MACHINE LEARNING METHODOLOGIES TO STUDY MOLECULAR INTERACTIONS

EDITED BY: Elif Ozkirimli, Tunca Dogan, Arzucan Ozgur
and Artur Yakimovich

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-121-2
DOI 10.3389/978-2-88974-121-2

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

MACHINE LEARNING METHODOLOGIES TO STUDY MOLECULAR INTERACTIONS

Topic Editors:
Elif Ozkirimli, Roche (Switzerland), Switzerland
Tunca Dogan, Hacettepe University, Turkey
Arzucan Ozgur, Boğaziçi University, Turkey
Artur Yakimovich, Roche (United Kingdom), United Kingdom

Dr. Elif Ozkirimli is a full time employee of F. Hoffmann-La Roche AG, Switzerland and Dr. Artur Yakimovich is a full time employee of Roche Products Limited, UK. All other Topic Editors declare no competing interests with regards to the Research Topic.

Citation: Ozkirimli, E., Dogan, T., Ozgur, A., Yakimovich, A., eds. (2022). Machine Learning Methodologies to Study Molecular Interactions.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-121-2

Table of Contents

frontiers in Molecular Biosciences

EDITORIAL
published: 03 December 2021
doi: 10.3389/fmolb.2021.806474

Editorial: Machine Learning Methodologies to Study Molecular Interactions

Artur Yakimovich[1], *Arzucan Özgür*[2], *Tunca Doğan*[3] *and Elif Ozkirimli*[4]*

[1]Pharma International Informatics, Roche Products Limited, Welwyn Garden City, United Kingdom, [2]Department of Computer Engineering, Bogazici University, Istanbul, Turkey, [3]Department of Computer Engineering, Hacettepe University, Ankara, Turkey, [4]Pharma International Informatics, F. Hoffmann-La Roche AG, Basel, Switzerland

Keywords: machine learning, molecular interactions, protein, biomolecule, DNA, interaction prediction

Editorial on the Research Topic

Machine Learning Methodologies to Study Molecular Interactions

The cell is a busy place with proteins, DNA, RNA, metabolites and other molecules interacting with each other with orchestral precision. Disease states arise when this precision is lost for intracellular interactions or when external entities, such as virus particles, interact with intracellular molecules and disrupt this precision. As such, the study of molecular interactions is a huge area of focus for experimental and computational biologists alike.

Recognising the ever increasing uptake of Machine Learning (ML) in biomedical research, in this research topic, our focus was on the use of computational methodologies and ML approaches to examine molecular interactions. While experimental approaches such as structure determination of multimolecular complexes using X-ray crystallography or cryoEM are often the gold standard in studying intermolecular interactions, computational approaches are advantageous both because they are faster and less costly than experimental approaches and because some molecular interactions are neither easy nor feasible to study experimentally. In this special issue, the questions that the authors aimed to address ranged from understanding interactions at the residue or atomic level Karakulak et al.; Wang et al. to the cellular level Kyrilis et al.

The authors used a multitude of data sources and their combinations highlighting the value of multimodal analysis. Both sequence and structure-based predictors of specificity-determining residues in protein complexes were evaluated in the study of Karakulak et al. The authors proposed that the use of either approach by itself is not sufficient to accurately identify these residues, and new methods combining the advantages of both sequence and structure centric approaches are required. Protein interaction sites were identified by combining protein sequence and structure based information Wang et al., a reduced representation of proteins was built by molecular dynamics simulation data Errica et al., and genomic data was used to build an alternative splicing gene signature for cancer prognosis Zhao et al. Additionally, Kutlay and Aydin Son combined microRNA, mRNA, and DNA methylation data to build a metastasis model for melanoma cancer.

The articles in this issue have used ML methodologies ranging from shallow Support Vector machines (SVM) to deep learning based Graph Neural Networks (GNN) with success in interaction prediction, molecular representations and disease modeling. Two articles used graph neural networks powered by GPU. Wang et al. proposed a GNN based docking decoy evaluation score to identify near-native complex structures. By using an attention and gate-augmented mechanism, they captured the interaction pattern at the interface. A deep graph network enhanced sampling approach was proposed by Errica et al. to identify the coarse grained representation of proteins with minimal information loss.

Edited and reviewed by:
Francesco Luigi Gervasio,
University College London,
United Kingdom

***Correspondence:**
Elif Ozkirimli
elif.ozkirimli@roche.com

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *31 October 2021*
Accepted: *10 November 2021*
Published: *03 December 2021*

Citation:
Yakimovich A, Özgür A, Doğan T and Ozkirimli E (2021) Editorial: Machine Learning Methodologies to Study Molecular Interactions.
Front. Mol. Biosci. 8:806474.
doi: 10.3389/fmolb.2021.806474

The mapping entropy provided information about the information loss due to mapping to a lower dimensional space. The authors used deep learning to accelerate mapping entropy calculation followed by Wang-Landau sampling to explore the mapping space of a molecule. This physics based coarse grained description of the molecular structure allowed the calculation of various properties by considering the dynamic nature of biomolecules.

Predicting interactions or interaction sites is not sufficient to predict the presence or absence of a potential disease state. Zhao et al. used alternative splicing signature as a predictor of non-small cell lung cancer prognosis using multivariate Cox regression. Going beyond the prediction of host-pathogen interactions, Karabulut et al. constructed an ML-based infection prediction model that predicts whether adenoviral infection can happen in a host, using multiple types of input features including host-pathogen PPIs and taxonomic preferences. Kutlay and Aydin Son built a prediction model for metastasis in melanoma. Arici and Tuncbag assessed the performance of various network reconstruction approaches over the use cases of reconstructing the Notch signaling and the glioblastoma (GBM) disease pathways, using different reference human interactome datasets, and showed that the performance is highly dependent on the source data. This study showed that the quality and coverage of the input data can be at least as important as the utilised algorithm when studying molecular interactions.

