which made this research possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2015.00080

[1] https://ncats.nih.gov/news/releases/2015/tox21-challenge-2014-winners.

[2] https://www.chemaxon.com.

REFERENCES

Ajmani, S., Jadhav, K., and Kulkarni, S. A. (2006). Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J. Chem. Inf. Model.* 46, 24–31. doi: 10.1021/ci0501286

Andersen, M. E., and Krewski, D. (2009). Toxicity testing in the 21st century: bringing the vision to life. *Toxicol. Sci.* 107, 324–330. doi: 10.1093/toxsci/kfn255

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* 5:4308. doi: 10.1038/ncomms5308

Bartkova, J., Hořejší, Z., Koed, K., Krämer, A., Tort, F., Zieger, K., et al. (2005). DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* 434, 864–870. doi: 10.1038/nature03482

Bender, A., Mussa, H., Glen, R. C., and Reiling, S. (2004). Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* 44, 170–178. doi: 10.1021/ci034207y

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics*

(COMPSTAT 2010), eds Y. Lechevallier and G. Saporta (Paris), 177–187. doi: 10.1007/978-3-7908-2604-3_16

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

Cao, D.-S., Huang, J.-H., Yan, J., Zhang, L.-X., Hu, Q.-N., Xu, Q.-S., et al. (2012). Kernel k-nearest neighbor algorithm as a flexible SAR modeling tool. *Chemometr. Intell. Lab.* 114, 19–23. doi: 10.1016/j.chemolab.2012.01.008

Cao, D.-S., Xu, Q.-S., Hu, Q.-N., and Liang, Y.-Z. (2013). ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29, 1092–1094. doi: 10.1093/bioinformatics/btt105

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi: 10.1023/A:1007379606734

Chawla, A., Repa, J. J., Evans, R. M., and Mangelsdorf, D. J. (2001). Nuclear receptors and lipid physiology: opening the X-files. *Science* 294, 1866–1870. doi: 10.1126/science.294.5548.1866

Cireşan, D. C., Meier, U., and Schmidhuber, J. (2012a). "Multi-column deep neural networks for image classification," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 3642–3649. doi: 10.1109/CVPR.2012.6248110

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). "Mitosis detection in breast cancer histology images with deep neural networks," in *16th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2013)*, eds K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab (Nagoya), 411–418. doi: 10.1007/978-3-642-40763-5_51

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2012b). "Deep big multilayer perceptrons for digit recognition," in *Neural Networks: Tricks of the Trade*, eds G. Montavon, G. B. Orr, and K.-R. Müller (Heidelberg: Springer), 581–598.

Clevert, D.-A., Mayr, A., Unterthiner, T., and Hochreiter, S. (2015). "Rectified factor networks," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, eds C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC), 1846–1854.

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. R. (2014). Multi-task neural networks for QSAR predictions. arXiv:1406.1231.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE T Audio Speech* 20, 30–42. doi: 10.1109/TASL.2011.2134090

Darnag, R., Mazouz, E. M., Schmitzer, A., Villemin, D., Jarid, A., and Cherqaoui, D. (2010). Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. *Eur. J. Med. Chem.* 28, 1075–1086. doi: 10.1016/j.ejmech.2010.01.002

Deng, L., Hinton, G. E., and Kingsbury, B. (2013). "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, BC), 8599–8603. doi: 10.1109/ICASSP.2013.6639344

Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.* 33, 933–940. doi: 10.1038/nbt.3299

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

Glorot, X., Bordes, A., and Bengio, Y. (2011). "Deep sparse rectifier neural networks," in *Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, eds G. J. Gordon, D. B. Dunson, and M. Dudík (Fort Lauderdale, FL), 315–323.

Graves, A., Mohamed, A. R., and Hinton, G. E. (2013). "Speech recognition with deep recurrent neural networks," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, BC), 6645–6649. doi: 10.1109/ICASSP.2013.6638947

Grün, F., and Blumberg, B. (2007). Perturbed nuclear receptor signaling by environmental obesogens as emerging factors in the obesity crisis. *Rev. Endocr. Metab. Dis.* 8, 161–171. doi: 10.1007/s11154-007-9049-x

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., and Zell, A. (2011). jCompoundMapper: an open source Java library and command-line tool for chemical fingerprints. *J. Cheminform.* 3:3. doi: 10.1186/1758-2946-3-3

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580.

Hochreiter, S. (1991). *Untersuchungen Zu Dynamischen Neuronalen Netzen*. Master's thesis, Institut für Informatik, Lehrstuhl Prof. Dr. Dr. h.c. Brauer, Technische Universität München.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2000). "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Networks*, eds J. Kolen and S. Kremer (New York, NY: IEEE), 237–244.

Huang, R., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., et al. (2014). Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep.* 4:5664. doi: 10.1038/srep05664

Jaeschke, H., McGill, M. R., and Ramachandran, A. (2012). Oxidant stress, mitochondria, and cell death mechanisms in drug-induced liver injury: lessons learned from acetaminophen hepatotoxicity. *Drug Metab. Rev.* 44, 88–106. doi: 10.3109/03602532.2011.602688

Kashima, H., Tsuda, K., and Inokuchi, A. (2003). "Marginalized kernels between labeled graphs," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, eds T. Fawcett and N. Mishra (Washington, DC), 321–328.

Kashima, H., Tsuda, K., and Inokuchi, A. (2004). "Kernels for graphs," in *Kernel Methods in Computational Biology*, eds B. Schölkopf, K. Tsuda, and J.-P. Vert (Cambridge, MA: MIT Press), 155–170.

Kauffman, G. W., and Jurs, P. C. (2001). QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.* 41, 1553–1560. doi: 10.1021/ci010073h

Kazius, J., McGuire, R., and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* 48, 312–320. doi: 10.1021/jm040835a

Klambauer, G., Wischenbart, M., Mahr, M., Unterthiner, T., Mayr, A., and Hochreiter, S. (2015). Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map. *Bioinformatics* 31, 3392–3394. doi: 10.1093/bioinformatics/btv373

Krewski, D., Acosta D. Jr., Andersen, M., Anderson, H., Bailar III, J. C., Boekelheide, K., et al. (2010). Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health* 13, 51–138. doi: 10.1080/10937404.2010.483176

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, eds F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Lake Tahoe), 1097–1105.

Labbe, G., Pessayre, D., and Fromenty, B. (2008). Drug-induced liver injury through mitochondrial dysfunction: mechanisms and detection during preclinical safety studies. *Fund. Clin. Pharmacol.* 22, 335–353. doi: 10.1111/j.1472-8206.2008.00608.x

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM* 54, 95–103. doi: 10.1145/2001269.2001295

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, eds A. P. Danyluk, L. Bottou, and M. L. Littman (Montreal, QC), 609–616. doi: 10.1145/1553374.1553453

Luco, J. M., and Ferretti, F. H. (1997). QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* 37, 392–401. doi: 10.1021/ci960487o

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

Mahé, P., Ralaivola, L., Stoven, V., and Vert, J.-P. (2006). The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.* 46, 2003–2014. doi: 10.1021/ci060138m

Mahé, P., Ueda, N., Akutsu, T., Perret, J.-L., and Vert, J.-P. (2005). Graph kernels for molecular Structure-Activity relationship analysis with support vector machines. *J. Chem. Inf. Model.* 45, 939–951. doi: 10.1021/ci050039t

Mohr, J. A., Jain, B. J., and Obermayer, K. (2008). Molecule kernels: a descriptor- and alignment-free quantitative Structure-Activity relationship approach. *J. Chem. Inf. Model.* 48, 1868–1881. doi: 10.1021/ci800144y

Mohr, J. A., Jain, B. J., Sutter, A., Laak, A. T., Steger-Hartmann, T., Heinrich, N., et al. (2010). A maximum common subgraph kernel method for predicting the chromosome aberration test. *J. Chem. Inf. Model.* 50, 1821–1838. doi: 10.1021/ci900367j

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, eds J. Fürnkranz and T. Joachims (Haifa), 807–814.

Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Le, Q. V., and Ng, A. Y. (2011). "On optimization methods for deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, eds L. Getoor and T. Scheffer (Bellevue, WA), 689–696.

Niu, B., Lu, W.-C., Yang, S.-S., Cai, Y.-D., and Li, G.-Z. (2007). Support vector machine for SAR/QSAR of phenethyl-amines1. *Acta Pharma. Sinica.* 28, 1075–1086. doi: 10.1111/j.1745-7254.2007.00573.x

Platt, J. C. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, eds A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Cambridge, MA: MIT Press), 61–74.

Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., and Kuzmin, V. E. (2009). Application of random forest approach to QSAR prediction of aquatic toxicity. *J. Chem. Inf. Model.* 49, 2481–2488. doi: 10.1021/ci900203n

Raina, R., Madhavan, A., and Ng, A. Y. (2009). "Large-scale deep unsupervised learning using graphics processors," in *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, eds A. P. Danyluk, L. Bottou, and M. L. Littman (Montreal, QC), 873–880. doi: 10.1145/1553374.1553486

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Netw.* 18, 1093–1110. doi: 10.1016/j.neunet.2005.07.009

Ranzato, M., Boureau, Y.-I., and LeCun, Y. (2008). "Sparse feature learning for deep belief networks," in *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Vancouver, BC), 1185–1192.

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

Rosenbaum, L., Hinselmann, G., Jahn, A., and Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. *J. Cheminform.* 3:11. doi: 10.1186/1758-2946-3-11

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

Rusyn, I., and Daston, G. P. (2010). Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ. Health Perspect.* 118, 1047–1050. doi: 10.1289/ehp.1001925

Sagardia, I., Roa-Ureta, R. H., and Bald, C. (2013). A new QSAR model, for angiotensin I-converting enzyme inhibitory oligopeptides. *Food Chem.* 136, 1370–1376. doi: 10.1016/j.foodchem.2012.09.092

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Shukla, S. J., Huang, R., Austin, C. P., and Xia, M. (2010). The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov. Today* 15, 997–1007. doi: 10.1016/j.drudis.2010.07.007

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13. doi: 10.18637/jss.v039.i05

Socher, R., and Manning, C. D. (2013). "Deep learning for NLP (without magic)," in *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, eds L. Vanderwende, H. D. III, and K. Kirchhoff (Atlanta, GA), 1–3.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958. doi: 10.1021/ci034160g

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Montreal, QC), 3104–3112.

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958.

Thalheim, T., Vollmer, A., Ebert, R.-U., Kuhne, R., and Schurmann, G. (2010). Tautomer identification and tautomer structure generation based on the InChI code. *J. Chem. Inf. Model.* 50, 1223–1232. doi: 10.1021/ci1001179

Unterthiner, T., Mayr, A., Klambauer, G., and Hochreiter, S. (2015). Toxicity prediction using deep learning. arXiv:1503.01445.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Ceulemans, H., Wegner, J. K., et al. (2014). "Deep learning as an opportunity in virtual screening," in *NIPS Workshop on Deep Learning and Representation Learning* (Montreal, QC).

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., The QSTAR Consortium, Shkedy, Z., Thas, O., et al. (2015). Using transcriptomics to guide lead optimization in drug discovery projects. *Drug Discov. Today* 20, 505–513. doi: 10.1016/j.drudis.2014.12.014

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *J. Mach. Learn. Res.* 11, 1201–1242.

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph.D. thesis, Harvard University.

Xia, X., Maliski, E. G., Gallant, P., and Rogers, D. (2004). Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.* 47, 4463–4470. doi: 10.1021/jm0303195

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 Environmental Science

ORIGINAL RESEARCH
published: 04 February 2016
doi: 10.3389/fenvs.2016.00002

Consensus Modeling for HTS Assays Using *In silico* Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge

Ahmed Abdelaziz[1,2]*, *Hilde Spahn-Langguth*[3,4], *Karl-Werner Schramm*[2,5] and *Igor V. Tetko*[6,7]

[1] Rosettastein Consulting UG, Freising, Germany, [2] Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, TUM-Technische Universität München, Freising, Germany, [3] Institute for Medical and Pharmaceutical Proficiency Assessment, Mainz, Germany, [4] Department of Pharmaceutical Sciences, Karl-Franzens-University Graz, Graz, Austria, [5] Molecular EXposomics, German Research Center for Environmental Health, Helmholtz Zentrum München, Neuherberg, Germany, [6] BigChem GmbH, Neuherberg, Germany, [7] Helmholtz Zentrum München - Research Center for Environmental Health (HMGU), Institute of Structural Biology, Neuherberg, Germany

The need for filling information gaps while reducing toxicity testing in animals is becoming more predominant in risk assessment. Recent legislations are accepting *in silico* approaches for predicting toxicological outcomes. This article describes the results of Quantitative Structure Activity Relationship (QSAR) modeling efforts within Tox21 Data Challenge 2014[1], which calculated the best balanced accuracy across all molecular pathway endpoints as well as the highest scores for ATAD5 and mitochondrial membrane potential disruption. Automated QSPR workflow systems, OCHEM (http://ochem.eu), the analytics platform, KNIME and the statistics software, CRAN R, were used to conduct the analysis and develop consensus models using 10 different descriptor sets. A detailed analysis of QSAR models for all 12 molecular pathways and the effect of underlying models' accuracy on the quality of the consensus model are provided. The resulting consensus models yielded a balanced accuracy as high as 88.1% ± 0.6 for mitochondrial membrane disruptors. Such high balanced accuracy and use of the applicability domain show a promising potential for *in silico* modeling to complement design HTS screening experiments. The comprehensive statistics of all models are publicly available online at https://github.com/amaziz/Tox21-Challenge-Publication while the developed consensus models can be accessed at http://ochem.eu/article/98009.

Keywords: computational toxicology, alternative testing, Quantitative structure activity relationship, high throughput screening, predictive toxicology, Tox21

OPEN ACCESS

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Guennady Ougolnitsky,
Southern Federal University, Russia
Nora Aptula,
Unilever, UK

****Correspondence:***
Ahmed Abdelaziz
contact@amaziz.com

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *16 October 2015*
Accepted: *12 January 2016*
Published: *04 February 2016*

Citation:
Abdelaziz A, Spahn-Langguth H, Schramm K-W and Tetko IV (2016) Consensus Modeling for HTS Assays Using In silico *Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. Front. Environ. Sci. 4:2. doi: 10.3389/fenvs.2016.00002*

INTRODUCTION

High-throughput screening (HTS) allows researchers to conduct millions of chemical, genetic, or pharmacological experiments with minimal intervention. Such procedures may quickly identify potentially active compounds, antibodies, or genes that control particular biochemical pathways. The results of such assays guide the research process. And thus this approach has become a valuable and viable tool for large-scale evaluation of chemicals (Kavlock and Dix, 2010; Judson et al., 2011; Wetmore et al., 2012). The large amounts of data generated by HTS available today may be used to correlate chemical structures to their biological activities. QSARs may support the identification

[1] Tox21 Data Challenge 2014—Data Available at: https://tripod.nih.gov/tox21/challenge/data.jsp

of key characteristics in chemical structures responsible for such activities. This knowledge is then used to provide predictions about the possible activity of test compounds in *virtual screening* settings for regulatory purposes. The quality of QSAR models based on large chemical libraries from HTS experiments varies. However, the accuracy is usually high enough to support prioritizing chemicals that are worth being subjected to experimental testing. This approach satisfies the imminent need to prioritize chemicals testing, filling information gaps, accelerating the chemical registration process and lowering the overall costs of testing (US EPA, OCSPP).[2]

Tox21 (Tice et al., 2009; Betts, 2013) represents a multi-agency effort that uses HTS assays for toxicity modeling and prediction in the US. The US Environmental Protection Agency (EPA), The National Institutes of Health (NIH), The National Center for Advancing Translational Sciences (NCATS), The National Institutes of Environmental Health Sciences/National Toxicology Program (NIEHS/NTP) and the Food and Drug Administration (FDA) cooperate in screening chemical substances for some selected potential toxic effects. The data may then be used, with the assistance of *in silico* techniques, for providing an alternative for expensive, time-consuming, and ethically-questioned animal testing. This implies the potential for providing an economical method for toxicity testing prioritization for thousands of yet untested compounds (Betts, 2013).

Similar efforts to reduce animal testing and utilize computational toxicity modeling are made in Europe. The European Chemical Agency (ECHA) described the role of animals in ensuring the safe use of chemical substances as being the last resort. This is one of the key principles for the REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) legislations. It encourages the use of so-called "alternative approaches" to reduce animal testing. QSAR modeling is one of the promoted mechanisms for alternative chemicals' risk assessment. Guiding documents exist that explain the best practices and the requirements for accepting QSAR models' predictions (Worth et al., 2005). These guidelines are essential for directing the stakeholders on how to utilize QSAR methodologies in a manner that gets accepted by the regulators. The guidelines warrant evaluating the human and environmental toxicity risks, complying with the regulatory requirements and reducing the need for animal testing at the same time.

The Tox21 Data challenge follows the open-innovation principles (Chesbrough, 2006) to crowdsource scientists' efforts in analyzing HTS data generated through the Tox21 project. It aspires to predict the pathways' interference of chemicals using only their chemical structures. Such predictions can therefore guide regulators and participating governmental agencies in identifying the chemicals (either drugs or industrial) that carry the highest concern for human and environmental risks. The aim of this study is to describe the methodologies used by the winning corresponding author during the challenge (team: AMAZIZ) and to extend the analysis on the chemical libraries beyond what was possible during the limited duration of the challenge. The study investigates a comprehensive approach on consensus modeling and analyzes multiple descriptor packages.

MATERIALS AND METHODS

Molecular Pathways Screening

In this study, 12 molecular pathway endpoints were investigated, which were selected on the basis of toxicological relevance. The targets were experimentally screened as part of the Tox21 program and the resulting data library made accessible for competitors by the Tox21 Data Challenge organizers (Tox21 Data Challenge 2014—Data).

Estrogen Receptor (ER) (AID 743077[3], AID 743079[4])

Tox21 compounds library was screened for potentially acting as agonist at the estrogen receptor alpha. Such activators could lead to reproductive dysfunction (Aop:30)[5]. Two different cell lines were used:

- ER-alpha-UAS-bla GripTiteTM cell line (ER-LBD): This cell line was developed by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an Upstream Activator Sequence (UAS) stably integrated into HEK293 cells.
- BG1-Luc-4E2 cell line (ER-full): Dr. Michael Denison from University of California provided the cell line. Cells endogenously express the full-length ER-alpha and are stably transfected with a plasmid containing four estrogen responsive elements (ERE) under the control of an upstream luciferase reporter gene.

Androgen Receptor (AR) (AID 743040[6], AID 743053[7])

Compounds that agonist the AR may cause reproductive dysfunction (Aop:23)[8]. The ability of Tox21 compounds to

[2] US EPA, OCSPP, O. Using Predictive Methods to Assess Hazard under TSCA. Available at: http://www2.epa.gov/tsca-screening-tools/using-predictive-methods-assess-hazard-under-tsca#models [Accessed October 15, 2015].

[3] AID 743077—qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743077 [Accessed July 10, 2015].

[4] AID 743079—qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743079 [Accessed July 10, 2015].

[5] Aop:30—Estrogen receptor antagonism leading to reproductive dysfunction-aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:30 [Accessed December 15, 2015].

[6] AID 743040—qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway using the MDA cell line—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743040#aDescription [Accessed July 10, 2015].

[7] AID 743053—qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743053 [Accessed July 10, 2015].

[8] Aop:23—Androgen receptor agonism leading to reproductive dysfunction—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:23 [Accessed December 15, 2015].

agonist the androgen receptor alpha was measured in two different cell lines.

- GeneBLAzer AR-UAS-bla-GripTite cell line (AR-LBD): This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an upstream activator sequence (UAS) stably integrated into HEK293 cells.
- MDA-kb2 AR-luc cell line (AR-full): This cell line was deposited by Wilson et al. It is a human breast carcinoma cell line that was stably transfected with a luciferase reporter gene under control of the MMTV promoter containing response elements for both androgen receptor (AR) and glucocorticoid receptor (GR).

Aryl Hydrocarbon Receptor (AHR) (AID 743122)[9]

AHR activation is thought to lead to multiple adverse outcomes including hepatic steatosis (Aop:57)[10], uroporphyria (Aop:131)[11], developmental abnormalities and embryolethality (in birds) (Aop:22)[12], and embryo toxicity in fish (Aop:21)[13] inter alia. A cell based HepG2-AhR-luc assay was used to assess the activation of AhR for Tox21 compounds. The HG2L7.5c1 cell line, as developed by Dr. Michael S. Denison (University of California at Davis), was utilized. The human hepatocellular carcinoma (HepG2) cells were stably transfected with an Ah receptor-responsive firefly luciferase reporter gene plasmid carrying 20 dioxin responsive elements and luciferase reporter gene. AhR activation leads to an increase in luciferase activity and therefore ligands can be detected.

Peroxisome Proliferator-Activated Receptor Gamma (PPAR-gamma) (AID 743140)[14]

PPAR-gamma activation has been associated with impaired fertility in adult females (Aop:7)[15]. GeneBLAzer PPAR gamma UAS-bla HEK293H cell line was used in this assay. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an upstream activator sequence (UAS) stably integrated into HEK293H cells.

Nuclear Factor (erythroid-derived 2)-Like 2/Antioxidant Responsive Element (Nrf2/ARE) (AID 743219)[16]

The CellSensor ARE-bla Hep-G2 assay was used to assess the activation of the report gene and thus identify chemicals that stimulate oxidative stress. The cells contain a beta-lactamase reporter gene controlled by the Antioxidant Response Element (ARE) stably integrated into HepG2 cells. Fluorescence intensity was measured to assess the activation of the responsive element.

Aromatase Enzyme Inhibitors (AID 743139)[17]

Aromatase inhibition is associated with reproductive dysfunction among other adverse outcomes (Aop:25)[18]. The MCF-7 aro ERE cell line (human breast carcinoma), as provided by Dr. Shiuan Chen (Beckman Research Institute of the City of Hope), was used in order to identify aromatase inhibitors. Cells were stably transfected with a promoter plasmid, pGL3-Luc, encompassing three repeats of the estrogen responsive element (ERE).

ATAD5 Receptor (ATAD5) (AID 720516)[19]

A cell-based assay using embryonic kidney cells (HEK293T) was used to screen the Tox21 compounds library. The assay was developed by Kyungjae Myung (NHGRI, NIH) to detect any enhanced Level of Genome Instability Gene 1 (ELG1; human ATAD5) protein, which increase in response to different kinds of DNA damage. The assay uses a luciferase reporter-gene tagged with ATAD5 to measure the induction of ELG1. Therefore, an increase in luciferase activity marks a chemically induced genetic stress.

Heat Shock Response (HSE) (AID 743228)[20]

HSE-bla HeLa cell line was utilized in this HTS assay. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by the heat shock response elements.

[9] AID 743122—qHTS assay to identify small molecule that activate the aryl hydrocarbon receptor (AhR) signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743122 [Accessed July 10, 2015].

[10] Aop:57—AhR activation leading to hepatic steatosis—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:57 [Accessed December 15, 2015].

[11] Aop:131—Aryl hydrocarbon receptor activation leading to uroporphyria—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:131 [Accessed December 15, 2015].

[12] Aop:22—AHR1 activation leading to developmental abnormalities and embryolethality (in birds)—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:22 [Accessed December 15, 2015].

[13] Aop:21—AhR activation leading to embryo toxicity in fish—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:21 [Accessed December 15, 2015].

[14] AID 743140—qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743140 [Accessed July 10, 2015].

[15] Aop:7—PPAR γ activation leading to impaired fertility in adult female- aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:7 [Accessed December 15, 2015].

[16] AID 743219—qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743219 [Accessed July 10, 2015].

[17] AID 743139—qHTS assay to identify aromatase inhibitors: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743139 [Accessed July 10, 2015].

[18] Aop:25—Aromatase inhibition leading to reproductive dysfunction (in fish)—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:25 [Accessed December 15, 2015].

[19] AID 720516—qHTS assay for small molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720516 [Accessed July 10, 2015].

[20] AID 743228—qHTS assay for small molecule activators of the heat shock response signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743228 [Accessed July 10, 2015].

TABLE 1 | Number of records and unique molecules in each dataset.

Molecular pathway endpoint	Training set records (unique molecules)	Test set records	Complete training set records (unique molecules)
NUCLEAR RECEPTOR SIGNALING PANEL			
Aryl hydrocarbon receptor (nr-ahr)	8169 (6716)	272	8441 (6988)
Androgen receptor MDA-kb2 AR-luc cell line (nr-ar)	9362 (7468)	292	9654 (7760)
Androgen receptor GeneBLAzer AR-UAS-bla-GripTite cell line (nr-ar-lbd)	8599 (6927)	253	8852 (7180)
Aromatase enzyme (nr-aromatase)	7226 (5966)	214	7440 (6180)
Estrogen receptor alpha BG1-Luc-4E2 cell line (nr-er)	7697 (6334)	265	7962 (6599)
Estrogen receptor alpha ER-alpha-UAS-bla GripTiteTM cell line (nr-er-lbd)	8753 (7138)	287	9040 (7425)
Peroxisome proliferator-activated receptor gamma (nr-ppar-gamma)	8184 (6607)	267	8451 (6874)
STRESS RESPONSE PANEL			
Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (Nrf2/ARE) (sr-are)	7167 (5959)	234	7401 (6193)
ATAD5 receptor (sr-atad5)	9091 (7256)	272	9363 (7528)
Heat shock factor response element (sr-hse)	8150 (6617)	267	8417 (6884)
Mitochondrial membrane potential (sr-mmp)	7320 (5941)	238	7558 (6179)
p53 signaling pathway (sr-p53)	8634 (6931)	269	8903 (7200)

Nuclear receptor (nr) assay panel contained seven assays while the stress response (sr) assay panel covered five assays.

Disruptors of the Mitochondrial Membrane Potential (MMP) (AID 720637)[21]

The mitochondrial dysfunction is considered a key event in multiple adverse outcomes (Event:177)[22] including neuroinflammation leading to neurodegeneration, excitotoxicity, and learning and memory impairment. A homogenous cell-based assay with a water-soluble mitochondrial membrane potential sensor (m-MPI, Codex Biosolutions, MD) was applied to the Tox21 compounds in order to identify those that can induce mitochondrial toxicity. In healthy cells, the water-soluble dye accumulates in the mitochondria as aggregates, causing red fluorescence. In case of a decrease in MMP, the dye cannot accumulate in the mitochondria and thus remains in the cytoplasm as monomers causing green fluorescence.

Agonists of the p53 Signaling Pathway (P53) (AID 720552)[23]

p53 gene has been identified as target of AFB1-induced adduction and subsequent mutation which is a key event leading to Hepatocellular Carcinoma (HCC; Aop:46)[24]. Using CellSensor p53RE-bla HCT-116 cell line, the Tox21 compounds were screened. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a stably integrated beta-lactamase (BLA) reporter gene controlled by the p53 response elements. Fluorescence intensity was measured to assess the activation of the responsive element.

Datasets and Data Cleaning

Data were downloaded from the Tox21 challenge website (NIH)[25] in both SDF and SMILES formats. The files contained the molecular representation (SDF or SMILES), a molecule name as well as the target response. In addition, SDF files contained few extra tags for the DSSTox compound ID (DSSTox_CID), the chemical formula and the average mass (FW). Both file formats were compared to examine consistency. KNIME (Berthold et al., 2007) was used to compare the structures and responses in both file formats. The data covered 12 pathway endpoints covering the "Nuclear Receptor Signaling Panel" (seven assays) and the "Stress Response Panel" (five assays). All assay endpoints are listed in **Table 1**.

For each molecular pathway endpoint, both training and leaderboard test sets were combined to form a whole training set. Some molecules were presented multiple times (i.e., exact SMILES representation in spite of different molecule names). The basis for such duplicated records may be the result of intentional repetitive testing for quality control purpose. The Online CHEmical database and Modeling environment platform (OCHEM; Sushko et al., 2011) was used to check records duplication. It calculates the INCHI (James et al., 1995) key structure hash to compare structures. Some records showed different experimental responses despite exhibiting the same

[21] AID 720637—qHTS assay for small molecule disruptors of the mitochondrial membrane potential: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720637 [Accessed July 10, 2015].

[22] Event:177—Mitochondrial dysfunction - aopwiki Available at: https://aopkb.org/aopwiki/index.php/Event:177 [Accessed December 15, 2015].

[23] AID 720552—qHTS assay for small molecule agonists of the p53 signaling pathway: Summary—PubChem BioAssay Summary Available at: https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720552 [Accessed July 10, 2015].

[24] Aop:46—Mutagenic Mode-of-Action leading to Hepatocellular Carcinoma (HCC)—aopwiki Available at: https://aopkb.org/aopwiki/index.php/Aop:46 [Accessed December 15, 2015].

[25] NIH Tox21 Data Challenge 2014. Available at: https://tripod.nih.gov/tox21/challenge/about.jsp

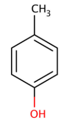
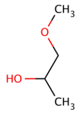
	Name: p-Kresol SMILES: CC1=CC=C(O)C=C1 NCGC00013272-01 Active NCGC00091519-04 Inactive NCGC00257956-01 Inactive NCGC00253980-01 Inactive NCGC00258667-01 Inactive
	Name: 1-methoxypropan-2-ol SMILES: COCC(C)O NCGC00256978-01 Active NCGC00259352-01 Inactive

FIGURE 1 | Example of conflicting training data. The examples shown were obtained from the Estrogen Nuclear Receptor dataset. In some cases, it could be reasonable to assume that p-Kresol would be inactive (four records shows inactive against only one active record). In other cases, such as methoxypropan-2-ol, it is not possible tell whether the compound was truly activating the Estrogen nuclear receptor (with one record in every class). Compounds are compared using their calculated INCHI keys generated from the SDF representation. All 12 targets showed similar cases.

molecular structures. **Figure 1** shows an example of such duplicates with conflicting experimental measurements. **Table 1** shows the number of records per dataset as well as the number of unique molecules.

Computational Methods

Software Tools

OCHEM (Sushko et al., 2011) offers an interactive web interface (http://www.ochem.eu) that may be used to explore the data, construct QSAR models and run predictions. It also offers the ability to interpret results using prediction-driven matched molecular pairs (Sushko et al., 2014). Handling large datasets and thousands of QSAR models is more convenient using workflow systems such as KNIME (Berthold et al., 2007). For that, OCHEM exposes a number of methods through SOAP web services (Using SOAP web-services—OCHEM user's manual—eADMET docs)[26]. These methods allow the user to login, upload data, create properties, create or delete QSAR models, download model statistics, and to run predictions on the constructed models. OCHEM implements an xml format that allows users to configure the QSAR modeling tasks with regard to all steps including descriptors calculation, descriptors pre-filtering, and configuring the machine learning algorithms.

Throughout this work, different KNIME (Berthold et al., 2007) workflows were used to explore the data, initialize the QSAR model building process on OCHEM and download the modeling results. All QSAR models were built using OCHEM. CRAN R (R Core Team, 2015) was used to build consensus models and analyze models' performance.

In silico Descriptors Calculation

Ten descriptor packages were selected from OCHEM to be used for constructing QSAR models. These packages were compiled from multiple academic and commercial sources. The selected packages are: GSFrag (Aires-de-Sousa and Gasteiger, 2001), ISIDA fragments (length 2–4; Varnek et al., 2008), Chemaxon descriptors (Introduction to Calculator Plugins—Calculator Plugins—ChemAxon - DOCS)[27], Estate indices (Hall et al., 1995; Huuskonen et al., 2000), and AlogPS (Tetko et al., 2001a,b), CDK (using all constitutional, topological, geometrical, electronic, and hybrid descriptors; Steinbeck et al., 2003), Inductive descriptors (Cherkasov et al., 2008), Dragon 6 (Todeschini and Consonni, 2009), Adriana.Code (ADRIANA.Code—Calculation of Molecular Descriptors |Inspiring Chemical Discovery)[28], Mera and Mesry (Grishina et al., 2002; Potemkin and Grishina, 2008; Potemkin et al., 2009), QNPR (using SMILES representations—length 1–3 and a threshold of 5; Thormann et al., 2007). Further details on these packages and their integration within OCHEM was reported earlier (Sushko et al., 2011).

The same structure-preprocessing protocol was used prior to the calculation of any descriptor package utilizing Chemaxon Standardizer that is integrated within OCHEM workflow. The standardization workflow consisted of salt counter-ion removal, charge neutralization and the standardizing of certain chemotype representations; such as nitro groups and aromatic rings. For 3D descriptor packages, structural coordinates were optimized using CORINA (Sadowski et al., 1994) starting from a clean SMILES representation. Descriptors calculation failed for some chemicals, the number of failed molecules depends on the nature of the descriptor package. Reasons for calculation failure could be large molecular sizes or undefined chemotypes. The Supplementary Materials (**Data Sheet 1**) include the count of failed molecules for each constructed model.

Machine Learning

The associative neural networks (ASNN; Tetko, 2002a,b) algorithm was used to construct all models. ASNN is a multilayered perceptron (Rosenblatt, 1957) neural networks algorithm that utilizes ensemble learning. As such, it can be represented by a multilayered graph in which all nodes in a certain layer are linked to the nodes of the preceding one. The resulting class membership is the output of a single neuron in the last layer of the network. ASNN uses a k-Nearest Neighbors (kNN) approach over the space of ensemble predictions to accommodate for a local correction for the ensemble of neural

[26] Using SOAP web-services—OCHEM user's manual—eADMET docs Available at: http://docs.ochem.eu/display/MAN/Using+SOAP+web-services [Accessed January 5, 2015].

[27] Introduction to Calculator Plugins—Calculator Plugins—ChemAxon—DOCS Available at: https://docs.chemaxon.com/display/CALCPLUGS/Introduction+to+Calculator+Plugins [Accessed January 9, 2015].

[28] ADRIANA.Code—Calculation of Molecular Descriptors |Inspiring Chemical Discovery Available at: http://www.molecular-networks.com/products/adrianacode [Accessed September 28, 2013].

networks. The kNN distance is based on the correlation between the vectors of predicted samples by the networks of the ensemble. All configurations for the algorithm were set to OCHEM defaults [i.e., three neurons in the hidden layer, 1000 iterations, using model ensemble size of 64, the method for neural network training was SuperSAB (Tollenaere, 1990)].

Performance Measures and Validation Protocol

Due to the unbalanced nature of the datasets, balanced accuracy was used throughout the study, as well as during the challenge, as the primary measure for comparing models' performance. It is important to notice that the challenge did not only account for the balanced accuracy but also the Area Under the Receiver Operating Characteristic (AUROC) curve (Hanley and McNeil, 1983).

Bagging (Breiman, 1996) was used to validate the accuracy of the training set. Bagging is a meta-algorithm that involves the aggregation of many models, each of which is based on its own training set ("bag"). Bagging utilizes the random sampling, with repetition, of many subsets of the training set. In each bagging meta-model constructed, an ensemble of 64 models was developed. For each model in the ensemble the training examples were selected randomly from the original training set allowing duplicates (i.e., resampling with replacement). The prediction of each classification was determined by majority voting among the ensemble members. Stratified bagging (Tetko et al., 2013) was used as the validation protocol. It also served to handle the unbalance of the training set (Kotsiantis et al., 2006). In the current implementation, for each of the 64 models in an ensemble, equal numbers of active and inactive compounds were randomly selected. Thus, the size of the training set was always double the size of the minority class.

The calculation of statistical measures was done only using the validation set (out of bag compounds). For molecules with conflicting experimental measurements (see **Figure 1**), the class with more experimental measurements (majority vote) was selected. Molecules that showed an equal number of active and inactive experimental measurements were excluded.

Consensus Modeling

For each endpoint, consensus models were built using all possible combinations of the underlying 10 models (each built using different *in silico* descriptor package), i.e., $\sum_{i=1}^{10} C_i^{10}$. In total, 12,276 models (1023 × 12 endpoints) were constructed. Simple averaging of the predictions was used for building each of the consensus models.

Two approaches for consensus model selection were investigated in this study. The first approach considers consensus models that show the highest validated balanced accuracy on the training set. The second approach considers consensus models which combine models built with all 10 descriptor packages regardless of the resulting validation balanced accuracy. Both approaches performed comparatively well.

Applicability Domain

In this study, a distance-based method was used to estimate the applicability domain for all models. The distance to model is

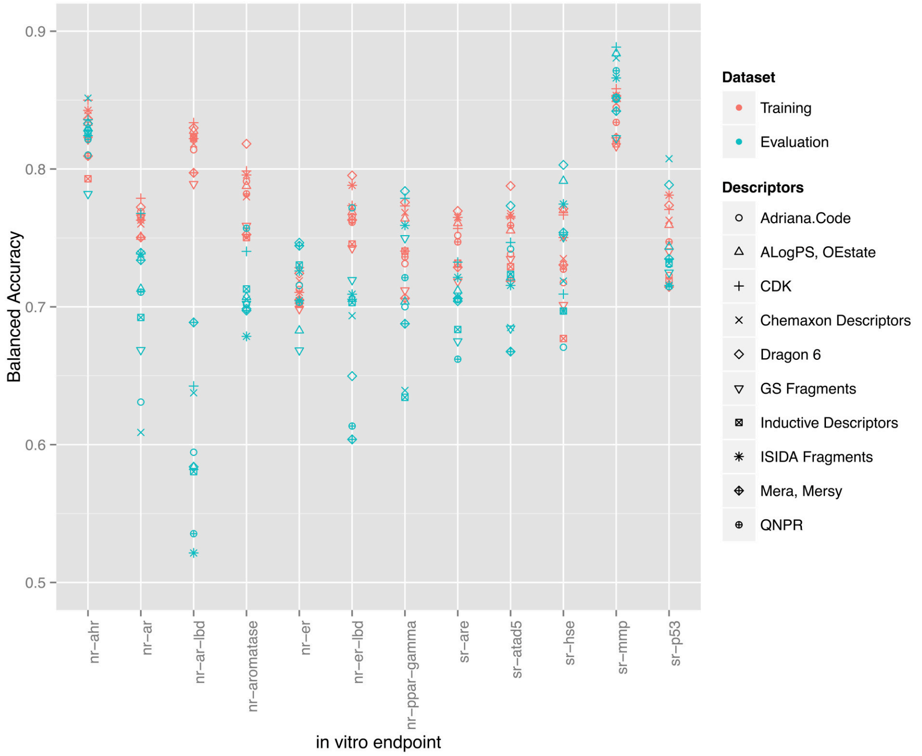


FIGURE 2 | Training set balanced accuracies for all 120 models as grouped by their respective endpoints. Red points represent the validated (through bagging) balanced accuracies calculated on the training set. Blue points represent the balanced accuracy on the evaluation set.

defined in the property space (rather than the descriptor space; Tetko et al., 2006). This approach uses the standard deviation between the predictions of an ensemble of models (generated through bagging) as a measure of distance.

RESULTS AND DISCUSSION

Individual Models

In total 10 descriptor packages were used to model 12 *in vitro* assay endpoints resulting in 120 QSAR models constructed with 64-bagging-validation. Different endpoints showed varying success. **Figure 2** shows the balanced accuracy of all 120 models as grouped by their respective targets with respect to both training and evaluation sets. Other statistical parameters such as specificity, sensitivity, Matthews's correlation coefficient (MCC), and overall accuracy are provided in the Supplementary Materials (**Data Sheet 1**). All models are published online and may be examined through http://www.ochem.eu/mode/[model-id] replacing [model-id] with the respective model identification number available in the results tables. Users can see a model's summary with performance statistics and applicability domain graphs as well as apply the model to new compounds.

To compare descriptor packages' success, each package was given a score from 1 to 10 according to its rank (a score of 10 was given to the descriptor package contributing to the model with the highest balanced accuracy and a score of 1 for the lowest). The scores were summed for all endpoints. The final rank of descriptors is summarized in **Table 2**. Dragon and CDK descriptor packages shared the top positions in both training and evaluation sets.

As shown in **Figure 2**, a direct correlation exists between the validated training and the evaluation sets' balanced accuracies with the exception of the nr-ar-lbd endpoint. This can also be seen by directly plotting the training set against the evaluation set balanced accuracies as shown in **Figure 3**.

Table 3 lists the performance of the single descriptor package models with the highest balanced accuracy for each pathway endpoint together with their corresponding performance on the

TABLE 2 | Comparison of the performance of different descriptor packages in constructing QSAR models for *in vitro* pathway disruption prediction.