Perhaps one of the best ways to illustrate the versatility of ML methodologies when applied to molecular interactions is to demonstrate that such application may be performed both in a bottom-up and a top-down fashion. Two examples of such demonstrations in our research topic are the review article by Zrimec et al. and a perspective article by Kyrilis et al. The former explored the representation learning application to the central molecular dogma, i.e. learning biological molecules and their interactions from the genetic code (DNA to RNA to protein sequences). The latter reviewed studies utilizing machine learning approaches to analyze native cell extract as a source of experimental data on higher order molecular interactions. While the problem at hand of Zrimec and colleagues seems much more well studied, they presented a convincing case of innovative approaches in the field. For example, the authors reviewed a body of literature demonstrating that deep neural networks can automatically learn regulatory grammar through utilization of convolutional or recurrent neural networks. While these methods are widely applied in fields like Computer Vision and Natural Language Processing, their application to the genetic code gained popularity only recently. At the same time, Kyrilis and co-authors discussed the top-down approaches, which look at relatively noisy data sources to provide rich information about the inner workings of the cell through techniques such as cryo-electron microscopy and structural proteomics. In their perspective article they made a convincing case for ML methods being necessary and sufficient to tie together these modalities. Authors noted that inspired by Computer Vision, the tool-of-choice for cryo-electron microscopy is convolutional neural networks. Hence, it is this family of algorithms that receives the most attention from the researchers working on cell extracts to devise higher order molecular interactions.

Altogether, these studies serve as a great demonstration of the level of ML penetration into the study of Molecular Interactions. With the significantly elevated performance of deep learning-based single-chain protein structure predictors such as AlphaFold (Jumper et al., 2021) and RosettaFold (Baek et al., 2021), the focus has now been shifting to the accurate prediction of protein complex structures (Evans et al., 2021). The advances in cryoEM imaging, single cell imaging, proteomics (Piazza et al., 2020) methodologies also open new avenues for analyzing interactions in their native environments. It is clear that ML approaches to study molecular interactions are rapidly gaining traction but we, the Editors of this Research Topic, believe that the most exciting applications of ML to this domain are yet to be published. We look forward to reading the future research in this field.

AUTHOR CONTRIBUTIONS

All authors conceived the idea, contributed to the topic editing, wrote the manuscript and approved it for publication.

REFERENCES

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* 373 (6557), 871–876. doi:10.1126/science.abj8754

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., et al. (2021). Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv.* doi:10.1101/2021.10.04.463034

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Piazza, I., Beaton, N., Bruderer, R., Knobloch, T., Barbisan, C., Chandat, L., et al. (2020). A Machine Learning-Based Chemoproteomic Approach to Identify Drug Targets and Binding Sites in Complex Proteomes. *Nat. Commun.* 11, 4200. doi:10.1038/s41467-020-18071-x

Conflict of Interest: AY was employed by the Roche Products Limited. EO was employed by the F. Hoffmann-La Roche AG.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

PERSPECTIVE
published: 15 April 2021
doi: 10.3389/fmolb.2021.660542

Detecting Protein Communities in Native Cell Extracts by Machine Learning: A Structural Biologist's Perspective

Fotis L. Kyrilis[1,2], *Jaydeep Belapure*[1] *and Panagiotis L. Kastritis*[1,2,3]*

[1] Interdisciplinary Research Center HALOmem, Charles Tanford Protein Center, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany, [2] Institute of Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany, [3] Biozentrum, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

OPEN ACCESS

Edited by:
Artur Yakimovich,
Roche (United Kingdom),
United Kingdom

Reviewed by:
Shruthi Viswanath,
National Centre for Biological Sciences, India
Leandro Estrozi,
UMR 5075 - Institut de Biologie Structurale Center for the National Scientific Research (CNRS)/CEA/UGA, France
Dimitry Tegunov,
Max-Planck-Gesellschaft (MPG), Germany
Bjoern Forsberg,
University of Oxford, United Kingdom
Amarda Shehu,
George Mason University Arlington Campus, United States
Deb Kelly,
Pennsylvania State University (PSU), United States
Jiawei Wang,
Tsinghua University, China
Charlotte Scarff,
University of Leeds, United Kingdom

***Correspondence:**
Panagiotis L. Kastritis
panagiotis.kastritis@bct.uni-halle.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *29 January 2021*
Accepted: *18 March 2021*
Published: *15 April 2021*

Citation:
Kyrilis FL, Belapure J and Kastritis PL (2021) Detecting Protein Communities in Native Cell Extracts by Machine Learning: A Structural Biologist's Perspective. Front. Mol. Biosci. 8:660542. doi: 10.3389/fmolb.2021.660542

Native cell extracts hold great promise for understanding the molecular structure of ordered biological systems at high resolution. This is because higher-order biomolecular interactions, dubbed as protein communities, may be retained in their (near-)native state, in contrast to extensively purifying or artificially overexpressing the proteins of interest. The distinct machine-learning approaches are applied to discover protein–protein interactions within cell extracts, reconstruct dedicated biological networks, and report on protein community members from various organisms. Their validation is also important, e.g., by the cross-linking mass spectrometry or cell biology methods. In addition, the cell extracts are amenable to structural analysis by cryo-electron microscopy (cryo-EM), but due to their inherent complexity, sorting structural signatures of protein communities derived by cryo-EM comprises a formidable task. The application of image-processing workflows inspired by machine-learning techniques would provide improvements in distinguishing structural signatures, correlating proteomic and network data to structural signatures and subsequently reconstructed cryo-EM maps, and, ultimately, characterizing unidentified protein communities at high resolution. In this review article, we summarize recent literature in detecting protein communities from native cell extracts and identify the remaining challenges and opportunities. We argue that the progress in, and the integration of, machine learning, cryo-EM, and complementary structural proteomics approaches would provide the basis for a multi-scale molecular description of protein communities within native cell extracts.

Keywords: cellular homogenates, random forest, convolutional neural network, cryo-EM, mass spectrometry, structural biology, protein–protein interactions, metabolons

INTRODUCTION

Since the dawn of biological research, humans are breaking-apart living systems to understand their structure and function. For example, in *Book VI of History of Animals*, Aristotle systematically addressed the processes of egg formation and chick embryo development by visual inspection. Nowadays, with the rapid technological advances in biochemical, biophysical, structural, and computational methods, cellular homogenates can be understood in great detail, providing network and structural information of the biomolecules within them. Crude extracts made by the lysis of cellular material possess operative aspects of cellular function, but in a context that is easier to manipulate. They are biotechnologically exploited for bioproduction (Karim and Jewett, 2016), cell-free gene expression, transcription, translation (Silverman et al., 2020), and, recently, molecular design (Hammerling et al., 2020). Probing the intrinsic structure of cell extracts is of paramount importance, so that their function is understood in detail. Until recently, the study of cell extracts was limited to low-resolution data (Han et al., 2009), but, with methodological advances, the resolution of 4.7 Å for the biomolecular complexes within those was reached (Kastritis et al., 2017).

Recent studies not only increased the achievable resolution (Arimura et al., 2020; Ho et al., 2020; Su et al., 2021), particularly in the membrane (Su et al., 2021) or nuclear extracts (Arimura et al., 2020) but also determined the snapshots of higher-order organization of in-extract flexible, functional metabolons (Kyrilis et al., 2021). The importance and challenges of integrative structural studies of native extracts and the correlation between structural disorder and function for in-extract metabolons were recently reviewed (Kyrilis et al., 2019; McCafferty et al., 2020; Skalidis et al., 2020). Reaching the milestone of near-atomic detail a few years ago proved that native cell extracts are amenable to structural studies and considerably broadened the structural proteomics field by expanding the concept of "protein communities" (Kastritis et al., 2017), primarily described by Gavin et al. (2006). Protein communities describe the associated molecules of several macromolecular complexes arranged in close proximity encoding functionally synchronized biomolecular entities. For example, they may efficiently transfer substrates along with enzymatic pathways [dubbed *metabolons*, reviewed in (Kastritis and Gavin, 2018)], effectively transduce signals, and regulate protein synthesis on local cellular demand. However, their inherent complexity limits probing their intrinsic structure to a few abundant biomolecular complexes, e.g., functional pyruvate dehydrogenase higher-order architecture (Kyrilis et al., 2021). The review of machine-learning approaches that are already applied in various intermediate analysis steps demonstrates an optimistic perspective in addressing this issue, and thus allowing a deeper understanding of protein communities in the future. In this study, by machine learning, we refer to the un-/supervised algorithms that are trained to learn the patterns in the scientific data retrieved from -omics, cryo-electron microscopy (cryo-EM), or any other method to predict the desired physically meaningful feature without human intervention.

HIGHER-ORDER COMPLEXITY OF PROTEIN COMMUNITIES: AN IDEAL TEST BED FOR MACHINE LEARNING

Protein communities (or, in general, biomolecular communities) are endogenously present in the cell and can be retrieved in native cell extracts. They are composed of biomolecular assemblies of varying compositional and chemical heterogeneity. A protein community comprises a functional cellular assembly and encodes localized functions (e.g., as in the case of metabolons). Protein communities also include interconnected protein complexes in variable stoichiometry and, therefore, represent a holistic view of cellular function beyond the description of their individual constituents. Due to their intricacy, communities must be characterized with an array of methods: (a) -omics methods, especially quantitative mass spectrometry (MS), to identify constituent molecules; (b) activity assays to probe their function; (c) cross-linking to find the interacting community biomolecules; (d) large-scale molecular modeling or cryo-EM characterization of community members to annotate complexes within the protein communities; and (e) cryo-EM characterization to visualize protein communities. This multi-scale, integrative characterization of protein communities can only be performed in native cell extracts and was previously discussed (Kyrilis et al., 2019). This integrative, systematic analysis was performed for eukaryotic communities involved in the synthesis of fatty acids (Kastritis et al., 2017) and in the metabolism of oxoacids (Kyrilis et al., 2021).

In this review, we outline the methods and challenges faced in such integrative studies of protein communities. Furthermore, we assess and discuss the state-of-the-art machine-learning methods applied in adjoint problems that could better aid investigations in this field. In the first two sections, we discuss the molecular characterization of protein communities, first in crude and then in simplified lysates. The next two sections describe the structural characterization of protein community members, since structural analysis of complete protein communities is a formidable task. This is because cryo-EM of complete protein communities can show ultrastructural features, but does not provide high-resolution three-dimensional (3D) reconstructions due to the highly complex and intricate structure of the community. We finally surveyed published machine-learning tools that are principally developed for diverse characterization of the biomolecular complexes. In each subsequent section, we discuss the applicability, promises, and limitations of machine-learning methods for deciphering protein communities.

PREDICTING PROTEIN COMMUNITIES IN CRUDE NATIVE CELL EXTRACTS

Cell extracts are amenable to biochemical treatment to probe the biomolecular content (**Figure 1A**), and methods were applied to study the retrieved homogenate directly (i.e., breaking the cellular material and subjecting it to an array of characterization tools). Proteins present in the cell extracts can be studied by MS, providing identification for thousands of protein sequences (Beck et al., 2011; Titeca et al., 2019). Unfortunately, this information offers a list of proteins, and, optimally, a report on their relative abundance, but not on their interactions. To predict communities, network analysis must then be performed by integrating the external interaction data for community members or their close homologs as, e.g., initially performed for the interconnected yeast complexes using tandem affinity purification (TAP) and MS (Gavin et al., 2002). In recent studies,