Descriptors package	Training total score	Training set rank	Evaluation total score	Evaluation set rank
Dragon 6	111	1	86	2
CDK	105	2	98	1
ISIDA Fragments	88	3	65	5
Chemaxon Descriptors	79	4	71	4
ALogPS, OEstate	73	5	79	3
Adriana.Code	55	6.5	55	8
QNPR	55	6.5	45	9
Inductive Descriptors	36	8	57	7
Mera, Mersy	30	9	62	6
GS Fragments	28	10	42	10

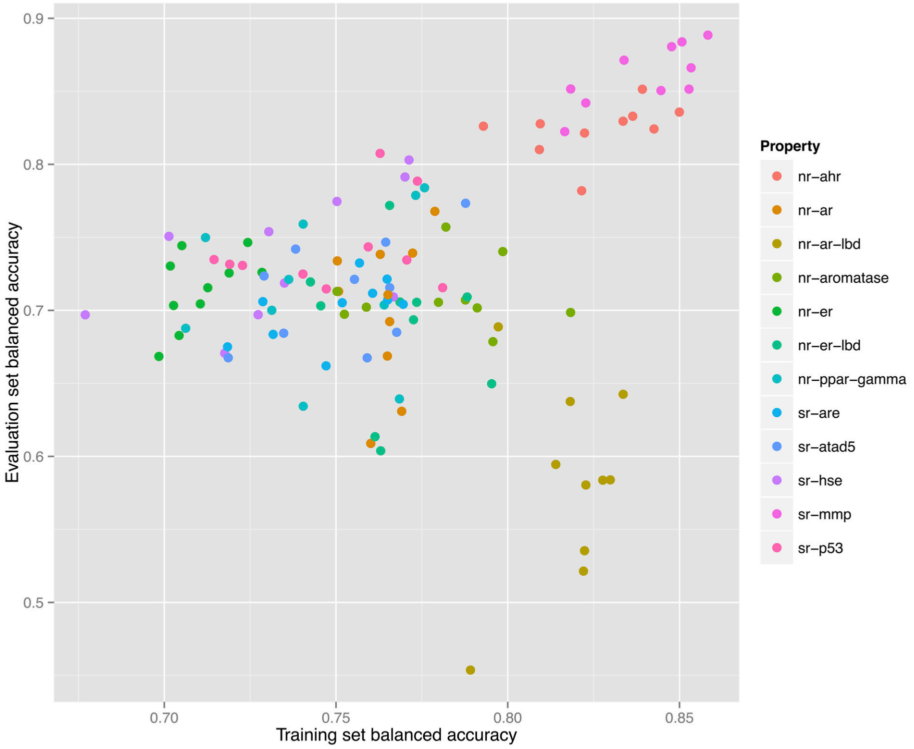


FIGURE 3 | Correlation between training and validation set balanced accuracies for 120 models constructed for 12 endpoints using 10 individual descriptor packages for each endpoint.

TABLE 3 | Performance of the single-descriptor-package models with the highest training set balanced accuracy for each pathway endpoint.

Molecular pathway endpoint	Descriptors package	Training balanced accuracy	Evaluation balanced accuracy	Wining balanced accuracy (evaluation set)
nr-ahr	CDK	0.850	0.836	0.853
nr-ar	CDK	0.779	0.768	0.736
nr-ar-lbd	CDK	0.834	0.643	0.650
nr-aromatase	Dragon 6	0.818	0.699	0.737
nr-er	CDK	0.728	0.726	0.749
nr-er-lbd	Dragon 6	0.795	0.650	0.715
nr-ppar-gamma	Dragon 6	0.776	0.784	0.785
sr-are	Dragon 6	0.770	0.704	0.729
sr-atad5	Dragon 6	0.788	0.773	0.741
sr-hse	Dragon 6	0.771	0.803	0.799
sr-mmp	CDK	0.858	0.888	0.904
sr-p53	ISIDA Fragments	0.781	0.716	0.765

*The balanced accuracies of winning models in the data challenge (Tox21 Data Challenge 2014 - Final Leaderboard) are shown for reference. Cases were models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. Supplementary Materials (**Data Sheet 1**) include the upper and lower boundaries for balanced accuracies as well as p-values.*

final evaluation set. The highest balanced accuracy achieved by any of the competing teams (measured on the evaluation set) during the challenge was reported online (Tox21 Data Challenge 2014—Final Leaderboard)[29]. It is also shown in **Table 3** (referred to as "winning balanced accuracy") for reference.

Consensus Modeling

Table 4 shows the consensus models with highest validated balanced accuracy based on the training set for each endpoint as well as their respective performance on the evaluation set. For all endpoints, consensus modeling was able to improve the performance on the training set. In six endpoints, the consensus models' predictive ability on the evaluation set would also result in a better than winning balanced accuracy.

For comparison, **Table 5** shows the performance of the consensus models involving all 10 underlying descriptor packages for each pathway endpoint. In seven endpoints, the predictive ability of these models on the evaluation set slightly exceeded those of the highest validated balanced accuracy (**Table 4**).

Descriptor packages differed in their success in representing the chemical structures. Some descriptor packages failed during the calculation phase for some of the molecules (e.g., reporting a chemical structure being too large for calculation). Therefore, models based on them would be deprived from any information gain from those failed molecules (i.e., will have a smaller training set size). A QSAR model built on such descriptors may show good statistics on the smaller training set but fail to perform similarly for an external evaluation set.

TABLE 4 | Performance of the consensus models with the highest training set balanced accuracy for each pathway endpoint.

Molecular pathway endpoint	Training set balanced accuracy	Evaluation set balanced accuracy	Wining balanced accuracy (evaluation set)	Ids for models used in building consensus
nr-ahr	0.865	0.859	0.853	512
nr-ar	0.785	0.752	0.736	515
nr-ar-lbd	0.838	0.592	0.650	516
nr-aromatase	0.824	0.715	0.737	513
nr-er	0.736	0.756	0.749	517
nr-er-lbd	0.810	0.726	0.715	518
nr-ppar-gamma	0.802	0.741	0.785	514
sr-are	0.799	0.730	0.729	534
sr-atad5	0.809	0.734	0.741	519
sr-hse	0.794	0.767	0.799	520
sr-mmp	0.882	0.900	0.904	521
sr-p53	0.795	0.783	0.765	522

*The balanced accuracies of winning models in the data challenge (Tox21 Data Challenge 2014 - Final Leaderboard) are shown for reference. Cases where models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. Supplementary Materials (**Data Sheet 1**) include the upper and lower boundaries for balanced accuracies as well as p-values.*

The second approach has the advantage of covering the largest number of molecules by compensating for the failure of some packages in descriptors calculation. It can also compensate for some packages bias by offering a wider range of molecular representations. However, it might suffer from the disadvantage of picking noise from descriptor packages with particularly bad performance. It also involves the highest computational expense, as applying such models to new molecules would require calculation of all descriptors from 10 packages. On the other hand, the first approach has the advantage of picking fewer descriptor packages with the highest performance.

DISCUSSION

The combination of the workflow tool (KNIME), the QSAR modeling platform (OCHEM), and the statistical package (CRAN R) allowed the creation and analysis of thousands of models with high efficiency. The use of HTS *in vitro* assays to construct QSAR models that are able to predict certain molecular pathways' perturbation paves the way toward a better understanding for the mode of chemical toxicity and allows for prioritization of testing efforts. This is in line with the vision of EPA and ECHA for replacing unnecessary animal toxicity testing, rapidly reducing information gaps, and achieving higher outcomes with available efforts and resources.

Due to the time constraint during the challenge, the consensus models selection for team AMAZIZ was based on expert

[29] Tox21 Data Challenge 2014—Final Leaderboard Available at: https://tripod.nih.gov/tox21/challenge/leaderboard.jsp [Accessed June 18, 2015].

TABLE 5 | Performance of the consensus models involving all 10 descriptor packages for each pathway endpoint.

Molecular pathway endpoint	Training set balanced accuracy	Evaluation set balanced accuracy	Wining balanced accuracy (evaluation set)
nr-ahr	0.850	0.858	0.853
nr-ar	0.770	0.754	0.736
nr-ar-lbd	0.824	0.599	0.650
nr-aromatase	0.811	0.760	0.737
nr-er	0.730	0.744	0.749
nr-er-lbd	0.794	0.756	0.715
nr-ppar-gamma	0.779	0.759	0.785
sr-are	0.789	0.707	0.729
sr-atad5	0.786	0.727	0.741
sr-hse	0.766	0.773	0.799
sr-mmp	0.875	0.903	0.904
sr-p53	0.784	0.759	0.765

*The balanced accuracies of winning models in the data challenge (Tox21 Data Challenge 2014 - Final Leaderboard) are shown for reference. Cases where models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. Supplementary Materials (**Data Sheet 1**) include the upper and lower boundaries for balanced accuracies as well as p-values.*

TABLE 6 | Models used for the final submission by team AMAZIZ during the Tox21 challenge.

Molecular pathway endpoint	Ids for models used in building consensus
nr-ahr	523
nr-ar	524
nr-ar-lbd	525
nr-aromatase	351
nr-er	526
nr-er-lbd	527
nr-ppar-gamma	528
sr-are	533
sr-atad5	529
sr-hse	530
sr-mmp	531
sr-p53	532

Consensus models involving all 10 descriptor packages (sr-are and sr-mmp) failed for the calculation of 23 molecules of the evaluation set and were replaced by simpler models, based on the consensus of three models only, predicting these molecules.

knowledge including the criteria discussed in this study, namely the performance of the model with regard to their balanced accuracy and to a lesser extent the AUROC, preference to descriptor packages, which show more success in representing a larger size of the training set and the simplicity of the underlying descriptor packages (e.g., 2D descriptors are simpler in calculation than 3D descriptors, as they do not need 3D optimization). **Table 6** shows the models that were used for the final submission of team AMAZIZ in the challenge. All models can be accessed through their identification numbers for further analysis and to run predictions on new compounds. This study represents a systemic approach to consensus models selection as well as a deeper analysis beyond the challenge.

The Androgen receptor GeneBLAzer AR-UAS-bla-GripTite cell line endpoint showed exceptional difficulty in modeling. Big discrimination exists between validated performance on the training set and the prediction ability on the evaluation set. Indeed, the endpoint has the lowest success in modeling in the challenge with the winning model being able to achieve a balanced accuracy of only 65% only (the lowest among all endpoints).

Further investigation of the models constructed for this endpoint shows multiple models that would have been able to achieve a higher predictive ability on the evaluation set (0.75–0.80) as shown in **Figure 4**. However, such models did not show the highest validated balanced accuracy and were thus not selected. The lack of direct correlation between validated balanced accuracy and predictive ability on the evaluation set (**Figure 3**) can be attributed to the statistical variation in the prediction performance of models for these sets and may also suggest that the split of the whole cluster of chemicals into training and evaluation sets may not have been completely random.

Although the alternative approaches for animal testing are highly encouraged, their proper use, and validity must be ensured. For QSAR model building, five **OECD principles** were established in 2004 (Directorate et al., 2007; OECD Quantitative Structure-Activity Relationships Project [(Q)SARs])[30]. The OECD principles were taken into consideration during the development of all QSAR models in this study as following:

- The first OECD principle is to have a defined endpoint to ensure the transparency in any physicochemical, biological, or environmental effect that a QSAR model is trying to assess. In this Tox21 challenge, 12 biological targets were well-defined by groups working on the experimental HTS part of the project - for assessment as listed in **Table 1**.
- The second principle is having an unambiguous algorithm. The “algorithm” refers to the form of relationship between the descriptors of chemical structure and the endpoint in the QSAR model. This can be mathematical/statistical methods or rule-based models defined by experts. Presenting a clear description of the algorithm ensures transparency and allows others to reproduce the model and explain how predictions are generated. In this study, all algorithms used for machine learning, descriptor packages, prefiltering criteria, validation as well as the chemical standardization procedures are described and can be reproduced using the online platform OCHEM. Indeed the process of building high quality QSAR models is tedious and complex. However, by documenting all steps, it is reproducible. Furthermore, by publishing all final models online, the scientific community has continuous access to perform predictions on the constructed QSAR models without a need to reproduce them.

[30] OECD Quantitative Structure-Activity Relationships Project [(Q)SARs] Available at: http://www.oecd.org/chemicalsafety/testing/oecdquantitativestructure-activityrelationshipsprojectqsars.htm [Accessed June 23, 2015].

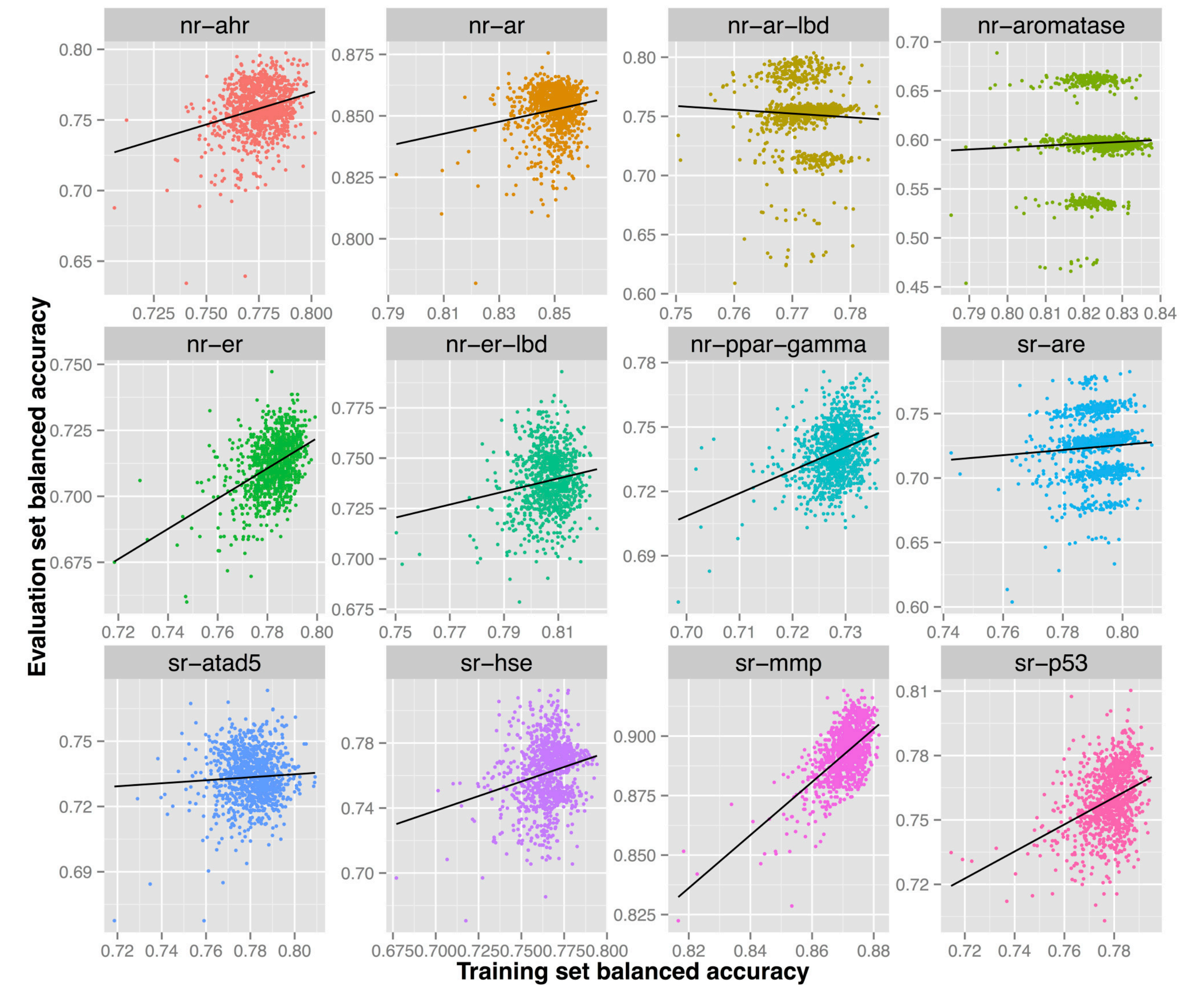


FIGURE 4 | Each sub-figure shows the performance of 1023 consensus models constructed for a single endpoint with x-axis representing the validated balanced accuracy on the training set and y-axis shows the balanced accuracy on the evaluation set. A positive trend line can be noticed with all endpoints except nr-ar-lbd.

- The third principle, defining domain of applicability, QSAR models are expected to give reliable predictions only for chemicals that are similar to the ones used in the model's training process. In this study, quantitative assessment of the model's confidence in prediction was estimated for all models. This reports the degree of similarity between the compound to be predicted and the model's training set (Sushko, 2011; Sahigara et al., 2012).
- The fourth principle is having appropriate measures of goodness-of-fit, robustness, and predictivity. This principle highlights the need for statistical validation of QSAR models in order to judge models' performance. Such performance validation can be either internal or external. In this study, bootstrap aggregation was used to estimate validation accuracy for the training set. The main statistical parameter applied for comparing all models was balanced accuracy. Performance of all models was also verified against an external test set.
- The fifth and last principle is having a mechanistic interpretation, if possible. The "if possible" phrase shows that the mechanistic interpretation is not mandatory for model acceptance by regulators. Sometimes, the iterative model building process and the involvement of data-mining techniques increase the complexity of the developed QSAR models through multiple training set refinements rendering the mechanistic interpretation hard to directly establish. A different approach for interpretation of complex models using matched molecular pairs was previously suggested (Sushko et al., 2014). All models in this study can be examined using this approach on the OCHEM platform.

The ultimate goal of QSAR models in predictive toxicology, ordinarily, is to forecast an adverse outcome rather than protein binding. In this sense, QSAR prediction of molecular pathways' perturbation is, in itself, an attempt to mechanistically understand toxicological risks. In the context of adverse outcome pathways (AOP), such perturbations are considered as molecular initiating events (MIE), or key events (KE) leading to certain adverse outcome. Such KEs are connected through key event relationships (KERs) to form the network of multiple AOPs. These AOPs form the functional prediction component for real-life circumstances (Villeneuve et al., 2014). In a joint effort between the European Commission—DG Joint Research Centre (JRC) and U.S. EPA, an AOP wiki is being developed. Among its goals is the accommodation of the worldwide efforts for AOP development. The wiki is one of the components of the OECD-sponsored AOP Knowledgebase. The investigated molecular pathways have been suggested to play a role in many adverse outcomes. A comprehensive analysis of the biological impact of the perturbation of these pathways is beyond the scope of this article.

CONCLUSIONS

Using QSAR for modeling the outcome of *in vitro* toxicity assays (representing different molecular pathways) showed promising success with balanced accuracies reaching up to more than 85% for several endpoints as shown in **Table 4**. The relatively high balanced accuracies among models confirmed the possibility of modeling HTS results from *in vitro* assays using *in silico* descriptors as reported in earlier studies (Abdelaziz et al., 2015).

Bagging validation provided a good indication for the models' predictive ability on external validation sets (**Figure 3**). Stratified bagging addressed the unbalanced nature of the training set and reduced bias toward the majority class. The stratified bagging contributed models, which were optimized toward the balanced accuracy. Moreover, the selection of consensus models also used balanced accuracy as the optimization criteria. This is one of the reasons why models developed in this study calculated the best balanced accuracy across all 12 analyzed targets and did not get the highest AUROC scores, which were used by competition organizers to rank the models. However, despite this, the used strategy allowed to calculate the highest AUROC scores for two targets. It is also important to realize that, due to the model prediction variances, selecting a model with the highest validated accuracy does not guarantee the highest predictive ability for an evaluation set.

Consensus modeling improved the predictive ability of models as signified by both validation and evaluation set accuracies. To a large degree this result was achieved thanks to the diversity of descriptor packages, which captured different aspects of the molecular structures. Use of different descriptors also compensated for failure of some descriptors to represent certain structures and thus covering the entire training set.

In summary, a computational methodology used to develop QSAR models was described. This methodology achieved the highest balanced accuracy for all of the Tox21 Data Challenge organized by the NIH. A similar strategy of consensus modeling was also successful to develop Rank-1 model for another Tox21 Challenge organized by the EPA and TopCoder (Novoratskyi et al., under review). Moreover, the developed models are made publicly available at http://ochem.eu/article/98009 thus allowing other researchers to use them for prospective and retrospective analyses.

AUTHOR CONTRIBUTIONS

AA designed and executed the study including the R scripts for model building and the KNIME workflows for data handling. He participated as the sole member of team AMAZIZ in the Tox21 data challenge. He is now CEO of Rosettastein Consulting. IT is CEO of BigChem GmbH, who supports and advances the OCHEM platform, which originated in his group in HMGU. He implemented calculation algorithms (ASNN, stratified bagging) and optimized OCHEM API interfaces employed in this study. HSL and KWS provided guidance on the conception and design of the overall study strategy. All authors revised the work, reviewed the manuscript and approved the final version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

We thank ChemAxon (http://www.chemaxon.com), Molecular Networks GmbH (http://www.molecular-networks.com), Talete Srl (http://www.talete.mi.it), and ChemoSophia (http://www.chemosophia.com) for contributing their software tools to OCHEM that were used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00002

Data Sheet 1 | Statistical parameters for 8296 models constructed using consensus between 120 models from 10 descriptor packages.

REFERENCES

Abdelaziz, A., Sushko, Y., Novotarskyi, S., Körner, R., Brandmaier, S. V., and Tetko, I. (2015). Using online tool (iPrior) for modeling toxcast™ assays towards prioritization of animal toxicity testing. *Comb. Chem. High Throughput Screen.* 18, 420–438. doi: 10.2174/1386207318666150305155255

Aires-de-Sousa, J., and Gasteiger, J. (2001). New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions. *J. Chem. Inf. Comput. Sci.* 41, 369–375. doi: 10.1021/ci000125n

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2007). "KNIME: the Konstanz Information Miner," in *Studies in Classification,* *Data Analysis, and Knowledge Organization (GfKL 2007)*, eds C. Preisach, P. D. H. Burkhardt, P. D. D. L. Schmidt-Thieme, and P. D. R. Decker (Freiburg: Springer), 319–326.

Betts, K. S. (2013). Tox21 to date: steps toward modernizing human hazard characterization. *Environ. Health Perspect.* 121:A228. doi: 10.1289/ehp.121-a228

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M., and Hammond, G. L. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *J. Med. Chem.* 51, 2047–2056. doi: 10.1021/jm7011485

Chesbrough, H. W. (2006). *Open Innovation: The New Imperative for Creating and Profiting from Technology.* Boston, MA: Harvard Business Press.

Directorate, E., Meeting, J., The, O. F., Committee, C., Working, T. H. E., and On, P. (2007). *OECD Environment Health and Safety Publications series on testing and assessment No. 69 GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP [(Q) SAR] MODELS Environment Directorate.*

Grishina, M. A., Bartashevich, E. V., Potemkin, V. A., and Belik, A., V (2002). Genetic algorithm for predicting structures and properties of molecular aggregates in organic substances. *J. Struct. Chem.* 43, 1040–1044. doi: 10.1023/A:1023663115138

Hall, L. H., Kier, L. B., and Brown, B. B. (1995). Molecular similarity based on novel atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* 35, 1074–1080. doi: 10.1021/ci00028a019

Hanley, J. A., and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148, 839–843. doi: 10.1148/radiology.148.3.6878708

Huuskonen, J. J., Livingstone, D. J., and Tetko, I. V. (2000). Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* 40, 947–955. doi: 10.1021/ci9904261

James, C. A., Weininger, D., and Delany, J. (1995). *Daylight Theory Manual*. Irvine, CA: Daylight Chemical Information Systems. Inc.

Judson, R. S., Kavlock, R. J., Setzer, R. W., Hubal, E. A. C., Martin, M. T., Knudsen, T. B., et al. (2011). Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem. Res. Toxicol.* 24, 451–462. doi: 10.1021/tx100428e

Kavlock, R., and Dix, D. (2010). Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *J. Toxicol. Environ. Health. B. Crit. Rev*. 13, 197–217. doi: 10.1080/10937404.2010.483935

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., and others (2006). Handling imbalanced datasets: a review. *GESTS Int. Trans. Comput. Sci. Eng.* 30, 25–36.

Potemkin, V. A., and Grishina, M. A. (2008). A new paradigm for pattern recognition of drugs. *J. Comput. Aided. Mol. Des.* 22, 489–505. doi: 10.1007/s10822-008-9203-x

Potemkin, V. A., Pogrebnoy, A. A., and Grishina, M. A. (2009). Technique for energy decomposition in the study of "receptor-ligand" complexes. *J. Chem. Inf. Model.* 49, 1389–1406. doi: 10.1021/ci800405n,

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Available online at: http://www.r-project.org

Rosenblatt, F. (1957). *The Perceptron, A Perceiving and Recognizing Automaton Project Para*. New York, NY: Cornell Aeronautical Laboratory.

Sadowski, J., Gasteiger, J., and Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* 34, 1000–1008. doi: 10.1021/ci00020a039

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., and Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17, 4791–4810. doi: 10.3390/molecules17054791

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The Chemistry Development Kit (CDK): an open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43, 493–500. doi: 10.1021/ci025584y

Sushko, I. (2011). *Applicability Domain of QSAR Models.* Available online at: http://mediatum.ub.tum.de?id=1004002 (Accessed August 31, 2014).

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A. K., Rupp, M., Teetz, W., et al. (2011). Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided. Mol. Des.* 25, 533–554. doi: 10.1007/s10822-011-9440-2

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, 48. doi: 10.1186/s13321-014-0048-0

Tetko, I. V. (2002a). Associative neural network. *Neural Process. Lett.* 16, 187–199. doi: 10.1023/A:1019903710291

Tetko, I. V. (2002b). Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* 42, 717–728. doi: 10.1021/ci010379o

Tetko, I. V., Bruneau, P., Mewes, H.-W., Rohrer, D. C., and Poda, G. I. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today* 11, 700–707. doi: 10.1016/j.drudis.2006.06.013

Tetko, I. V., Novotarskyi, S., Sushko, I., Ivanov, V., Petrenko, A. E., Dieden, R., et al. (2013). Development of dimethyl sulfoxide solubility models using 163,000 molecules: using a domain applicability metric to select more reliable predictions. *J. Chem. Inf. Model.* 53, 1990–2000. doi: 10.1021/ci400213d

Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N., and Villa, A. E. P. (2001a). Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* 41, 1488–1493. doi: 10.1021/ci000392t

Tetko, I. V., Tanchuk, V. Y., and Villa, A. E. P. (2001b). Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* 41, 1407–1421. doi: 10.1021/ci010368v

Thormann, M., Vidal, D., Almstetter, M., and Pons, M. (2007). Nomen est omen: quantitative prediction of molecular properties directly from IUPAC names. *Open Appl. Informatics J.* 1, 28–32. doi: 10.2174/1874136300701010028

Tice, R., Kavlock, R., and Christopher Austin (2009). *The U.S. "Tox21 Community" and the Future of Toxicology.* Available online at: http://www.epa.gov/ncct/bosc_review/2009/posters/1-08_Tice_CompTox_BOSC09.pdf [Accessed January 15, 2014].

Todeschini, R., and Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. 2nd, Rev. Milano: John Wiley & Sons.

Tollenaere, T. (1990). SuperSAB: fast adaptive back propagation with good scaling properties. *Neural Netw.* 3, 561–573. doi: 10.1016/0893-6080(90)90006-7

Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., et al. (2008). ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput. Aided. Drug Des.* 4, 191–198. doi: 10.2174/157340908785747465

Villeneuve, D. L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T. H., LaLone, C. A., et al. (2014). Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicol. Sci.* 142, 312–320. doi: 10.1093/toxsci/kfu199

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S., Sochaski, M. A., Rotroff, D. M., Freeman, K., et al. (2012). Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol. Sci. An Off. J. Soc. Toxicol.* 125, 157–174. doi: 10.1093/toxsci/kfr254

Worth, A. P., Bassan, A., Gallegos, A., Netzeva, T. I., Patlewicz, G., Pavan, M., et al. (2005). *The Characterisation of (quantitative) Structure-Activity Relationships: Preliminary Guidance*. Institute for Health and Consumer Protection, Toxicology and Chemical Substances Unit, European Chemical Bureau.

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 Environmental Science

ORIGINAL RESEARCH
published: 15 August 2016
doi: 10.3389/fenvs.2016.00052

Identifying Biological Pathway Interrupting Toxins Using Multi-Tree Ensembles

Gergo Barta *

Data Mining Group, Data and Content Technologies Laboratory, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

The pharmaceutical industry constantly seeks new ways to improve current methods that scientists use to evaluate environmental chemicals and develop new medicines. Various automated steps are involved in the process as testing hundreds of thousands of chemicals manually would be infeasible. Our research effort and the Toxicology in the Twenty First Century Data Challenge focused on cost-effective automation of toxicological testing, a chemical substance screening process looking for possible toxic effects caused by interrupting biological pathways. The computational models we propose in this paper successfully combine various publicly available substance fingerprinting tools with advanced machine learning techniques. In our paper, we explore the significance and utility of assorted feature selection methods as the structural analyzers generate a plethora of features for each substance. Machine learning models were carefully selected and evaluated based on their capability to cope with the high-dimensional high-variety data with multi-tree ensemble methods coming out on top. Techniques like Random forests and Extra trees combine numerous simple tree models and proved to produce reliable predictions on toxic activity while being nearly non-parametric and insensitive to dimensionality extremes. The Tox21 Data Challenge contest offered a great platform to compare a wide range of solutions in a controlled and orderly manner. The results clearly demonstrate that the generic approach presented in this paper is comparable to advanced deep learning and domain-specific solutions. Even surpassing the competition in some nuclear receptor signaling and stress pathway assays and achieving an accuracy of up to 94 percent.

Keywords: Classification, random forest, toxicity, Tox21, challenge, competition

OPEN ACCESS

Edited by:
Ruili Huang,
National Center for Advancing Translational Sciences (NIH), USA

Reviewed by:
Luis Gomez,
University of Las Palmas de Gran Canaria, Spain
Mohamed Diwan M. AbdulHameed,
Department of Defense Biotechnology High Performance Computing Software Applications Institute, USA

****Correspondence:***
Gergo Barta
barta@tmit.bme.hu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *22 March 2016*
Accepted: *29 July 2016*
Published: *15 August 2016*

Citation:
Barta G (2016) Identifying Biological Pathway Interrupting Toxins Using Multi-Tree Ensembles. Front. Environ. Sci. 4:52. doi: 10.3389/fenvs.2016.00052

1. INTRODUCTION

Traditional toxicity testing protocols using animal experiment-based models have many drawbacks; they are expensive, time-consuming (Shukla et al., 2010) and might raise ethical or reliability concerns. The urgent need to involve alternative methods in chemical risk assessment drove the National Research Council (NRC) in the U.S. to project a new vision and strategy for the increased use of *in vitro* technologies in toxicity screening studies (Krewski et al., 2010). European measures soon followed as the European Chemical Agency (ECHA) issued similar guidelines. These guidelines promoted quick and cost effective computational methods and described the role of animal testing as the last resort. Over the years, this lead to the development and wide-scale implementation of high-throughput screening (HTS) techniques. A process that is capable of screening thousands of compounds using a quick and standardized protocol, furthermore, it may be combined with robotic methods (Malo et al., 2006).

The popularity of HTS opened up chemical toxicity research to machine learning and the big data era. The need of novel techniques in data handling, data transformation, and data mining sparked substantial research efforts throughout the years. This new emerging trend brought about the convergence of toxicity screening protocols and conventional graphical data mining tools (e.g., RapidMiner[1], KNIME[2]) or popular scripting languages in data science (R[3], Python[4]). With various modules, libraries, and extensions available to read, transform and analyze HTS assay data, it really comes down to a choice of preference.

Over the years, Random forests (Svetnik et al., 2003), projection pursuit, partial least squares and Support vector machines (Si et al., 2007) have been applied successfully to the Quantitative Structure-Activity Relationship (QSAR) task. Each of these methods has different advantages and disadvantages (see Liu and Long, 2009 for a detailed review). Judson et al. (2008) also carried out an extensive review of conventional machine learning methods applied in HTS; methods included Nearest neighbors, Nave Bayes, Regression trees, Support vector machines, Artificial neural networks. The comparison showed that most models provide comparable performance when suitable data preparation is carried out. The authors identified careful feature selection as the most crucial step in preparing the data. Furthermore, Dahl et al. successfully applied multi-task neural networks to exploit task inter-dependencies (Dahl et al., 2014).

The usage of Random forests in HTS applications was first suggested by Svetnik et al. (2003). Svetnik et al. demonstrated superior performance compared to other methods at the time and described additional useful features of the proposed method. The main strengths were identified as high classification performance, aggressive regularization to capture sparsity and useful services such as built-in performance assessment and feature importance.

The following document describes in detail team Dmlab's approach to solving the Tox21 Data Challenge[5]. The challenge offered a compound toxicity screening classification problem on two panels [Nuclear Receptor Signaling (NR) and Stress Response (SR)] and 12 different assays: Androgen Receptor (AR, AR-LBD), Aryl Hydrocarbon Receptor (AhR), Estrogen Receptor (ER, ER-LBD), Aromatase Inhibitors (aromatase), Peroxisome Proliferator-activated receptor gamma (ppar-gamma), Antioxidant Response Element (ARE), luciferase-tagged ATAD5 (ATAD5), Heat Shock Response (HSE), Mitochondrial Membrane Potential (MMP), and Agonists Of The P53 Signaling Pathway (P53). For further details on the competition, see Huang et al. (2016).

Our general approach was to utilize the vast machine learning features offered by Python's scikit-learn library[6] and prepare the dataset for analysis by combining data manipulating tools (RapidMiner and KNIME) with domain specific structure analyzers in order to provide high-accuracy toxicity screening.

This article contains three major sections:

1. Materials and Methods shows the underlying models in detail with references, introduces the software used, provides data description, and basic statistics. The second part of this section describes how the substance screening framework works and how to reproduce contest results.
2. Results contains the thorough evaluation of the proposed methods in the competition context
3. Conclusions and discussions are provided in the last section with an indication of future research directions.

2. MATERIALS AND METHODS

The Tox21 Data Challenge portal contains helpful guidance and a multitude of materials to start working on the problem. The challenge organizers even generously provided a simple benchmark solution to kickstart the process, comparison of the benchmark and team Dmlab's approach can be found in **Table 1**. While the Naïve Bayes classifier utilized in the benchmark is a good initial approach, it falls behind when it comes to parameter tuning options and accuracy in general. Finding a more suitable classifier was chief among the goals of this competition. The same goes for replacing the lower level components of the stack; using the same inputs as other challengers gives no edge in a competitive environment.

The flow chart in **Figure 1** gives a high-level overview of the solution process and the techniques combined. The process involves 3 major steps: data preparation, modeling, and post-processing. The data preparation step includes deriving descriptors from structural information, transforming the data to suit modeling purposes and finalizing the set of descriptors to be used. Modeling involves model selection, parameter tuning and generating predictions. The post-processing step covers the optimal threshold selection and application process to generate toxicity decisions.

2.1. Data Description

The Tox21 Data Challenge provided a dataset with the structural information of 11,737 distinct molecules. The different assays contained results for between 7143 and 9068 of the molecules. The respective activity flag was used as the target variable of analysis for each individual track.

While challenge tracks are intended to be independent a quick correlation and clustering analysis shows signs of

TABLE 1 | Building the solution stack.

	Benchmark solution	Dmlab solution
Molecular descriptors	Library synthesizer	PaDel descriptor/RDKit
Fingerprinting	PCFP (PubChem)	PubChem/Avalon
Structure standardizer	LyChI	PaDel descriptor/RDKit
Classifier model	Naïve bayes	Random forest/extra trees

[1] https://rapidminer.com/products/studio/
[2] https://www.knime.org/
[3] https://www.r-project.org/
[4] https://www.python.org/
[5] https://tripod.nih.gov/tox21/challenge/
[6] http://scikit-learn.org/

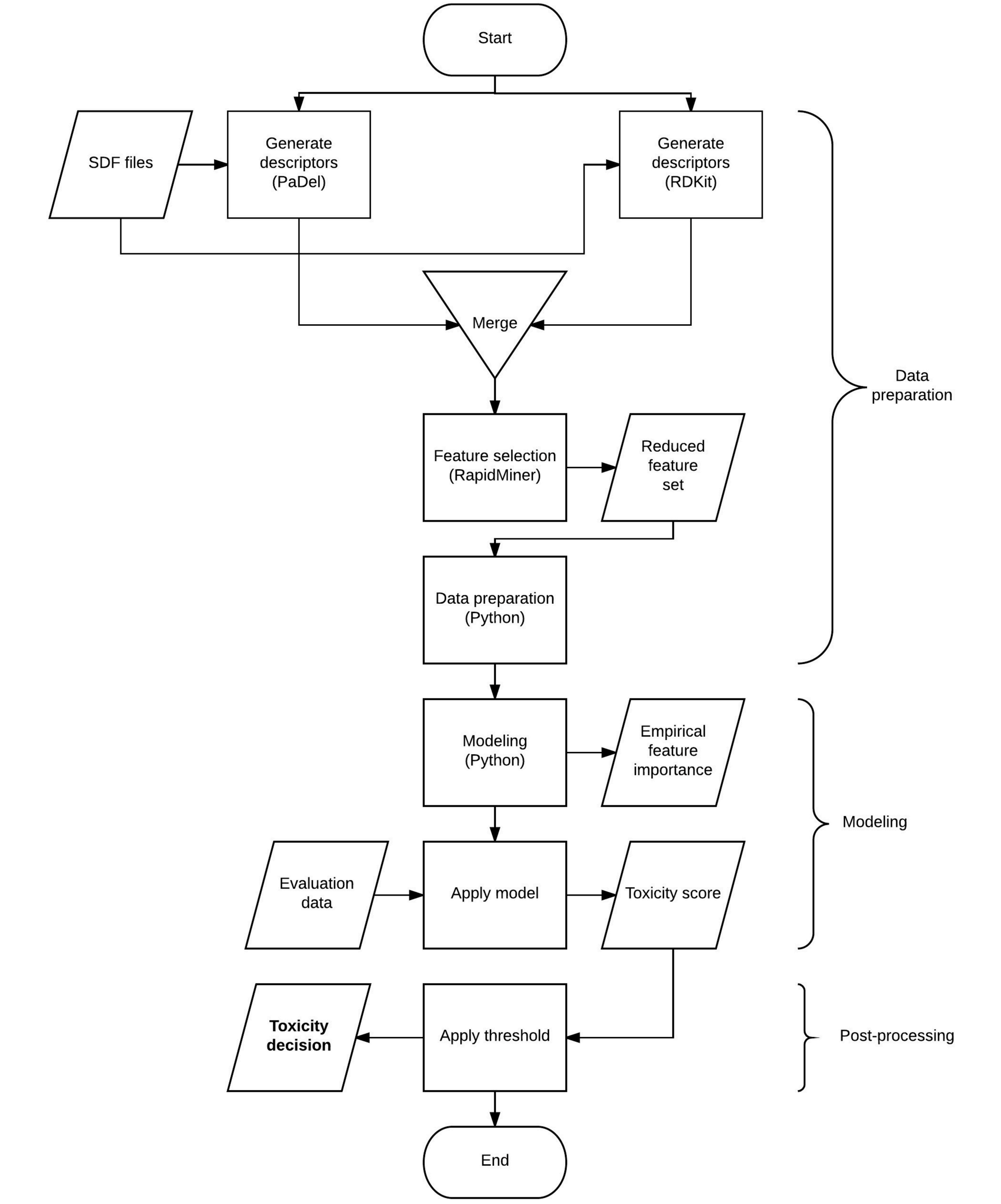


FIGURE 1 | Detailed overview of the proposed solution.

positive correlation between track activities in various cases (see **Figure 2**). Notably, the closest relationship is between NR-AR, NR-ER and their LBD counterparts. Surprisingly enough, NR and SR assays mix in the two other clusters, one containing AhR, Aromatase, ARE, and MMP, while the other includes the remaining assays; PPAR-gamma, ATAD5, HSE, and p53. Correlation coefficients hint at possible inter-track information gain that could be harnessed to achieve better classifier performance, but no such action was taken during the challenge. A promising direction for future research.

2.1.1. Generating Descriptors

At the beginning of the analysis, the structural information of the molecules in the training and test set has to be processed to generate descriptive attributes for data analysis. During the challenge, our team used 2 different versatile tools to generate the descriptive attributes; PaDel Descriptor and RDKit cheminformatics toolkit. Other tools, like the CDK Descriptor Calculator[7], were also experimented with but failed to generate conclusive results.

2.1.2. PaDel Descriptor

PaDel Descriptor[8] was developed by the Pharmaceutical Data Exploration Laboratory at the National University of Singapore. The tool has the capabilities to generate 1-dimensional, 2-dimensional structural information and many fingerprints as seen in Yap (2011), and also operates in a multi-core

[7] http://www.rguha.net/code/java/cdkdesc.html
[8] http://www.yapcwsoft.com/dd/padeldescriptor/

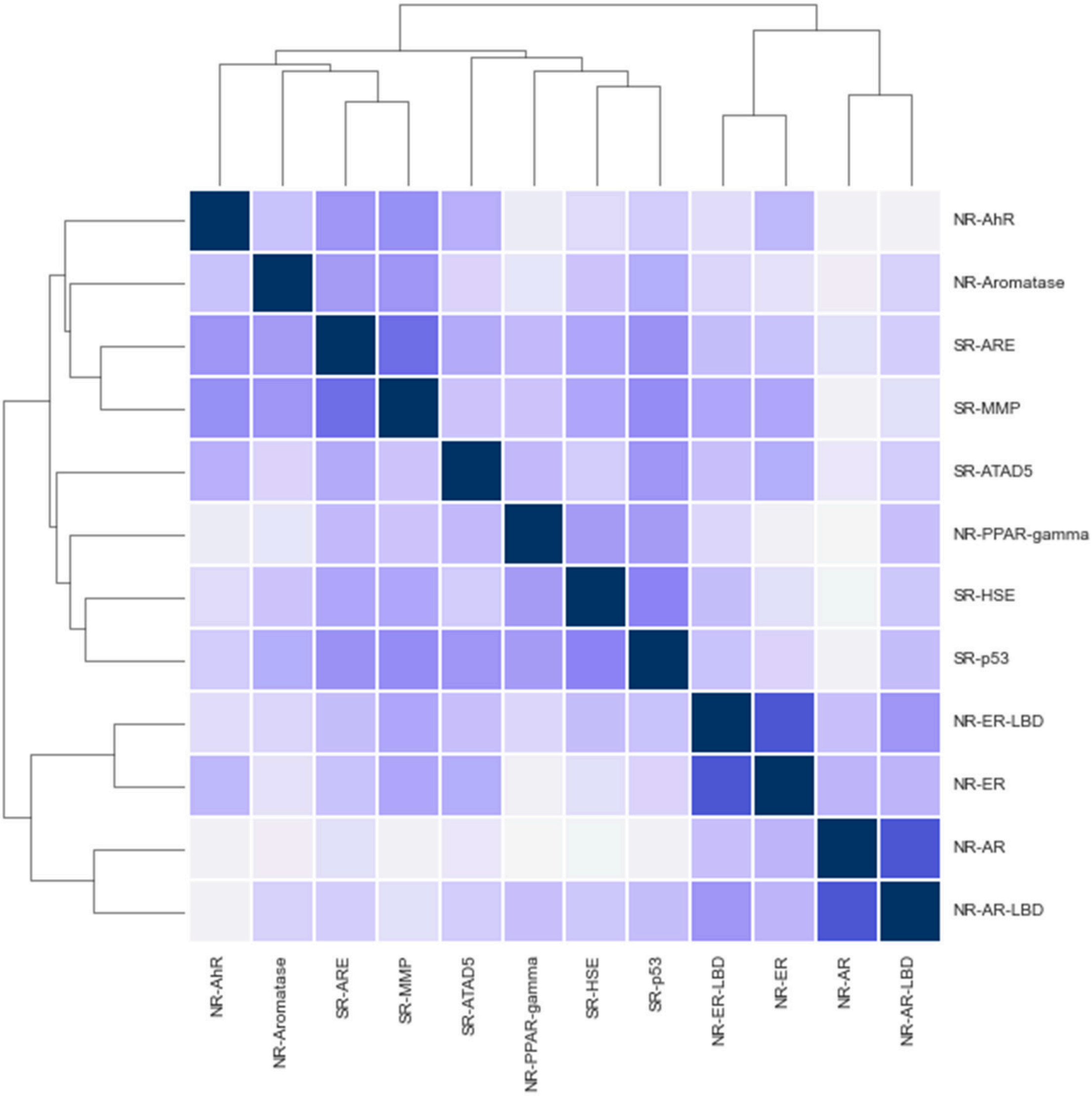


FIGURE 2 | Correlation and potential clustering of challenge tracks.

fashion to reduce computational times. It also acts as a structure standardizer; removes salts, detects aromaticity, and standardizes nitro groups. As a result, 1444 2-dimensional attributes were extracted from the structures. 3-dimensional descriptors were also experienced with, but failed to be generated for many molecules, and ultimately were discarded from the analysis.

In addition, the tool also offers 12 different fingerprint versions; CDK fingerprint, CDK extended fingerprint, Estate fingerprint, CDK graph only fingerprint, MACCS fingerprint, PubChem fingerprint, Substructure fingerprint, Substructure fingerprint count, Klekota-Roth fingerprint, Klekota-Roth fingerprint count, 2D atom pairs, and 2D atom pairs count. Out of those the PubChem Substructure Fingerprint (see Bolton et al., 2008) was selected based on its empirical performance widespread use, which is also the default fingerprinting method in the PaDel Descriptor. It is a 2-dimensional chemical structure fingerprint that consists of an 881-dimension binary vector. Each bit represents a boolean determination of the absence or presence of a specific structural element as can be seen in the PubChem Substructure Fingerprint manual[9].

2.1.3. RDKit Cheminformatics Toolkit

RDKit, an open source toolkit for cheminformatics[10], was also utilized in the descriptor generating process. There are many wrappers available to work with the tool in different environments; our team chose the KNIME extension, which inherently works with SDF and SMILES files (see **Figure 3**).

Descriptive attributes were generated by the descriptor calculation feature. In addition Gasteiger charges were also calculated by the calculate charges feature, see Gasteiger and Marsili (1980) for details. The toolkit also offers 8 distinct fingerprints to be generated; Morgan, FeatMorgan, AtomPair, Torsion, RDKit, Avalon, Layered, and MACCS. Empirical results showed the Avalon fingerprint[11] as the most promising, and was selected as the final fingerprinting option to work with.

Similarly to the PaDel scenario, salts were removed at the beginning of the process using the salt stripper feature. In the end, 117 descriptors, 118 charges, and 1024 fingerprint flags were extracted as new features.

2.2. Data Preparation

The resulting dataset used in data analysis combines two main sources. PaDel descriptor provides 2-dimensional features and the PubChem fingerprint, while RDKit adds its respective structural descriptors along with the Avalon fingerprint.

The whole dataset contains 3418 attributes; this means a relatively wide data table that makes feature selection a top priority. The many descriptors represent a high dimensional sparsely inhabited feature space. Cautious measures have to be

[9] ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt

[10] http://www.rdkit.org

[11] http://sourceforge.net/projects/avalontoolkit

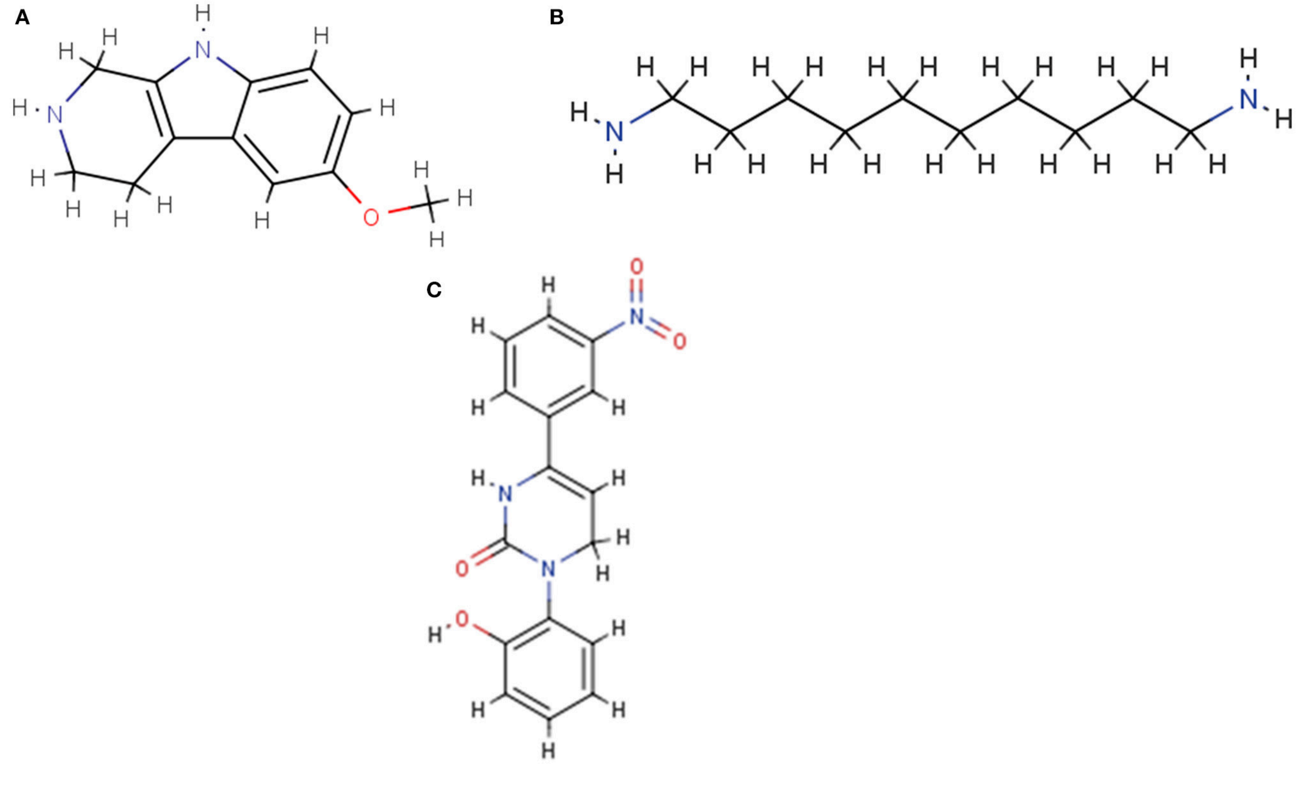


FIGURE 3 | Sample molecule structures displayed by the RDKit cheminformatics toolkit. (A) NCGC00260687. **(B)** NCGC00261143. **(C)** NCGC00261111.

taken as this kind of classification problem is particularly prone to overfitting.

Additionally, some of these attributes overlap, as different sources provide the same functionality. First, correlated attributes were removed to avoid the effect of multicollinearity as suggested by Chong and Jun (2005). In this step, attributes were filtered where a pairwise correlation was above 0.95. Other attributes were deemed useless and removed, based on their low variance (below 0.1) or high ratio of missing values (above 10%). Note, this step removes many of the Gasteiger charges.

Literature review underlined the importance of feature selection in QSAR protocols. To carry out careful filtering of the feature space, the functionality of conventional methods were combined in a novel way. RapidMiner, a reliable data analysis software, offers various feature selection operators (Schowe, 2011), and also comes with a powerful extension[12] to further extend options. In this attempt, 5 basic feature selector operators were combined to generate a versatile ranking of individual attributes, thus creating a flexible attribute filtering scheme. The basic operators include calculating feature relevance by computing the value of correlation with respect to the target attribute, based on the information gain ratio, based on the Gini impurity index, by measuring the symmetrical uncertainty with respect to the class, and according to how well their values distinguish between the instances of the same and different classes that are near each other.

Summing the aforementioned ranks represents the universal scoring for the given input variable provided by a committee of experts, thus creating a more reliable ranking. Using the universal scoring, 681 features were selected for further analysis, meaning depending on the assay the analytical base table contained roughly 10–13 times more observations than features, a data table size much less prone to overfitting. Further details on the final set of descriptors and most important input features for all 3 winning tracks are provided in the Supplementary Materials.

Many of the structural descriptor features contain missing values; we decided that attributes with excessive missing values are to be entirely removed. Some molecule structures are prone to fail to generate descriptors in PaDel and/or RDKit, and thus missing values are generated. Classification models implemented in Python do not handle missing values well, so all rows in the training set including such values were removed entirely. On the validation sets (test set and final evaluation set), where dropping a molecule was no option, such values were imputed with a fixed 0 value, which in the case of fingerprints, represents the absence of a specific pattern and is considered a safe option.

2.3. Random Forests and Extra Trees

The Random forest is perhaps the best-known of ensemble methods, thus it combines simple models called base learners for

TABLE 2 | Searching the parameter space.

Model	Parameter	Options tested
Random forest classifier Extra trees classifier	Splitting criterion	Gini, entropy
	Number of estimators	499, 799, 999, 1200
Support vector classifier	Kernel	Radial basis function, linear
	Gamma parameter	0.01, 0.001, 0,0001
	C parameter	1, 10, 100,1000
	Class weight	auto, none
Gradient boosting classifier	Learning rate	0.01, 0.1, 0.3
	Number of estimators	250, 500,1200
	Max tree depth	2, 3, 5
	Subsampling	0.75, 0.9, 1.0

[12] http://sourceforge.net/projects/rm-featselext/

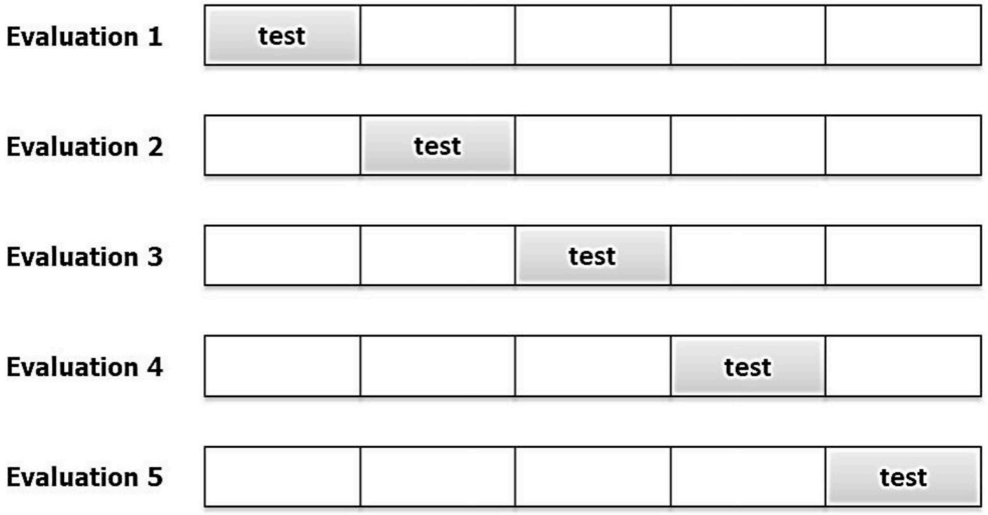


FIGURE 4 | Illustrative example of 5-fold cross-validation.

TABLE 3 | Comparison of leaderboard and final performance on all assays.

Panel	Assay	Modeling method	Tuned model parameters	LB AUC	Eval. AUC	Balanced acc.
NR	AR	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.71	0.83	0.61
NR	Ahr	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.85	0.78	0.56
NR	AR-LBD	RandomForestClassifier	No. estimators: 499, criterion = "entropy"	0.86	0.82	0.49
NR	ER	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.70	0.77	0.66
NR	ER-LBD	RandomForestClassifier	No. estimators: 799, criterion = "entropy"	0.79	0.77	0.59
NR	Aromatase	ExtraTreesClassifier	No. estimators: 999, criterion = "entropy"	0.85	0.84	0.56
NR	PPAR-gamma	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.83	0.83	0.55
SR	ARE	SupportVectorClassifier	Kernel type: ANOVA	0.82	0.77	0.52
SR	ATAD5	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.80	0.80	0.61
SR	HSE	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.88	0.86	0.56
SR	MMP	ExtraTreesClassifier	No. estimators: 799, criterion = "entropy"	0.93	0.95	0.69
SR	p53	ExtraTreesClassifier	No. estimators: 499, criterion = "entropy"	0.74	0.88	0.58

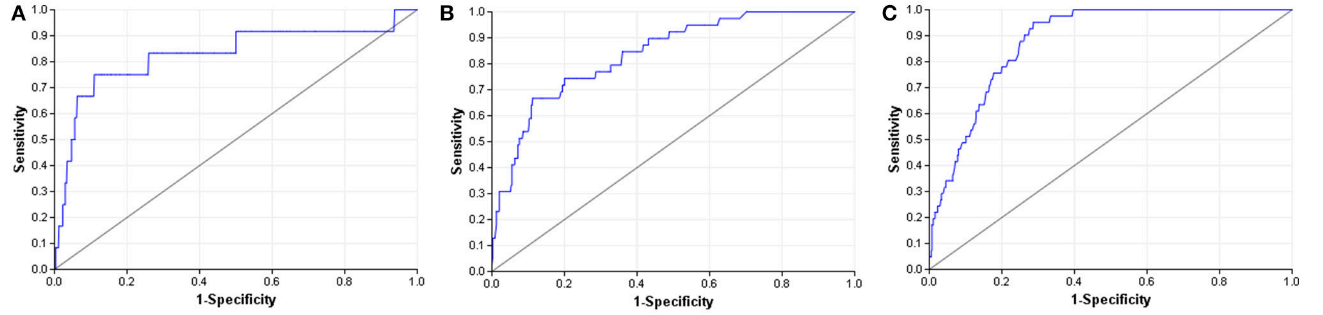


FIGURE 5 | ROC curves of the winning models. (A) NR-AR. **(B)** NR-Aromatase. **(C)** SR-P53.

increased performance. In this case, multiple tree models are used to creating a forest as introduced by Breiman (2001).

There are three key factors of forest creation:

1. bootstrapping the dataset
2. growing unpruned trees
3. limiting the candidate features at each split

These steps ensure that reasonably different trees are grown in each turn of iteration, which is key to the effective model combination.

The bootstrapping step of the model creation carries out a random sampling of a dataset with N observations with a replacement that results in N rows, but only ca. 63% of the data used as stated in (1) (Efron and Tibshirani, 1993). The probability that an observation x does not get into the sample S equals

$$P(x \notin S) = (1 - \frac{1}{n}) \approx e^{-1} = 0.368 \tag{1}$$

Pruning the trees would reduce variance between trees and thus considered inessential as the overfitting of individual trees is balanced anyway by the ensemble.

When growing trees a different set of features is proposed as candidates in finding the best split based on information criteria like Gini or entropy. The subset of features is selected randomly further increasing the variance between trees.

The output of the trees is then combined by averaging the results based on some weights or by performing a majority vote in the case of classification problems.

Random forests have very few vital parameters to tune, they are effectively non-parametric. The unique architecture provides many benefits and is widely recognized as a good initial approach to most problems. Unlike decision trees, the ensemble method's averaging property inherently finds a balance between high variance and bias. It is insensitive to many data related issues such as the large number and heterogeneity of features, outliers, missing data, and even an unbalanced target. Other than being a great out-of-the-box tool it offers various useful services. Random forest gives an intrinsic evaluation of the results based on the data discarded by bootstrapping (called out-of-bag error), it also gives estimates what variables are important.

Extra Trees is a slightly different Random forest variant suggested by Pierre Geurts, Damien Ernst and Louis Wehenkel

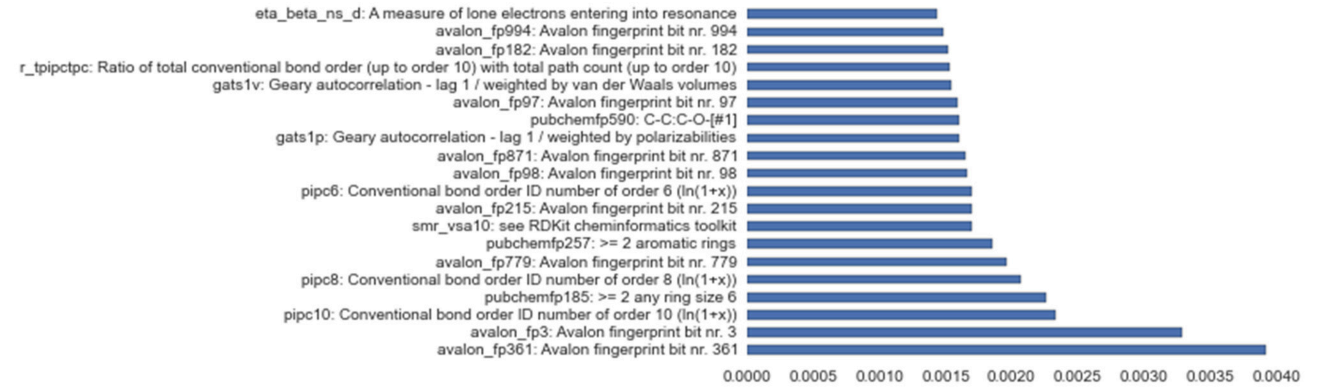
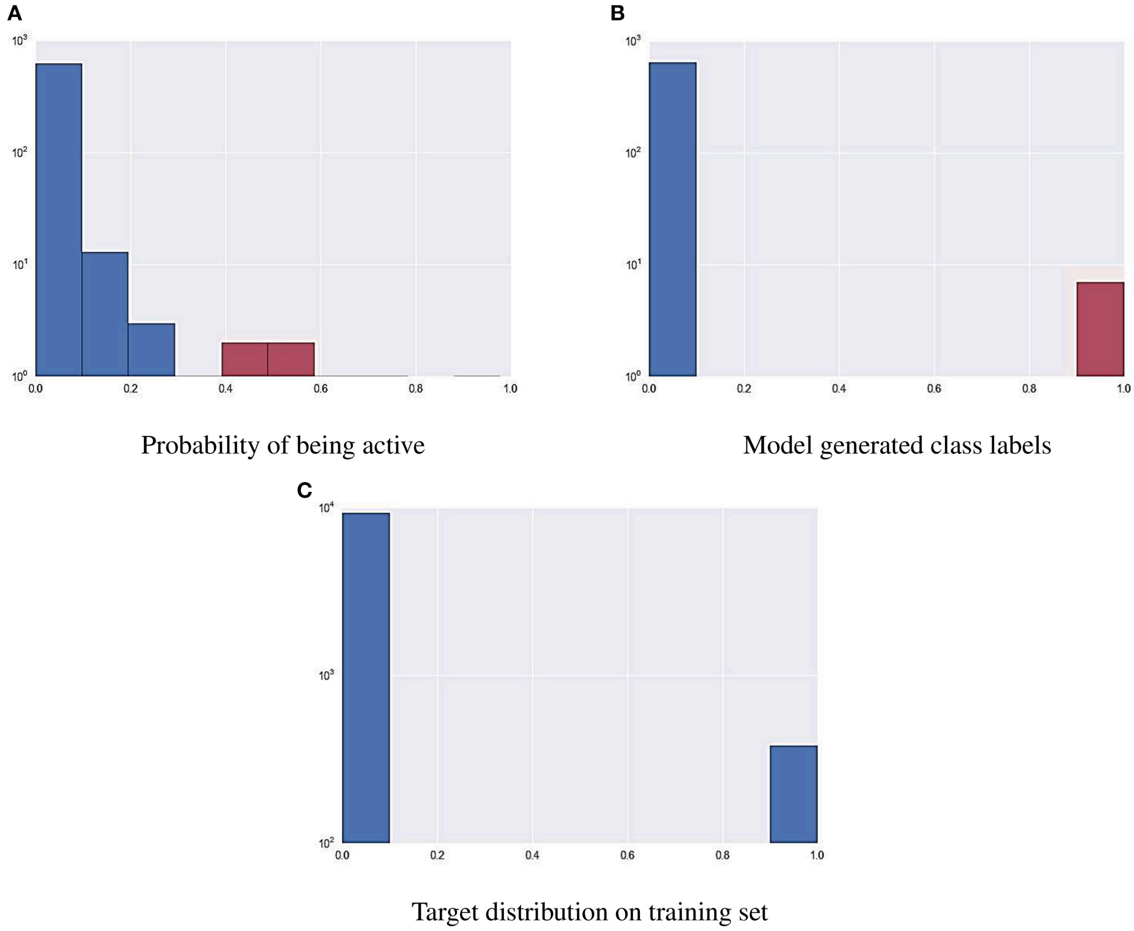


FIGURE 6 | Top 20 empirical feature importance assessment on assay SR-P53.



Probability of being active

Model generated class labels

Target distribution on training set

FIGURE 7 | Using a cutoff threshold on assay NR-AR to transform probabilistic predictions (A) to actual activity (B) that resembles training distribution (C).

in the article "Extremely randomized trees" in 2006 (Geurts et al., 2006). The extreme randomization comes from the fact that the variable splitting in each node is no longer based on finding the best split, but done in a completely random manner. This causes the trees grown to be even less data dependent, thus introducing extra variance between them.

TABLE 4 | Performance comparison of final solution and winning solution on all assays.

Panel	Assay	Modeling method	Cutoff point	Evaluation AUC	Position	Best AUC	Perf. ratio (%)
NR	AR	ExtraTreesClassifier	0.50	0.83	1	0.83	100
NR	Ahr	ExtraTreesClassifier	0.40	0.78	28	0.93	84.20
NR	AR-LBD	RandomForestClassifier	0.50	0.82	7	0.88	93.11
NR	ER	ExtraTreesClassifier	0.35	0.77	11	0.81	94.61
NR	ER-LBD	RandomForestClassifier	0.35	0.77	12	0.83	93.26
NR	Aromatase	ExtraTreesClassifier	0.45	0.84	1	0.84	100
NR	PPAR-gamma	ExtraTreesClassifier	0.50	0.83	6	0.86	96.58
SR	ARE	SupportVectorClassifier	0.60	0.77	10	0.84	91.43
SR	ATAD5	ExtraTreesClassifier	0.35	0.80	4	0.83	96.65
SR	HSE	ExtraTreesClassifier	0.50	0.86	7	0.86	98.93
SR	MMP	ExtraTreesClassifier	0.50	0.95	2	0.95	99.54
SR	p53	ExtraTreesClassifier	0.35	0.88	1	0.88	100

2.4. K-Fold Cross-Validation

Cross-validation is the primary method of model evaluation. In this technique, multiple models are trained using the same tuning parameters and subsequently tested on a different subset of data. The results are more reliable than performing the simple holdout method that could be misleading when a not-so-fortunate split is used.

During cross-validation the data is partitioned into K disjoint subsamples; typical K values lie between 5 and 10. Model training is then carried out using K-1 folds and testing on the last fold, as seen in **Figure 4**. The process is performed until all the folds have been used for testing and the cross-validation error equals

$$E_{CV} = \frac{1}{K}\sum_{i=1}^{K} E_i, \qquad (2)$$

where E_i is the error measured at each iteration. A 3-fold cross-validation scheme was used in the evaluation phase to ensure honest performance assessment. In general, local cross-validation scores were close to the leaderboard but slightly overestimated accuracy in some cases. K-fold cross-validation also ensured that the modeling has been executed using all data.

3. RESULTS

3.1. Model Implementation and Evaluation

The distribution of the target variable for all assays is highly skewed (target event between 3 and 16%). This causes difficulties for conventional modeling methods when it comes to predicting target values. Model alternatives were preselected based on their ability to handle the characteristics of the specific classification problem; having highly imbalanced target and a high dimensional feature space. Out of the many modeling methods Python's scikit-learn provides, the following were tested thoroughly:

1. Random Forest Classifier
2. Extra Trees Classifier
3. Gradient Boosting Classifier
4. Support Vector Classifier

Results clearly showed that neither Gradient boosting classifier (GBC) nor Support vector classifier (SVC) was able to handle target imbalance properly. Literature suggests balancing of target (Zakharov et al., 2014), which takes either substantial modification of the original method (Chen et al., 2004) or re-sampling of the whole dataset (Zhang et al., 2013). None of these advanced approaches was pursued in depth, as the random forest class was able to deliver convincing results in most cases without any further transformation needed. GBR and SVC approaches were subsequently discarded.

All models were evaluated using the K-fold cross-validation paradigm, using 3 folds to perform honest performance assessment. As the number of observations compared to the number of features is relatively low, this represents a crucial step in involving all observations in both the training and the testing phase. Model parameters were tuned using the grid search method; a combination of cross-validation and an exhaustive search in parameter space. Results were evaluated based on the area under the receiver operating characteristics curve score (ROC-AUC) as designated by the challenge organizers. **Table 2** shows the parameter settings that were tested.

The final evaluation models were trained on the combined training and testing dataset to encapsulate all the information available.

Table 3 shows the parameter settings found optimal for each track along with leaderboard and final evaluation performance. All solutions were ranked based on the challenge criteria: ROC-AUC, but balanced accuracy scores are also provided. As a sole exception, modeling for the SR-ARE assay was carried out completely in RapidMiner using Support vector classifier and is not discussed in this article. Any parameters not mentioned in **Table 3** were set to their respective default values (see the scikit-learn documentation for details). Although a wide spectrum of models was experimented with, all

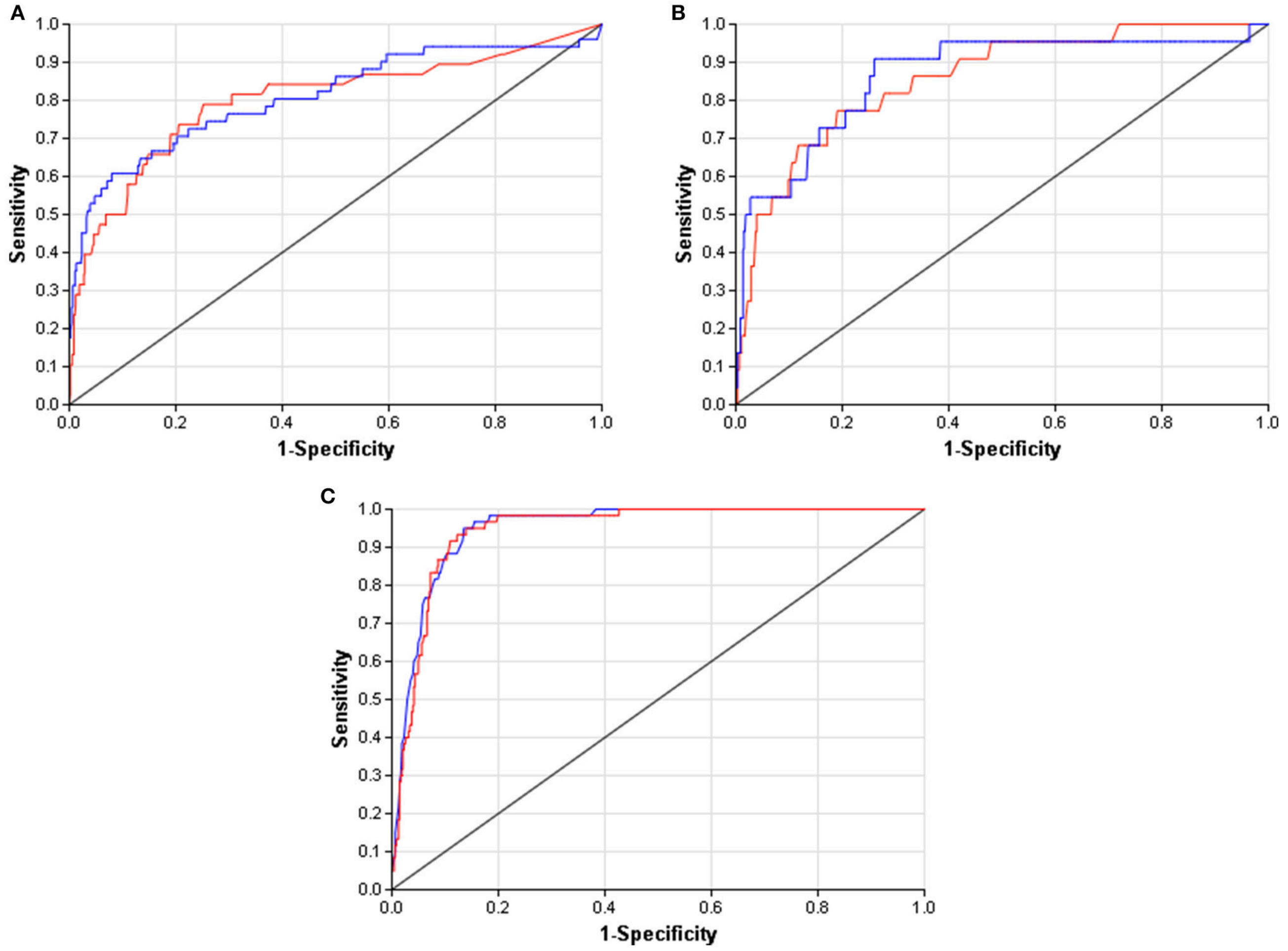


FIGURE 8 | ROC curves of best models (blue) and selected Dmlab models ranked 2–10th (red). (A) SR-ATAD5. **(B)** SR-MMP. **(C)** SR-HSE.

optimal solutions came from the Random forest class of models with very similar parameter settings, meaning this approach proved to be a relatively robust that worked well on all assays. The only real difference is the number of estimators employed, that varies in a broader spectrum from 499 to 999 depending on the assay. Performance discrepancy between leaderboard and final evaluation was also minimal, 4.73 ± 0.04 percent for NR and 4.62 ± 0.05 percent for SR panel.

Besides successfully avoiding overfitting and working reliably on all assays, the solution stack discussed in this paper also provides useful insights into variable importance, a feature crucial to the deeper understanding of complex problems like toxicity screening. **Figure 6** shows empirical feature importance assessment for the assay SR-P53, and underlines the significance of specific patterns to this problem, such as conventional bond orders and the presence of particular ring patterns. Further details on the most important input features for all 3 winning tracks are provided in the Supplementary Materials.

3.2. Post-Processing the Results

As part of the final evaluation task, molecule activity decisions had to be submitted instead of simple activity probabilities. As seen previously, the distribution of the target variable for all assays is highly skewed. This made cutting at the conventional 0.5 probability threshold impractical. The output of each model was further tuned to better represent the expected distribution of the target using a flexible cutoff point. **Figure 7** contains details of the process; do note the logarithmic scale on the figure. The optimal cutoff point per assay was calculated to closely resemble the target distribution observed on the training set published by the organizers. A strong assumption was made that evaluation and training data was sampled in a nearly stratified manner. The optimized cutoff point used in each assay can be found in **Table 4**.

4. DISCUSSION

The Tox21 Data Challenge offered a novel way of mass chemical assay classification. Much of our team's efforts were focused on developing accurate predictions with the help of well-established domain specific descriptors and finding the right approach to feature selection. Modeling was carried out using the cutting edge of open source data science tools available. This approach was highly capable of capturing toxicity driving factors while also avoiding overfitting on the training data. In the competition context, the proposed solution achieved a winning position in 3 of the Tox21 Data Challenge 2014 tracks and delivered highly comparable results on the rest.

The solution's robustness and competitiveness are proven through empirical results. The model evaluation shows empirical evidence that Random forest class predictors suit the particular classification problem well. When built on carefully preselected features they offer extremely high performance in the chemical assay classification domain. Model performance, however, greatly depends on the feature set used and the cutoff threshold applied; the proposed approach for both issues worked convincingly in 11 out of the 12 challenge tracks. The Random forest method is found to be insensitive to most modeling parameters; the number of estimators has a slight effect on performance, but overfitting is rarely an issue.

When compared to other challenger's solutions the Random forest stack offers convincing performance with 3 assay wins and 6 more places among the top 10. **Figure 5** shows the graphical representation of the winning solution performances. Even when the achieved ranking is not so prominent, ROC-AUC scores show a promising performance ratio compared to the assay winning solutions proving the approach's versatility (see **Table 4**). Average performance ratios were found to be 94.54 ± 4.99 percent for NR and 97.31 ± 3.16 percent for SR panel. **Figure 8** offers additional graphical comparison of performance ratios on selected assays SR-ATAD5 (96.65%), SR-MMP (99.54%), and SR-HSE (98.93%) respectively.

All computations were carried out on a quad core PC with Intel Core i5 CPU @ 3.20 GHz processor and 16 GB of RAM. Depending on the assay, single thread model building on the full dataset took between 28.3 and 42.2 s. Random Forests also possess the capability for multi-thread execution; using scikit-learn's parallelization feature reduces model building time between 9.6 and 13.7 s. Model application is generally quick; predictions are generated within seconds regardless of the data size.

In summary, the article provides a detailed description of the solution stack used to develop high accuracy QSAR models. This approach was able to achieve the highest accuracy in 3 different tracks of Tox21 Data Challenge. This accurate modeling approach also provides useful services, such as intrinsic feature importance that gives immediate feedback and further facilitates understanding the proposed toxicology screening method. The methodology used in the competition may be applied in other problems in cheminformatics as well. Furthermore, winning models are made publicly available for comparison and further research.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

The author would like to thank the reviewers for their valuable comments and various suggestions to improve the quality of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00052

To ensure reproducibility, all processes, scripts, configurations and miscellaneous supplementary materials for this article are publish online at https://github.com/themrbarti/tox21-challenge-publication.

REFERENCES

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). Pubchem: integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.* 4, 217–241. doi: 10.1016/S1574-1400(08)00012-1

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

Chen, C., Liaw, A., and Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*. Berkeley, CA: University of California.

Chong, I.-G., and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chem. Intell. Lab. Syst.* 78, 103–112. doi: 10.1016/j.chemolab.2004.12.011

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Gasteiger, J., and Marsili, M. (1980). Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* 36, 3219–3228. doi: 10.1016/0040-4020(80)80168-2

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

Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., et al. (2016). Tox21 Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.* 3:85. doi: 10.3389/fenvs.2015.00085

Judson, R., Elloumi, F., Setzer, R. W., Li, Z., and Shah, I. (2008). A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics* 9:241. doi: 10.1186/1471-2105-9-241

Krewski, D., Acosta, D. Jr., Andersen, M., Anderson, H., Bailar III, J. C., Boekelheide, K., et al. (2010). Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health B* 13, 51–138. doi: 10.1080/10937404.2010.483176

Liu, P., and Long, W. (2009). Current mathematical methods used in qsar/qspr studies. *Int. J. Mol. Sci.* 10:1978. doi: 10.3390/ijms10051978

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 24, 167–175. doi: 10.1038/nbt1186

Schowe, B. (2011). "Feature selection for high-dimensional data with rapidminer," in *Proceedings of the 2nd RapidMiner Community Meeting And Conference (RCOMM 2011)*, Aachen.

Shukla, S. J., Huang, R., Austin, C. P., and Xia, M. (2010). The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov. Today* 15, 997–1007. doi: 10.1016/j.drudis.2010.07.007

Si, H., Wang, T., Zhang, K., Duan, Y.-B., Yuan, S., Fu, A., et al. (2007). Quantitative structure activity relationship model for predicting the depletion percentage of skin allergic chemical substances of glutathione. *Anal. Chim. Acta* 591, 255–264. doi: 10.1016/j.aca.2007.03.070

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inform. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

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

Zakharov, A. V., Peach, M. L., Sitzmann, M., and Nicklaus, M. C. (2014). Qsar modeling of imbalanced high-throughput screening data in pubchem. *J. Chem. Inform. Model.* 54, 705–712. doi: 10.1021/ci400737s

Zhang, L., Fourches, D., Sedykh, A., Zhu, H., Golbraikh, A., Ekins, S., et al. (2013). Discovery of novel antimalarial compounds enabled by qsar-based virtual screening. *J. Chem. Inform. Model.* 53, 475–492. doi: 10.1021/ci300421n

Conflict of Interest Statement: The author declares 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 Environmental Science

ORIGINAL RESEARCH
published: 15 February 2016
doi: 10.3389/fenvs.2016.00009

Rigorous Selection of Random Forest Models for Identifying Compounds that Activate Toxicity-Related Pathways

Yoshihiro Uesawa *

Department of Clinical Pharmaceutics, Meiji Pharmaceutical University, Tokyo, Japan

Random forest (RF) is a machine-learning ensemble method with high predictive performance. Majority voting in RF uses the discrimination results in numerous decision trees produced from bootstrapping data. For the same dataset, the bootstrapping process yields different predictive capacities in each generation. As participants in the Toxicology in the Twenty-first Century (Tox21) DATA Challenge 2014, we produced numerous RF models for predicting the structures of compounds that can activate each toxicity-related pathway, and then selected the model with the highest predictive ability. Half of the compounds in the training dataset supplied by the competition organizer were allocated to the validation dataset. The remaining compounds were used in model construction. The charged and uncharged forms of each molecule were calculated using the molecular operating environment (MOE) software. Subsequently, the descriptors were computed using MOE, MarvinView, and Dragon. These combined methods yielded over 4,071 descriptors for model construction. Using these descriptors, pattern recognition analyses were performed by RF implemented in JMP Pro (a statistical software package). A hundred to two hundred RF models were generated for each pathway. The predictive performance of each model was tested against the validation dataset, and the best-performing model was selected. In the competition, the latter model selected a best-performing model from the 50% test set that best predicted the structures of compounds that activate the estrogen receptor ligand-binding domain (ER-LBD).

Keywords: random forest model, estrogen receptor ligand-binding domain, model-selection, Tox21 DATA Challenge 2014, pattern recognition

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Luis Gomez,
University of Las Palmas de Gran Canaria, Spain
Minjun Chen,
Food and Drug Administration, USA

****Correspondence:***
Yoshihiro Uesawa
uesawa@my-pharm.ac.jp

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *07 September 2015*
Accepted: *28 January 2016*
Published: *15 February 2016*

Citation:
Uesawa Y (2016) Rigorous Selection of Random Forest Models for Identifying Compounds that Activate Toxicity-Related Pathways. Front. Environ. Sci. 4:9. doi: 10.3389/fenvs.2016.00009

INTRODUCTION

The Toxicology in the Twenty-first Century (Tox21) challenge, launched in the United States in 2008, is the largest study of toxic substances to date (Shukla et al., 2010). The Tox21 project is promoted as a collaborative research among the National Institute of Health (NIH), Environmental Protection Agency (EPA), and Food and Drug Administration (FDA), and accords with the Memorandum of Understanding, which outlines the legal requirements of collaboration among U.S. public institutions (http://epa.gov/ncct/Tox21/; Ettlin, 2013; Tice et al., 2013). Tox21 is a far-reaching plan embracing the understanding of toxicities, establishment of evaluation systems, comprehensive experimental analyses, and constructions of prediction methods. Tox21 provides a chemical library of approximately 10,000 typical toxic compounds (the Tox21 10K library) as proper objects for toxicity evaluations (Hsieh et al., 2015). Activation and inhibition of various nuclear receptors such as the androgen receptor (AR), aryl hydrocarbon receptor (AhR), Ligand-binding domain of the androgen receptor (AR-LBD), estrogen receptor α (ER), ligand-binding domain of the estrogen receptor α (ER-LBD), aromatase and peroxisome proliferator-activated receptor γ (PPAR-γ), and stress response pathways such as nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (ARE), ATP-ase family AAA domain containing 5 (ATAD5), heat shock factor response element (HSE), mitochondrial membrane potential (MMP), and p53 were selected as investigation items in toxicity evaluation for the Tox21 10K library based on a published research for key regulatory pathways that integrate genetic and environmental modulators (Gohlke et al., 2009). One goal of Tox21 is to construct prediction models of toxicity-related biological activities from chemical structures. NIH's National Center for Advancing Translational Sciences (NCATS) organized the Tox21 Data Challenge 2014, in which participants compete in predicting biological toxic responses using computational toxicology technologies (https://tripod.nih.gov/tox21/challenge/index.jsp). Competitors predicted the existence of unpublished activities of various "evaluation set compounds" from the chemical structures and known activities of "training set compounds" extracted from the Tox21 10K library. The competition targets were the 12 proteins/pathways mentioned above, including PPARγ and p53 (Shukla et al., 2010). On a homepage designed for the challenge, the organizers uploaded the chemical structures of approximately 8,000 compounds in the Tox21 10K library as the target data, and their assay results of the proteins/pathways as the training set. As the evaluation set, they also offered the chemical structures of 647 compounds without the assay results. Each model submitted by the registered teams was ranked by its ability to predict the activities in the evaluation set. We registered the prediction results of 7 proteins/pathways based on random forest (RF) models (Breiman, 2001) with elaborate technologies, which yielded excellent prediction results. In particular, among the submitted models, our model best predicted the active compounds in the ER-LBD assay system (http://ncats.nih.gov/news/releases/2015/tox21-challenge-2014-winners). The RF method constructs many decision-trees and averages the predicted values to obtain the final ones. Trees are grown by bootstrapping samples of observations, and each split on each tree considers a randomly sampled descriptor. The current study describes the construction of our model and the characteristics of the prediction results in the Tox21 data challenge 2014.