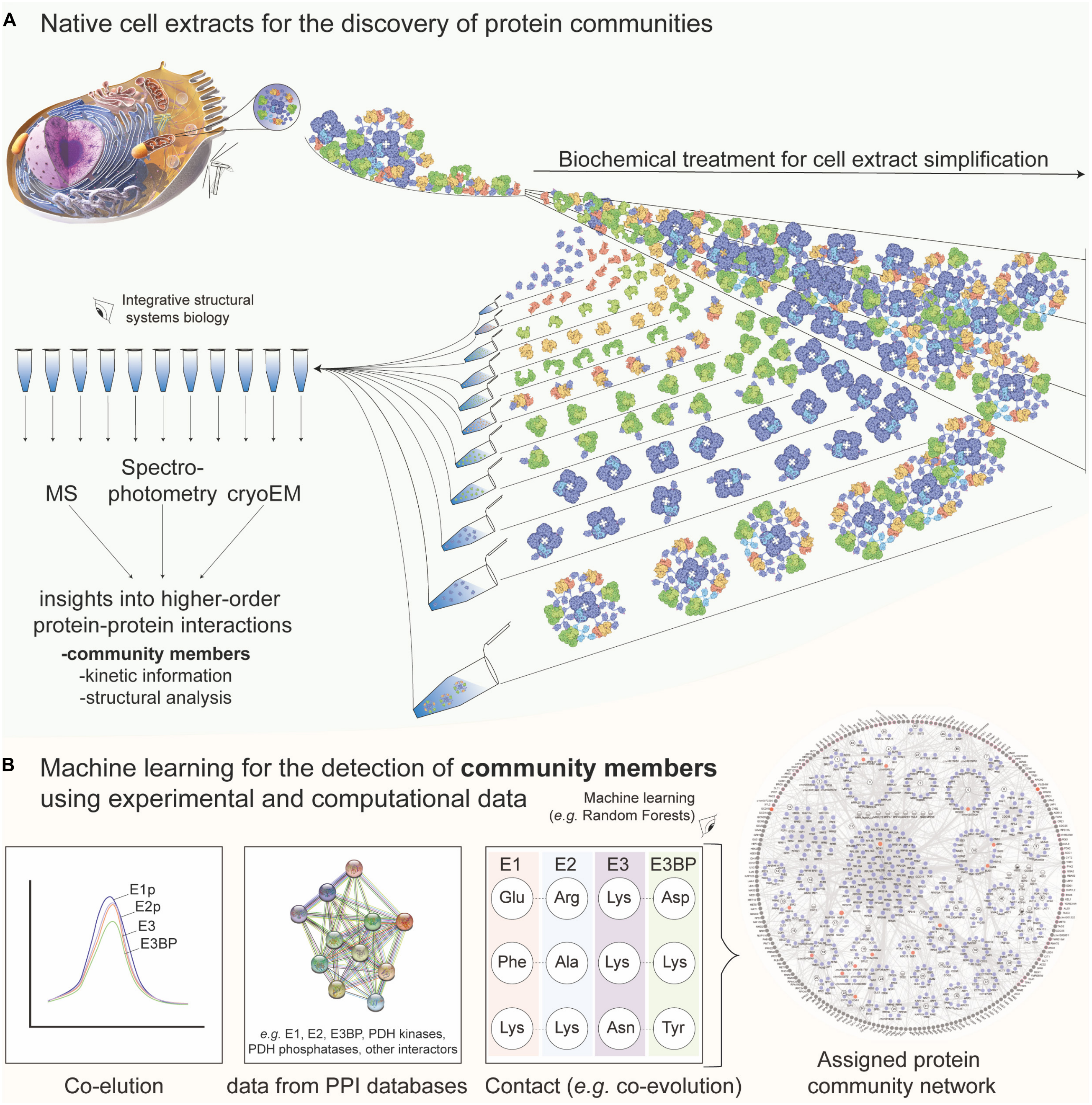

A
Native cell extracts for the discovery of protein communities
Biochemical treatment for cell extract simplification
Integrative structural systems biology
MS
Spectro-photometry
cryoEM
insights into higher-order protein-protein interactions
-community members
-kinetic information
-structural analysis
B
Machine learning for the detection of community members using experimental and computational data
Machine learning (e.g. Random Forests)
E1p
E2p
E3
E3BP
e.g. E1, E2, E3BP, PDH kinases, PDH phosphatases, other interactors
E1
E2
E3
E3BP
Glu
Arg
Lys
Asp
Phe
Ala
Lys
Lys
Lys
Lys
Asn
Tyr
Co-elution
data from PPI databases
Contact (e.g. co-evolution)
Assigned protein community network

FIGURE 1 | Native cell extracts as a tool for discovering protein communities with the aid of machine learning. **(A)** Methods to experimentally extract identity, structure, and dynamics information of protein communities. In short, the cell is lysed and the subsequent fractionation is applied to recover co-eluting protein material. In a large-scale manner, mass spectrometric, kinetic, and cryo-EM analysis of the fractions leads to the characterization of protein communities in native cell extracts. The example of the pyruvate dehydrogenase complex (PDHc) metabolon is shown. Molecular representations for PDHc are retrieved and further edited from Protein Data Bank "Molecule of the Month" section [Source: Image from the RCSB PDB September 2012 Molecule of the Month feature by David S. Goodsell (doi: 10.2210/rcsb_pdb/mom_2012_9)]. The cell representation on the top left was retrieved from Microsoft PowerPoint 2019 v16.47. **(B)** Combined data regarding protein–protein interactions stemming from fractionation (co-elution), external database information (network data), and contact information prediction (e.g., from co-evolution analysis, chemical cross-linking or mutagenesis experiments) among community members are used for machine learning, e.g., using a random forest. Finally, a network with interconnected protein communities is derived and insights into community members can be retrieved. External data shown are extracted from STRING (https://string-db.org/) and network shown from Kastritis et al. (2017). E1, E2, E3, and E3BP are the proteins structuring the 10-MDa complex of the PDHc metabolon, all involved in the complex reaction of pyruvate oxidation.

experimental and/or computational methods for characterizing protein–protein interactions (PPI) are included, connecting *in vivo*, *in vitro,* and *in silico* data (Rao et al., 2014). By meticulous data integration, considering the strengths and limitations of each approach that was applied to discover PPIs (Rao et al., 2014), a network is then constructed using the machine-learning (Havugimana et al., 2017) method. In particular, interesting computational approaches for PPI prediction include, but are not limited to, a combination of different machine-learning models to take a majority vote for final prediction (Saha et al., 2014), a game theory-based approach inspired by a non-cooperative sequential game (Maulik et al., 2017), and deep neural networks that either incorporate physical/chemical properties and graph theory (Zhang and Kabuka, 2019) or combine with decision-tree classifiers for the final PPI prediction (Wang et al., 2019).

Naturally, training sets are of vital importance for reconstructing a biological network and are mostly extracted from the PPI databases such as CORUM (Giurgiu et al., 2019), IntAct (Hermjakob et al., 2004), and GO (Harris et al., 2004). The availability of a high-confidence set of PPIs is often limited, especially when it comes to organisms that lack genome, transcriptome, and/or proteome data. Even in well-studied organisms, the construction of a *confusion matrix* (*error matrix*) for PPIs is not an easy task. Proteins dynamically interact, change localization, and can even alter their function due to moonlighting (Jeffery, 2014), and therefore, according to the cellular state and environmental conditions, PPIs may differ. Such discoveries revealed localized variations in interaction networks of disease phenotypes (Vidal et al., 2011), and, recently, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cellular interactors (Gordon et al., 2020). Protein networks are, therefore, commonly employed in biotechnological and medical applications because the cellular function is probed in a holistic approach, complementing mechanistic investigations into molecular recognition. Traditionally, reconstruction of protein networks is not only essential for characterizing protein complexes, but also for their higher-order interactions present in their communities (Gavin et al., 2002, 2006).