MATERIALS AND METHODS

Conformations

The training dataset includes both chemical structures and activities; the evaluation dataset includes only the chemical structures. Both datasets were downloaded from the homepage set up for the challenge (https://tripod.nih.gov/tox21/challenge/index.jsp). The prediction model for discriminating between active and inactive compounds in the assays of each sub-challenge was constructed from the training dataset in the SDF files. The SDF files were changed to mdb file format by molecular operating environment (MOE) 2013.08 (Chemical Computing Group Inc., Quebec, Canada), which also cleaned up the chemical structures, removing smaller molecules such as counter-ions. Meanwhile, the charged and uncharged forms of each molecule were calculated using the protonation function in MOE. In other words, if there were chargeable functional groups in a chemical structure, both charged and uncharged forms were generated for that structure. Partial charges with force field (MMFF94x) parameters were allocated to the atoms in each molecule (Halgren, 1996). Next, the local-minimum 3D conformations of the charged and uncharged forms were computed by the MOE's *Rebuild3D minimization* function. The charged and uncharged forms might introduce different descriptors with different 3D/topological conformations and counts of functional groups. Therefore, acquisition of the descriptors should provide detailed molecular information on each chemical structure.

Descriptors

A variety of descriptors were computed by 3 software packages: MOE, MarvinView 5.12.4 (ChemAxon Kft., Budapest, Hungary), and Dragon 6 (Talete srl., Milano, Italy). Excluding the overlapping descriptors, 4071 descriptors were selected for constructing the prediction models. These descriptors are summarized in **Supplemental Table 1**. However, some of the descriptors could not be calculated in lithium, which was included in the evaluation datasets. Because the structural and physicochemical properties in lithium were dissimilar to all other compounds in the datasets, the activity of lithium was evaluated with the lowest probability among the compounds in the evaluation datasets.

Datasets

In seven of the sub-challenge targets (ER, ARE, p53, PPAR-γ, ATAD5, ER-LBD, and AhR), half of the compounds in each training dataset were randomly selected as the validation dataset (50% test set). The remaining compounds in the training datasets (50% training set) were used to construct the prediction models. That is, the 50% test set was used to externally validate the constructed models but not the internal data generating the RF resampling processes. The evaluation process is outlined in **Figure 1**. The evaluation set in **Figure 1** is the final evaluation set prepared by the competition organizer. This set included 647 chemical structures without their assay results during the period of the competition.

Pattern Recognition and Rigorous Selection

Pattern recognition analyses of these descriptors were performed by the RF method (Breiman, 2001) using the *bootstrap-forest* function in the statistical software package JMP Pro 10.02 (SAS Institute Inc., Cary, NC, USA).

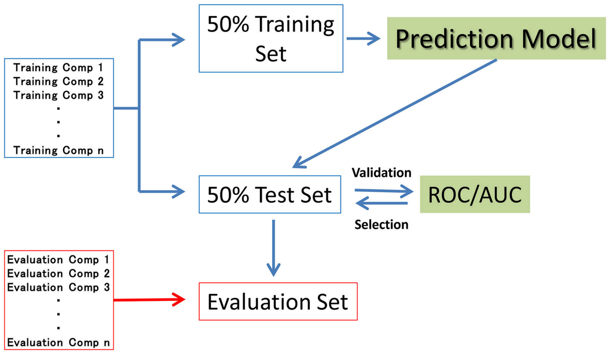


FIGURE 1 | Datasets used in model construction.

Because they perform bootstrapping, RF models have different predictive performances for the same combination of hyperparameters. The contribution of a combination can be checked by comparing the performance of that combination with the average performance of plural models. Therefore, to select the best combination of hyperparameters, we estimated at least 10 models with each combination. The hyperparameter decision step was based on data sets with ER-LBD. Specifically, we estimated the discrimination abilities of the models for predicting compound activities by the area under the receiver operating characteristic curve (ROC–AUC), which scores the probability of activity of the compound. The predictive performance of each model contributed from 50% of the ER-LBD training set was evaluated on the 50% test set in the assay result. Based on the investigations, the hyperparameters were set as follows: *Number of Trees*, specifying the number of trees to grow before averaging (100), *Number of Terms*, denoting the number of columns specified as predictors (1032), *Bootstrap Sample Rate*, specifying the proportion of observations sampled in each tree growth (1), *Maximum Splits Per Tree*, defining the minimum number of splits in each tree (2000), and *Minimum Size Split*, defining the minimum number of observations required for candidate splitting (2). The same hyperparameters were applied in all predictions of all targets. Using combinations of these hyperparameters, from 100 to 200 RF models were generated for each target (the numbers of the models, decided a priori, were 192, 100, 150, 109, 151, 200, and 132 in ER, ARE, p53, PPAR-γ, ATAD5, ER-LBD, and AhR, respectively), and their predictive performances were evaluated on the 50% test set in each sub-challenge. After the competition, the assay results of the evaluation set were available for viewing. Therefore, we could compute the ROC-AUC values in the evaluation set and the prediction results of the numerous RF models for ER-LBD constructed by the above method. The modeling process was validated in ROC-AUC comparisons of the 50% training set and 50% test set (**Figure 2**). In this validation, prediction values in the 50% training set and 50% test set as well as in the evaluation set were recalculated for the different values of the hyperparameters, Number of Terms (1–1000), and Maximum Splits Per Tree (2–400) during the construction of 190 models.

Computational Environment

All simulations were performed on a desktop personal computer: Endeavor MR7200-M (Epson Direct Corporation, Nagano, Japan) with Windows 7 sp1 (64 bit), an Intel® Core™i7-4790 CPU (3.6 GHz), and 32.0 GB RAM.

RESULTS

Rigorous Selection

Figure 2 presents the relationships among the ROC-AUC values of ER-LBD obtained by the RF models in the 50% training set, 50% test set, and evaluation set for different values of the hyperparameters *Number of Terms* (1–1000) and *Maximum Splits Per Tree* (2–400). The AUC values in the 50% training set were well-correlated with those in the 50% test set. However, the AUC values in the 50% test set and the final evaluation set were not simply correlated; rather, there was an optimal point at which the AUC of the 50% test set corresponded to the highest AUC of the evaluation model.

Targets

The RF models described in the Methods section were constructed for 7 targets; namely, ER, ARE, p53, PPAR-γ, ATAD5, ER-LBD, and AhR. Using these models, we predicted the activities of the compounds included in the final evaluation set of the targets. The model performances were evaluated by their ROC–AUC values and their rankings in the Tox21 data challenge 2014 (**Table 1**; https://tripod.nih.gov/tox21/challenge/index.jsp). In the competition, 125 participants were registered from 18 countries, and finally, 40 teams from 11 countries submitted prediction models (Huang et al., 2015). Most of our models were within the top 10 of the registered sub-challenges. In particular, our models achieved the highest ROC–AUC in the ER-LBD sub-challenge. The estimated scores of the compounds in the evaluation set of each target are plotted against the assay results in **Figure 3**.

DISCUSSION

In constructing the prediction models, we considered the various factors discussed below.

Charged and Uncharged Forms

For chemical structures with chargeable functional group(s), we generated both the charged form under neutral pH conditions and the uncharged form. The structures of these forms should differ in their numbers of functional groups and optimized 3D-conformations. Therefore, by constructing both forms, we can extract more structural information from each molecule because of the greater variety of generated descriptors. Actually, a previous investigation confirmed that including the descriptors from both forms improved the predictive ability of the RF models, relative to descriptors from unilateral forms (data not shown).

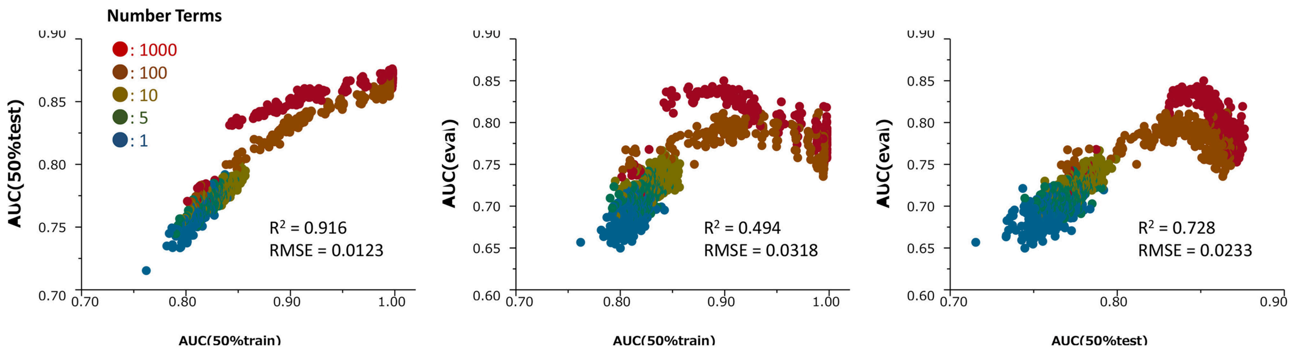


FIGURE 2 | Relationships among the ROC-AUC values of numerous prediction models in the ER-LBD test, training and evaluation datasets for different values of the hyperparameter *Number of Terms* (1–1000) and *Maximum Splits Per Tree* (2–400) during the construction of 190 models. R^2 and RMSE means determination coefficient and root mean squared error, respectively.

TABLE 1 | ROC–AUC values of the RF models.

Target assay	Category	ROC-AUC	Best ROC-AUC	Ranking
ER-LBD	Nuclear receptor	0.827	0.827	1
ATAD5	Stress response pathway	0.812	0.828	3
ARE	Stress response pathway	0.802	0.84	7
ER	Nuclear receptor	0.783	0.81	7
AhR	Nuclear receptor	0.901	0.928	8
p53	Stress response pathway	0.826	0.88	10
PPAR-gamma	Nuclear receptor	0.718	0.861	15

ROC–AUC values are the results of the author's study. Best ROC-AUC denotes the best result in the Tox21 data challenge 2014. Ranking refers to the authors' results.

Usage of Numerous Descriptors

To maximize the information on the chemical structures, we calculated numerous descriptors including structural and physicochemical features using the MOE, Marvin, and Dragon software packages. Approximately 10,000 descriptors were generated from the above-described charged and uncharged forms. After discarding the overlapping descriptors, we obtained 4,000 descriptors for the modeling studies.

RF Model

The model construction was based on the RF algorithm (Breiman, 2001), which has advantage of high prediction potency, low computational cost, handling of large and prejudiced data, and resistance to effects containing outliers (Bruce et al., 2007). On account of its high cost performance, model construction by RF can proceed in standard computing environments. Because it ranks the contributions of the model descriptors, the RF algorithm can also estimate the importance of physicochemical features of the compounds during interactions with biopolymers such as proteins. For instance, the importance of the descriptors in constructing the ER-LBD model was estimated from the entropic changes and the descriptor-usage numbers at the split points. In this model, the most significant descriptors were the number of aromatic hydroxyls (nArOH) and the smallest eigenvalue (n.1) of the mass-weighted Burden matrix (SpMin1_Bh_m) (sup. **Table 1**) (Burden, 1989; Todeschini and Consonni, 2009). Such characteristics of the RF algorithm might provide useful knowledge for understanding the interactions.

Rigorous Model Selection

A high-performance model was selected among numerous RF models. When constructing the RF models, we performed bootstrap sampling of compound–descriptor combinations. Because the bootstrap process is repeated randomly, each constructed model has a unique predictive ability. For example, the average, highest and lowest ROC-AUC values among numerous ER-LBD prediction models in the trial shown in **Figure 2** were respectively 0.869, 0.998, and 0.762 in the 50% training set, 0.802, 0.876, and 0.715 in the 50% test set, and 0.743, 0.850, and 0.650 in the evaluation set. Selection of the submitted models was based on their prediction potentials in the 50% test set. The competition rules allowed free selection of the target pathways for submission. Our 7 targets were selected from the 12 pathways for no special reason. Although we did not attempt to predict the active compounds in the remaining five pathways, we expect that our method would predict these with equal efficacy.

Furthermore, our rigorous model selection strategy successfully predicted the ATAD5, ARE, ER, AhR, and p53 pathways. The ROC–AUC values obtained for

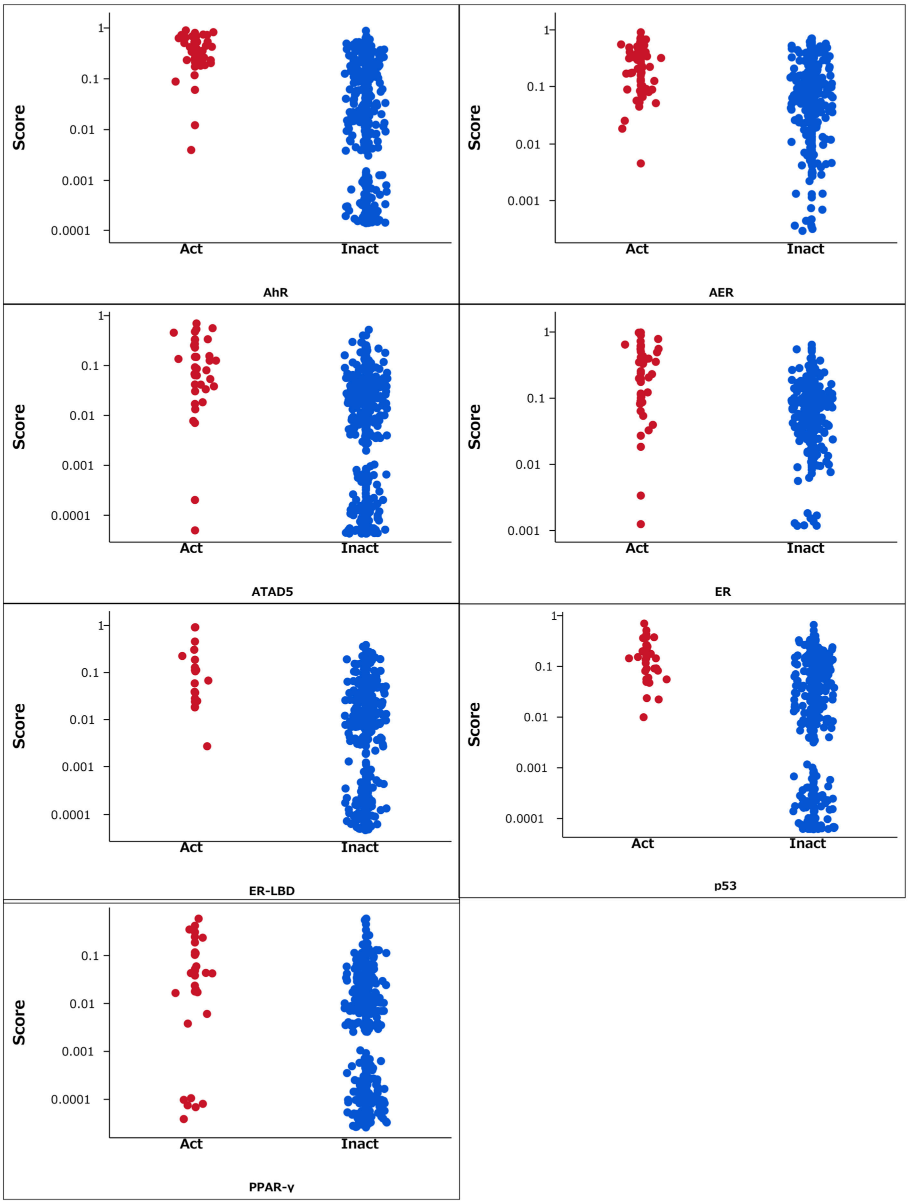


FIGURE 3 | Estimated scores of compounds in the evaluation set. Act and Inact denotes the active and inactive results in each assay.

these pathways ranked among the top 10 registered by competitors (**Table 1**). In contrast, the predictive performance of PPAR-γ (ROC–AUC = 0.718) was ranked at 15. Such varying performance for different targets might reflect varying compatibility with the hyperparameter combination in the RF modeling. The hyperparameter combination was determined from the training data for ER-LBD, and might be markedly suboptimal for PPAR-γ prediction.

After the competition, the assay results of the compounds in the evaluation set were opened for viewing, so we could validate the current construction method of the RF models. Scatter plots between the ROC-AUC values in the 50% test set and evaluation set revealed a strong correlation between the 50% test set and the evaluation dataset, confirming the prediction potential of the models for this dataset (**Figure 2**). This result supports the strategy of selecting the best model from numerous RF models, based on their ROC-AUC values of the 50% test set. However, the optimal combinations of hyperparameters for predicting the evaluation set and the 50% test set were non-identical. The best-performing models for the test set may be overfitted. To improve the model selection, we should use the current results to refine the hyperparameter combinations.

CONCLUSIONS

We have constructed a high-performance single RF model for biological pathway prediction. The method succeeded by rigorously selecting the best model among numerous previous models. Each of the previous models has a unique performance because of the bootstrap data sampling used in model construction. Increasing the number of previous models and refining the hyperparameter combinations might improve the final model. In other words, the generalizability of the prediction models can be influenced by the number of generated RF models, which depends on the performance of the computational environment.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

The author is grateful to Mr. Yuki Asako, Department of Clinical Pharmaceutics, Meiji Pharmaceutical University for skillful technical assistance. This work was partially supported by KAKENHI from the Japan Society for the Promotion of Science (JSPS) (15K08111 and 25460227). The annual license of the statistical software, JMP Pro, was supported by the grant-in-aid of the oncology specialists' promotion program by the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00009

Supplemental Table 1 | Descriptors.

REFERENCES

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

Bruce, C. L., Melville, J. L., Pickett, S. D., and Hirst, J. D. (2007). Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* 47, 219–227. doi: 10.1021/ci600332j

Burden, F. R. (1989). Molecular identification number for substructure searches. *J. Chem. Inf. Comp. Sci.* 29, 225–227.

Ettlin, R. A. (2013). Toxicologic pathology in the 21st century. *Toxicol. Pathol.* 41, 689–708. doi: 10.1177/0192623312466192

Gohlke, J. M., Thomas, R., Zhang, Y., Rosenstein, M. C., Davis, A. P., Murphy, C., et al. (2009). Genetic and environmental pathways to complex diseases. *BMC Syst. Biol.* 3:46. doi: 10.1186/1752-0509-3-46

Halgren, T. A. (1996). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* 17, 490–519.

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

Huang, R., Xia, M., Nguyen, D., Zhao, T., Sakamuru, S., Zhao, J., et al. (2015). Tox21 Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.* 3:85. doi: 10.3389/fenvs.2015.00085

Shukla, S. J., Huang, R., Austin, C. P., and Xia, M. (2010). The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov. Today* 15, 997–1007. doi: 10.1016/j.drudis.2010.07.007

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* 121, 756–765. doi: 10.1289/ehp.1205784

Todeschini, R., and Consonni, V. (2009). "Molecular descriptors for chemoinformatics," in *Methods and Principles in Medicinal Chemistry*, Vol. 41, eds R. Mannhold, H. Kubinyi, and G. Folkers (Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA), 886–925.

Conflict of Interest Statement: The author declares 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 Environmental Science

ORIGINAL RESEARCH
published: 30 July 2015
doi: 10.3389/fenvs.2015.00054

Molecular similarity-based predictions of the Tox21 screening outcome

Malgorzata N. Drwal[1], *Vishal B. Siramshetty*[1], *Priyanka Banerjee*[1,2], *Andrean Goede*[1], *Robert Preissner*[1,3] *and Mathias Dunkel*[1]*

[1] Structural Bioinformatics Group, Institute for Physiology, Charité – University Medicine Berlin, Berlin, Germany, [2] Graduate School of Computational Systems Biology, Humboldt-Universität zu Berlin, Berlin, Germany, [3] BB3R – Berlin Brandenburg 3R Graduate School, Freie Universität Berlin, Berlin, Germany

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Luis Gomez,
University of Las Palmas de Gran Canaria, Spain
Ijaz Hussain,
Quaid-i-Azam University, Islamabad, Pakistan

***Correspondence:**
Mathias Dunkel,
Structural Bioinformatics Group,
Institute for Physiology, Charité –
University Medicine Berlin,
Lindenberger Weg 80,
13125 Berlin, Germany
mathias.dunkel@charite.de

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: 04 May 2015
Accepted: 14 July 2015
Published: 30 July 2015

Citation:
Drwal MN, Siramshetty VB, Banerjee P, Goede A, Preissner R and Dunkel M (2015) Molecular similarity-based predictions of the Tox21 screening outcome. Front. Environ. Sci. 3:54. doi: 10.3389/fenvs.2015.00054

To assess the toxicity of new chemicals and drugs, regulatory agencies require *in vivo* testing for many toxic endpoints, resulting in millions of animal experiments conducted each year. However, following the Replace, Reduce, Refine (3R) principle, the development and optimization of alternative methods, in particular *in silico* methods, has been put into focus in the recent years. It is generally acknowledged that the more complex a toxic endpoint, the more difficult it is to model. Therefore, computational toxicology is shifting from modeling general and complex endpoints to the investigation and modeling of pathways of toxicity and the underlying molecular effects. The U.S. Toxicology in the twenty-first century (Tox21) initiative has screened a large library of compounds, including approximately 10K environmental chemicals and drugs, for different mechanisms responsible for eliciting toxic effects, and made the results publicly available. Through the Tox21 Data Challenge, the consortium has established a platform for computational toxicologists to develop and validate their predictive models. Here, we present a fast and successful method for the prediction of different outcomes of the nuclear receptor and stress response pathway screening from the Tox21 Data Challenge 2014. The method is based on the combination of molecular similarity calculations and a naïve Bayes machine learning algorithm and has been implemented as a KNIME pipeline. Molecules are represented as binary vectors consisting of a concatenation of common two-dimensional molecular fingerprint types with topological compound properties. The prediction method has been optimized individually for each modeled target and evaluated in a cross-validation as well as with the independent Tox21 validation set. Our results show that the method can achieve good prediction accuracies and rank among the top algorithms submitted to the prediction challenge, indicating its broad applicability in toxicity prediction.

Keywords: molecular fingerprints, molecular similarity, machine learning, toxicity prediction, Tox21 Data Challenge 2014

Introduction

The U.S. Toxicology in the twenty-first century (Tox21) initiative has been established in 2008 with the vision to support the transformation of toxicology into a predictive science (Krewski et al., 2010). In order to achieve this goal, a large library of compounds, including approximately 10K environmental chemicals and drugs, was screened for different mechanisms responsible for eliciting toxic effects. Among the screens were high-throughput assays for two important pathways, the nuclear receptor and the stress response pathway, which were the subject of the Tox21 Data Challenge 2014.

Interactions of chemicals with nuclear receptors represent a major health concern. In particular, binding of chemicals to steroid receptors can cause the disruption of the normal endocrine function and have an adverse effect on development, reproduction and metabolic homeostasis (Huang et al., 2014). A famous example of an endocrine disrupting chemical is bisphenol A, a compound which has been widely used, e.g. in plastic bottles and metal cans, but has only recently been associated with impairments of neurobehavioral development (Weiss, 2012). Bisphenol A and its derivatives have been shown to exhibit a promiscuous binding behavior involving, for instance, estrogen receptors (ER), androgen receptors (AR) and peroxisome proliferator-activated receptors (PPAR) of the γ subtype (Delfosse et al., 2014), all of which are subject of the Tox21 screening. Another current focus of the Tox21 screening is aromatase, an enzyme involved in the conversion of androgen to estrogen and therefore a target of endocrine disrupting chemicals (Chen et al., 2014), as well as the aryl hydrocarbon receptor (AhR), a nuclear receptor involved in the mediation of tumorgenesis induced by dioxin (Murray et al., 2014). Similarly, mechanisms related to cellular stress also play a role in toxicological pathways. For example, recent studies have shown that the impairment of mitochondrial function is associated with drug-induced adverse effects on the liver and cardiovascular system (Nadanaciva and Will, 2011; Attene-Ramos et al., 2015).

To assess the risks of new chemical entities, *in vivo* animal studies are required by regulatory agencies to evaluate various toxicological endpoints. However, *in silico* toxicology is gaining acceptance as an alternative method which can help to reduce the number of animal experiments performed. Computational predictions often rely on the observation or assumption that similar molecules manifest a similar biological effect. Similarity-based methods have been successfully applied to solve various research questions including predictions of targets (Campillos et al., 2008), therapeutic indications (Nickel et al., 2014) or side-effects (Lounkine et al., 2012). In particular, machine learning approaches such as k-nearest neighbors, naïve Bayes models, support vector machines, random forests or ensembles of different classification methods can use the similarity defined the molecular structure and properties to make predictions for novel compounds. This concept has also been frequently and successfully applied to predictions of various toxicological endpoints (Drwal et al., 2014; Gadaleta et al., 2014; Li et al., 2014; Liu et al., 2015).

Here, we describe the development of a fast and successful method for the prediction of different outcomes of the nuclear receptor and stress response pathway screening from the Tox21 Data Challenge 2014. The method is based on the combination of a simple molecular similarity calculation with a naïve Bayes machine learning algorithm. Three different two-dimensional (2D) molecular representation methods as well as their combination were compared and the prediction methods were optimized individually for every target. The evaluation of each model showed that all models can achieve good performance and prediction accuracies as well as rank among the top submissions among the Tox21 challenge participants.

Materials and Methods

Overview

An overview of the workflow used in this study is given in **Figure 1**. In the first step, all molecular structures were standardized and the duplicates as well as compounds with ambiguous activity values were removed. The training and test set provided by the Tox21 Data Challenge 2015 organizers were merged and used in a 13-fold cross-validation to optimize parameters for the classification algorithms. The optimized models were then used to predict the activities of the evaluation set compounds. All steps are described in detail in the following sections. For the majority of tasks, the open pipeline generation platform KNIME v.2.10.0 (Knime.com AG) was used.

Data Preparation

Standardization

All molecular structures were downloaded from the Tox21 Data Challenge 2014 website (https://tripod.nih.gov/tox21/challenge/index.jsp) and their molecular structures were standardized using the Instant JChem software (version 6.2, Chemaxon) with the following settings: Water molecules were removed, molecules were aromatized, adjacent positive and negative charges transformed into double/triple bonds, explicit hydrogens were added and the 3D conformation was generated and cleaned. After the standardization, InChIKeys were calculated using RDKit (http://www.rdkit.org) nodes in KNIME in order to identify and remove duplicates. In case duplicate molecules were found to have different activities (1 and 0) for a particular target, they were marked as ambiguous and removed from the training set of this target.

Additional Data

For each target, a search for additional known ligands was performed in the ChEMBL bioactivity database v.19 (Bento et al., 2014). A search was performed for the target name and EC_{50} or IC_{50} values in case of agonists or antagonists, respectively.

Abbreviations: 2D, two-dimensional; AhR, aryl hydrocarbon receptor; AR, androgen receptor; ARE, antioxidant response element; ATAD5, genotoxicity induction; AUC, area under the curve; BAC, balanced accuracy; ER, estrogen receptor 1; HSE, heat shock response; LBD, ligand binding domain; MMP, mitochondrial membrane potential; PPAR, peroxisome proliferator-activated receptor; ROC, receiver operating characteristic; Tox21, U.S. Toxicology in the twenty-first century initiative.

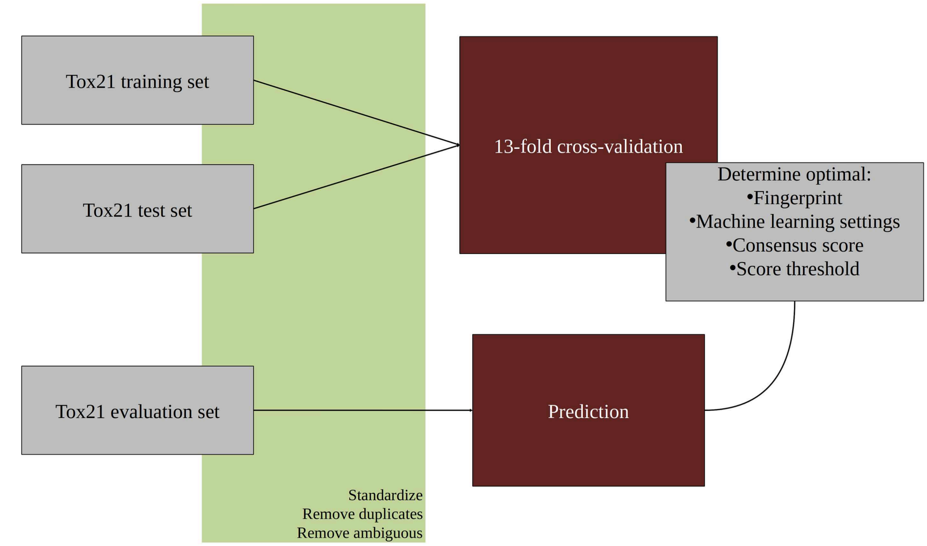


FIGURE 1 | Workflow overview.

Additional datasets were standardized and checked for duplicates as described above.

Calculation and Combination of Fingerprints

Different types of molecular representations were calculated for each compound: ToxPrint fingerprints were calculated using the ChemoTyper software (version 1.0, Molecular Networks GmbH). Extended-connectivity fingerprints (Rogers and Hahn, 2010) of the ECFP4 type were calculated using RDKit nodes in KNIME. 960-bit MACCS keys were calculated using the Discovery Studio 3.1 program (Accelrys Inc./BIOVIA). In addition, several topological properties indicating the three-dimensional (3D) structure were calculated using RDKit and CDK nodes in KNIME. The use of topological descriptors has been previously reported in a structure-toxicity relationship study (Pasha et al., 2009). Furthermore, topological descriptors have several advantages compared to 3D descriptors, including conformational independency, simplicity and low computational resources. A number of topological descriptors were calculated, but only those displaying values with considerable difference between active and inactive molecules were used further. These included the Chi0V, Chi1N, Kappa1 and HallKierAlpha descriptors (Hall and Kier, 1991) as well as the topological polar surface area. The descriptors were transformed into a binary vector by binning. For each descriptor, a number of "bins" (and bits in the fingerprint) was defined, representing different descriptor value ranges. Whenever the descriptor value was found in a specific range, the bit at the respective position was set to 1. Therefore, it was ensured that close values exhibited high fingerprint similarity. The combined fingerprint consisted of a concatenation of all four binary fingerprints with a length of 2929 bits—960 bits for MACCS keys, 1024 bits for ECFP4, 729 bits for ToxPrint and 216 bits for the property-based fingerprint, as indicated in **Figure 2**.

Toxicity Prediction Methods

Cross-validation

In order to validate the prediction models, a 13-fold cross-validation was implemented in KNIME. The KNIME workflows are presented in Supplementary Figures S1, S2. A 13-fold validation was chosen in order to produce a test set similar in size to the final validation set of the Tox21 challenge. It was investigated whether the addition of external data (known ligands from the ChEMBL database, see Section Additional data) was able to improve the prediction rate. Different activity cut-offs for the ChEMBL compounds were considered for this purpose. Furthermore, it was also investigated whether reducing the actives in the training set to the most diverse compounds was able to increase the performance of the model. In this case, the RDKit Diversity Picker node was used using different thresholds. Finally, the effect of the removal of highly correlated fingerprint bits on the model performance was explored using the Correlation Filter node. To determine the best settings, the performance was evaluated using a receiver operating characteristic (ROC) analysis. The area under the curve (AUC) was calculated using the ROC curve node.

Naïve Bayes Learning

Naïve Bayes is a commonly applied stochastic classifier based on the Bayes theorem of conditional probability (Nidhi et al., 2006). The major characteristic of the classifier is the naïve assumption that all input features are independent. Main advantages of the method compared to other machine learning algorithms are fast computational time during training and prediction as well as a low parameter complexity and insusceptibility to irrelevant features. Furthermore, it has been suggested that the combination of molecular fingerprints with descriptors can be beneficial in the context of Bayesian modeling (Vogt and Bajorath, 2008).

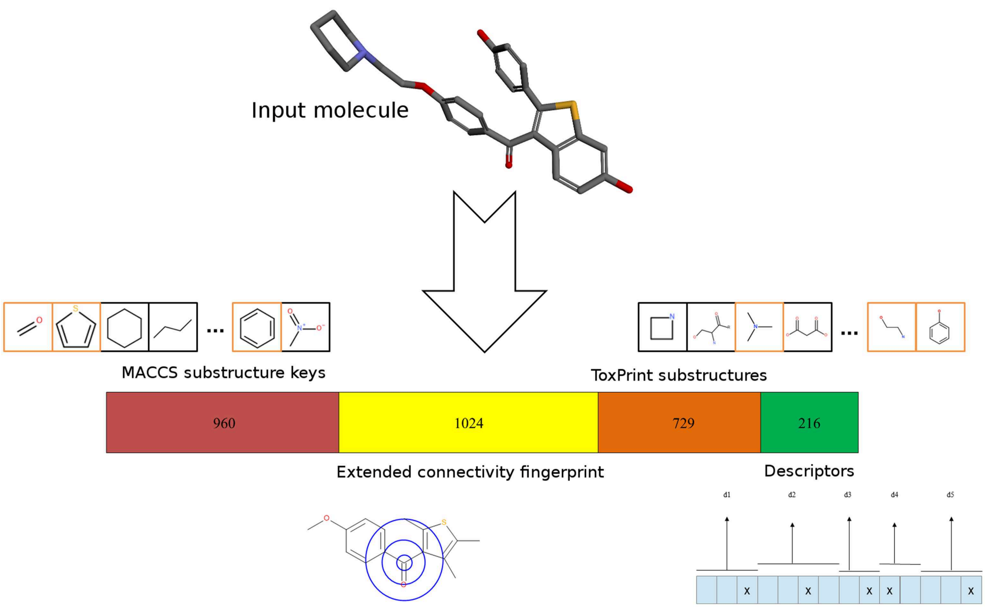


FIGURE 2 | Molecular representation. For every input molecule from the Tox21 data set, different 2D-fingerprints are calculated and combined. The concatenation consists of MACCS keys (960 bits), the extended-connectivity fingerprint ECFP4 (1024 bits), ToxPrint (729 bits) and a fingerprint developed from topological descriptors (216 bits). Both MACCS as well as ToxPrint fingerprints encode the presence of specific substructures. Examples of MACCS and ToxPrint substructures are shown in boxes. Substructures present in a sample molecule taken from the Tox21 dataset are highlighted in orange boxes. ECFP4 encodes the connections of each atom within a 4-atom radius. The property-fingerprint encodes the presence of descriptor values in specific bins representing value ranges.

Thus, we implemented a naïve Bayes predictor with the Tox21 training sets. The Fingerprint Bayesian Learner and Predictor nodes in KNIME were used for this purpose. The predictor received an input of active and inactive molecules and their fingerprints. The output consisted of two scores for each molecule, a score for being active (B_1) and a score for being inactive (B_0).

Molecular Similarity

The Tanimoto index is one of the most common metrics for fingerprint-based molecular similarity calculations and has recently been shown to be among the best choices for this purpose (Bajusz et al., 2015). For the comparison of molecular similarity, three Tanimoto coefficients were computed: the maximum Tanimoto coefficient to actives in the training set (T_1), the average Tanimoto coefficient to actives in the training set (T_2), and the maximum Tanimoto coefficient to all inactives in the training set (T_3).

Combination of Methods

All scores and Tanimoto coefficients were normalized in KNIME using *Z*-score normalization to obtain scores following a Gaussian distribution and MinMax-normalization to obtain values between 0 and 1. Different combinations of the naïve Bayes scores B_1 and (1-B_0) as well as the Tanimoto scores T_1, T_2 and (1-T_3) were examined, including the minimum, maximum and mean of the scores.

Determination of Score Threshold

For every target, a threshold of the final score was determined which was used to classify the compounds into active and inactive molecules. The score threshold was determined by choosing the threshold which resulted in the maximal balanced accuracy ((sensitivity+specificity)/2) over all rounds of cross-validation.

Results

The Tox21 Data Challenge 2014 consisted of the prediction of 12 different screening outcomes (*targets*): the activation or inhibition of nuclear receptors AhR, PPARγ, aromatase, ER and AR (full length and ligand binding domain, LBD) as well as the effect on stress response pathways consisting of the activation of the antioxidant response element (ARE), heat shock response (HSE) and p53 signaling, the disruption of mitochondrial membrane potential (MMP) and the induction of genotoxicity (ATAD5). Before building predictive models, all chemical structures were normalized as described in the Methods section and duplicates were removed. Only compounds explicitly marked as active or inactive were used for model development. Wherever available, additional active molecules were extracted from the ChEMBL database (Bento et al., 2014) and used for model development. As summarized in Supplementary Table S1, the proportion of unique active and inactive molecules as well as the presence of external actives differed considerably between targets.

Choice of Molecular Representation

How well a prediction model performs does not only depend on the underlying algorithm, but also the features used as input. In the case of predictions of small molecule toxicities and other biological activities, the performance thus depends on the molecular representation which ultimately influences the computed similarity between molecules (Floris et al., 2014). Here we compared the performance of three common molecular fingerprints as well as their combination. ECFP4 is a member of the extended-connectivity fingerprint type often used to analyze structure-activity relationships of small molecules (Rogers and Hahn, 2010). MACCS keys are another frequently used fingerprint type which encodes the presence of specific substructures and has been successfully used for predictions of acute oral toxicity (Li et al., 2014). The ToxPrint fingerprint (Yang et al., 2015a) is based on a library of more than 700 chemotypes which represent molecules in public chemical and toxicity databases and cover substructures associated with toxic effects and thus may be of particular importance for *in silico* toxicity predictions. We also evaluated the addition of a property-based fingerprint as has been suggested previously (Xue et al., 2003). Here, descriptors encoding the topology of the Tox21 compounds were calculated and translated into a binary fingerprint.

In order to determine the optimal fingerprint for the prediction, fingerprints were used individually as well as in combination and evaluated in cross-validation on one of the targets, namely ER-LBD. As summarized in **Table 1**, all three types of fingerprints showed a good performance using both the Bayesian classifier as well as the similarity search approach. In the majority of cases models built with individual fingerprints exhibited AUC values above 0.75 and a concatenation of all three fingerprints led to a slight increase in performance. Furthermore, a combination of the concatenated fingerprints with a property-based fingerprint encoding the topology of the molecules demonstrated the best prediction results and was thus used as a descriptor for all targets of the challenge.

Model Optimization and Validation

In the preliminary evaluation of descriptors for ER-LBD, a common observation was that a consensus score consisting of a machine learning score and a similarity coefficient usually resulted in the best model performance (**Table 1**). Therefore, it was investigated which combination of scores led to the best prediction. In particular, the scores from the Bayesian classifier and the similarity search were combined into a consensus score using either a mean, maximum or minimum value. Since the optimal settings might differ depending on the target and its active and inactive molecules, the best parameters were determined individually for every target in a cross-validation study. The optimization involved the variation of the following parameters: the addition of active molecules from external sources (ChEMBL database) using different activity value thresholds, the addition of a correlation filter to remove highly correlated fingerprint features as well as the incorporation of a diversity picker to restrict the number of active to train a naïve Bayes model to the ones with highest diversity.

The best settings found for every Tox21 target are shown in **Table 2**. As indicated, similarity search gave the best performance for 4/12 targets when an average Tanimoto was calculated from the T_1, T_2, and $(1\text{-}T_3)$ scores indicating the similarity to active as well as the dissimilarity to inactive molecules (see Methods). For all other targets, a combination of the machine learning algorithm and a similarity scoring showed the best results. In most cases, a mean function was used to generate a consensus score combining the naïve Bayes and Tanimoto coefficients.

The performance of each model was evaluated using ROC-AUC values as well as balanced accuracies. The cross-validation results for the best settings as well as the external validation results provided by the challenge organizers are summarized in **Figure 3**. In cross-validation, all models exhibited excellent performance with AUC values between 0.78 and 0.9, with the best three models obtained for the targets

TABLE 1 | Performance of different fingerprints in cross-validation of predictions for ER-LBD.

Score[a]	ROC-AUC				
	MACCS	ECFP4	Toxprint	Combined[b]	All[c]
naïve Bayes B_1	0.7664	0.7870	0.7744	0.7833	0.7874
naïve Bayes $1-B_0$	0.7720	0.7716	0.7818	0.8031	0.8021
Similarity T_1	0.7805	0.7773	0.7840	0.7957	0.8008
Similarity T_2	0.6660	0.6873	0.7223	0.6697	0.7023
Similarity $1-T_3$	0.5455	0.6228	0.5751	0.5831	0.6299
Mean Bayes score	0.7718	0.7823	0.7813	0.7968	0.7991
Mean tanimoto	0.7752	0.8014	0.8034	0.7901	0.8173
Mean consensus[d]	0.7951	0.8145	0.8148	0.8134	0.8240

[a] *Scores have been calculated as follows: B_1, naïve Bayes score for actives; B_0, naïve Bayes score for inactives; T_1, maximum Tanimoto score to actives; T_2, average Tanimoto score to actives; T_3, maximum Tanimoto score to inactives.*

[b] *Combination of MACCS, ECFP4 and Toxprint fingerprints.*

[c] *Combination of all fingerprints with property-based fingerprint calculated from topological descriptors.*

[d] *Mean of the average Bayes score and the average Tanimoto score.*

TABLE 2 | Parameters of the most successful prediction models.

Target	External compounds	Correlation filter	Diversity Picker	Naïve Bayes[a]	Similarity[b]	Consensus score
AhR	≤ 5000 nM	–	19% actives	Mean	Mean	Mean
AR	≤ 5 nM	–	–	Mean	Mean	Mean
AR-LBD	≤ 5 nM	–	–	–	Mean	–
Aromatase	–	–	58% actives	–	Mean	–
ER	≤ 5 nM	–	–	Max	Mean	Mean
ER-LBD	≤ 5 nM	0.9	44% actives	Min	Mean	Mean
PPARγ	–	–	47% actives	Max	Max	Min
ARE	–	–	–	–	Mean	–
ATAD5	≤9200 nM	–	9% actives	1–B_0	T_1	Mean
HSE	≤160 nM	–	43% actives	Max	Mean	Mean
MMP	–	–	17% actives	1–B_0	T_1	Mean
P53	–	0.9	54% actives	–	Mean	–

[a] *Combination of the Naive Bayes scores for active (B_1) and inactive (1-B_0) compounds.*

[b] *Combination of the Tanimoto similarity scores: maximum Tanimoto score to actives (T_1), average Tanimoto score to actives (T_2), 1—maximum Tanimoto score to inactives (T_3).*

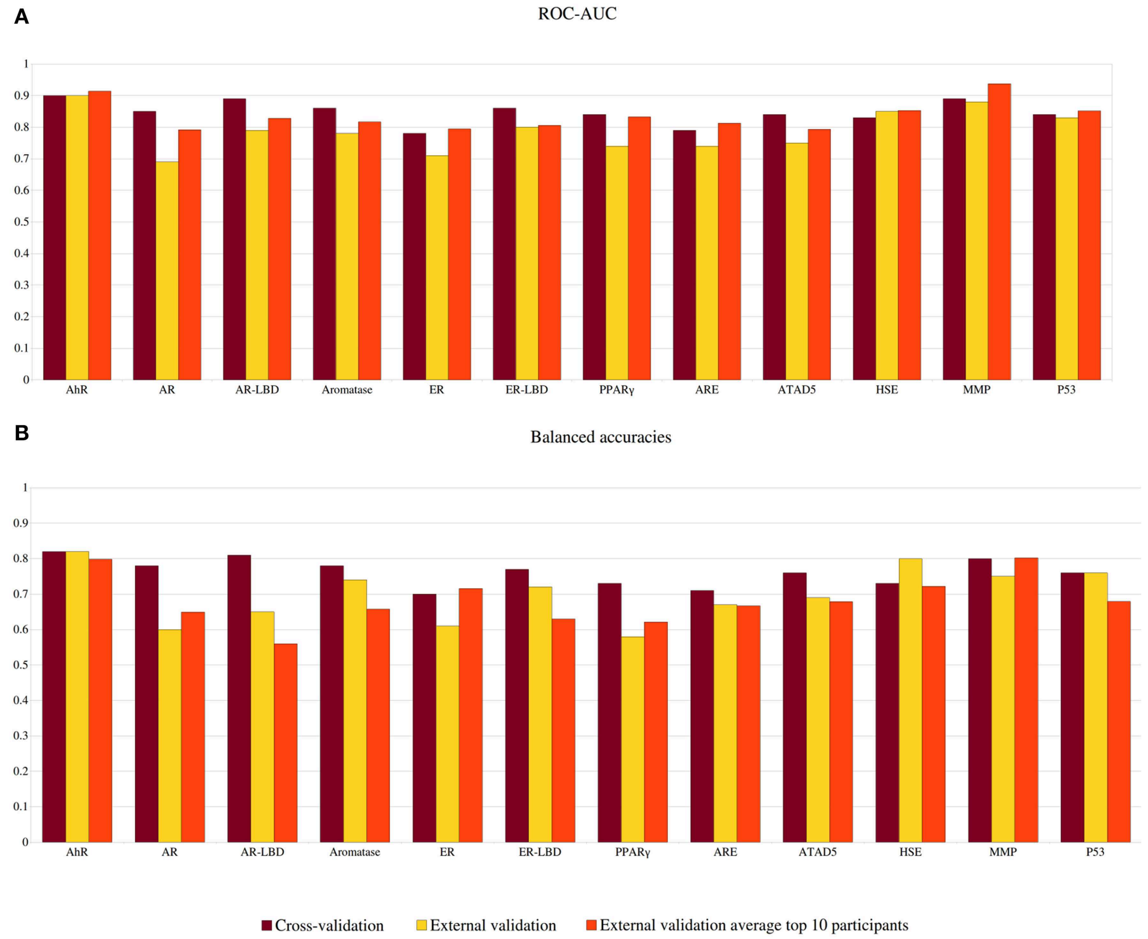


FIGURE 3 | Performance of models predicting the outcome of the Tox21 screening outcomes. (A) Area under the curve (AUC) calculated in a ROC analysis. **(B)** Balanced accuracies (BAC). Results are shown for our models in cross-validation (dark red) and external validation (yellow) as well as the average external validation results among the top 10 challenge participants.

AhR, AR-LBD, and MMP. For AhR, MMP, and p53, the results of the external validation set showed a very similar performance to the cross-validation, indicating good and universal models and scores. In the cross-validation, the balanced accuracies of the individual models ranged between 70 and 82% (see **Figure 3**). For several targets, including AhR, HSE, and p53, the balanced accuracy obtained in external validation remained constant or increased in comparison to the cross-validation results, illustrating broadly applicable models.

Comparison to Other Challenge Participants

All models submitted to the challenge were evaluated by the challenge organizers and ranked according to their AUC values for the external validation set. The prediction values for the top 10 participating teams are publicly available (https://tripod.nih.gov/tox21/challenge/leaderboard.jsp) and summarized in **Figure 3**, Supplementary Tables S2, S3. Taken together, 7 out of 12 models we submitted were found in the top 10 leaderboard. While our models were not nominated as the sub-challenge winners, in many cases their AUC value was found very close to the winning model. This was for instance observed for the target HSE, where the top 9 ranking models showed AUC values differing only by 0.02, suggesting that similarly good models can be obtained with various approaches. As indicated in **Figure 3**, our models for the targets AhR, ER-LBD and p53 were also very close to the average AUC of the leading models. Although most leaderboard models showed AUC values within a small range, large differences were observed for the prediction accuracies (between 49 and 90%). Interestingly, four of our models (targets: AR-LBD, ER-LBD, aromatase, and HSE) were the determined to be the most accurate amongst all submissions (see **Figure 3** and Supplementary Table S3). Four additional models, developed for the targets AhR, ARE, ATAD5, and p53, displayed accuracies higher or equal to the average of the top 10 submitted models.

Discussion

Here, we describe a successful machine learning method for the prediction of different outcomes of the nuclear receptor and stress response pathway screening from the Tox21 Data Challenge 2014. The key to our method is the combination of different molecular fingerprints and descriptors as well as the integration of two different algorithms, a similarity-based approach and a naïve Bayes machine learning technique.

Combination of Features and Algorithms

The selection of features is a crucial and non-trivial part of development of predictive models. The features should be able to describe the differences between actives and inactives in the training set and allow extrapolating to other, yet untested compounds. Although several molecular fingerprints, such as extended-connectivity, substructure-based or path-based fingerprints are standards in the chemoinformatics field and have been successfully applied to prediction tasks, the results are dependent on the data and none of the methods is able to clearly outperform the others (Duan et al., 2010). To avoid the choice of the wrong descriptor, the combination of (independent) fingerprints has been suggested (Duan et al., 2010) and several studies have successfully applied combinations of path- and substructure-based fingerprints (Drwal et al., 2014; Banerjee et al., 2015). As we report here, the combination of different fingerprint types has also been of advantage for the prediction of estrogen receptor ligands. An associated problem, however, is that a combined fingerprint is likely to contain highly correlated features. We have thus investigated the use of a correlation filter to remove fingerprint bits with high correlation, but the filter was able to increase the prediction performance only for two targets. A more effective approach proved to be the use of a diverse subset of active molecules in the training set, though the size of the diverse subset giving the best results had to be optimized individually for every target. As the active molecules of the different Tox21 sub-challenges might contain different important molecular characteristics, the use of extensive cross-validation to optimize the feature selection for every sub-challenge could further improve the prediction performance. Automated feature selection using deep neural networks, as suggested by one of the other teams participating in the Tox21 challenge (Unterthiner et al., 2015), offers an alternative way to determine the most relevant features in the input molecules which can be advantageous for large sets of molecules, but is obviously associated with large computational costs.

Combinations of multiple machine learning algorithms, also referred to as hybrid or ensemble learning, are a well-described approach and have been applied to solve diverse research questions (Yang et al., 2015b). It is usually assumed that the use of multiple models can increase the prediction accuracy as compared to the use of a single model and help to manage high-dimensional and complex data sets. Similarly to our approach, several other studies have proven that merging a naïve Bayes classifier with a similarity-based approach such as k-nearest neighbors can result in highly predictive models for various applications including the prediction of molecular targets (Ferdousy et al., 2013; Liu et al., 2013). Future investigations could focus on the evaluation of other classification methods (logistic regression, random forests, etc.) and larger model ensembles for the purposes of toxicity prediction.

Conclusions

Our models use a combination of molecular fingerprints and algorithms and show consistently good performance for the 12 outcomes of the Tox21 screen, four of the models being the most accurate amongst the challenge participants. We are planning to make our models publicly available by incorporating them into our toxicity prediction platform ProTox (http://tox.charite.de) in the future.

The Tox21 Data Challenge 2014 has provided an excellent opportunity for academic and industrial groups to assess and directly compare the quality of their toxicity prediction methods. The results will be of great value to the scientific community and can help to pave the way toward the use of more *in silico* toxicity models as decision-making tools to evaluate potential health hazards of environmental chemicals and drugs.

Author Contributions

Data preparation and analysis: MND, VS, PB, MD, AG; Generation and validation of predictive models: MD, MND, VS, PB; Calculation and selection of descriptors: PB, VS, MD, MND; Writing of manuscript: MND; Project coordination: RP, MD, MND.

Acknowledgments

The authors kindly acknowledge the following funding sources: German Cancer Consortium (DKTK); Berlin-Brandenburg research platform BB3R (BMBF) [031A262C]; Immunotox project (BMBF) [031A268B]; research training group "Computational Systems Biology" [GRK1772].

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2015.00054

References

Attene-Ramos, M. S., Huang, R., Michael, S., Witt, K. L., Richard, A., Tice, R. R., et al. (2015). Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ. Health Perspect.* 123, 49–56. doi: 10.1289/ehp.1408642

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7:20. doi: 10.1186/s13321-015-0069-3

Banerjee, P., Erehman, J., Gohlke, B. O., Wilhelm, T., Preissner, R., and Dunkel, M. (2015). Super Natural II–a database of natural products. *Nucleic Acids Res.* 43, D935–D939. doi: 10.1093/nar/gku886

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090. doi: 10.1093/nar/gkt1031

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266. doi: 10.1126/science.1158140

Chen, S., Zhou, D., Hsin, L. Y., Kanaya, N., Wong, C., Yip, R., et al. (2014). AroER tri-screen is a biologically relevant assay for endocrine disrupting chemicals modulating the activity of aromatase and/or the estrogen receptor. *Toxicol. Sci.* 139, 198–209. doi: 10.1093/toxsci/kfu023

Delfosse, V., Grimaldi, M., Le Maire, A., Bourguet, W., and Balaguer, P. (2014). Nuclear receptor profiling of bisphenol-A and its halogenated analogues. *Vitam. Horm.* 94, 229–251. doi: 10.1016/B978-0-12-800095-3.00009-2

Drwal, M. N., Banerjee, P., Dunkel, M., Wettig, M. R., and Preissner, R. (2014). ProTox: a web server for the *in silico* prediction of rodent oral toxicity. *Nucleic Acids Res.* 42, W53–W58. doi: 10.1093/nar/gku401

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 29, 157–170. doi: 10.1016/j.jmgm.2010.05.008

Ferdousy, E. Z., Islam, M., and Matin, M. (2013). Combination of naive bayes classifier and K-Nearest Neighbor (cNK) in the classification based predictive models. *Comput. Inf. Sci.* 6, 48. doi: 10.5539/cis.v6n3p48

Floris, M., Manganaro, A., Nicolotti, O., Medda, R., Mangiatordi, G. F., and Benfenati, E. (2014). A generalizable definition of chemical similarity for read-across. *J. Cheminform.* 6, 39. doi: 10.1186/s13321-014-0039-1

Gadaleta, D., Pizzo, F., Lombardo, A., Carotti, A., Escher, S. E., Nicolotti, O., et al. (2014). A k-NN algorithm for predicting the oral sub-chronic toxicity in the rat. *ALTEX* 31, 423–432. doi: 10.14573/altex.140 5091s

Hall, L. H., and Kier, L. B. (1991). "The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling," in *Reviews in Computational Chemistry,* eds K. B. Lipkowitz and D. B. Boyd (Hoboken, NJ: John Wiley & Sons, Inc.), 367–422.

Huang, R., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., et al. (2014). Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep.* 4:5664. doi: 10.1038/srep05664

Krewski, D., Acosta, D. Jr., Andersen, M., Anderson, H., Bailar, J. C. 3rd, Boekelheide, K., et al. (2010). Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health B Crit. Rev.* 13, 51–138. doi: 10.1080/10937404.2010.483176

Li, X., Chen, L., Cheng, F., Wu, Z., Bian, H., Xu, C., et al. (2014). *In silico* prediction of chemical acute oral toxicity using multi-classification methods. *J. Chem. Inf. Model.* 54, 1061–1069. doi: 10.1021/ci5000467

Liu, J., Mansouri, K., Judson, R. S., Martin, M. T., Hong, H., Chen, M., et al. (2015). predicting hepatotoxicity using toxcast *in vitro* bioactivity and chemical structure. *Chem. Res. Toxicol.* 28, 738–751. doi: 10.1021/tx500501h

Liu, X., Vogt, I., Haque, T., and Campillos, M. (2013). HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics* 29, 1910–1912. doi: 10.1093/bioinformatics/btt303

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, 361–367. doi: 10.1038/nature11159

Murray, I. A., Patterson, A. D., and Perdew, G. H. (2014). Aryl hydrocarbon receptor ligands in cancer: friend and foe. *Nat. Rev. Cancer* 14, 801–814. doi: 10.1038/nrc3846

Nadanaciva, S., and Will, Y. (2011). Investigating mitochondrial dysfunction to increase drug safety in the pharmaceutical industry. *Curr. Drug Targets* 12, 774–782. doi: 10.2174/138945011795528985

Nickel, J., Gohlke, B. O., Erehman, J., Banerjee, P., Rong, W. W., Goede, A., et al. (2014). SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.* 42, W26–W31. doi: 10.1093/nar/gku477

Nidhi, Glick, M., Davies, J. W., and Jenkins, J. L. (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46, 1124–1133. doi: 10.1021/ci060003g

Pasha, F. A., Neaz, M. M., Cho, S. J., Ansari, M., Mishra, S. K., and Tiwari, S. (2009). *In silico* quantitative structure-toxicity relationship study of aromatic nitro compounds. *Chem. Biol. Drug Des.* 73, 537–544. doi: 10.1111/j.1747-0285.2009.00799.x

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

Unterthiner, T., Mayr, A., Klambauer, G., and Hochreiter, S. (2015). Toxicity Prediction Using Deep Learning. *Machine Learning.* Available online at:http://arxiv.org/abs/1503.01445

Vogt, M., and Bajorath, J. (2008). Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chem. Biol. Drug Des.* 71, 8–14. doi: 10.1111/j.1747-0285.2007.00602.x

Weiss, B. (2012). The intersection of neurotoxicology and endocrine disruption. *Neurotoxicology* 33, 1410–1419. doi: 10.1016/j.neuro.2012.05.014

Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J. (2003). Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* 43, 1151–1157. doi: 10.1021/ci030285+

Yang, C. H., Tarkhov, A., Marusczyk, J., Bienfait, B., Gasteiger, J., Kleinoeder, T., et al. (2015a). New publicly available chemical query language, csrml, to support chemotype representations for application to data mining and modeling. *J. Chem. Inf. Model.* 55, 510–528. doi: 10.1021/ci500667v

Yang, P., Yang, Y. H., Zhou, B. B., and Zomaya, A. Y. (2015b). A Review of Ensemble Methods in Bioinformatics. *Curr. Bioinform.* 5, 296–308. doi: 10.2174/157489310794072508

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.

ORIGINAL RESEARCH
published: 04 February 2016
doi: 10.3389/fenvs.2016.00003

QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays

Stephen J. Capuzzi†, *Regina Politi*†, *Olexandr Isayev*†, *Sherif Farag and Alexander Tropsha* *

Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

The ability to determine which environmental chemicals pose the greatest potential threats to human health remains one of the major concerns in regulatory toxicology. Computational methods that can accurately predict a chemical's toxic potential *in silico* are increasingly sought-after to replace *in vitro* high-throughput screening (HTS) as well as controversial and costly *in vivo* animal studies. To this end, we have built Quantitative Structure-Activity Relationship (QSAR) models of 12 stress response and nuclear receptor signaling pathways toxicity assays as part of the 2014 Tox21 Challenge. Our models were built using the Random Forest, Deep Neural Networks and various combinations of descriptors and balancing protocols. All of our models were statistically significant for each of the 12 assays with the balanced accuracy in the range between 0.58 and 0.82. Our results also show that models built with Deep Neural Networks had higher accuracy than those developed with simple machine learning algorithms and that dataset balancing led to a significant accuracy decrease.

Keywords: Tox21, machine-learning, stress response signaling pathways, nuclear receptor signaling pathways, endocrine disrupting chemicals, QSAR, deep learning

OPEN ACCESS

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Patrick D. McMullen,
The Hamner Institutes for Health Sciences, USA
Zhichao Liu,
National Center for Toxicological Research, Food and Drug Administration, USA

Correspondence:
Alexander Tropsha
alex_tropsha@email.unc.edu

†*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *30 September 2015*
Accepted: *12 January 2016*
Published: *04 February 2016*

Citation:
Capuzzi SJ, Politi R, Isayev O, Farag S and Tropsha A (2016) QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. Front. Environ. Sci. 4:3. doi: 10.3389/fenvs.2016.00003

INTRODUCTION

The ability to determine which environmental chemicals pose the greatest potential threats to human health remains one of the major concerns in regulatory toxicology. In addition, the inability to recognize potentially toxic substances during the initial steps of drug development contributes to the failure of promising pharmaceutical leads in more than 30% of human clinical trials (Kola and Landis, 2004). Historically, the estimated human health impact of these chemicals has been assessed through *in vivo* animal studies. Animal studies, however, are costly, laborious, impractical for evaluating large numbers of chemicals, and are being progressively eliminated due to their controversial nature (Anastas et al., 2010). However, over the past several years, the focus has switched to high-throughput *in vitro* screening (HTS) in order to identify chemical hazards and prioritize chemicals for additional *in vivo* testing (O'Brien et al., 2006).

Abbreviations: AR, androgen receptor; AR-LBD, androgen receptor—ligand binding domain; AhR, aryl hydrocarbon receptor; ER, estrogen receptor alpha—full; ER-LBD, estrogen receptor alpha—ligand binding domain; PPAR-gamma, peroxisome proliferator-activated receptor gamma; ARE, nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element; HSE, heat shock factor response element; MMP, mitochondrial membrane potential; p53, tumor suppressor p53; QSAR, Quantitative Structure-Activity Relationship; HTS, High-Throughput Screening; AUC, Area under the curve; BA, Balanced accuracy; DNN, Deep Neural Network.

The ToxCast project and the Tox21 consortium have used high-throughput screening to characterize the *in vitro* biological activity of chemicals across multiple cellular pathways and biochemical targets (Dix et al., 2007). HTS campaigns, however, can also be costly and time-consuming because every new series of chemicals must be screened against multiple toxicity endpoints and at various concentrations. Therefore, *in silico* methods that can accurately predict toxicity toward the prioritization of chemicals for experimental testing are in-demand. To this end, the 2014 Tox21 Challenge sought to "crowdsource" predictive models from various researchers across the globe to assess how well their models can predict the toxic potential of a compound in several biological pathways screened against the Tox21 10,000 compound library (https://tripod.nih.gov/tox21/challenge/about.jsp).

Quantitative Structure-Activity Relationship (QSAR) models provide such a computational method toward the *in silico* prediction of chemical toxicity. QSAR models utilize complex machine learning algorithms to establish a relationship between chemical structure and the modeled endpoint (toxicity). Robust and rigorously-validated QSAR models are then used to provide *in silico* predictions of the endpoint-of-interest for yet-untested chemicals (Tropsha, 2010). Thus, the Tox21 program aimed to identify new methods for assessing chemical toxicity in the form of QSAR models in order to improve the identification of chemicals that may affect the functions of seven nuclear receptors (AR, AR-LBD, ER, ER-LBD, AhR, Aromatase, PPAR-gamma) and five stress response pathways (ARE, ATAD5, HSE, MMP, p53) in the human body.

Several of these pathways of interest regulate normal endocrine function. Endocrine disrupting chemicals (EDCs) interfere with the endocrine system through interactions with nuclear receptors (Diamanti-Kandarakis et al., 2009). EDCs engender myriad adverse developmental, reproductive, neurological, and immunological effects in both humans and wildlife. Unfortunately, both humans and wildlife are ubiquitously exposed to EDCs, as EDCs have widespread industrial applications, resulting in endocrine toxicity (Casals-Casas and Desvergne, 2011). For instance, bisphenol-A and its analogs—EDCs which are used heavily in the manufacturing of polycarbonate plastics and epoxy resins (Bae et al., 2002)—have been shown to bind to the estrogen receptor (ER), androgen receptor (AR), and peroxisome proliferator-activated receptor (PPAR) gamma (Han et al., 2003). Moreover, there is ample evidence that EDCs also interact with stress response pathways, such as mitochondrial membrane potential (MMP) and tumor suppressor p53 (Min et al., 2003; Chandra, 2013). For these reasons, the identification of endocrine disrupting chemicals (EDCs) is of particular interest to the Tox21 program and environmental chemical hazard screening in general.

The overall goal of the Tox21 Challenge was to predict compound activity (toxic or non-toxic) in pathway assays provided by the Challenge organizers using only chemical structure data. The data provided was generated from seven nuclear receptor and five stress response pathway assays run against the Tox21 compound library. We performed various permutations of curation and balancing protocols to generate Random Forest (RF) and deep neural net (DNN) models employing either Dragon or SiRMS descriptors.

METHODS

Datasets

All datasets (training and test sets) of compound toxicity in 12 different pathway assays were downloaded from the Tox21 Challenge website (https://tripod.nih.gov/tox21/challenge/index.jsp). The training set included 11,764 compounds with activities 0 (non-toxic) and 1 (toxic) in each of the 12 assays. Test set 1 comprised 296 compounds with various activities in each of the 12 assays. This test set, initially used to evaluate model performance, was subsequently merged into the training set. Test set 2 included 647 compounds with various activities in each of the 12 assays. This set was used to evaluate model performance and to rank model submissions of various participants. For all datasets in each assay, a compound was active (1), inactive (0), or untested.

Dataset Curation

Each dataset was curated according to our well-established protocol (Fourches et al., 2010). Structural standardization, the cleaning of salts, and the removal of mixtures, inorganics, and organometallics was performed using Instant JChem software (version 6.2, ChemAxon).

In the case of replicate compounds, InChI Keys were generated using Instant JChem software. For replicates with the same activities in a given assay, a single representative compound was selected for inclusion into the training set. For replicates with the different activities in a given assay, all compounds were excluded.

After curation, the sizes of the training set, test set 1, and test set 2 were reduced to 9323 compounds, 291 compounds, and 641 compounds, respectively.

Dataset Balancing

For each pathway assay, only compounds that were explicitly tested (active or inactive) were used. Inactive (non-toxic) compounds were the predominant majority (ratio 10:1 or higher) as compared to active (toxic) compounds in the training sets for each of the 12 assays. Inactive compounds were down-sampled such as to make the remaining number of inactives similar to the respective number of active compounds in each of the individual assays either (a) randomly or (b) according to highest Tanimoto similarity to compounds in test set 2. In a separate study (c), the training set was left unbalanced (see Supplemental for individual assay counts).

Molecular Descriptors

Dragon Descriptors

An ensemble of 2489 molecular descriptors was computed with the Dragon software (version 5.4) for all compounds (with explicit hydrogen atoms) in every dataset.

SiRMs

2D Simplex Representation of Molecular Structure (SiRMS) descriptors (Muratov et al., 2010) were generated by the HiT QSAR software (Kuz'min et al., 2008). At the 2D level, the connectivity of atoms in a simplex, atom type, and bond nature (single, double, triple, or aromatic) have been considered. SiRMS descriptors account not only for the atom type, but also for other atomic characteristics that may impact biological activity of molecules, e.g., partial charge, lipophilicity, refraction, and atom ability for being a donor/acceptor in hydrogen-bond formation (H-bond). Detailed description of HiT QSAR and SiRMS can be found elsewhere (Kuz'min et al., 2008; Muratov et al., 2010).

Model Building and Evaluation

Random Forest (RF)

QSAR models were built using an in-house implementation on Chembench (http://chembench.mml.unc.edu) of the original RF algorithm (Breiman, 2001).

External 5-fold Cross Validation

All RF models were evaluated using external 5-fold cross validation (Tropsha et al., 2003). Every training set for each of the 12 assays was randomly partitioned into five equal parts with the same active (toxic)/inactive (non-toxic) ratio before modeling. In turn, each of the five parts was "left out" to form an external set used to validate the model developed on the remaining four parts that collectively amounted to the modeling set.

Score Threshold

The ensemble of selected RF models outputs a continuous consensus score (RF score) ranging from 0 (non-toxic) to 1 (chemical predicted to be toxic by all models). When there is a disagreement between those individual RF models, the consensus RF score can thus take any value between 0 and 1. When computed for a set of chemicals, RF scores can be used to rank those chemicals based on their increasing RF-evaluated likelihood of being toxic. For all assays, a RF score threshold was arbitrarily set to 0.5, with scores ≥ 0.5 being active (toxic) and scores <0.5 being inactive (non-toxic).

Y-Randomization

Models were further validated through Y-randomization, wherein activities (i.e., the response variable Y) observed for the original training set are randomly assigned to the training set compounds multiple times and the models are built for all datasets generated by these multiple permutations of the response variable. This procedure ensures that the models built for the original datasets do not reflect a chance correlation between multiple independent variables (i.e., chemical descriptors) and the dependent variable.

Deep Learning Models

We trained deep neural net (DNN) (Schmidhuber, 2015) models with the rectified linear units (ReLU) activation function (Nair and Hinton, 2010) instead of typical sigmoidal units. The rectified linear unit computes the function $f(x) = \max(0,x)$. In other words, the activation is simply thresholded at zero when $x < 0$ and then linear with a of slope 1 when $x > 0$.

Neural networks can have many hyperparameters. Therefore, in order to choose the best network architecture, we performed a grid search over the parameters based on the 10% randomly selected validation set from the training data. The parameter space include number of hidden layers {2, 3}, number of neurons {100, 200, 400, 800, 1600}, amount of dropout {0, 0.25, 0.5}, and L2 regularization.

All networks were trained using mini-batched stochastic gradient descent (SGD) and AdaGrad (Duchi et al., 2011). AdaGrad is an Adaptive Gradient Method that utilizes different adaptive learning rates for every feature. It was shown to significantly accelerate convergence and slightly improve performance of DNNs (Dean et al., 2012). The output layer is a standard softmax classifier and cross entropy objective function. For every endpoint, DNN models were trained independently.

In addition, we also investigated the performance of a multitask network (one model for all 12 tasks trained jointly) using the identical training approach. Learning several tasks at the same time is performed with the aim of mutual benefits between different tasks. The similarity (and dissimilarity) between the tasks is exploited to enrich a model (Caruana, 1997).

All models were trained using in-house software based on Theano framework (Bastien et al., 2012). We also used normalized DRAGONH descriptors as our input vectors.

Data Visualization

We use a multidimensional scaling (MDS) approach (Borg and Groenen, 1997), implemented in Python, to seek a low-dimensional representation of the data that conserves the distances in the original high-dimensional space. ECFP6 fingerprints are used to calculate the similarity matrix between the chemicals. MDS applied on this similarity matrix attempts to model the similarity or dissimilarity of data as distances in geometric space. In this way, higher similarity between the chemicals results in shorter distances between the chemicals in the projection.

RESULTS

Overview

We have developed several Random Forest models using different descriptors and balancing approaches as described in Methods; these models are summarized in **Table 1**. Models 1, 2, and 3 were submitted for final evaluation and ranking; whereas, Model 4 was built after the Tox21 Challenge had closed (**Table 1**).

Evaluation and Ranking

The performance of all submitted models was evaluated by AUC-ROC resulted from predictions made for test set 2. Results for all of our models in comparison with the winning model for each assay are summarized in **Figure 1**. None of our submitted models (Model 1, Model 2, and Model 3) were ranked in the top 10. Additionally, differences in balancing protocol and descriptor type in our submitted models had little effect on the overall performance. Model 4 was built using unbalanced data. It was not submitted for evaluation by the organizers, and therefore was ineligible for ranking.

TABLE 1 | Description of models implemented using Random Forest.

Model name	Descriptor	Balancing protocol
Model 1	DRAGONH	1:1 Randomly
Model 2	DRAGONH	1:1 to test set 2
Model 3	SiRMS	1:1 Randomly
Model 4	DRAGONH	Unbalanced

TABLE 2 | Post-challenge assessment of the accuracy (AUC) of different models and their comparison with the wining solution.

Subchallenge	Model 4	Model 4/CV5	DNN/1 task	DNN/12 tasks	Winner
AhR	0.91	0.91	0.90	0.87	0.93
AR	0.73	0.82	0.83	0.89	0.83
AR-LBD	0.72	0.91	0.89	0.88	0.88
ARE	0.78	0.83	0.81	0.76	0.84
aromatase	0.80	0.82	0.86	0.76	0.84
ATAD5	0.81	0.83	0.85	0.72	0.83
ER	0.79	0.79	0.81	0.74	0.81
ER-LBD	0.78	0.86	0.83	0.90	0.83
HSE	0.79	0.80	0.79	0.77	0.86
MMP	0.93	0.92	0.95	0.85	0.95
p53	0.85	0.82	0.84	0.77	0.88
PPAR-gamma	0.79	0.81	0.70	0.80	0.86
Average AUC	0.81	0.84	0.84	0.81	0.86

The color gradient is a heat map for each model. The highest AUC for each subchallenge is darkest green, etc.

Nevertheless, Model 4 showed a greater AUC value for 10 of the 12 assays over our three submitted models. Model 4 also showed comparable predictive performance to the winning models: seven of the 12 AUCs (AhR, Aromatase, ATAD5, ER, ER-LBD, MMP, p53) differ from the winner by 0.05. Interestingly, when comparing the external balanced accuracy, defined as (Sensitivity + Specificity)/2, of our models to those of winning models (based on AUC), a different trend emerges (**Figure 2**). For nine out of the 12 assays, the external balanced accuracy of at least one of our models is higher than that of the winning model. Indeed, for all submitted models in the Challenge, our Model 2 had the highest external balanced accuracy for AR (0.74); Model 4 had an even higher external balanced accuracy (0.82).

Figure 3 visualizes the distribution of active and inactive compounds from the training dataset and test set 2 based on fingerprint similarity (see Section Model Building and Evaluation for details). **Figures 3A,B** show the distribution of compounds in the training sets of Models 2 and 4, respectively, as well as in test set 2 for one of our most accurately predicted endpoints, AhR. **Figures 3C,D** show the same type of distribution for one of the least accurately predicted endpoints, HSE, that has largest increase in AUC as a result of using unbalanced dataset for modeling (Model 4). This analysis reveals that balanced training dataset used in Model 2 for AhR (**Figure 3A**) has tight clustering of active compounds in addition to broad coverage for the compounds to be predicted in test set 2. Thus, an increase in the number of compounds in the training dataset when unbalanced dataset is used does not result in a significant gain in AUC. However, as opposed to AhR, no distinct clusters are observed in the balanced dataset for HSE. Active and inactive compounds are widely dispersed, which calls into question the assay quality of this endpoint. This dispersion results in the misclassification of inactive compounds in test set 2. **Figure 3D**, however, shows that using unbalanced data increases the chemical diversity, which provides better coverage of test set 2, and enhances representation of inactives found in the test set 2. This expansion reflected in an increase of AUC (see **Figure 1**).

After the results of the challenge were announced, we also decided to evaluate the limit of model performance even further. We used Model 4 as our base line (see **Table 2**). We combined all three datasets and retrained Model 4 with the same RF parameters using 5-fold external cross validation (Model 4/5CV column in **Table 2**). Unexpectedly, we obtained significant performance boost. AUCs for three endpoints, AR, AR-LBD, and ER-LBD are significantly higher as compared to AUC values achieved by Model 4. Accuracy for the other nine assays were approximately on par with the balanced models. It is not clear why such discrepancy is observed, most likely it is due to the small size of the test set and very small number of active compounds in each of them.

Given that the overall challenge winner used DNN (Mayr et al., 2015), we decided to investigate the utility of DNN after the completion of the challenge. Due to very limited technical information released by the winning team, however, we were not able to independently verify their models. Instead, we trained DNN according to our own protocol (See Section Model Building and Evaluation). **Table 2** also reports performance of DNN models in single task and multitask regimes. On average, both approaches were not able to match the winning model, AUCs 0.84 and 0.81 vs. 0.86. However, the difference between DNN/1task (our best overall model) and winning team is small with a notable exception of PPAR-gamma, ΔAUC = 0.16. The single task DNN model was also significantly better than Model 4 for AR and AR-LBD. Very recently, models of drug-induced liver injury (DILI) with DNN were also found to provide better performance than previously described "shallow" prediction models (Xu et al., 2015). Therefore, DNN architectures seems to be beneficial for toxicity prediction. In strike contrast, performance of the multitask model was poor for five assays (ARE, Aromatase, ATAD5, ER, and p53). Due to the limited dataset size, we were not able to reliably train all DNN models. In order to take full advantage of deep learning methods at least an order of magnitude larger number of training examples is necessary.

DISCUSSION

The results of our submitted models (Model 1, Model 2, and Model 3) indicate that for these data no combination of descriptors or balancing protocol outperforms any other combination. Intriguingly, our unbalanced (and un-submitted) Model 4 outperformed our submitted models and had AUC values comparable to the winning models. This observation demonstrates that for these assays balancing actually decreases model performance. This may be because balancing restricts the

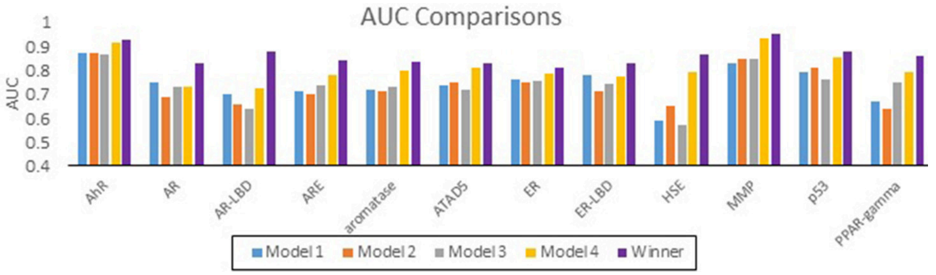


FIGURE 1 | Comparison of AUC values for the Tox21 assays. AUC values of our models (blue, orange, gray, and yellow) as well as the AUC values of the winning model (purple).

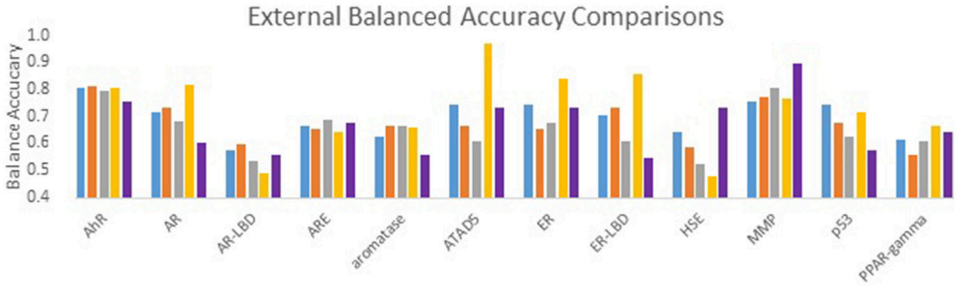


FIGURE 2 | Comparison of external balanced accuracy (BA) values for the Tox21 assays. BA values of our models (blue, orange, gray, and yellow) as well as the BA values of the winning model (purple).

chemical space covered *in toto* by inactives. Since the number of actives in test set 2 is much smaller than the number of inactives (between 1 and 14% of test set 2 compounds are actives for a given assay), reducing the chemical space of inactives through balancing may have resulted in the misclassification of inactives in test set 2. In general, when training set compounds are highly imbalanced toward the inactive class, QSAR classification will favor the majority (inactive) class, resulting in low sensitivity for the minority (active) class (Chen et al., 2005). For this reason, datasets are usually balanced as to maximize the sensitivity and specificity of the training set. In the current challenge, however, models were evaluated on an external dataset that was highly populated with the inactive class. Therefore, for future challenges and/or modeling efforts regarding these assay endpoints, using unbalanced data may be preferable.

Conflicting performance trends obtained in **Table 2** also emphasizes the following community needs:

1. Judging model performance using a very small test set can be suboptimal.
2. Deep Learning can provide some accuracy improvement compared to regular machine learning methods. However, model reproducibility is very hard to achieve, especially for this rapidly emerging field.
3. Further methodological developments are required to investigate applicability and methods of training multitask-DNN method. There is a significant room for model improvement and exploiting information about assay relations as well as target features and other biological information.

CONCLUSION

In this work, we investigated the use of different QSAR approaches for toxicity assays prediction in the 2014 Tox21 challenge. We carefully curated all datasets according the well-established protocol. We used Random Forest and Deep Neural Nets to train models. In addition we also explored several balancing strategies. The model performance was evaluated by the area under the receiver operating characteristic curve (AUC-ROC) and by the balanced accuracy (BA). The values for AUC-ROC were in the range of 0.55–0.87 and those for BA were in the range of 0.58–0.82; the highest predictive power was achieved for the AR pathway assay. No significant difference in respective model performance was found when using different curation protocols or different descriptors. Marginal increase in AUC-ROC as well as in BA was observed for some of the pathways when the dataset was balanced based on the similarity to the external test set (test set 2). Moreover, a significant increase in the balanced accuracy of prediction for external datasets was found once the unbalanced datasets were used to build the model. Our results show that overall neural networks achieved improvement over simple machine learning algorithms and that balancing lead to a significant accuracy decrease.

The Tox21 Challenge was evaluated using the AUC metric. Interestingly we noticed, when evaluated using BA, at least one of our models outperformed the winning model in 10/12 assays. Furthermore, our Model 2 had the highest balanced accuracy for AR (0.74) against all submissions. Our models, therefore, can

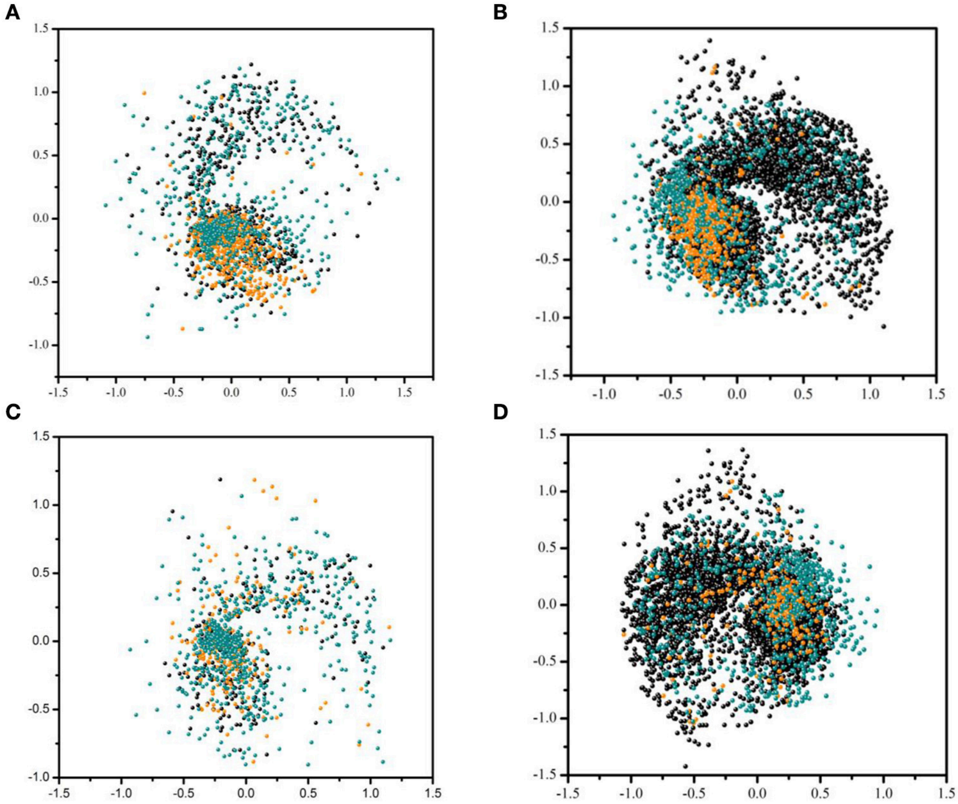


FIGURE 3 | Distribution of active and inactive compounds from the training dataset and test set 2 based on fingerprint similarity. (A) Balacned training set used for endpoint AhR in Model 2 and test set 2. **(B)** Unbalanced training set used for AhR in Model 4 and test set 2. **(C)** Unbalanced training set used for endpoint HSE in Model 2 and test set 2. **(D)** Unbalanced training set used for HSE in Model 4 and test set 2. Each point represent either compounds from the test set 2 (cyan) and training set inactives (black) and actives (orange).

be used for future screening of compounds for toxicity in these pathways. Our models have the additional advantage of being freely and publicly available through our Chembench platform (https://chembench.mml.unc.edu/; Walker et al., 2010).