SIMPLIFYING PROTEIN COMMUNITY DETECTION WITHIN CELL EXTRACTS BY INTEGRATING CO-ELUTION DATA AND CHEMICAL CROSS-LINKING

The increased complexity of cellular homogenates brings various limitations in the study of their biomolecular content, mainly because of the well-known bias toward the identification of high-abundant proteins and complexes (Fursch et al., 2020). An idea to confidently annotate proteins in cell extracts, retrieve more interactors, and optimize the robust identification of protein communities is to subject the extracted homogenate to a subsequent biochemical treatment that would coarsely separate the biomolecular complexes on a certain biophysical property (termed protein co-fractionation, e.g., using the hydrodynamic radius as performed *via* size-exclusion chromatography (SEC) of the native cell extract). Mapping fractionated extracts with various proteomics methods was recently reviewed (Salas et al., 2020). The application of co-fractionation to monitor protein associations (Havugimana et al., 2012; Kristensen et al., 2012) perhaps stems from previous works that measured the enzymatic activities across retrieved cellular fractions, e.g., in the fractionated extracts of *Escherichia coli*, where interactions of Krebs cycle enzymes were probed (Barnes and Weitzman, 1986). Nowadays, the high-resolution separation of cell extracts is mostly performed by using high-resolution SEC coupled to MS (Salas et al., 2020). This method (a) simplifies the cell extract according to an intrinsic physical property of the contained biomolecules; (b) provides per-fraction quantitative data regarding protein abundance and co-detection; and (c) offers robust per-protein elution profiles across the studied fractions, which may be used for subsequent integration into a PPI network. Protein co-fractionation can be used to identify interactors within protein communities (Kastritis et al., 2017) and compare PPI networks across species, highlighting evolutionary implications (Wan et al., 2015). An example of data integration to derive a PPI network, highlighting protein communities, is shown in **Figure 1B**.

As with the previously described PPI networks, the application of machine-learning approaches is crucial, not only to integrate the protein co-elution data but also to discriminate random co-elution events from true (interacting) protein complexes. The machine-learning-based tools to probe the complexes within cell extracts of different organisms were developed (Kastritis et al., 2017; Stacey et al., 2017; Hu et al., 2019; Fossati et al., 2020). EPIC (Hu et al., 2019), an open-source software tool, may specifically use co-elution data to predict protein complexes found in cell extracts after training and validating a random forest algorithm (Tin Kam, 1995) or a support vector machine algorithm (Boser et al., 1992). The random forest algorithm showed superior performance when applied to predict co-eluting complexes and their communities after cross-validation from *Caenorhabditis elegans* (Hu et al., 2019), *Chaetomium thermophilum* (Kastritis et al., 2017), and HeLa cells (Fossati et al., 2020). Recently, PCprophet incorporated Bayesian inference to identify altered protein profiles across experiments that probe phenotypic changes (Fossati et al., 2020). Predicting protein communities from co-fractionation data rely on complex inference from the resulting network after reconstructing it with identified PPIs. Due to the density of the network, partitioning methods to recover protein complexes are limited, and often graph clustering algorithms that handle weighted graphs to generate overlapping clusters are applied [e.g., ClusterONE (Nepusz et al., 2012), or the more recent, ONCQS (Zhao and Lei, 2019)]. High-density chemical cross-linking can, therefore, offer complementary data to enrich and validate true protein co-elution and protein complex/community member data (Sinz, 2018). Cross-linking was applied to soluble extracts (Liu et al., 2015; Gotze et al., 2019), membrane complexes (Larance et al., 2016), large macromolecular complexes to dissect conformational flexibility (Tuting et al., 2020), and, importantly, directly within the SEC fractions where proteins are determined to co-elute for the characterization of protein communities

(Kastritis et al., 2017). Algorithms to detect co-eluting PPIs (Elias and Gygi, 2007; Havugimana et al., 2012) or cross-links (Ji et al., 2016; Huang et al., 2020) can include machine-learning tools to probe the complexity of high data dimensionality.

PROCESSING (CRYO-)EM IMAGES FROM NATIVE EXTRACTS WITH A FOCUS ON MACHINE LEARNING

Using cryo-EM imaging of native cell extracts to structurally analyze protein communities is essential. This is because proteomics methods discover the sequences of the community members or their interactions but do not provide information on their higher-order structure within their communities. Even if high-density cross-linking retrieves interacting proteins and their relative interacting distances, the community structure is unknown, including stoichiometry. It is noted that deriving stoichiometry for protein communities is not trivial, and a combination of cryo-EM, immunoblotting data, MS, and cross-linking MS in fractionated extracts was recently performed to derive approximate stoichiometry for the higher-order structure of the endogenous pyruvate dehydrogenase complex (Kyrilis et al., 2021). Direct methods, such as electron microscopy, can, therefore, be applied to observe cell extracts and were previously used in combination with MS at low resolution to visualize protein complexes (Han et al., 2009). However, recently, with advances in cryo-EM (Kuhlbrandt, 2014), native cell extracts delivered high-resolution data (Kastritis et al., 2017) and the first images of protein communities involving fatty acid synthase (FAS) together with other megadalton complexes (Kastritis et al., 2017). Recent results in the field also showed that abundant complexes can be reconstructed *de novo* (Ho et al., 2020), but not as members of protein communities. We also recently communicated the structural and functional characterization of communities involved in oxo acid metabolism by integrative methods (Kyrilis et al., 2021). Despite these advances, the high complexity of the imaged cell extract hinders proper quantification and 3D reconstruction of the interacting molecules within the extracts, and this is because of multiple issues regarding the specimen complexity. Therefore, most of the algorithms that were developed are applied to protein complexes and not to their higher-order assemblies in their native communities.

Cryo-EM micrographs contain two-dimensional (2D) projections of the particles in different orientations but are inherently of low contrast and often include contamination or undesirable features (see, e.g., **Figure 2A**). The signal-to-noise ratio in typical cryo-EM tomographs is ~0.1, perhaps comparable to imaging in astronomy. Except in cryo-EM, multiple short exposures are recorded. The traditional methods, such as bandpass, or Wiener filtering (Jain and Seung, 2008; Sindelar and Grigorieff, 2011; Xie et al., 2012), to improve the contrast are insensitive to the underlying noise properties. The cryo-EM field recently witnessed a surge in machine-learning models that are trained to learn the noise characteristics and offer better denoising [(Bepler et al., 2020) and references. therein].