The goal of the 2014 Tox21 Challenge was to predict toxicity in the various biological pathways using chemical structure data only. The availability of these chemical structures and their associated biological activity in the pathways of interest affords the opportunity to build pathway-based hybrid QSAR models. Hybrid QSAR models utilize *in vitro* bioactivity as biological descriptors in conjunction with chemical descriptors in order to improve the predictivity of QSAR models (Liu et al., 2015). These hybrid QSAR models could be employed toward the prediction of *in vivo* toxic effects, which is a considerable challenge for predictive toxicology.

In sum, the 2014 Tox21 Challenge successfully enabled academic groups, industrial teams, and fans of machine-learning from around the world to compare and contrast various *in silico* methodologies toward the prediction of toxicity in several different assays. These modeling efforts and their associated findings will be of great use to the scientific community and will enhance the quality of toxicity prediction going forward.

AUTHOR CONTRIBUTIONS

SC, RP, and OI contributed equally to this work, collaborating on all sections of the manuscript. SF provided information related to pathways. AT provided final edits.

ACKNOWLEDGMENTS

The authors kindly thank the following funding source: National Institutes of Health (1R01GM096967-01A1). The authors would like to acknowledge ChemAxon for providing an academic license and the organizers of the 2014 Tox21 Challenge. OI gratefully acknowledge the support and hardware donation of NVIDIA Corporation and personally Mark Berger. OI also acknowledges partial support from 2015 CSA Trust grant for Deep Learning work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00003

REFERENCES

Anastas, P., Teichman, K., and Hubal, E. C. (2010). Ensuring the safety of chemicals. *J. Expo. Sci. Environ. Epidemiol.* 20, 395–396. doi: 10.1038/jes.2010.28

Bae, B., Jeong, J. H., and Lee, S. J. (2002). The quantification and characterization of endocrine disruptor bisphenol-a leaching from epoxy resin. *Wat. Sci. Technol.* 46, 381–387. Available online at: http://wst.iwaponline.com/content/46/11-12/381

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., et al. (2012). *Theano: New Features and Speed Improvements. Symbolic Computation; Learning.* Available online at: http://arxiv.org/abs/1211.5590

Borg, I., and Groenen, P. (1997). *Modern Multidimensional Scarles. Springer Series in Statistics*. New York, NY: Springer-Verlag.

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

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi: 10.1023/A:1007379606734

Casals-Casas, C., and Desvergne, B. (2011). Endocrine disruptors: from endocrine to metabolic disruption. *Annu. Rev. Physiol.* 73, 135–162. doi: 10.1146/annurev-physiol-012110-142200

Chandra, D. (ed.). (2013). *Mitochondria as Targets for Phytochemicals in Cancer Prevention and Therapy.* New York, NY: Springer.

Chen, J. J., Tsai, C. A., Young, J. F., and Kodell, R. L. (2005). Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR QSAR Environ. Res.* 16, 517–529. doi: 10.1080/10659360500468468

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., et al. (2012). "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 1223–1231. Available online at: http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks

Diamanti-Kandarakis, E., Bourguignon, J.-P., Giudice, L. C., Hauser, R., Prins, G. S., Soto, A. M. R., et al. (2009). Endocrine-disrupting chemicals: an endocrine society scientific statement. *Endocr. Rev.* 30, 293–342. doi: 10.1210/er.2009-0002

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Woodrow Setzer, R., Kavlock, R. J., et al. (2007). The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* 95, 5–12. doi: 10.1093/toxsci/kfl103

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159. Available online at: http://dl.acm.org/citation.cfm?id=1953048.2021068

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. Model.* 50, 1189–1204. doi: 10.1021/ci100176x

Han, W.-D., Mu, Y.-M., Lu, X.-C., Xu, Z.-M., Li, X.-J., Yu, L., et al. (2003). Up-regulation of LRP16 mRNA by 17beta-estradiol through activation of estrogen receptor alpha (ERalpha), but not ERbeta, and promotion of human breast cancer MCF-7 cell proliferation: a preliminary report. *Endocr. Relat. Cancer* 10, 217–224. doi: 10.1677/erc.0.0100217

Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715. doi: 10.1038/nrd1470

Kuz'min, V. E., Artemenko, A. G., and Muratov, E. N. (2008). Hierarchical QSAR technology based on the simplex representation of molecular structure. *J. Comput. Aided Mol. Des.* 22, 403–421. doi: 10.1007/s10822-008-9179-6

Liu, J., Mansouri, K., Judson, R. S., Martin, M. T., Hong, H., Chen, M., et al. (2015). Predicting hepatotoxicity using ToxCast *in vitro* bioactivity and chemical structure. *Chem. Res. Toxicol.* 28, 738–751. doi: 10.1021/tx500501h

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2015). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3:80. doi: 10.3389/fenvs.2015.00080

Min, J., Lee, S.-K., and Bock Gu, M. (2003). Effects of endocrine disrupting chemicals on distinct expression patterns of estrogen receptor, cytochrome P450 aromatase and p53 genes in oryzias latipes liver. *J. Biochem. Mol. Toxicol.* 17, 272–227. doi: 10.1002/jbt.10089

Muratov, E. N., Artemenko, A. G., Varlamova, E. V., Polischuk, P. G., Lozitsky, V. P., Fedchuk, A. S., et al. (2010). Per Aspera Ad Astra: application of simplex QSAR approach in antiviral research. *Fut. Med. Chem.* 2, 1205–1226. doi: 10.4155/fmc.10.194

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* 807–814. Available online at: http://citeseer.ist.psu.edu/viewdoc/summary?

O'Brien, P. J., Irwin, W., Diaz, D., Howard-Cofield, E., Krejsa, C. M., Slaughter, M. R., et al. (2006). High concordance of drug-induced human hepatotoxicity with *in vitro* cytotoxicity measured in a novel cell-based model using high content screening. *Arch. Toxicol.* 80, 580–604. doi: 10.1007/s00204-006-0091-3

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* 29, 476–488. doi: 10.1002/minf.201000061

Tropsha, A., Gramatica, P., and Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 22, 69–77. doi: 10.1002/qsar.200390007

Walker, T., Grulke, C. M., Pozefsky, D., and Tropsha, A. (2010). Chembench: a cheminformatics workbench. *Bioinformatics* 26, 3000–3001. doi: 10.1093/bioinformatics/btq556

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

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 Environmental Science

ORIGINAL RESEARCH
published: 01 December 2015
doi: 10.3389/fenvs.2015.00077

Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors

Filip Stefaniak *

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

OPEN ACCESS

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Keith Shockley,
National Institute of Environmental Sciences, USA
Denis Fourches,
North Carolina State University, USA
Costica Nitu,
University "Politehnica"Bucharest, Romania

***Correspondence:**
Filip Stefaniak
fstefaniak@genesilico.pl

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *01 September 2015*
Accepted: *16 November 2015*
Published: *01 December 2015*

Citation:
Stefaniak F (2015) Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors. Front. Environ. Sci. 3:77. doi: 10.3389/fenvs.2015.00077

Toxicity evaluation of newly synthesized or used compounds is one of the main challenges during product development in many areas of industry. For example, toxicity is the second reason—after lack of efficacy—for failure in preclinical and clinical studies of drug candidates. To avoid attrition at the late stage of the drug development process, the toxicity analyses are employed at the early stages of a discovery pipeline, along with activity and selectivity enhancing. Although many assays for screening *in vitro* toxicity are available, their massive application is not always time and cost effective. Thus, the need for fast and reliable *in silico* tools, which can be used not only for toxicity prediction of existing compounds, but also for prioritization of compounds planned for synthesis or acquisition. Here I present the benchmark results of the combination of various attribute selection methods and machine learning algorithms and their application to the data sets of the Tox21 Data Challenge. The best performing method: Best First for attribute selection with the Rotation Forest/ADTree classifier offers good accuracy for most tested cases. For 11 out of 12 targets, the AUROC value for the final evaluation set was =0.72, while for three targets the AUROC value was = 0.80, with the average AUROC being 0.784 ± 0.069. The use of two-dimensional descriptors sets enables fast screening and compound prioritization even for a very large database. Open source tools used in this project make the presented approach widely available and encourage the community to further improve the presented scheme.

Keywords: toxicity prediction, machine learning, molecular descriptors, molecular fingerprints, Tox21 Data Challenge 2014

INTRODUCTION

Toxicity evaluation of newly synthesized or used chemicals (pharmaceuticals and its metabolites, cosmetic ingredients, biocides, or anthropogenic pollutants) is one of the main challenges during product development in many areas of industry. For example, it has been estimated that in the pharmaceutical industry the toxicology and clinical safety is accounting for 30% of failures in clinical trials (Kola and Landis, 2004). The risk of attrition can be substantially reduced by the introduction of toxicity testing at the early stages of product development. Such evaluation, especially when performed on a large scale, is neither time/cost effective, nor—in case of tests performed on animals—ethically justified. It is estimated that the introduction of a new pesticide to the market requires testing on 7000 animals and costs tens of millions of dollars (Erickson, 2011). Moreover, animal models are frequently poorly correlated with response on humans (Knight, 2007; Shanks et al., 2009). Although *in vivo* testing seems to be inevitable at the late stage of a product development, many efforts to shift from traditional *in vivo* tests to higher-throughput and less expensive cell-based assays have been made. For example "The Toxicology in the 21st Century" (Tox21) program, is aimed at developing more reliable toxicity assessment methods as well as developing and validating cellular (*in vitro*) toxicity assays. The Tox21 10 K chemical library consists of ∼10,500 plated compound solutions, consisting of 8311 unique chemical substances, including pesticides, industrial chemicals, food-use additives and drugs (Huang et al., 2014). Acquired activity data can serve not only as *in vitro* signatures that could be used to predict *in vivo* toxicity endpoints (Martin et al., 2011; Sipes et al., 2011) and to prioritize chemicals for extensive toxicity testing (Judson et al., 2010), but also to provide the scientific community with training data sets for developing reliable *in silico* toxicity models (Sun et al., 2012). Also, many attempts toward development of new computational methods for high-throughput toxicity prediction have been made and many techniques and algorithms have been proposed (Deeb and Goodarzi, 2012; Bakhtyari et al., 2013; Cheng et al., 2013; Valerio, 2013; Low et al., 2014; Omer et al., 2014; Toropov et al., 2014; Rouquie et al., 2015). In recent years, machine learning methods are gaining more attention as robust and accurate tools for Quantitative structure–activity relationship (QSAR) and Quantitative structure–property relationships (QSPR) modeling (Durrant and Amaro, 2015; Freitas et al., 2015; Liu, 2015). The key to success in building predictive models are: (a) the quality of a training data set, (b) the descriptive power of molecular descriptors, and (c) selecting and tuning machine learning algorithms. Here I present a detailed description of creating activity prediction models using the Tox21 Data Challenge data set (Subchallenges 1–12). It consists of activity data for two panels playing important roles in toxicological pathways. Nuclear Receptor Signaling Panel (nr) included activity data for seven targets: aryl hydrocarbon receptor (ahr), androgen receptor—full length (ar) and Ligand Binding Domain (ar-lbd), aromatase, estrogen receptor alpha—full length (er) and Ligand Binding Domain (er-lbd) and peroxisome proliferator-activated receptor gamma (ppar-gamma). Stress Response Panel (sr) included data for five targets: nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (are), ATAD5, heat shock factor response element (hse), the disruption of mitochondrial membrane potential (mmp) and p53. Great emphasis is laid upon the initial performance benchmark of the various combinations of attribute selection methods and classification algorithms. Two-dimensional molecular descriptors set and dictionary-based fingerprints enable fast screening and compound prioritization even for very large databases. All software used during this study is freely available and open source, making the presented approach widely available for the scientific community.

MATERIALS AND METHODS

The training dataset provided by the Challenge organizers (https://tripod.nih.gov/tox21/challenge/data.jsp) consisted of the activity data for ∼10 k compounds (Tox21 10 K compound library, structures provided as SMILES) on 12 targets, with the activity class assigned "Active" or "Not active" (for discussions of activity call procedures, see Shockley, 2012; Tice et al., 2013). The Testing dataset, provided later by the Challenge organizers consisted of activity data for 269 compounds. The final predictions were performed on the evaluation set of 647 compounds with unknown activity.

All calculations were performed on the desktop computer with Intel Core i7-4770 K CPU processor (eight cores) and 16 GB RAM, running Ubuntu 12.04.5 LTS.

Structures Standardization and Preprocessing

The chemical structures in the provided Tox21 Challenge data sets were standardized using the LyChI (Layered Chemical Identifier) program (version 20141028, https://github.com/ncats/lychi). Compounds with ambiguous structure (compound identifier with more than one chemical structure assigned) or activity (compound identifier with activity labels "Active" and "Not active" on a single target) were excluded using KNIME GroupBy node (KNIME 2.10.4, http://www.knime.org/; Berthold et al., 2007). For each compound, only the biggest component was preserved (KNIME component Separator node). For each target, data set was downsized such that the activity values were evenly distributed—all records from the minority class were retained and a random sample from the majority class was added (KNIME Row Sampling node). Standardized and downsized datasets used for modeling are available as Supplementary Materials.

Descriptors Generation

For standardized data sets, two-dimensional molecular descriptors were calculated using KNIME nodes: RDKit (http://rdkit.org/, 117 descriptors), CDK (Beisken et al., 2013; http://sourceforge.net/projects/cdk/, 97 descriptors) and fingerprints [PubChem (881 bits) and MACCS (167 bits)], giving 1262 descriptors for each compound. For the list of used descriptors and literature references see Supplementary Table S5. For each target, Arff weka file was created using KNIME Arff Writer node.

Classification Algorithms Screen

Preprocessing and classification algorithms screen was performed in the Weka Experiment Environment (Weka 3.6.6, Hall et al., 2009), with 10-fold cross validation with 10 repetitions. In each run, data was preprocessed with Remove Useless filter (all constant attributes are deleted, along with any that exceed the maximum percentage of variance, set to 99%) and Standardize filter (standardizes all numeric attributes to have zero mean and unit variance). Attribute selection was performed with two search methods: Best First and Rank Search, with CfsSubset attribute evaluator. Machine learning algorithms tested were: ADTree (alternating decision tree), FT (functional trees), FURIA (Fuzzy Unordered Rule Induction Algorithm), IBk (*k*-nearest neighbors), J48, Naïve Bayes, REPTree, and SMO (sequential minimal optimization for training a support vector classifier). Ensemble methods tested in the second step of the screen were: Rotation Forest, Decorate, Dagging, Bagging and AdaBoost M1. Unless otherwise stated, all algorithms were used with default settings. The performance of the models was measured using area under the receiver operating characteristic (ROC) curve metrics (AUROC).

Predictions

The final models were built in KNIME with Weka 3.6 nodes, using the Best First attribute selection method with Rotation Forest/ADTree classifier (for parameters of the classifier see Supplementary Table S6). For each target, 10 models were built using randomly selected subset of 95% of training set. Each model was evaluated on the remaining 5% of the training set and on the testing set. The model with the best AUROC value was selected for the final predictions. The estimation of probability of a chemical being active was rounded to three decimal places.

RESULTS AND DISCUSSION

The data processing workflow is shown in **Figure 1**. It involved six main steps: data preprocessing, descriptors calculation, feature selection and classification algorithms screen, training, testing, and predictions.

Data Preprocessing

The first stage of data preprocessing included data sanitization. First, SMILES were standardized with the LyChi program. For the training dataset, out of 11,764 unique input compounds, 9231 (78%) had fixed structure. Among the most frequent modifications were: unifying aromaticity model, neutralization and small counterions removal. Next, structures containing more than one component were separated and only the biggest component was preserved. This was the most vague reduction of the initial data, but this step was necessary for proper descriptors calculations. Also, an analysis of the most frequently removed components showed that these were mainly inorganic acids, metal ions and water molecules (see **Table 1**), which are frequent components of pharmaceutical mixtures and should not be treated as a factors determining activity on investigated targets. Finally, each subset of the training data set was downsized such that the activity values are equally distributed. The selection of the majority class members (inactives) was random (see Sections Structures Standardization and Preprocessing: Materials and Methods), which means that the output from this step could influence the results of further predictions. Here, the downsizing was a single-time procedure and the influence of various sets of majority class on models' performance was not investigated. For the initial and final compositions of the training data set (see **Table 2**).

Molecular Descriptors Calculation

Generation of higher-dimensional molecular descriptors (3D, 4D, 5D) is time consuming and may be prone to conformer generation errors. To avoid these shortcomings, low-dimensional (0D, 1D, 2D) descriptors and dictionary-based fingerprints were

TABLE 1 | Top 10 most frequently removed minor components from an initial training data set.

Removed component	Count	% of all removed components
HCl	955	32.6
Na^+	533	18.2
H_2O	254	8.7
Cl^-	157	5.4
Br^-	110	3.8
Sulphuric acid	83	2.8
Methylsulfonic acid	54	1.8
K^+	50	1.7
Maleic acid	47	1.6
I^-	41	1.4

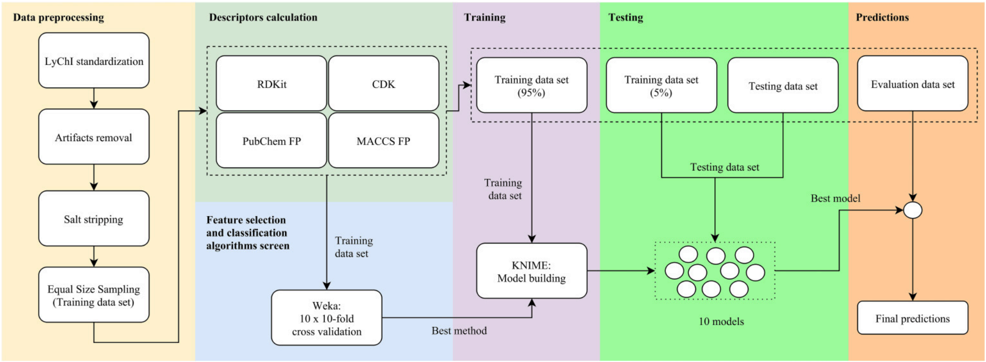


FIGURE 1 | Activity prediction workflow.

TABLE 2 | Initial and final training data sets composition.

Target	Initial training data set			Preprocessed training data set		
	Data set size	Actives count	% actives	Data set size	Actives count	% actives
nr-ahr	8169	950	11.6	1900	950	50.0
nr-ar	9362	380	4.1	756	378	50.0
nr-ar-lbd	8599	303	3.5	604	302	50.0
nr-aromatase	7226	360	5.0	712	356	50.0
nr-er	7697	937	12.2	1866	933	50.0
nr-er-lbd	8753	446	5.1	882	441	50.0
nr-ppar-gamma	8184	222	2.7	442	221	50.0
sr-are	7167	1098	15.3	2188	1094	50.0
sr-atad5	9091	338	3.7	674	337	50.0
sr-hse	8150	428	5.3	850	425	50.0
sr-mmp	7320	1142	15.6	2246	1123	50.0
sr-p53	8634	537	6.2	1064	532	50.0

used here. It was shown earlier that such descriptors may carry the similar information-level to higher dimensional ones (Estrada et al., 2001; Oprea, 2002; Roy and Das, 2014) and can be successfully used in building predictive QSAR models (Roy and Roy, 2009; Garcia et al., 2011; Chavan et al., 2014; Su et al., 2015).

Feature Selection and Classification Algorithms Screen

Various attribute selection, data preprocessing and classification algorithms are available (Witten et al., 2011). It is not known *a priori* which combination of the above is optimal for the problem under consideration, as for different data sets the accuracy of algorithms varies (Smusz et al., 2013). This is why an initial methods assessment was conducted, evaluating the performance (expressed as the AUROC value) of the combination of:

- Attribute selection methods: two search methods were evaluated: Best First and Rank Search
- Classifiers: 14 classifiers setups were evaluated

Most classifiers were used with default settings. For IBk, four values of k were probed (1, 3, 5, and 10), as this parameter may significantly influence the performance of this classifier. SMO algorithm was probed with three kernels (RBF kernel, polynomial kernel, and normalized polynomial kernel). To validate various modeling approaches, a 10-fold cross validation with 10 repetitions was used. In each run, training data were preprocessed independently (removal of a constant attribute, data standardization, attribute selection). This allowed an estimation of how the procedures under the investigation will generalize to an independent data set. Results of the initial evaluation are summarized in **Figure 2** (for values obtained in the initial methods evaluation see Supplementary Table S1).

As expected, the performance of evaluated classifiers varied. For the tested set of the descriptors, among the best performing ones were ADTree, IBk, and Naïve Bayes. Performance of IBk classifier varied slightly for various values of k, with better AUROC values for the higher k (5 and 10). The worst performance was observed for SMO (Sequential Minimal Optimization). However, the parameters for these methods (C, gamma) were not optimized and certainly such optimization would increase their performance. As for the attribute selection methods, in most cases there were no significant differences in performance between algorithms. The exception is the Naïve Bayes classifier, where the differences are substantial. Generally, the Best First method was slightly better than Rank Search (mean AUROC for all experiments: 0.778 ± 0.056 and 0.768 ± 0.055, respectively). In the studied descriptors space, the overall "target predictability" also varied. The sr-mmp and nr-ar-lbd are "the most predictable" targets while sr-hse and nr-er are "the least predictable" ones. The latter observation may be caused by the insufficient descriptive power of calculated molecular features to describe the nature of binding small molecule ligands to these targets.

After initial algorithms screen, the four best performing methods (Naïve Bayes, ADTree, and IBk) were evaluated in combination with ensemble methods: Rotation Forest, Decorate, Dagging, Bagging, and AdaBoost. The SMO classifier was treated as the "negative control." The Best First attribute selection method was used. Results are summarized in **Figure 3** (for AUROC values obtained in this experiment see Supplementary Table S2).

The application of the ensemble methods in most cases caused increase of the obtained AUROC values. The average AUROC for all targets for Naïve Bayes classifier increased from 0.79 to 0.80 (when combined with Bagging, Dagging, Decorate and Rotation Forest) but decreased to 0.78 in case of AdaBoostM1. For ADTree, the AUROC values increased from the initial 0.79–0.82 (in combination with Decorate) and 0.83 (for Rotation Forest). For comparison of the performance of the selected ensemble classifiers see Supplementary Table S4. The best and most stable performance for all targets was observed for Rotation Forest ensemble method with two classifiers: ADTree and IBk ($k = 10$) (Mean AUROC for all experiments: 0.831 ± 0.038 and 0.820 ± 0.038 respectively). Based on these results, the Best First attribute selection method with Rotation Forest/ADTree classifier was used for the final activity predictions for all targets.

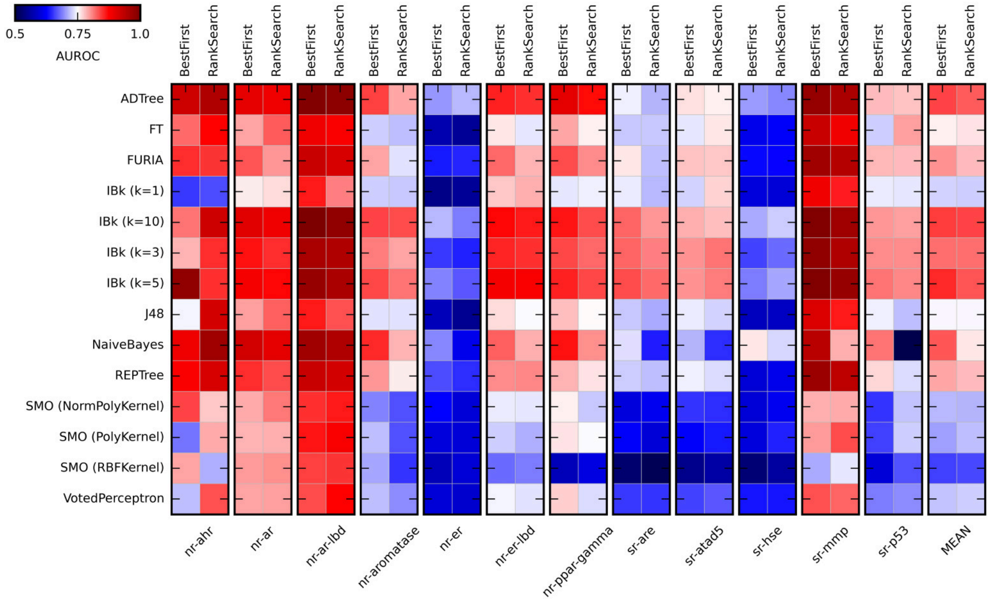


FIGURE 2 | Heat maps presenting results of the initial methods evaluation. Color coded AUROC values are presented for 14 classifiers (Y axis) in combination with two attribute selection methods (top X axis), grouped by the target (down X axis). Additional group presenting mean AUROC values is added for classifiers comparison.

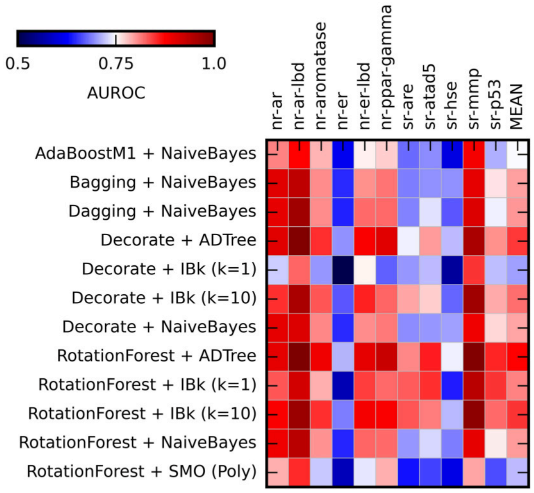


FIGURE 3 | Heat maps presenting results of the ensemble methods assessment. Color coded AUROC values for 12 classifiers (Y axis) for each challenge target (X axis) are shown. Aditional column presenting mean AUROC values is added for classifiers comparison.

TABLE 3 | AUROC values obtained for the best models selected for final predictions.

Target	AUROC testing		AUROC evaluation set
	Training set 5%	Testing set	
nr-ahr	0.92	0.84	0.89
nr-ar	0.76	0.50	0.73
nr-ar-lbd	0.91	0.82	0.79
nr-aromatase	0.92	0.79	0.78
nr-er	0.85	0.67	0.77
nr-er-lbd[a]	0.95	0.70	0.78
nr-ppar-gamma[a]	0.97	0.71	0.67
sr-are	0.87	0.80	0.72
sr-atad5[a]	0.91	0.65	0.76
sr-hse	0.90	0.74	0.80
sr-mmp	0.92	0.86	0.93
sr-p53	0.88	0.72	0.79

[a] *These models were not submitted to the final evaluation of the Tox21 Challenge.*

Training, Testing, and Final Predictions

For each target, 10 models were built using randomly selected subsets of 95% of the training set. Each model was tested on two sets: the remaining 5% of the training set and the provided testing set. The use of the 5%-random subset, apart from the constant testing set, helped to assure that the performance of the selected model is obtained not due to chance, but by merit inherent to the method. The model with the highest AUROC value was selected for the final predictions on the evaluation set. The performance on the testing and evaluation data sets of selected best models is summarized in **Table 3**. For AUROC statistics of all generated models see Supplementary Table S3.

The average AUROC value for the final predictions for all 12 targets was 0.784 ± 0.069. The best results were obtained for nr-ahr and sr-mmp (AUROC values: 0.89 and 0.93, respectively). The lowest AUROC value was obtained for nr-ppar-gamma (0.67), despite good performance of the model on the testing sets. As stated earlier, lower performance for some targets may be caused by the insufficient descriptive power of calculated molecular features to describe the complex nature of binding small molecule ligands to these targets.

In general, one can observe the correlation between AUROC values for testing and evaluation data sets. Most prominent examples include sr-mmp and nr-ahr (good performance in both testing and final evaluation) and nr-ar (moderate performance in both cases). On the other hand, for nr-ppar-gamma, the results obtained on the testing data sets are very good, while the final performance is moderate. In this case, one of the reasons could be that the chemical space of the evaluation set is out of the applicability domain of the selected model.

Computational Performance

Descriptors Calculation

The choice of low-dimensional descriptors guaranteed a high speed of calculations. A test run, carried for randomly selected 50 k clean drug like compounds fetched from ZINC database, showed a calculation rate at 12.65 s/1000 compounds (±1.33 s). The workflow for the descriptors calculation may be further optimized by applying a better parallelization scheme and by using all available CPUs on all stages of calculations.

Classification Performance

The biggest influence on the training time has the attribute selection step. Results from initial algorithms assessment (10-fold cross validation with 10 repetitions) shows that, for Best First, the average time of a single run was 10.197 ± 6.359 s, while for Rank Search it was 80.983 ± 66.302 s. Although the differences between these algorithms are high, in many cases training is a one-time procedure and training time is not a main factor for consideration. The average testing time for Best First method was 0.034 ± 0.063 s, while for Rank Search it was 0.119 ± 0.242 s. For the setup used for final evaluation (Best First attribute selection method with Rotation Forest/ADTree classifier) the average training time for all targets was 13.084 ± 8.627 s, while the testing time was 0.042 ± 0.033 s. For training and testing time values see Supplementary Tables S1, S2).

Related Works

Recently, a few papers describing various classification methods applied to the Tox21 dataset have been published. Drwal et al. described a successful approach of applying similarity comparison and machine learning for activity prediction (Drwal et al., 2015). These authors also used two dimensional descriptors sets in the form of 2929 bit-long bitvector, encoding molecular features, properties and connectivity information. The training dataset was enriched by adding activity data fetched from the literature (when available). Various parameters of similarity searching (Tanimoto fingerprint similarity to active or inactive compounds), of machine learning (Naïve Bayes) and of the combination of these methods were evaluated. The established methodology applied to the Tox21 dataset gave comparable results to the ones shown in this work (for four targets, the methods presented here gave better AUROC values, for two, the values were equal).

Deep learning methods were also applied to the Tox21 classification challenge. Unterthiner et al. used deep neural network with 40,000 input features describing molecules (Unterthiner et al., 2015). The presented scheme allowed the team to get the highest AUROC values in most of the Tox21 sub-challenges. The drawback of this methodology is the high demand for computational resources. Ramsundar et al. used simple two-dimensional descriptors and fingerprints in connection with Massively Multitask Networks (Ramsundar et al., 2015). Comparison to other classification algorithms (logistic regression, random forest) showed better performance for the deep learning method. Again, this methodology is computationally very expensive.

CONCLUSIONS

The presented method uses fast to calculate, two-dimensional descriptors and, in most cases, shows good predictive performance. Moreover, the use of free and open source tools makes the presented approach widely available for the community. To further improve the described workflow, a wider set of descriptors may be used, including fingerprints basing on connectivity information (like ECFP4 or Morgan fingerprints) or recently presented ToxPrint fingerprint, which cover substructures associated with toxicity (Yang et al., 2015). Also, other classification methods, including ensemble methods and deep learning techniques, should be investigated.

The Tox21 Data Challenge 2014 has offered the opportunity to compare and benchmark various approaches for toxicity prediction. The results clearly show that the very accurate *in silico* methods are now, or soon will be, at our fingertips. However, there is still a lot of work to be done to improve the quality of models to fully supersede traditional, *in vitro* assays.

ACKNOWLEDGMENTS

FS was supported by the European Research Council (ERC, StG grant RNA+P=123D project number 261351, to Janusz M. Bujnicki.). I would like to thank Krzysztof Szczepaniak, Grzegorz Lach, Catarina Almeida and Janusz M. Bujnicki for carefully reading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2015.00077

REFERENCES

Bakhtyari, N. G., Raitano, G., Benfenati, E., Martin, T., and Young, D. (2013). Comparison of *in silico* models for prediction of mutagenicity. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* 31, 45–66. doi: 10.1080/10590501.2013.763576

Beisken, S., Meinl, T., Wiswedel, B., De Figueiredo, L. F., Berthold, M., and Steinbeck, C. (2013). KNIME-CDK: workflow-driven cheminformatics. *BMC Bioinformatics* 14:257. doi: 10.1186/1471-2105-14-257

Berthold, M. C. N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., et al. (2007). "KNIME: the konstanz information miner," in *Studies in Classification, Data Analysis, and Knowledge Organization*, eds C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Heidelberg: Springer), 319–326.

Chavan, S., Nicholls, I. A., Karlsson, B. C., Rosengren, A. M., Ballabio, D., Consonni, V., et al. (2014). Towards global QSAR model building for acute toxicity: munro database case study. *Int. J. Mol. Sci.* 15, 18162–18174. doi: 10.3390/ijms151018162

Cheng, F., Li, W., Liu, G., and Tang, Y. (2013). *In silico* ADMET prediction: recent advances, current challenges and future trends. *Curr. Top. Med. Chem.* 13, 1273–1289. doi: 10.2174/15680266113139990033

Deeb, O., and Goodarzi, M. (2012). *In silico* quantitative structure toxicity relationship of chemical compounds: some case studies. *Curr. Drug Saf.* 7, 289–297. doi: 10.2174/157488612804096533

Drwal, M., Siramshetty, V., Banerjee, P., Goede, A., Preissner, R., and Dunkel, M. (2015). Molecular similarity-based predictions of the Tox21 screening outcome. *Front. Environ. Sci.* 3:54. doi: 10.3389/fenvs.2015.00054

Durrant, J. D., and Amaro, R. E. (2015). Machine-learning techniques applied to antibacterial drug discovery. *Chem. Biol. Drug Des.* 85, 14–21. doi: 10.1111/cbdd.12423

Erickson, B. E. (2011). Modernizing toxicity tests. *Chem. Eng. News* 89, 25–26. doi: 10.1021/cen-v089n029.p025

Estrada, E., Molina, E., and Perdomo-Lopez, I. (2001). Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? *J. Chem. Inf. Comput. Sci.* 41, 1015–1021. doi: 10.1021/ci000170v

Freitas, A. A., Limbu, K., and Ghafourian, T. (2015). Predicting volume of distribution with decision tree-based regression methods using predicted tissue:plasma partition coefficients. *J. Cheminform.* 7, 6. doi: 10.1186/s13321-015-0054-x

Garcia, I., Fall, Y., Garcia-Mera, X., and Prado-Prado, F. (2011). Theoretical study of GSK-3 alpha: neural networks QSAR studies for the design of new inhibitors using 2D descriptors. *Mol. Divers.* 15, 947–955. doi: 10.1007/s11030-011-9325-2

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations* 11, 10–18. doi: 10.1145/1656274.1656278

Huang, R. L., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., et al. (2014). Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep.* 4, 1664–1673. doi: 10.1038/srep05664

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., et al. (2010). *In vitro* screening of environmental chemicals for targeted testing prioritization: the toxcast project. *Environ. Health Perspect.* 118, 485–492. doi: 10.1289/ehp.0901392

Knight, A. (2007). Systematic reviews of animal experiments demonstrate poor human clinical and toxicological utility. *Altern. Lab. Anim.* 35, 641–659. Available online at: http://www.atla.org.uk/systematic-reviews-of-animal-experiments-demonstrate-poor-human-clinical-and-toxicological-utility/

Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715. doi: 10.1038/nrd1470

Liu, Y. (2015). Machine learning for drug design. *Int. J. Comput. Inf. Technol.* 4, 1–7.

Low, Y. S., Sedykh, A. Y., Rusyn, I., and Tropsha, A. (2014). Integrative approaches for predicting *in vivo* effects of chemicals from their structural descriptors and the results of short-term biological assays. *Curr. Top. Med. Chem.* 14, 1356–1364. doi: 10.2174/1568026614666140506121116

Martin, M. T., Knudsen, T. B., Reif, D. M., Houck, K. A., Judson, R. S., Kavlock, R. J., et al. (2011). Predictive model of rat reproductive toxicity from toxcast high throughput screening. *Biol. Reprod.* 85, 327–339. doi: 10.1095/biolreprod.111.090977

Omer, A., Singh, P., Yadav, N. K., and Singh, R. K. (2014). An overview of data mining algorithms in drug induced toxicity prediction. *Mini Rev. Med. Chem.* 14, 345–354. doi: 10.2174/1389557514666140219110244

Oprea, T. I. (2002). On the information content of 2D and 3D descriptors for QSAR. *J. Braz. Chem. Soc.* 13, 811–815. doi: 10.1590/s0103-50532002000600013

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery. *arXiv*:1502.02072. Available online at: http://arxiv.org/abs/1502.02072

Rouquié, D., Heneweer, M., Botham, J., Ketelslegers, H., Markell, L., Pfister, T., et al. (2015). Contribution of new technologies to characterization and prediction of adverse effects. *Crit. Rev. Toxicol.* 45, 172–183. doi: 10.3109/10408444.2014.986054

Roy, K., and Das, R. N. (2014). A review on principles, theory and practices of 2D-QSAR. *Curr. Drug Metab.* 15, 346–379. doi: 10.2174/1389200215666140908102230

Roy, P. P., and Roy, K. (2009). QSAR Studies of CYP2D6 inhibitor aryloxypropanolamines using 2D and 3D descriptors. *Chem. Biol. Drug Des.* 73, 442–455. doi: 10.1111/j.1747-0285.2009.00791.x

Shanks, N., Greek, R., and Greek, J. (2009). Are animal models predictive for humans? *Philos. Ethics Hum. Med.* 4:2. doi: 10.1186/1747-5341-4-2

Shockley, K. R. (2012). A three-stage algorithm to make toxicologically relevant activity calls from quantitative high throughput screening data. *Environ. Health Perspect.* 120, 1107–1115. doi: 10.1289/ehp.1104688

Sipes, N. S., Martin, M. T., Reif, D. M., Kleinstreuer, N. C., Judson, R. S., Singh, A. V., et al. (2011). Predictive models of prenatal developmental toxicity from toxcast high-throughput screening data. *Toxicol. Sci.* 124, 109–127. doi: 10.1093/toxsci/kfr220

Smusz, S., Kurczab, R., and Bojarski, A. J. (2013). A multidimensional analysis of machine learning methods performance in the classification of bioactive compounds. *Chemometr. Intel. Lab. Syst.* 128, 89–100. doi: 10.1016/j.chemolab.2013.08.003

Su, B. H., Tu, Y. S., Lin, O. A., Harn, Y. C., Shen, M. Y., and Tseng, Y. J. (2015). Rule-based classification models of molecular autofluorescence. *J. Chem. Inf. Model.* 55, 434–445. doi: 10.1021/ci5007432

Sun, H. M., Xia, M. H., Austin, C. P., and Huang, R. L. (2012). Paradigm shift in toxicity testing and modeling. *Aaps J.* 14, 473–480. doi: 10.1208/s12248-012-9358-1

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* 121, 756–765. doi: 10.1289/ehp.1205784

Toropov, A. A., Toropova, A. P., Raska, I. Jr., Leszczynska, D., and Leszczynski, J. (2014). Comprehension of drug toxicity: software and databases. *Comput. Biol. Med.* 45, 20–25. doi: 10.1016/j.compbiomed.2013.11.013

Unterthiner, T., Mayr, A., Klambauer, G., and Hochreiter, S. (2015). Toxicity prediction using deep learning. *arXiv*. Available online at: http://arxiv.org/abs/1503.01445

Valerio, L. G. Jr. (2013). Predictive computational toxicology to support drug safety assessment. *Methods Mol. Biol.* 930, 341–354. doi: 10.1007/978-1-62703-059-5_15

Witten, I. H., Frank, E., and Hall, M. A. (2011). "Data mining practical machine learning tools and techniques," in *Morgan Kaufmann Series in Data Management Systems, 3rd Edn.,* (Burlington, MA: Morgan Kaufmann Publishers).

Yang, C., Tarkhov, A., Marusczyk, J., Bienfait, B., Gasteiger, J., Kleinoeder, T., et al. (2015). New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J. Chem. Inf. Model.* 55, 510–528. doi: 10.1021/ci500667v

Conflict of Interest Statement: The author declares 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 Environmental Science

ORIGINAL RESEARCH
published: 02 March 2016
doi: 10.3389/fenvs.2016.00011

Predictive Toxicology: Modeling Chemical Induced Toxicological Response Combining Circular Fingerprints with Random Forest and Support Vector Machine

*Alexios Koutsoukas, Joseph St. Amand, Meenakshi Mishra and Jun Huan**

Department of Electrical Engineering and Computer Science, Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS, USA

OPEN ACCESS

Edited by:
Ruili Huang,
NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Xiaoqing Chang,
Integrated Laboratory System, USA
Jui-Hua Hsieh,
Kelly Government Solutions, USA

****Correspondence:***
Jun Huan
jhuan@ittc.ku.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *01 October 2015*
Accepted: *05 February 2016*
Published: *02 March 2016*

Citation:
Koutsoukas A, St. Amand J, Mishra M and Huan J (2016) Predictive Toxicology: Modeling Chemical Induced Toxicological Response Combining Circular Fingerprints with Random Forest and Support Vector Machine. Front. Environ. Sci. 4:11. doi: 10.3389/fenvs.2016.00011

Modern drug discovery and toxicological research are under pressure, as the cost of developing and testing new chemicals for potential toxicological risk is rising. Extensive evaluation of chemical products for potential adverse effects is a challenging task, due to the large number of chemicals and the possible hazardous effects on human health. Safety regulatory agencies around the world are dealing with two major challenges. First, the growth of chemicals introduced every year in household products and medicines that need to be tested, and second the need to protect public welfare. Hence, alternative and more efficient toxicological risk assessment methods are in high demand. The Toxicology in the 21st Century (Tox21) consortium a collaborative effort was formed to develop and investigate alternative assessment methods. A collection of 10,000 compounds composed of environmental chemicals and approved drugs were screened for interference in biochemical pathways and released for crowdsourcing data analysis. The physicochemical space covered by Tox21 library was explored, measured by Molecular Weight (MW) and the octanol/water partition coefficient (cLogP). It was found that on average chemical structures had MW of 272.6 Daltons. In case of cLogP the average value was 2.476. Next relationships between assays were examined based on compounds activity profiles across the assays utilizing the Pearson correlation coefficient r. A cluster was observed between the Androgen and Estrogen Receptors and their ligand bind domains accordingly indicating presence of cross talks among the receptors. The highest correlations observed were between NR.AR and NR.AR_LBD, where it was $r = 0.66$ and between NR.ER and NR.ER_LBD, where it was $r = 0.5$. Our approach to model the Tox21 data consisted of utilizing circular molecular fingerprints combined with Random Forest and Support Vector Machine by modeling each assay independently. In all of the 12 sub-challenges our modeling approach achieved performance equal to or higher than 0.7 ROC-AUC showing strong overall performance. Best performance was achieved in sub-challenges NR.AR_LBD, NR.ER_LDB and NR.PPAR_gamma, where ROC-AUC of 0.756, 0.790, and 0.803 was achieved accordingly. These results show that computational methods based on machine learning techniques are well suited to support and play critical role in toxicological research.

Keywords: *in silico*, toxicology, tox21 data challenge 2014, machine learning, data-mining, cheminformatics, predictive toxicology

INTRODUCTION

The average person is exposed to hundreds of chemicals not found naturally in the human organism during his lifespan. Xenobiotic man-made products can be found in wide range of cleaning and healthcare products, as food additives or drugs ingredients among others in various concentrations and mixtures. Advances in modern combinatorial chemistry have led to an unprecedented growth of synthetic chemicals availability on the market. Over the course of the last five decades the number of registered organic and inorganic substances in Chemical Abstract Service (CAS) Registry database grew well over 33 million, when in the 1965 the number was barely exceeding that of 200 thousands (Binetti et al., 2008).

Chemical toxicity may cause life-threating adverse effects on human health, therefore it is necessary to conduct regular risk assessments to ensure and protect public safety (Landrigan and Goldman, 2011). Hazardous toxicological effects on human health that may result due to short or chronic exposure to toxic chemicals include acute toxicity, toxicity to reproduction, mutagenicity and carcinogenicity (Binetti et al., 2008).

The traditional paradigm in toxicity testing consists of *in vivo* toxicology, where compounds are tested in various and usually high concentrations against tens or even hundreds of rodents or other animals (Merlot, 2010). This paradigm in toxicity testing is not feasible in modern toxicological research due to the large number of chemicals that need to be tested, the high cost of animal models, low throughput readouts, ethical issues, often contradictory findings and poor extrapolability to humans among others and have been extensively discussed in literature (Sun et al., 2012; Calafat et al., 2015).

Safety regulatory agencies are currently dealing with two major challenges. First, the increased number of chemicals that need to be tested for potential harmful effects on human health and second, the time and cost required to evaluate those chemicals (Hartung, 2009). Hence, novel and more efficient assessment methods for evaluation of potential toxicological effects are in high demand. Alternative avenues are currently being explored for chemical risk assessment using *in-vivo* and *in-vitro* approaches, such as human cell-based assays and high-throughput screening technologies (HTS; Ekins et al., 2005; Inglese et al., 2006; Shukla et al., 2010). Quantitative high-throughput screening (qHTS) technology has emerged as powerful and efficient way to alleviate limitations of single-point concentration HTS screening and allow to study complex toxicological mechanisms to specific pathways of targeted organs that may lead to disease (Inglese et al., 2006; Lock et al., 2012). qHTS is a titration-based screening approach that utilizes modern screening technologies, such as high-sensitivity detectors, low-volume dispensing and robotic plate handle (Inglese et al., 2006). As opposed to single-point concentration HTS screening, which typically suffers from large number of false positives and false negative readouts, qHTS is capable of identifying and efficiently elucidating structure-activity relationships (SARs) from primary screens. Furthermore, qHTS screening allows thousands compounds ($>10^4$ compounds) to be evaluated in different concentrations in cell models in an unprecedented rate (Schmidt, 2009; Attene-Ramos et al., 2013).

Computational approaches for modeling pharmacological and toxicological data combined with powerful data mining algorithms have been steadily gaining popularity by public and private bodies over the last decades (Muster et al., 2008; Kavlock and Dix, 2010). *In-silico* approaches utilize experimental data generated by *in-vivo* and *in-vitro* screening technologies and combined with cutting-edge data mining and cheminformatic techniques are capable of developing powerful predictive models. Such models could be applied to "virtually screen" thousands of chemicals for potential unwanted reactions early on during development cycles or to re-evaluate existing ones. *In silico* approaches can be applied to generate testable hypothesis for chemicals and direct experimentation toward the most likely unwanted interactions, which can then be validated or invalidated. Hence, *in-silico* approaches could become the "next big thing" as decision-making tools during the development and risk assessment stages. Therefore, computational approaches could provide more efficient utilization of the limited experimental resources.

The Toxicology in the 21st Century (Tox21) consortium is a major collaborative effort involving several agencies, the National Institutes of Health (NIH), the Environmental Protection Agency (EPA), and the Food and Drug Administration (FDA), was formed to develop and evaluate alternative risk assessment methods (Dix et al., 2007; Judson et al., 2009). A collection of 10,000 compounds composed of environmental chemicals and approved drugs was screened for interference in biochemical pathways of Nuclear and Stress receptor pathways and released for crowdsourcing data analysis.

The datasets released as part of the data challenge were generated by qHTS screening assays and contained compounds activity data against 12 assays, seven of which were part of the Nuclear Receptors (NR) and five of Stress Response (SR) pathways. Nuclear Receptors (NR) are an important family of transcription factors responsible for regulating gene expression and have a wide range of key roles in organisms' cell growth and proliferation, metabolism and homeostasis (Olefsky, 2001). Chemical interference by environmental pollutants or other xenobiotic chemicals can disturb homeostasis and lead to severe toxicities (Janošek et al., 2006). *In vivo* effects may range from male feminization to reproduction disorders and have been linked with chemical interference of NR (Baker, 2001). Structurally members of NR family present common features, which consist of a DNA binding domain (DBD), which recognize and bind to specific DNA sequences, and a ligand binding domain (LBD), which is located at the C-terminal half, and is responsible for recognizing and interacting with hormone molecules (Wurtz et al., 1996; Moras and Gronemeyer, 1998; Bourguet et al., 2000). NRs included in the challenge were Androgen Receptor (NR.AR), Androgen Receptor Ligand Binding Domain (NR.AR_LBD), Estrogen Receptor (NR.ER) and Estrogen Receptor Ligand Binding Domain (NR.ER_LBD), Aryl hydrocarbon Receptor (NR.AhR) and Peroxisome Proliferator-Activated Receptor gamma (NR.PPAR-γ). Aromatase (NR.Aromatase), member of the Cytochrome P450 protein family, responsible for the biosynthesis of estrogens, was the last included assay part of the NR pathway group (Simpson et al., 1994, 1997).

Cells respond to environmental stress factors, such as elevated and extreme temperature ranges, DNA damages, environmental and chemical toxicants and mechanical damages through a number of mechanisms that belong to Stress Response (SR) pathways (Fulda et al., 2010). Stress Response pathways are responsible for maintaining cell and tissue homeostasis. Five such biochemical assays were included in the challenge namely the ATPase family AAA domain-containing protein 5 (SR.ATAD5), which is involved in DNA damage response (Fox et al., 2012). Heat Shock response Elements (SR.HSE), which are proteins responsible for regulating the expression of heat shock genes (Wu, 1995). Mitochondrial Membrane Potential (SR.MMP) assays are used to evaluate chemically induced mitochondrial toxicity (Varga et al., 2015). Mitochondrial membrane potential changes are commonly measured using fluorescent dyes tools and are linked with cell capacity to generate ATP (Perry et al., 2011). Tumor suppressor protein (SR.p53), typically the p53 pathway is "off" and is activated when cells are under stress or damaged, hence being a good indicator of DNA damage and other cellular stresses (Vogelstein et al., 2000). Tumor suppressor protein p53 is activated by inducing DNA repair, cell cycle arrest and apoptosis (Levine, 1997). The fifth and last SR assay was the antioxidant response element (SR.ARE) signaling pathway. SR.ARE is responsible for regulating the expression of genes in cells exposed to oxidative stress that can change the cellular redox statues (Nguyen et al., 2003).

First the distribution of physicochemical space covered by the Tox21 library by utilizing simple molecular descriptors, the molecular weight (MW) and the octanol/water coefficient (cLogP) were examined. This analysis was performed to obtain an overview of the physicochemical space covered by the Tox21 library and the overlap between the training and testing datasets released during the competition.

It's been shown that chemicals can be active against multiple targets simultaneously, which has been termed as "polypharmacology" (Keiser et al., 2007; Klabunde, 2007). One of the major limitations when analyzing public bioactivity datasets is data incompleteness, which results to sparse bioactivity matrices (Mestres et al., 2008). On the contrary the Tox21 dataset provides a less incomplete bioactivity matrix across the 12 tested assays allowing such analysis to be carried out. The goal here was to investigate relationships between assays in bioactivity space based on the reported chemicals activities across the assays.

Our approach to model the Tox21 data consisted of utilizing circular molecular fingerprints combined with Random Forest and Support Vector Machines by modeling each assay independently. Circular fingerprints were selected for the study as they have been previously shown to perform well in virtual screening applications (Bender, 2010; Hu et al., 2012; Cereto-Massagué et al., 2015). As machine learning techniques two well-established algorithms in the field of cheminformatics were selected and applied, namely the Random Forest (RF) and the Support Vector Machine (SVM). Since their introduction to the field of molecular modeling they have both been successfully applied for a wide range of modeling tasks ranging from virtual screening (Koutsoukas et al., 2011), QSARs/QSPRs (Dudek et al., 2006; Guha, 2008) and to more recent proteocheometric modeling tasks (van Westen et al., 2011). Random Forest (RF), developed by Breiman, is an ensemble of unpruned classification or regression tress formed by applying bootstrap samples of the training data and random features selection in tree induction (Breiman, 2001). On the other hand, the Support Vector Machine (SVM), developed by Cortes and Vapnik, is a non-probabilistic kernel-based supervised learning method that maps input vectors into high-dimensional feature space where the decision hyperplane is constructed (Cortes and Vapnik, 1995). Our main hypothesis was that utilizing circular fingerprints combined with supervised machine learning methods would allow us to develop fast and accurate predictive models well suited for predictive toxicology.

MATERIALS AND METHODS

In total three datasets were released by the Tox21 data challenge team during the competition: The training set which was designated to serve for model development and hyper-parameters tuning, which from now on will be referred as Tox21_10k, and contained initially 11,764 structures covering activity measurements against 12 assays. The first released test set, which was used to rank teams submissions during the early phase of the competition, which from now on will be referred as Tox21_LDB, and contained 296 structures. The final released dataset was the external validation set, this dataset was used for the final phase of the competition for model evaluation and ranking teams' submissions, which from now on will be referred as Tox21_Ext_Valid, and contained 647 structures. Compounds activities for the external dataset were made publicly available only after the completion of the competition. Final teams submissions were evaluated based on the generated predictions on the external set Tox21_Ext_Valid set.

Data Preprocessing

Prior to modeling steps the datasets were pre-processed and chemical structures standardized with the aim of retaining only suitable structures for the following modeling steps. The importance of data curation prior to modeling steps has been extensively discussed in Fourches et al. (2010). Chemical structures were standardized using the ChemAxon Standardizer software package and stored in SDF (ChemAxon Standardizer, 2014) with the options on: (i) remove salts and solvents, (ii) disconnect metal atoms, (iii) remove fragments (keep largest ones), (iv) add explicit hydrogens, (v) aromatize, (vi) neutralize, (vii) tautomerize, (viii) mesomerize, the protocol utilized is provided in the Supplementary Material named "Stand_Prot.xml."

The number of unique structures in the Tox21_10k was measured to be 7,502 from the initial 11,764 and 295 for the Tox21_LBD following the standardization process. Following the structure standardization steps compounds activities were normalized by applying the majority rule based on standardized SMILES strings on a per assay basis. In cases where multiple activities were reported against a assay the activity with the most occurrences was retain, else were discarded as ambiguous. Instances where only a single activity was present were retained. Those cases could be attributed to variances in experimental conditions, concentrations, levels of purity and different vendors used, as was also stated by the Tox21 Team during the competition. The number of total instances, the number of active and inactive compounds as also the ratio of inactive/active per assay is shown on **Table 1**. The number of total instances per assay ranged from 5,747 for NR.Aromatase and up to 6,950 for NR.AR. The ratio of inactive/active instances per assay ranged from 5.5, relatively imbalanced, for SR.ARE and SR.MMP and up to 30.2, highly imbalanced, for NR.AR_LBD. The final datasets that resulted from the above described process are provided in the Supplementary Files "Sup_Tox21_10k" and "Sup_Tox21_LDB."

Molecular Descriptors

As molecular descriptors the Morgan Fingerprints (Circular Fingerprints) with radius 3 were utilized, which are equivalent to the extended connectivity fingerprints ECFP_6 (Rogers and Hahn, 2010), with diameter 6. The open source RDKit library (version 2014.09.1) was used to generate the molecular fingerprints from the standardized chemical structures (Landrum, 2015). The descriptors were generated as hashed binary vectors of 1,024 bits length. Morgan fingerprints were the only descriptor utilized during the modeling steps. Molecular Weight (MW) and the octanol/water partition coefficient (cLogP) were calculated using the MOE software package and used to examine and visualize the physicochemical space covered by the Tox21 library (Chemical Computing Group Inc., 2015).

Modeling Approach

Following the data pre-processing the two datasets Tox21_10k and Tox21_LDB were merged to form one larger dataset that was used for model development and hyper-parameters tuning, shown in **Figure 3**. No external data outside of those provided by the Tox21 Challenge team were utilized in any step of the modeling process. RF and SVM were utilized as implemented in the open-source machine learning library Scikit-learn (Pedregosa et al., 2011).

Each assay/sub-challenge was modeled independently following a single-task approach. 10-fold cross-validation was applied to tune the hyper-parameters for each algorithm. As performance metric the area under the ROC curve (AUC) was used. Receiver Operating Characteristics (ROC) graphs are commonly used in machine learning to compare and visualize the performance of binary classifiers (Fawcett, 2006). The area under the ROC curve (ROC-AUC) of a classifier is a single scalar values represents expected performance, and is equal to the probability that the classifier will rank a random chosen positive instance higher than a negative instance (Bradley, 1997). AUC takes values between (0,1), where values equal to or smaller than 0.5 show that a classifier performs no better or worse than random, instead for values greater than 0.5 a classifier is expected to perform better than random.

In case of SVM the radial basis function "rbf" kernel was considered with values for Cost $\{10^3,10^2,10^1,1,10^{-1}\}$ and *gamma* $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. In case of Random Forest the values considered for the number of trees was {50, 100, 300, 500, 1000, 1500} and number of features in each split {log2, sqrt}. The best average AUC and the standard deviation observed over 10-fold cross validation per assay by RF and SVM during the hyper-parameter tuning are shown in **Figure 4**, named "Best RF 10-CV" and "Best SVM 10-CV" accordingly. The implementations used to tune RF and SVM using ROC-AUC as evaluation metric based on the Scikit-learn are provided in the Supplementary Material "RF_tune.py" and "SVM_tune.py."

RESULTS

First the chemical space covered by the Tox21 chemical library was examined by calculating and analyzing the distribution of Molecular Weight (MW) and cLogP for the library, shown in **Figure 1**. As mentioned earlier the total number of structures counted was 7,502 in Tox21_10k, 295 in Tox21_LBD and 647 in Tox21_Ext_Valid following the pre-processing steps. Here it was found that compounds had on average MW of 271.2 and a median of 244.3 Dalton with 50% of compounds having MW between 166 and 337 Dalton. In case of cLogP the average value was 2.41 and median 2.39, with 50% of compounds having values between 0.98 and 3.753. This analysis indicates that a large portion of compounds included in the Tox21 library represent chemicals with drug-like properties, although compounds with MW and cLogP values outside of those typically occupied by drug-like molecules are not rare, e.g., compounds with MW over 1,000 Daltons and cLogP lower than -1 or higher than 6.

Next the relationships between assays based on bioactivity profiles of the tested chemicals were examined. Relationships between assays were calculated utilizing the Pearson correlation coefficient *r* based on compounds activities across tested assays (Todeschini et al., 2012), shown in **Figure 2**. The analysis was generated using the R programming language (Ihaka and Gentleman, 1996) and the "corrplot" package for visualization, the R script is provided in the Supplementary Material "CorrelationAssaysPlot.R" (Wei, 2013). A cluster was formed between the Androgen and Estrogen Receptors and

TABLE 1 | Number of data points per assay obtained following the standardization process.

	NR-AR	NR-AR	NR-ER	NR-ER	NR	NR-AhR	NR-PPAR	SR-ARE	SR-MMP	SR-p53	SR-HSE	SR-
		-LBD		-LBD	-Aromatase		-gamma					ATAD5
Total	7202	6714	6107	6912	5747	6493	6429	5790	5770	6739	6430	7027
Active	252	215	650	290	274	733	175	896	890	412	316	263
Inactive	6950	6499	5457	6622	5473	5760	6254	4894	4880	6327	6114	6764
Ratio of Inactive/active	27.6	30.2	8.4	22.8	20.0	7.9	35.7	5.5	5.5	15.4	19.3	25.7

The dataset utilized for model development and hyper-parameter tuning resulted by merging the Tox21_10k and Tox_LDB datasets. The ratio of inactive/active instances per assay ranged from 5.5 in SR.ARE and SR.MMP and up to 30.2 in NR.AR_LBD.

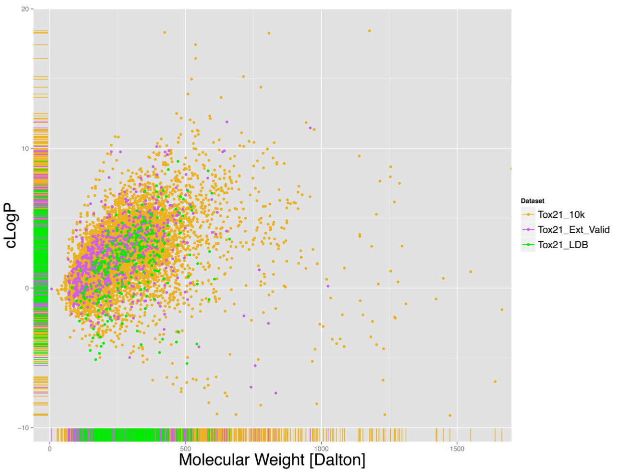


FIGURE 1 | Molecular Weight (MW) and cLogP distribution for the Tox21 chemical library, (gold) Tox21_10k, (green) Tox21_LDB and (purple) Tox21_Ext_Valid. For each color, the rug on x and y-axis represent the areas that are more densely populated. On average the compounds in the Tox21 library present MW of 272.6 and a median of 240.4 Daltons with 50% of compounds having MW between 166 and 339. In case of cLogP the average value was 2.476 and median 2.439, with 50% of compounds being between 1.045 and 3.811.

their ligand bind domains accordingly, which can be seen on the top-left corner of the **Figure 2**, indicating presence of cross-talks between the two receptors. The correlation between NR.AR and NR.AR_LBD was found to $r = 0.66$ and between NR.ER and NR.ER_LBD $r = 0.5$. Furthermore, correlation of $r = 0.39$ between the NR.AR_LBD and the NR.ER_LBD was observed, indicating that the two ligand binding domains share some structural similarities that can accommodate similar ligands. On the contrary, weak correlations were measured between the NR.AR and the NR.ER_LBD as also between the NR.ER and the NR.AR_LBD, where it was measured to be $r = 0.33$ and $r = 0.22$ accordingly. The rest receptors didn't show any correlation between them as the highest observed correlation didn't exceed of $r = 0.23$ between NR.ER_LBD and SR.p53 and of $r = 0.21$ between SR.p53 and SR.HSE.

Our group participated in the competition under team aliases frozenarm and ToxFit, where the first submission was based on the results obtained modeling the data using SVM and the latter based on RF independently. In all of the 12 sub-challenges our modeling approaches achieved performance of at least 0.7 ROC-AUC, only for the assay (SR.HSE) the results achieved by SVM were below 0.7 (0.689), showing strong overall performance, shown in **Table 2** and **Figure 4**. As expected the performance achieved by both algorithms during cross validation on the training set and on the external set were different, as shown in **Figure 4**, with the results achieved on the external dataset being lower. These observed differences could be attributed to several factors, e.g., structural differences between chemical space included in the training and test set, imbalances among inactive/active instances per assay and limitations of utilized molecular descriptors to capture complex chemical features responsible for the bioactivities. When comparing the results achieved by SVM and RF on the Tox21_Ext_Valid, as shown in **Table 2**, it can be seen that both algorithms achieved comparable results, with RF achieving slightly better ROC-AUC in 7 out of 12 tasks, while SVM in 4 out of 12, and in 1 task (NR.AR) where both algorithms achieved the same ROC-AUC of 0.744.

It worth noting that in our modeling approach no external data besides of those provided during the competition were utilized and only a single molecular descriptor was used, mainly due to time constrains during the competition. Utilizing external bioactivity data, e.g., from ChEMBL (Gaulton et al., 2012) or PubChem (Wang et al., 2009) databases, and additional molecular descriptors could potentially improve the performance of the models on the external evaluation set.

DISCUSSION

Chemical toxicological risk assessment is a necessary step to ensure public safety and to promote well-being. Potential hazardous side-effects should be detected as early as possible in order to allow informed decisions to be made regarding the future fate of those products. Computational approaches that combine experimental data generated by next generation of high-throughput screening technologies, such as qHTS, and powerful data mining techniques could provide valuable predictive systems for the identification of potential safety alerts for yet untested chemicals, while simultaneously reducing unnecessary animal testing. Furthermore, collaborative research initiatives such as the Toxicology in the 21st Century (Tox21)

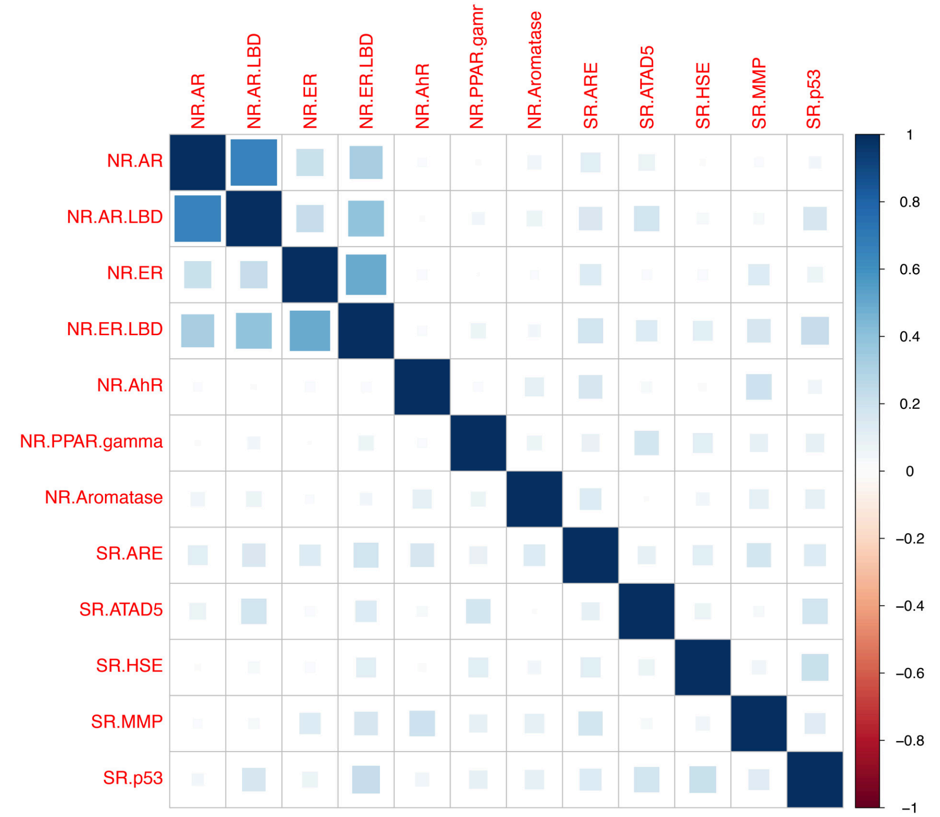


FIGURE 2 | Pearson correlation coefficient *r* between assays based on compounds activity profiles across the assays. A cluster was formed between the Androgen receptor (NR.AR) and the Androgen receptor ligand-binding domain (NR.AR_LBD), where *r* was measured 0.66. The Estrogen receptor (NR.ER) and the Estrogen receptor ligand-binding domain (NR.ER.LBD) had $r = 0.5$. Weaker correlation between the NR.AR_LBD and the NR.ER_LBD was observed (Pearson $r = 0.39$) showing possible crosstalk between the ligand binding domains. The rest measured correlation were low and did not exceed $r = 0.23$ between the NR.ER.LBD and SR.p53 and $r = 0.21$ between SR.HSE and SR.p53.

TABLE 2 | Performance achieved by our modeling approach using Random Forest (RF) and Support Vector Machine (SVM) measured by ROC-AUC per sub-challenge on the external To21_Ext_Val.

Algorithm	NR-AhR	NR-AR	NR-AR	NR	NR-ER	NR-ER	NR-PPAR	SR.ARE	SR.ATAD5	SR.HSE	SR.MMP	SR.p53
			-LBD	-Aromatase		-LBD	-gamma					
RF	**0.865**	**0.744**	0.722	**0.739**	**0.745**	**0.790**	**0.803**	**0.700**	0.726	**0.752**	0.859	0.802
SVM	0.861	**0.744**	**0.756**	0.738	0.729	0.752	0.791	0.697	**0.729**	0.689	**0.862**	**0.803**

Our team participated in the Tox21 data challenge 2014 under team aliases frozenarm and ToxFit. Performance achieved per assay ranged from 0.7 for the SR.ARE and up to 0.865 ROC-AUC for NR.AhR. Our best achieved performance per assay is indicated in bold.

consortium with the support of the research community could contribute toward the development of novel and powerful approaches for predictive toxicological research. These *in-silico* approaches could direct experimentation toward the most likely toxic chemicals first, hence providing a far better utilization of the limited experimental resources and ultimately leading to safer chemical products reaching the market or hazardous ones being removed from circulation.

The modeling approach devised by our team to model the Tox21 data challenge 2014 was based on simple circular molecular fingerprints and supervised machine-learning algorithms Random Forest and Support Vector Machine. Here a single task approach was followed, where each assay was modeled independently by RF and SVM. Overall the modeling approach achieved decent performance with results achieving strong performance measured by ROC-AUC equal to or higher than 0.7. The described approach has the advantage of being fast as it is based on simple circular descriptors, which can be generated efficiently for large number of chemical structures and utilized open-source software packages for the main modeling steps. As expected both algorithms selected for the study, RF

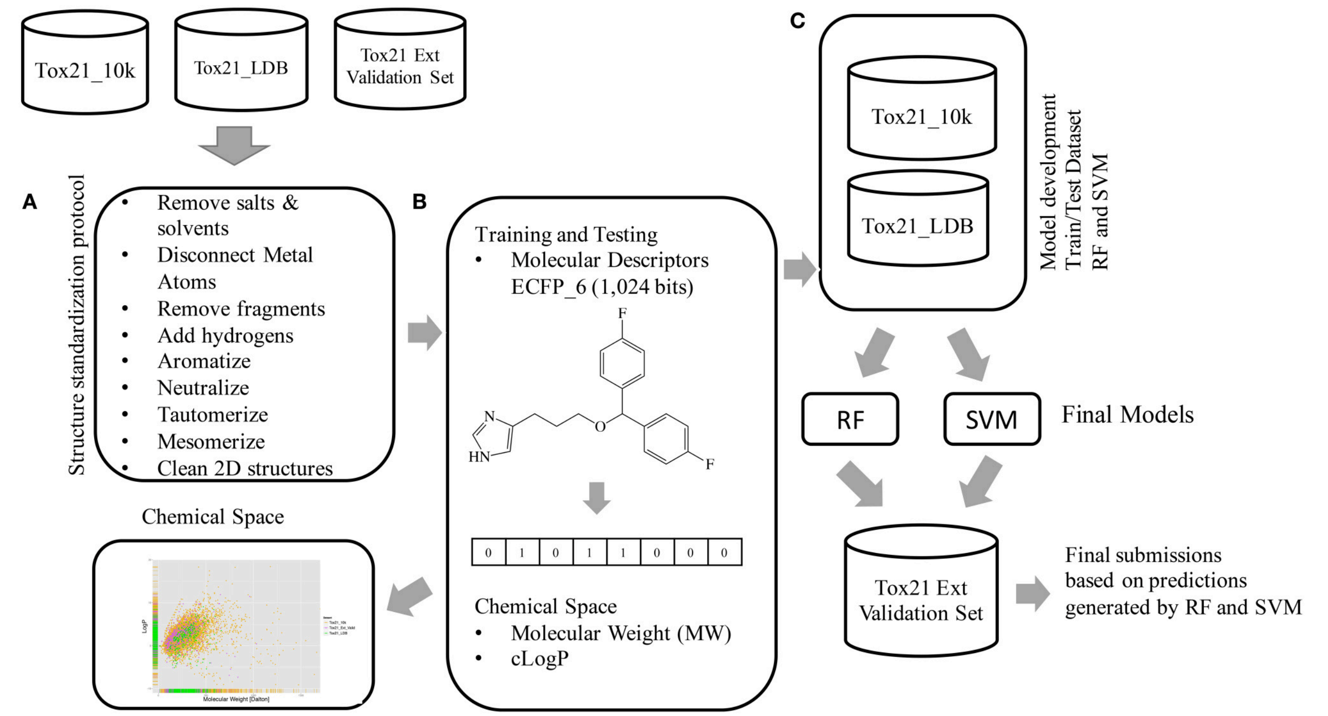


FIGURE 3 | Workflow followed to model the Tox21 dataset (the Tox21_10k dataset containing the initially released data for model development, the Tox21_LDB used for leaderboard ranking and the blind dataset Tox21 External Validation Set for final model evaluation). (A) Structure standardization step, **(B)** Activities normalization and molecular descriptors calculations. **(C)** Datasets Tox21_10k and Tox21_LDB were merged into one larger dataset that was subsequently used for model development and hyper-parameters tuning. Final models were evaluated based on the generated predictions on the blind Tox21_Ext_Val Set.

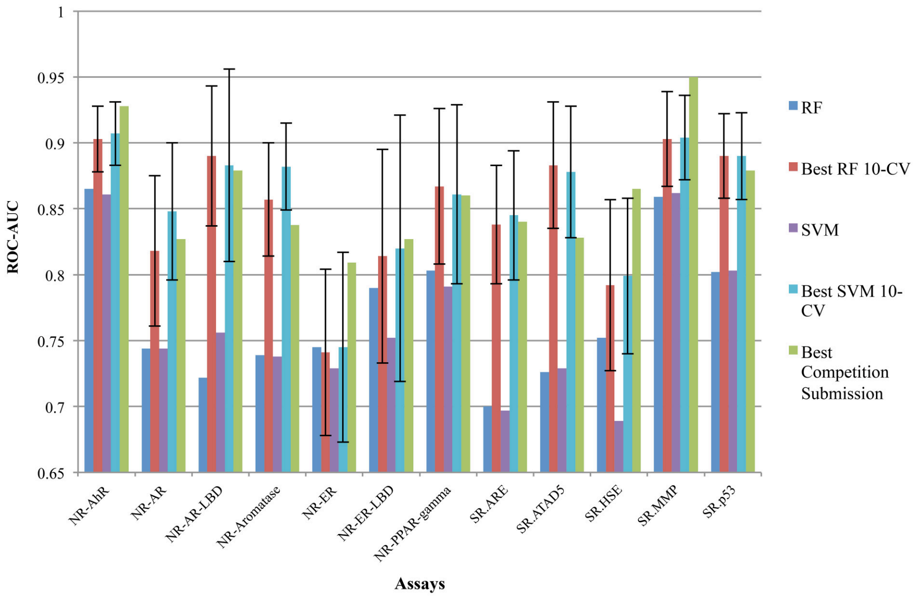


FIGURE 4 | Performance measured by ROC-AUC achieved by RF and SVM accordingly (shown as RF and SVM) per sub-challenge compared to the top-1 reported performance. The best-reported results of the competition per sub-challenge are shown as Best Competition Submission. Best performance achieved during hyper-parameters optimization for RF and SVM over 10-fold cross validation, both the best mean AUC and the standard deviation over the 10-folds are shown as Best RF 10-CV and Best SVM 10-CV, respectively. Our submissions for the To21_Ext_Valid ranged from 0.69 AUC in SR.HSE achieved by SVM, worse performance, and up to 0.865 for SR.AhR by RF, best performance.

and SVM, showed good performance and achieved comparable results on the external set.

AUTHOR CONTRIBUTIONS

The authors AK, JS, and MM designed and ran the experiments, analyzed the data and wrote the manuscript contributing equally. JH is the PI and contributed to experiment design, data analysis and writing the manuscript.

ACKNOWLEDGMENTS

This project was supported by NSF grant 1337899.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00011

REFERENCES

Attene-Ramos, M. S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R. J., et al. (2013). The Tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. *Drug Discov. Today* 18, 716–723. doi: 10.1016/j.drudis.2013.05.015

Baker, V. A. (2001). Endocrine disrupters—testing strategies to assess human hazard. *Toxicol. In vitro* 15, 413–419. doi: 10.1016/S0887-2333(01)00045-5

Bender, A. (2010). How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discov.* 5, 1141–1151. doi: 10.1517/17460441.2010.517832

Binetti, R., Costamagna, F. M., and Marcello, I. (2008). Exponential growth of new chemicals and evolution of information relevant to risk control. *Annali dell'Istituto Superiore di Sanita* 44, 13–15.

Bourguet, W., Germain, P., and Gronemeyer, H. (2000). Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* 21, 381–388. doi: 10.1016/S0165-6147(00)01548-0

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patt. Recognit.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

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

Calafat, A. M., Valentin-Blasini, L., and Ye, X. (2015). Trends in exposure to chemicals in personal care and consumer products. *Curr. Environ. Health Rep.* 2, 348–355. doi: 10.1007/s40572-015-0065-9

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71, 58–63. doi: 10.1016/j.ymeth.2014.08.005

ChemAxon Standardizer (2014). *ChemAxon Standardizer 14.10.6.0.*

Chemical Computing Group Inc. (2015). *Molecular Operating Environment (MOE)*, 2013.08, 1010, Montreal, QC.

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

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci. Off. J. Soc. Toxicol.* 95, 5–12. doi: 10.1093/toxsci/kfl103

Dudek, A., Arodz, T., and Galvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combin. Chem. High Through. Screen.* 9, 213–228. doi: 10.2174/138620706776055539

Ekins, S., Nikolsky, Y., and Nikolskaya, T. (2005). Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol. Sci.* 26, 202–209. doi: 10.1016/j.tips.2005.02.006

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

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. Model.* 50, 1189–1204. doi: 10.1021/ci100176x

Fox, J. T., Sakamuru, S., Huang, R., Teneva, N., Simmons, S. O., Xia, M., et al. (2012). High-throughput genotoxicity assay identifies antioxidants as inducers of DNA damage response and cell death. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5423–5428. doi: 10.1073/pnas.1114278109

Fulda, S., Gorman, A. M., Hori, O., and Samali, A. (2010). Cellular stress responses: cell survival and cell death. *Int. J. Cell Biol.* 2010, 1–23. doi: 10.1155/2010/214074

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–D1107. doi: 10.1093/nar/gkr777

Guha, R. (2008). On the interpretation and interpretability of quantitative structure–activity relationship models. *J. Comput. Aided Mol. Des.* 22, 857–871. doi: 10.1007/s10822-008-9240-5

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature* 460, 208–212. doi: 10.1038/460208a

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

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* 5, 299–314.

Inglese, J., Auld, D. S., Jadhav, A., Johnson, R. L., Simeonov, A., Yasgar, A., et al. (2006). Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11473–11478. doi: 10.1073/pnas.0604348103

Janošek, J., Hilscherová, K., Bláha, L., and Holoubek, I. (2006). Environmental xenobiotics and nuclear receptors—Interactions, effects and *in vitro* assessment. *Toxicol. In vitro* 20, 18–37. doi: 10.1016/j.tiv.2005.06.001

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., et al. (2009). *In vitro* screening of environmental chemicals for targeted testing prioritization: the toxcast project. *Environ. Health Pers.* 118, 485–492. doi: 10.1289/ehp.0901392

Kavlock, R., and Dix, D. (2010). Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *J. Toxicol. Environ. Health B* 13, 197–217. doi: 10.1080/10937404.2010.483935

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

Klabunde, T. (2007). Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* 152, 5–7. doi: 10.1038/sj.bjp.0707308

Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P. J., Whitmore, A. V., Zimmer, S., et al. (2011). From *in silico* target prediction to multi-target drug design: current databases, methods and applications. *J. Proteomics* 74, 2554–2574. doi: 10.1016/j.jprot.2011.05.011

Landrigan, P. J., and Goldman, L. R. (2011). Children's vulnerability to toxic chemicals: a challenge and opportunity to strengthen health and environmental policy. *Health Aff.* 30, 842–850. doi: 10.1377/hlthaff.2011.0151

Landrum, G. (2015). *RDKit: Open-Source Cheminformatics*. Available online at: http://www.rdkit.org

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell* 88, 323–331. doi: 10.1016/S0092-8674(00)81871-1

Lock, E. F., Abdo, N., Huang, R., Xia, M., Kosyk, O., O'Shea, S. H., et al. (2012). Quantitative high-throughput screening for chemical toxicity in a population-based *in vitro* model. *Toxicol. Sci. Off. J. Soc. Toxicol.* 126, 578–588. doi: 10.1093/toxsci/kfs023

Merlot, C. (2010). Computational toxicology–a tool for early safety evaluation. *Drug Discov. Today* 15, 16–22. doi: 10.1016/j.drudis.2009.09.010

Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. (2008). Data completeness—the Achilles heel of drug-target networks. *Nat. Biotechnol.* 26, 983–984. doi: 10.1038/nbt0908-983

Moras, D., and Gronemeyer, H. (1998). The nuclear receptor ligand-binding domain: structure and function. *Curr. Opin. Cell Biol.* 10, 384–391. doi: 10.1016/S0955-0674(98)80015-X

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Muller, L., and Pahler, A. (2008). Computational toxicology in drug development. *Drug Discov. Today* 13, 303–310. doi: 10.1016/j.drudis.2007.12.007

Nguyen, T., Sherratt, P. J., and Pickett, C. B. (2003). Regulatory mechanisms controlling gene expression mediated by the antioxidant response element. *Annu. Rev. Pharmacol. Toxicol.* 43, 233–260. doi: 10.1146/annurev.pharmtox.43.100901.140229

Olefsky, J. M. (2001). Nuclear receptor minireview series. *J. Biol. Chem.* 276, 36863–36864. doi: 10.1074/jbc.R100047200

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.

Perry, S., Norman, J., Barbieri, J., Brown, E., and Gelbard, H. (2011). Mitochondrial membrane potential probes and the proton gradient: a practical usage guide. *BioTechniques* 50, 98–115. doi: 10.2144/000113610

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

Schmidt, C. W. (2009). TOX21 new dimensions of toxicity testing. *Environ. Health Pers.* 117, A348–A353. doi: 10.1289/ehp.117-a348

Shukla, S. J., Huang, R., Austin, C. P., and Xia, M. (2010). The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov. Today* 15, 997–1007. doi: 10.1016/j.drudis.2010.07.007

Simpson, E. R., Mahendroo, M. S., Means, G. D., Kilgore, M. W., Hinshelwood, M. M., Graham-Lorence, S., et al. (1994). Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. *Endocr. Rev.* 15, 342–355.

Simpson, E. R., Zhao, Y., Agarwal, V. R., Michael, M. D., Bulun, S. E., Hinshelwood, M. M., et al. (1997). Aromatase expression in health and disease. *Recent Prog. Horm. Res.* 52, 185–213. discussion: 213–214.