The traditional template-based approaches [e.g., (Huang and Penczek, 2004)] pick particle candidates by estimating the similarity of an image region to a reference, also known as a template, through cross-correlation techniques. The template-matching methods are prone to introduce template-based bias and are known for a high rate of false positives. This stems from the fact that, if matching is performed over enough number of random regions (e.g., noise only), then meaningless noise can be perceived as a pattern, a phenomenon dubbed as "Einstein-from-noise" (Shatsky et al., 2009). For the purpose of selecting desirable regions without a reference, deep learning algorithms were developed (Wang et al., 2016; Zhu et al., 2017; Punjani et al., 2017; Bepler et al., 2018; Tegunov and Cramer, 2019; Wagner et al., 2019; Zhang et al., 2019; Sanchez-Garcia et al., 2020b). Inspired by computer vision applications, using convolutional neural networks (CNNs) (Tegunov and Cramer, 2019; Sanchez-Garcia et al., 2020b), per pixel-image segmentation of particle/non-particle regions was demonstrated (**Figure 2B**). Many of these architectures are explicitly designed to eliminate undesirable features or implicitly learn to avoid them (Wang et al., 2016; Zhu et al., 2017; Bepler et al., 2018; Wagner et al., 2019; Zhang et al., 2019). Recent machine-learning and deep learning-based methods demonstrated improved accuracy and low false-positive rates (Wang et al., 2016; Punjani et al., 2017; Zhu et al., 2017; Bepler et al., 2018; Tegunov and Cramer, 2019; Wagner et al., 2019; Zhang et al., 2019; Sanchez-Garcia et al., 2020b). Since templates can be essentially seen as filters, CNNs are the most successful models for the task of image classification and particle picking, as they are trained to learn thousands of 2D filters (Rawat and Wang, 2017). We speculate that these algorithms if trained in the heterogeneous mixtures of cell extracts instead of single-particle datasets, are expected to effectively detect particles of varying shapes and sizes and separate them from the artifacts in the micrographs of cellular extracts to systematically retrieve members of protein communities. However, the learning algorithm would still need to address the subsequent challenging step of segregating and clustering the particles into correctly assigned classes and yet incorporate rotational as well as contrast transfer function (CTF) invariance. Another important aspect is how multiple distinct 3D reconstructions stemming from heterogeneous 2D projections can be achieved. This can be generally performed by the conventional cryo-EM classification methods, but here we refer to a more specific challenge of faithfully representing the true variability in the data sufficiently well to be used for protein community discovery. This is in contrast to current classification methods that only aim to homogenize the data subset to yield the highest possible resolution. This notion in the data analysis would eventually lead to average densities of the particles that may or may not participate in the same communities. Recently, Verbeke et al. (2020) applied the projection-slice theorem principles to group the particles into consistent subsets prior to 3D classification and, therefore, avoid guessing the number of underlying 3D shapes present in the data. Still, current methods, during the reconstruction of cryo-EM data, assume that sample heterogeneity originates from a small number of

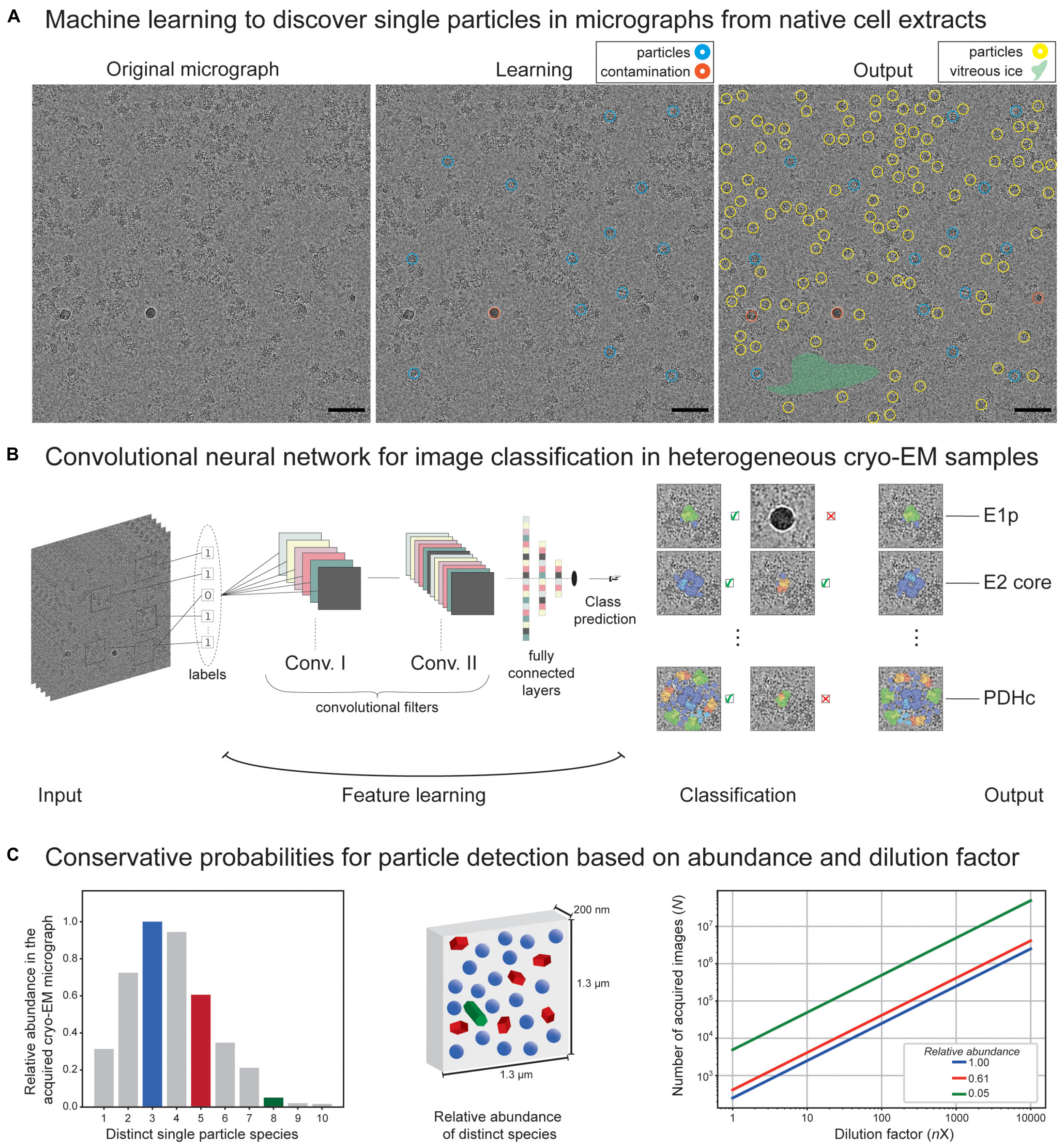

A
Machine learning to discover single particles in micrographs from native cell extracts
Original micrograph
Learning
particles
contamination
Output
particles
vitreous ice
B
Convolutional neural network for image classification in heterogeneous cryo-EM samples
labels
1
1
0
1
1
Conv. I
Conv. II
convolutional filters
fully connected layers
Class prediction
E1p
E2 core
PDHc
Input
Feature learning
Classification
Output
C
Conservative probabilities for particle detection based on abundance and dilution factor
Relative abundance in the acquired cryo-EM micrograph
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Distinct single particle species
200 nm
1.3 μm
1.3 μm
Relative abundance of distinct species
Number of acquired images (N)
10^7
10^6
10^5
10^4
10^3
Relative abundance
1.00
0.61
0.05
1
10
100
1000
10000
Dilution factor (nX)

FIGURE 2 | Application of machine learning on cryo-EM images derived from native cell extracts. **(A)** A cryo-electron micrograph from *C. thermophilum* fractionated cell extracts is shown. During machine learning, the algorithm is being trained to discriminate particles from contamination, vitreous ice, aggregation, and noise. At the end, the algorithm optimally picks and selects learned features that were not previously recognized during learning. Red circles indicate contamination, and blue and yellow circles indicate learned and predicted particles. Size of the circle does not match particle size but represents a correctly picked particle. Green highlighted area signifies empty regions of vitreous ice recognized by the algorithm. **(B)** Structure of a convolutional neural network algorithm frequently used to detect signal in cryo-EM micrographs. Input micrographs are used for feature learning during the convolution step of algorithm training. Optimal training would lead to efficient classification of the single particles and/or their higher-order assemblies and discriminate those from noise, contamination, and aggregates. A final output is achieved with metabolon members in their unbound and bound states as recognized by the convolutional neural networks in heterogeneous cryo-EM micrographs of native cell extracts. **(C)** Conservative probabilities for particle detection based on abundance and dilution factor. In the left panel, an example of 10 distinct single-particle species is shown with their relative abundance following an assumed T-squared distribution. In the middle panel, an illustration of relative particle abundance for three distinct particles (blue, green, and red, representing high, medium, and low abundant species in a calculated 4K × 4K micrograph with a pixel size of 3.17 Å and thickness of 200 nm) is shown. In the right panel, dependency of the number of images required to reach ~5,000 single particles on the dilution factor is shown (assuming no biochemical manipulation for particle enrichment).

independent, distinct states; however, in reality, the number of distinct states is (often) unknown. This issue becomes more important when other specimens of increased complexity are considered. A method that addresses this issue by approximating the continuous 3D density function of a single particle is CryoDRGN (Zhong et al., 2021), a deep neural network-based algorithm. Recent machine-learning methods may improve the protein density of experimental cryo-EM maps, while the use of generative adversarial networks (GANs) trained on pairs of 3D atomic models and their noise-free cryo-EM maps is shown to generate a more realistic ground-truth 3D density map (Sanchez-Garcia et al., 2020a). An excellent discussion by the Scheres laboratory covers these aspects through the implementation of neural networks for simulated cryo-EM 3D reconstructions (Kimanius et al., 2021). Finally, for post-processing of cryo-EM maps, new machine-learning algorithms were developed to account for resolution anisotropy (Ramirez-Aportela et al., 2019; Sanchez-Garcia et al., 2020a).

For machine-learning models to work in the context of data stemming from cryo-EM micrographs of native cell extracts, it is reasonable to assume that they may efficiently be trained to pick and sort the community members by their heterogeneity. However, to construct the corresponding *de novo* 3D cryo-EM maps, novel *ab initio* algorithms should be developed to tackle this complexity. Moreover, the proximity calculations by accounting the Cartesian coordinates of the derived single particles in the cryo-EM micrographs can aid in understanding the protein complex interconnectivity within communities. It would further aid the detection and structural analysis of protein communities and their members.

MODEL BUILDING IN CRYO-EM MAPS FROM NATIVE CELL EXTRACTS COMBINED WITH STRUCTURE PREDICTION

Traditionally, protein complexes from the high-resolution cryo-EM reconstructions can be built because the purified constructs are used. Such approaches are well-established for cryo-EM, but, again, become a challenge for native cell extracts, where the identity of the reconstructed protein complexes and their interactors can be unknown. It is even more difficult to reconstruct such complexes when they are participating in higher-order assemblies, and therefore additional heterogeneity is manifested. cryo-EM may be used to visualize protein communities but, without complementary data, it cannot characterize their structure at a reasonable resolution. It is extremely challenging to determine the 3D models of isolated flexible complexes, but not their native interactions within protein communities. cryo-EM is unlikely to provide discovery or evidence of protein communities by itself without correlating the image information to proteomic, literature, and other sources of data. Interestingly, abundant, rigid complexes within communities can be retrieved at sub-nanometer resolution from native cell extracts, as in the cases of FAS (Kastritis et al., 2017) and pyruvate dehydrogenase complex (PDHc) (Kyrilis et al., 2021).

If high resolution is achieved for a given protein complex, and side-chain resolution is realistic, then multiple methods can be used to model the density, including, for example, cryoID (Ho et al., 2020), that may perform *de novo* model building, assuming that the proteome of the organism is available. However, if the resolution is more than ∼4.0 Å, then side-chain resolution is unattainable, and modeling methods must be ultimately employed [e.g., (Russel et al., 2012; van Zundert et al., 2016)]. In this case, only orthogonal identification methods may be applied to recover the map identity. This information can then be used for subsequent model building. To resolve this unknown density, the previously mentioned proteomic methods for network construction and community detection are of vital importance. Prior to the protein modeling methods, fold recognition should be the primary consideration for structural analysis and implementation of fast-fold search algorithm into the cryo-EM map is important, as proposed by Saha and Morais (2012). Of course, if complexes include other, non-protein components, the identification is laborious. For such scenarios, neural networks are developed to localize nucleotides as well (Mostosi et al., 2020), but machine learning should be expected to resolve cryo-EM densities stemming from multiple types of biological (macro-) molecules. To localize different chemical molecules in a cryo-EM map, a ground truth is required, i.e., the training set as pairs of cryo-EM maps and coordinates of chemical molecules in it. The hydrogen bonding patterns could then be recovered by calculating the geometrical properties of the modeled biomolecule(s) which are used to correlate chemical structure with portions of the cryo-EM maps and, ultimately, serve as input for machine learning.