Sun, H., Xia, M., Austin, C. P., and Huang, R. (2012). Paradigm shift in toxicity testing and modeling. *AAPS J.* 14, 473–480. doi: 10.1208/s12248-012-9358-1

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J. Chem. Inf. Model.* 52, 2884–2901. doi: 10.1021/ci300261r

van Westen, G. J. P., Wegner, J. K., Ijzerman, A. P., van Vlijmen, H. W. T., and Bender, A. (2011). Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.* 2, 16–30. doi: 10.1039/C0MD00165A

Varga, Z. V., Ferdinandy, P., Liaudet, L., and Pacher, P. (2015). Drug-induced mitochondrial dysfunction and cardiotoxicity. *Am. J. Physiol. Heart Circul. Physiol.* 309, H1453–H1467. doi: 10.1152/ajpheart.00554.2015

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. *Nature* 408, 307–310. doi: 10.1038/35042675

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

Wei, T. (2013). *corrplot: Visualization of a Correlation Matrix*. R package version 0.73. Available online at: http://CRAN.R-project.org/package=corrplot

Wu, C. (1995). Heat shock transcription factors: structure and regulation. *Annu. Rev. Cell Dev. Biol.* 11, 441–469. doi: 10.1146/annurev.cb.11.110195.002301

Wurtz, J.-M., Bourguet, W., Renaud, J.-P., Vivat, V., Chambon, P., Moras, D., et al. (1996). A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Biol.* 3, 87–94. doi: 10.1038/nsb0196-87

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 Environmental Science

ORIGINAL RESEARCH
published: 08 March 2016
doi: 10.3389/fenvs.2016.00012

Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data

*Kathryn Ribay[1], Marlene T. Kim[1, 2], Wenyi Wang[2], Daniel Pinolini[2] and Hao Zhu[1, 2]**

[1] Department of Chemistry, Rutgers University, Camden, NJ, USA, [2] The Rutgers Center for Computational and Integrative Biology, Camden, NJ, USA

Estrogen receptors (ERα) are a critical target for drug design as well as a potential source of toxicity when activated unintentionally. Thus, evaluating potential ERα binding agents is critical in both drug discovery and chemical toxicity areas. Using computational tools, e.g., Quantitative Structure-Activity Relationship (QSAR) models, can predict potential ERα binding agents before chemical synthesis. The purpose of this project was to develop enhanced predictive models of ERα binding agents by utilizing advanced cheminformatics tools that can integrate publicly available bioassay data. The initial ERα binding agent data set, consisting of 446 binders and 8307 non-binders, was obtained from the Tox21 Challenge project organized by the NIH Chemical Genomics Center (NCGC). After removing the duplicates and inorganic compounds, this data set was used to create a training set (259 binders and 259 non-binders). This training set was used to develop QSAR models using chemical descriptors. The resulting models were then used to predict the binding activity of 264 external compounds, which were available to us after the models were developed. The cross-validation results of training set [Correct Classification Rate (CCR) = 0.72] were much higher than the external predictivity of the unknown compounds (CCR = 0.59). To improve the conventional QSAR models, all compounds in the training set were used to search PubChem and generate a profile of their biological responses across thousands of bioassays. The most important bioassays were prioritized to generate a similarity index that was used to calculate the biosimilarity score between each two compounds. The nearest neighbors for each compound within the set were then identified and its ERα binding potential was predicted by its nearest neighbors in the training set. The hybrid model performance (CCR = 0.94 for cross validation; CCR = 0.68 for external prediction) showed significant improvement over the original QSAR models, particularly for the activity cliffs that induce prediction errors. The results of this study indicate that the response profile of chemicals from public data provides useful information for modeling and evaluation purposes. The public big data resources should be considered along with chemical structure information when predicting new compounds, such as unknown ERα binding agents.

Keywords: QSAR modeling, estrogen receptor α, bioassay profiling, endocrine disrupting chemicals, biosimilarity

OPEN ACCESS

Edited by:
*Juergen Pilz,
Alpen-Adria Universitaet Klagenfurt,
Austria*

Reviewed by:
*Ijaz Hussain,
Quaid-i-Azam University, Pakistan
Venkata Krishna Jandhyala,
Washington State University, USA*

***Correspondence:**
*Hao Zhu
hao.zhu99@rutgers.edu*

Specialty section:
*This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Environmental Science*

***Received:** 30 September 2015*
***Accepted:** 19 February 2016*
***Published:** 08 March 2016*

Citation:
*Ribay K, Kim MT, Wang W, Pinolini D
and Zhu H (2016) Predictive Modeling
of Estrogen Receptor Binding Agents
Using Advanced Cheminformatics
Tools and Massive Public Data.
Front. Environ. Sci. 4:12.
doi: 10.3389/fenvs.2016.00012*

INTRODUCTION

Estrogen receptors are cellular proteins that are activated when bound to estrogen molecules. When activated, estrogen receptors trigger the expression of gene products crucial to the endocrine system (Hall et al., 2001). These receptors can also be activated by certain endocrine disrupting chemicals (EDC), resulting in a disruption of normal estrogen signaling (Shanle and Xu, 2011). There are two unique estrogen receptors: ERα and ERβ. These two receptors are highly similar in the DNA binding domain, but differ more significantly in other regions. While there are many EDC that interact with both receptors, the difference between these two receptors allows some ligands specifically bind to only one receptor as well. Among all known binding agents, the ERα binders are much better characterized than ERβ binders (Hall et al., 2001; Shanle and Xu, 2011). Due to the nature of available data, this study focuses solely on ligands binding to ERα.

When estrogen receptors are activated by small molecules other than estrogens, the expression of the associated genes is deregulated leading to neurological, developmental, and reproductive toxicity (Mueller and Korach, 2001). There are many small molecules with different chemical structures which exhibit interaction with the ligand binding domain of the estrogen receptor (Blair et al., 2000; Schug et al., 2011). Considering the large number of compounds which needs to be evaluated for their estrogen receptor binding potentials, traditional experimental toxicology protocols can be costly and time-consuming. As a result, there is a strong need to effectively pre-screen and prioritize small molecules for potential endocrine disruption prior to more costly animal testing. In a 2007 publication, the U.S. National Research Council identified both high-throughput screening (HTS) and computational models as critical chemical toxicity evaluation tools in Twenty-First century toxicology (Committee on Toxicity Testing and Assessment of Environmental Agents N.R.C., 2007). HTS has been viewed as a potential alternative to animal models due to the ability to test many molecules at a rapid pace and lower cost. The large number of HTS studies has resulted in publically available bioassay databases which are a rich source of *in vitro* data (Zhu et al., 2014). Motivated by these available data, computational modeling, which costs even less than HTS, has been used as another important evaluation protocols for EDCs (Ding et al., 2010).

Quantitative structure-activity relationship (QSAR) modeling has been applied to develop estrogen receptor binding models in the past decade, as shown in **Table 1** (Hong et al., 2002; Serafimova et al., 2007; Liu et al., 2008; Li and Gramatica, 2010; Taha et al., 2010; Vedani et al., 2012; Zang et al., 2013; Zhang et al., 2013, 2014; Deng et al., 2014; Ng et al., 2015). These studies have covered a wide range of modeling approaches and data set sizes, from a descriptor-based decision tree (Hong et al., 2002) to 3-D docking and multi-dimensional QSAR (Vedani et al., 2012). The number of compounds used for modeling purpose in these studies range from less than 100 to more than 8000. The QSAR modeling of estrogen receptor binding agents has also been reviewed (Lo Piparo and Worth, 2010).

TABLE 1 | A sampling of QSAR studies on estrogen receptor interaction.

Year	Receptor studied	Data set size	Method	References
2005	α	232 training/ 463 test	Decision Tree	Hong et al., 2002
2007	α	645	COREPA	Serafimova et al., 2007
2008	α	108	OLS/GA-VSS	Liu et al., 2008
2010	β	119	GA-MLR	Taha et al., 2010
2010	α	132	GA-MLR/kNN	Li and Gramatica, 2010
2012	α	106α/96β	Docking/mQSAR (VirtualToxLab)	Vedani et al., 2012
2013	α/β	546α/137β	kNN (STL and MTL)	Zhang et al., 2013
2013	α	8147	SVM	Zang et al., 2013
2014	α/β	81	MLR/RBFNN	Deng et al., 2014
2015	α	3308	Decision forest	Ng et al., 2015

Although, there have been many promising models developed to predict ER binding data, these QSAR models are all based on data derived from chemical structure alone. As a result, there is increasing evidence that the applicability of these models is limited to certain compounds (Johnson, 2008; Scior et al., 2009). In certain cases, compounds with similar structures may show significantly different activities, leading to prediction errors in QSAR models. These pairs of molecules are known as "activity cliffs" in QSAR studies (Maggiora, 2006). QSAR models predict the activity of compounds only based on their chemical structure information, but the presence of activity cliffs can lead to unavoidable prediction errors if there is no other information than chemical structures (Cruz-Monteagudo et al., 2014).

Inspired by the biosimilarity study reported by Low and her coworkers (Low et al., 2013), in this study, we developed enhanced computational models for estrogen receptor binding agents using both QSAR approaches and a biosimilarity search, which is based on publically available bioassay data. The initial QSAR models developed using the combination of various chemical descriptors and modeling approaches, were integrated with the biosimilarity information to generate hybrid predictions. Using the resulting hybrid models, the new compounds can be directly predicted for their estrogen receptor binding potential. The incorporation of a biosimilarity search based on additional bioassay data can solve the activity cliffs issue of QSAR modeling and improve the prediction accuracy of new compounds.

MATERIALS AND METHODS

Data Curation

The original dataset used in this study was obtained in two parts separately from the National Center for Advancing the Translational Science (NCATS) via the Tox21 Challenge project. The dataset (PubChem assay AID 743077) consisted of the results of the quantitative High Throughput Screening (qHTS) to identify agonists of the ERα signaling pathway by measuring the expression of a beta lactamase reporter gene controlled by an ERα ligand binding domain (ER-LBD) fusion protein

(National Center for Biotechnology Information, 2015). This dataset was used as the training set in the Tox21 Challenge. The original dataset consisted of 8753 compounds, of which 446 were categorized as active (ERα binders) and 8307 were categorized as inactive (non-binders). The compounds were processed by the CaseUltra® (www.multicase.com) structure checker tool to remove duplicates and inorganic compounds, resulting in 5647 unique organic compounds (259 actives and 5388 inactives). All the active compounds were selected for the training set and combined with a randomly selected 259 inactive compounds to produce a balanced training set of 518 compounds. An additional but much smaller set of compounds not included in the original qHTS data was provided by the Tox21 Challenge project as an external test set to validate the resulting models (see **Figure 1** for modeling workflow). This external test set of 297 compounds (25 actives and 272 inactives) was also processed by the CaseUltra® structure checker to remove duplicates and inorganics, resulting in 264 unique compounds (24 actives and 240 inactives).

Chemical Descriptors

Once the datasets were curated, chemical descriptors were calculated using two commercial descriptor generators. A total of 192 2-D Molecular Operating Environment® (MOE) (www.chemcomp.com) descriptors were generated using MOE version 2013, which include physical properties, atom and bond counts, connectivity and shape indices, adjacency and distance matrix descriptors, etc. Dragon® (www.talete.mi.it/) version 6 was used to generate 1259 descriptors including constitutional indices, drug-like indices, connectivity indices, functional group counts, etc. All descriptors were normalized to (0,1) and any redundant descriptors were removed by deleting those with low variance (standard deviation <0.01 for the whole training set) and randomly keeping one of any pairs of descriptors that had high correlation ($R^2 > 0.95$ between two descriptor values for the training set compounds), leaving 132 unique MOE descriptors and 594 unique Dragon descriptors for both data sets. In order to calculate the chemical similarity among compounds, MOE 2013 was used to calculate 166 MACCS fingerprints of each compound. These fingerprints were used as descriptors to calculate the Tanimoto coefficient of each compound pair to determine their chemical similarity (Willett, 2006).

QSAR Model Development and Model Validation

Three machine learning algorithms were used to develop QSAR models: support vector machines (SVM), random forest (RF), and *k* nearest neighbor (*k*NN; Mitchell, 2014). In this study, the RF (Breiman, 2001) and SVM (Vapnik, 2000) algorithms available in R® 3.0.2 using the packages “e1071” and “randomForest” (Dalgaard, 2008) were implemented. The available SVM algorithm was tuned to identify the optimal inputs for model performance. The *k*NN models (Zheng and Tropsha, 2000) were built using in-house modeling tools, also available at Chembench (http://chembench.mml.unc.edu; Walker et al.,

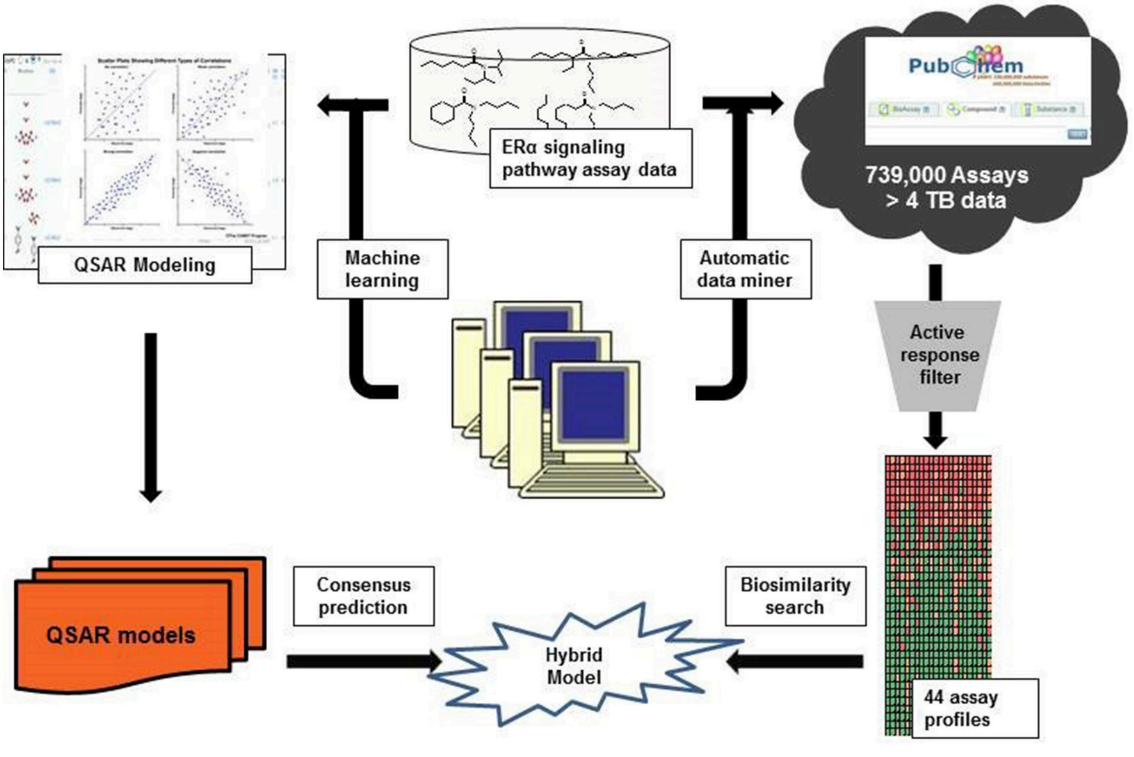


FIGURE 1 | The hybrid modeling workflow.

2010). This model uses a genetic algorithm selection procedure to predict the activity of a target compound by identifying the k most similar compounds within the chemical descriptor space and using their activity to predict that of the target compound. The best model of each run is kept, while inferior models are discarded. In our modeling process, a random selection of 50 chemical descriptors was used in each iteration of the algorithm. Each method was performed with both MOE and Dragon descriptors, as shown in the modeling workflow in **Figure 1**. The six resulting models (SVM-Dragon; SVM-MOE; RF-Dragon; RF-MOE; *k*NN-Dragon; and *k*NN-MOE) were averaged to give a consensus prediction, as described in previous publications (Solimeo et al., 2012; Kim et al., 2014). All models were validated using a five-fold cross validation. In this procedure, the training set was randomly split into five equal selected subsets. Four subsets (80%) were used as a training set and the compounds in the fifth subset (20%) were used as a test set. The training set was used to develop QSAR models and the resulting models were used to predict the test set. This procedure was repeated five times until all compounds were used in the test set once (Golbraikh et al., 2003; Tropsha and Golbraikh, 2007).

Biosimilarity Calculation

An in-house profiling tool (Zhang et al., 2014) was used to extract relevant bioassay data from PubChem for each compound in both the training and test sets. The PubChem assays were ranked by the numbers of active responses for the compounds in our training set. The resulting PubChem bioassay profile consisted of 44 bioassays, which contain the largest number of active responses in the training set, and was then used to calculate the biosimilarity between pairs of two compounds using the following formula:

$$\textit{Weighted Estimate of Biological Similarity}\ (\textit{WEBS}) = \frac{\sum(p + (\omega)n)}{\sum(p + (\omega)\, n + d)}$$

where p is the number of assays in which both compounds show active results, n is the number of assays in which both compounds show inactive results, and d is the number of assays in which the two compounds show opposite results. Inconclusive data were not considered in the calculation. The negative response data (inactives) are weighted less than positive responses (actives) in the biosimilarity calculation. In this study, the weight parameter ω was given the value of 0.06. The resulting WEBS values range from 0 to 1 and were used to determine the nearest neighbors in the training set for each test set compound. Any compound with WEBS similarity score over 0.6 was considered as a potential nearest neighbor for the target compound. The ERα binding activities of up to the top five nearest neighbors were used to calculate the predicted activity of the relevant test set compound. When fewer than five nearest neighbors existed within the training set, all nearest neighbors were used.

In order to form a hybrid model, the biosimilarity prediction was averaged with the QSAR prediction for each compound. For compounds which were not able to be predicted by the biosimilarity tool due to missing data, the QSAR consensus prediction was used as the predicted value. Compounds with opposite results from QSAR consensus models and biosimilarity search were considered as inconclusive and removed. This method returned a prediction for 192 of the 264 test set compounds.

RESULTS

QSAR Results

The modeling set was used to develop six individual QSAR models and their predictions were averaged as a consensus prediction. The model performance was indicated by five-fold cross validation of the modeling set itself and external prediction of a set of 264 unknown compounds. The performance was evaluated by calculating the sensitivity, specificity, and CCR for all models, as shown in **Figure 2**.

$$\textit{sensitivity} = \frac{\textit{true positives}}{\left(\textit{true positives} + \textit{false negatives}\right)}$$

$$\textit{specificity} = \frac{\textit{true negatives}}{\left(\textit{true negatives} + \textit{falsepositives}\right)}$$

$$\textit{CCR} = \frac{\textit{sensitivity} + \textit{specificity}}{2}$$

For the five-fold cross-validation procedures, the predictivity was similar across all the models (CCR = 0.642−0.749). However, the external predictions of the 264 unknown compounds showed a significant decrease in accuracy (CCR = 0.544−0.627), as observed in previous QSAR studies (Zhu et al., 2008a; Solimeo et al., 2012; Ng et al., 2015). Compared to individual models, the consensus model gave similar performance to the best individual models for both five-fold cross validation (sensitivity = 0.730, specificity = 0.704, and CCR = 0.717) and external predictions (sensitivity = 0.500, specificity = 0.683, and CCR = 0.592). Applying an applicability domain (AD), as described in previous studies (Zhu et al., 2008a, 2009), to both validation procedures did not show an improvement in predictive ability, so all predictions (100%) were retained when analyzing the QSAR models.

Bio-Assay Profile and Predictions

Our previous studies have shown improvements of QSAR models by incorporating biological data as extra descriptors into the modeling procedure (Sedykh et al., 2011; Kim et al., 2014). Relevant bioassay activity has been shown to be useful for the bioactivity predictions (Zhu et al., 2008b; Wang et al., 2015; Kim et al., 2016). In this study, the in-house profiling tool was used to automatically extract and optimize a biological profile containing 44 PubChem assays for 518 modeling set compounds. Using the WEBS score to calculate the biological similarity of each two compounds, those most similar compounds with WEBS scores over the nearest neighbor cut-off were identified for each test set compound and then used to predict the ERα binding

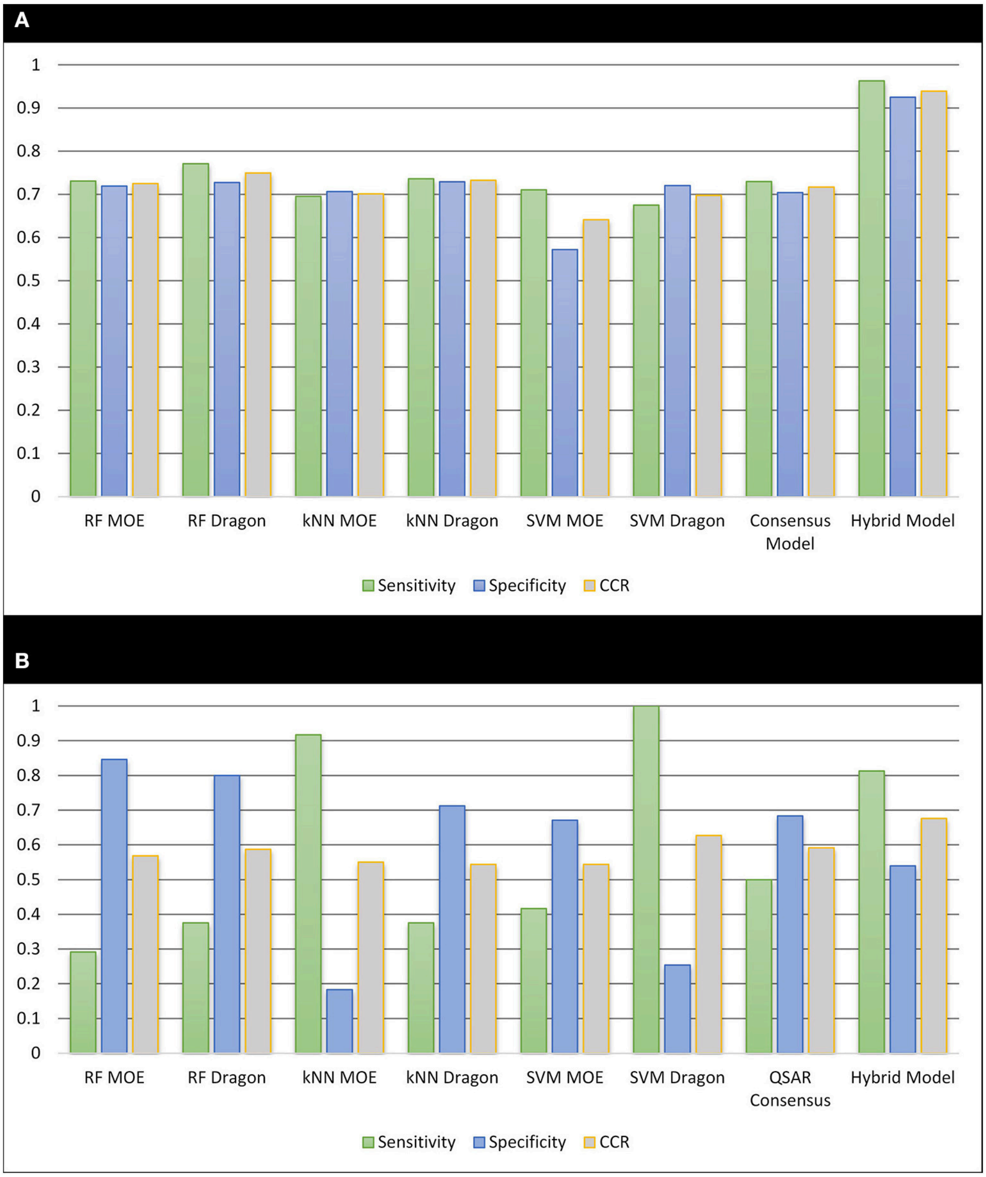


FIGURE 2 | The performance of all resulting models. (A) Cross-validation of the 518 training set compounds; **(B)** external validation of 264 unknown compounds.

potential. When combining the biosimilarity search with the QSAR consensus model as a hybrid model, the cross validation demonstrated a significant improvement of the accuracy over traditional QSAR modeling only based on chemical descriptors. Compared to the QSAR consensus model, the sensitivity, specificity and CCR of the hybrid model increased from 0.730 to 0.963, from 0.704 to 0.925, and from 0.717 to 0.939, respectively.

The external test set was also predicted by including up to five of the most biosimilar compounds in the training set. These hybrid predictions showed a noticeable improvement over the QSAR based solely on chemical descriptors. The external test set predictions returned a sensitivity = 0.813, specificity = 0.540, and CCR = 0.676 with a coverage of 73% (192 out of 264). The increase of sensitivity in both cross validation and external predictions brings considerable benefit when prioritizing potential EDCs for experimental testing.

DISCUSSION

The estrogen receptor has been the target of many modeling studies due to the effects of endocrine disruption that occur when a compound present in the environment or in a consumer product activates the receptor. While recent modeling studies (Ng et al., 2015) have demonstrated impressive relative balanced accuracy and specificity based on only chemical structures, these models are still challenged by the high prevalence of false negative results when testing an external set, leading to a low sensitivity.

There is a need for methods that can quickly and effectively screen a wide range of chemicals to correctly identify potential EDCs before a product is brought to market. This is a particular challenge when screening new compound sets, such as that used as an external test set in this study, where only a small fraction of the new compounds may be active binders. The attempt to use QSAR models based on only chemical descriptors to fill this need has been hindered by the structural diversity of the estrogen receptor binders and has reached a bottleneck due to the existence of activity cliffs. In this study, the noticeable improvement of the sensitivity of the model when predicting an external test set using the hybrid model suggests that the use of biological response data may be of particular importance in lowering the rate of false negative predictions from a model. Although this study focuses on activation of ERα only, there is a wide variety of chemical structures that are able to activate this receptor due to its large ligand binding domain (Shanle and Xu, 2011). The lack of experimental data, especially for active compounds (ERα binders), has resulted in activity cliffs in QSAR models based solely on chemical structures and limited the applicability of traditional QSAR modeling methods.

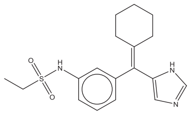
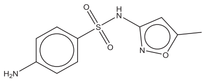
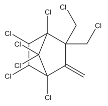
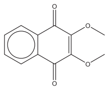
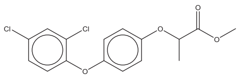
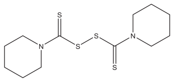
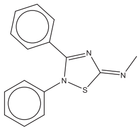
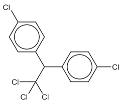
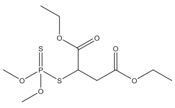
The QSAR models all showed acceptable predictivity when considering the cross validation of the training set. However, the external prediction of 264 unknown compounds had significantly decreased prediction accuracy, especially for individual models. Although the consensus model shows relatively stable performance, the sensitivity of its external test set prediction is much lower than the cross validation results due to the high proportion of false negatives. **Table 2** displays examples of compounds that were consistently predicted incorrectly by the original QSAR models along with both their chemical nearest neighbor and biological nearest neighbor in the training set. The first active compound, A-315456 (PubChem CID 6603710), an α-1D-adrenoceptor antagonist, is an ERα binder that was incorrectly predicted as inactive by all QSAR models. This compound's chemical nearest neighbor in the training set is the inactive compound sulfamethoxazole (PubChem CID 5329). Dimethoxynaphtoquinone (PubChem CID 3136) is also an active ERα binder that was incorrectly predicted by the QSAR consensus model. Its chemical nearest neighbor dichlofop-methyl (PubChem CID 39985) is an inactive compound in this assay. Similarly, the compound N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine (PubChem CID 682802) is an inactive compound. However, its chemical nearest neighbor, in the training set, dichlorodiphenyltrichloroethane (DDT) (PubChem CID 3036), is an ERα binder. These prediction errors cannot be avoided if only chemical structure information is used for modeling.

The prediction of the test set compounds improved when biosimilarity results were combined with the QSAR consensus model to form a hybrid model. Of particular note, the sensitivity of the external test set prediction increased from 0.500 for the QSAR consensus model alone to 0.813 for the hybrid model. In these examples, the biological nearest neighbors, as determined by WEBS score, provide more useful information for the predictions of external compounds. For example, the biological nearest neighbor in the training set of A-315456 (PubChem CID 6603710), an ERα binder, is toxaphene (PubChem CID 5284469), also an active compound (**Table 2**). For the other external test set compounds in **Table 2**, their biological nearest neighbors show the same ERα binding activities as the relevant target compounds. Furthermore, the WEBS scores for these test set compounds show dissimilarity to their chemical nearest neighbors. For example, the inactive compound N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine (PubChem CID 682802) has a biological nearest neighbor, malathion (PubChem CID 4004), a widely used insecticide that also showed inactive response in the ERα binding assay. Its chemical nearest neighbor, DDT (PubChem CID 3036), a now-banned insecticide, has a very low biosimilarity (WEBS = 0.0169) to N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine. Seven PubChem assays with testing data for both compounds show opposite results between these two compounds. The above analysis indicates that the activity cliffs are chemically similar compounds but have different biological effects (i.e., ERα binding). The hybrid model, using biosimilarity search as additional information in the modeling process, was able to differentiate them.

The bioassay response profile of the compounds shows promising potential to improve traditional QSAR models. Furthermore, when examining the PubChem assays used in the profile of this study, many targets of the assays regulate or are regulated by ERα. This provides additional useful information as to the types of bioassays which may be most useful in developing hybrid prediction models for ERα. The highest ranked assay, which consists of the highest number of active responses for our training set compounds, was used to screen potential inhibitors of histone lysine methyltransferase G9a (PubChem AID 504332). This assay acts as a co-regulator in the estradiol-induced activation or repression of gene transcription by ERα (Métivier et al., 2003; Purcell et al., 2011). Several other assays used in this profile specifically target enzymes in the cytochrome P450 (CYP450) family. These assays include screening inhibitors for CYP1A2 (PubChem AID 410) and CYP3A4 (PubChem AID 884), and a composite screening results for various CYP450 inhibitors (PubChem AID 1851). These proteins modulate ERα signaling by helping to maintain the androgen/estrogen balance (Tsuchiya et al., 2005). By analyzing the bioassays within the response profile, it indicates the future direction of gathering useful data for evaluating potential ERα binders.

The biosimilarity methodology used in this project shows a promising way to improve the predictivity of traditional QSAR modeling, particularly for increasing the sensitivity of the prediction results. However, since many compounds may not have been tested and have no data available in public resources, the usefulness of biosimilarity is limited by its coverage. A potential strategy to address the limitation of missing data is by using "read-across" methods (Patlewicz et al., 2014) to fill gaps in bioassay data for unknown compounds. Another pitfall of using the public data is the presence of experimental errors and the redundancy between various assay results. Currently, we are developing multiple novel data mining approaches to address this issue and will report them in future studies.

TABLE 2 | Three test set compounds (the first compound in each group) with their chemical nearest neighbor (the second compound) and biological nearest neighbor (the third compound).

	Compound	Activity	WEBS Score	Bioprofiles*
1	 CID= 6603710	Active	–	🟥🟥🟥🟦🟥🟥 *
	 CID= 5239	Inactive	0.117	🟦🟦🟦🟦🟦🟦
	 CID= 5284469	Active	1.00	⬜⬜🟥🟦🟥⬜
2	 CID= 3136	Active	–	🟥🟥🟥🟥🟥🟥 **
	 CID= 39985	Inactive	N/A	N/A
	 CID= 7188	Active	1.00	⬜⬜⬜🟥🟥🟥
3	 CID=682802	Inactive	–	🟥🟥🟥🟥🟥🟥🟥 ***
	 CID= 3036	Active	0.0169	🟦🟦🟦🟦🟦🟦🟦
	 CID= 4004	Inactive	1.00	🟥⬜⬜🟥🟥🟥🟥

**In the selected bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available. The bioprofiles only consist of the assays out of 44 PubChem assays that have the data for the three compounds in each group:*

**First group bioprofile assays: PubChem AID 410, 883, 884, 893, 504832, 686978.*

***Second group bioprofile assays: AID 410, 884, 504847, 686978, 686979, 743244.*

****Third group bioprofile assays: AID 884, 886, 887, 893, 504847, 686978, 686979.*

N/A indicates there is no data available for this compound within these assays.

CONCLUSION

In this study, we first developed QSAR models for the qHTS assay data, which identify agonists for the ERα signaling pathway, provided in the Tox21 challenge. The external test set prediction of all QSAR models, including the consensus model, is lower than the cross validation results of the training set. However, by combining the biosimilarity search, developed using the bioassay response profile automatically extracted from PubChem, with the QSAR consensus predictions, a hybrid model was created. The resulting hybrid model showed a noticeable improvement in both cross-validation and external prediction results compared to QSAR models based only on chemical descriptors. This result demonstrated that integrating extra biological data in the modeling process can improve traditional QSAR models when predicting ERα binding potentials for unknown compounds. This strategy can be used to develop enhanced models to evaluate other types of toxicity for compounds of interest.

AUTHOR CONTRIBUTIONS

Substantial contributions to the conception or design of the work: HZ. Acquisition, analysis, or interpretation of data for the work: KR, MK, WW, and DP. Drafting the work: KR and HZ. Final approval of the version to be published: HZ. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: HZ.

FUNDING

This research was supported in part by National Institutes of Health grants P30ES005022 and R15ES023148, and the Colgate-Palmolive Grant for Alternative Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

The authors want to thank Dr. Ruili Huang and Dr. Menghang Xia at NCATS to help with the data curation and result discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00012

REFERENCES

Blair, R. M., Fang, H., Branham, W. S., Hass, B. S., Dial, S. L., Moland, C. L., et al. (2000). The estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol. Sci.* 54, 138–153. doi: 10.1093/toxsci/54.1.138

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

Committee on Toxicity Testing and Assessment of Environmental Agents N.R.C. (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press.

Cruz-Monteagudo, M., Medina-Franco, J., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N., and Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* 19, 1069–1080. doi: 10.1016/j.drudis.2014.02.003

Dalgaard, P. (2008). *Introductory Statistics with R*. New York, NY: Springer Science & Business Media.

Deng, C. L., Chen, X. X., Lu, H. Y., Yang, X., Luan, F., and Cordeiro, M. (2014). Prediction of the Estrogen Receptor Binding Affinity for both hER(alpha) and hER(beta) by QSAR Approaches. *Lett. Drug Des. Disc.* 11, 265–278. doi: 10.2174/15701808113109990067

Ding, D., Xu, L., Fang, H., Hong, H., Perkins, R., Harris, S., et al. (2010). The EDKB: an established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics* 11(Suppl 6):S5. doi: 10.1186/1471-2105-11-S6-S5

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee, K. H., and Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* 17, 241–253. doi: 10.1023/A:1025386326946

Hall, J. M., Couse, J. F., and Korach, K. S. (2001). The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J. Biol. Chem.* 276, 36869–36872. doi: 10.1074/jbc.r100029200

Hong, H., Tong, W., Fang, H., Shi, L., Xie, Q., Wu, J., et al. (2002). Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* 110, 29–36. doi: 10.1289/ehp.0211029

Johnson, S. R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* 48, 25–26. doi: 10.1021/ci700332k

Kim, M., Huang, R., Sedykh, A., Zhang, J., Xia, M., and Zhu, H. (2016). Mechanism profiling of hepatotoxicity caused by oxidative stress using the antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* doi: 10.1289/ehp.1509763. [Epub ahead of print].

Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., and Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm. Res.* 31, 1002–1014. doi: 10.1007/s11095-013-1222-1

Li, J., and Gramatica, P. (2010). The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. *Mol. Divers.* 14, 687–696. doi: 10.1007/s11030-009-9212-2

Liu, H., Papa, E., and Gramatica, P. (2008). Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. *Chemosphere* 70, 1889–1897. doi: 10.1016/j.chemosphere.2007.07.071

Lo Piparo, E., and Worth, A. (2010). *Review of QSAR Models and Software Tools for Predicting Developmental and Reproductive Toxicity*. Luxemborg: Publications Office of the European Union. doi: 10.2788/9628

Low, Y., Sedykh, A., Fourches, D., Golbraikh, A., Whelan, M., Rusyn, I., et al. (2013). Integrative chemical-biological read-across approach for chemical hazard classification. *Chem. Res. Toxicol.* 26, 1199–1208. doi: 10.1021/tx400110f

Maggiora, G. M. (2006). On outliers and activity cliffs–why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535–1535. doi: 10.1021/ci060117s

Métivier, R., Penot, G., Hübner, M. R., Reid, G., Brand, H., Kos, M., et al. (2003). Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* 115, 751–763. doi: 10.1016/S0092-8674(03)00934-6

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4, 468–481. doi: 10.1002/wcms.1183

Mueller, S. O., and Korach, K. S. (2001). Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice. *Curr. Opin. Pharmacol.* 1, 613–619. doi: 10.1016/S1471-4892(01)00105-9

National Center for Biotechnology Information (2015). *PubChem BioAssay Database; AID=743077*. (Accessed September 15, 2015).

Ng, H. W., Luo, H., Ye, H., Ge, W., Tong, W., Hong, H., et al. (2015). Development and validation of decision forest model for estrogen receptor binding prediction of chemicals using large data sets. *Chem. Res. Toxicol.* 28, 2343–2351. doi: 10.1021/acs.chemrestox.5b00358

Patlewicz, G., Ball, N., Becker, R. A., Booth, E. D., Cronin, M. T. D., Kroese, D., et al. (2014). Read-across approaches - Misconceptions, promises and challenges ahead. *Arch. Med. Vet.* 46, 387–396. doi: 10.14573/altex.1410071

Purcell, D. J., Jeong, K. W., Bittencourt, D., Gerke, D. S., and Stallcup, M. R. (2011). A distinct mechanism for coactivator versus corepressor function by histone methyltransferase G9a in transcriptional regulation. *J. Biol. Chem.* 286, 41963–41971. doi: 10.1074/jbc.m111.298463

Schug, T. T., Janesick, A., Blumberg, B., and Heindel, J. J. (2011). Endocrine disrupting chemicals and disease susceptibility. *J. Steroid Biochem. Mol. Biol.* 127, 204–215. doi: 10.1016/j.jsbmb.2011.08.007

Scior, T., Medina-Franco, J., Do, Q. T., Martínez-Mayorga, K., Rojas, J., and Bernard, P. (2009). How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* 16, 4297–4313. doi: 10.2174/092986709789578213

Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., et al. (2011). Use of *in vitro* HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of *in vivo* toxicity. *Environ. Health Perspect.* 119, 364–370. doi: 10.1289/ehp.1002476

Serafimova, R., Todorov, M., Nedelcheva, D., Pavlov, T., Mekenyan, O., Akahori, Y., et al. (2007). QSAR and mechanistic interpretation of estrogen receptor binding. *SAR QSAR Environ. Res.* 18, 389–421. doi: 10.1080/10629360601053992

Shanle, E. K., and Xu, W. (2011). Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem. Res. Toxicol.* 24, 6–19. doi: 10.1021/tx100231n

Solimeo, R., Kim, M., Zhu, H., Zhang, J., and Sedykh, A. (2012). Predicting chemical ocular toxicity using a combinatorial QSAR approach. *Chem. Res. Toxicol.* 25, 2763–2769. doi: 10.1021/tx300393v

Taha, M. O., Tarairah, M., Zalloum, H., and Abu-Sheikha, G. (2010). Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and *in silico* search for new hits. *J. Mol. Graph. Model.* 28, 383–400. doi: 10.1016/j.jmgm.2009.09.005

Tropsha, A., and Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* 13, 3494–3504. doi: 10.2174/138161207782794257

Tsuchiya, Y., Nakajima, M., and Yokoi, T. (2005). Cytochrome P450-mediated metabolism of estrogens and its regulation in human. *Cancer Lett.* 227, 115–124. doi: 10.1016/j.canlet.2004.10.007

Vapnik, V. (2000). *The Nature of Statistical Learning theory.* New York, NY: Springer Science & Business Media.

Vedani, A., Dobler, M., and Smieško, M. (2012). VirtualToxLab — A platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol. Appl. Pharmacol.* 261, 142–153. doi: 10.1016/j.taap.2012.03.018

Walker, T., Grulke, C. M., Tropsha, A., and Pozefsky, D. (2010). Chembench: a cheminformatics workbench. *Bioinformatics* 26, 3000–3001. doi: 10.1093/bioinformatics/btq556

Wang, W., Kim, M., Sedykh, A., and Zhu, H. (2015). Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm. Res.* 32, 3055–3065. doi: 10.1007/s11095-015-1687-1

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 11, 1046–1053. doi: 10.1016/j.drudis.2006.10.005

Zang, Q., Rotroff, D. M., and Judson, R. S. (2013). Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. *J. Chem. Inf. Model.* 53, 3244–3261. doi: 10.1021/ci400527b

Zhang, J., Zhu, H., and Hsieh, J. H. (2014). Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. *PLoS ONE* 9:e99863. doi: 10.1371/journal.pone.0099863

Zhang, L., Sedykh, A., Tripathi, A., Zhu, H., Afantitis, A., Mouchlis, V. D., et al. (2013). Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol. Appl. Pharmacol.* 272, 67–76. doi: 10.1016/j.taap.2013.04.032

Zheng, W., and Tropsha, A. (2000). Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* 40, 185–194. doi: 10.1021/ci980033m

Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M., and Tropsha, A. (2009). Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* 22, 1913–1921. doi: 10.1021/tx900189p

Zhu, H., Rusyn, I., Richard, A., and Tropsha, A. (2008b). Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* 116, 506–513. doi: 10.1289/ehp.10573

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatical, P., et al. (2008a). Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis. J. Chem. Inf. Model.* 48, 766–784. doi: 10.1021/ci700443v

Zhu, H., Zhang, J., Kim, M. T., Boison, A., Sedykh, A., and Moran, K. (2014). Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol.* 27, 1643–1651. doi: 10.1021/tx500145h,

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
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on