The abundance of protein complexes within sequential fractions may be correlated to the corresponding structural signatures that were recovered by negative staining or cryo-EM, and therefore assign an identity to recovered structural signatures, which are also members of their respective communities (Kastritis et al., 2017). This was previously performed for *C. thermophilum* complexes using simple cross-correlation functions (Kastritis et al., 2017) but was limited to assigning abundant species. Theoretically, if the abundance of distinct single particles is expected to follow a T-squared distribution (**Figure 2C**, left panel) within a particular thick micrograph (1,300 nm × 1,300 nm × 200 nm, pixel size of 3.17 Å), then their relative abundance can be estimated (**Figure 2C**, middle panel). Without cell lysis (e.g., by cryo-electron tomography of a cell), a surprisingly high number of tilt series is required for less abundant particles to reach ∼5,000 single particles [e.g., enough for efficiently retrieving structural signatures of FAS (Kastritis et al., 2017) or PDHc (Kyrilis et al., 2021)]. After cell lysis and without biochemical enrichment, this effect further magnifies due to dilution (**Figure 2C**). It is important to note that, using cell extracts, protein complexes can be selectively biochemically enriched, and their conservative estimates are shown in **Figure 2C**. Nevertheless, rare species will be difficult to capture, and an extremely high amount of data will be required. In addition, capturing rare species will be algorithmically challenging. Therefore, we expect only abundant complexes to be captured and the abundant community members to be structurally characterized [as in the case of communities involved in oxo acid metabolism (Kyrilis et al., 2021)]. The availability of data for heterogeneous mixtures is still highly scarce. A possible bottleneck is the availability of both MS data and negative staining/cryo-EM data for sequential cellular fractions, preferentially from the same experiment because alterations in the organism biology can drastically alter recovered profiles. Another idea is to generate all possible protein folds from the sequences identified in the fraction using automated 3D structure prediction algorithms and, then, systematically fit those 3D models in the reconstructed densities. Such work has not been performed to date, mainly because current methods are limited to the study of a few abundant protein complexes present in the fractions (Kastritis et al., 2017; Verbeke et al., 2018; Arimura et al., 2020; Ho et al., 2020; Kyrilis et al., 2021; Su et al., 2021) and, sometimes, their communities (Kastritis et al., 2017; Kyrilis et al., 2021).

Protein structure prediction, in particular, recently witnessed advances, not only in traditional structure prediction methods [e.g., ROSETTA (Leman et al., 2020), I-TASSER (Roy et al., 2010)], but also in methods that are based on machine/deep learning (Torrisi et al., 2020), such as basic feed-forward neural network, CNN, recurrent neural network (RRN), and generative adversarial networks (GAN) (Torrisi et al., 2020). A recent example that excelled in the Critical Assessment of protein Structure Prediction [CASP, (Moult et al., 1995)], which is a blind protein structure prediction experiment, is AlphaFold2 developed by DeepMind. AlphaFold2 is based on an attention-based neural network system (Jumper et al., 2020) and was trained on all publicly available experimental 3D structures in the Protein Data Bank (PDB). Even if a fold can be recognized [and, currently thousands of those were predicted *via* machine-learning-based ROSETTA functions (Yang et al., 2020) and added in Pfam (Mistry et al., 2021)], it is still far from explaining the higher-order interactions captured within the cryo-EM map. For understanding the molecular recognition, large protein complex assembly and community function are still out of reach: only methods that include experimental data to drive the modeling process with physics-based potentials [e.g., HADDOCK (van Zundert et al., 2016), IMP (Russel et al., 2012)] can provide physically realistic models. It is noted that the Critical Assessment of PRotein–protein Interactions (CAPRI) (Janin et al., 2003) is a blind experiment where algorithms are tested in their ability to solve the biomolecular recognition problem. To date, in CAPRI, the top-performing algorithms are physics-based which integrate experimental data from various targets.

DISCUSSION: ASPIRING DEEPER STRUCTURAL CHARACTERIZATION OF PROTEIN COMMUNITIES

Machine/deep learning is applied to a multitude of optimization problems that are related with the recovery and characterization of protein communities at high resolution. In each step toward their multi-scale molecular characterization, distinct approaches are applied, fitted to answer diverse questions arising from experimentally measured multidimensional data. Unambiguous and large training sets, avoiding overfitting and careful cross-validation, true test sets, and, overall, systematic benchmarking are all required to accurately predict the desirable outcome. However, the complex nature of native cell extracts has not yet been fully explored systematically from a structural perspective, especially in (a) deriving 3D reconstructions out of the cryo-EM data in an un-/supervised manner, (b) model building in the recovered 3D maps, and (c) interconnecting multi-scale structural information from (a) and (b) to discover structural data about protein communities. As of note, cryo-electron tomography of complex specimen and associated image processing methods for in-tomogram particle detection and classification (Xu et al., 2011, 2019; Chen et al., 2013; Zhou et al., 2020) may also inspire methods for chemically heterogeneous single-particle datasets (and vice versa) for future applications in the characterization of protein communities. Structural biology of native cell extracts, therefore, provides an ideal test bed for the development and application of artificial intelligence. It is of paramount importance to note that the studies of native cell extracts and the structural characterization of protein communities that reside within should not simply focus on retrieving high resolution. The extreme flexibility and heterogeneity of the participating biomolecules pose a practical limitation on the resolution; even if high resolution is achieved, it will be non-uniform and will be prohibitive for a deeper understanding of function. Instead, the studies should aim to characterize components, stoichiometry, and, *via* cryo-EM, to utilize structural data in the discovery of PPIs within communities. We expect that, in the years to come, more datasets for heterogeneous specimen will be available through dedicated databases [e.g., UNIPROT (UniProt Consortium, 2019), PRIDE (Perez-Riverol et al., 2019), CORUM (Giurgiu et al., 2019), EMDB (Lawson et al., 2011), EMPIAR (Iudin et al., 2016), and PDB (Berman et al., 2000)]. Given the exponential increase of open-source data, and significant advancement in computational hardware over the past decade, machine/deep learning algorithms will become more efficient. The machine-learning methods will be eventually able to tackle some of the aforementioned limitations in the analysis of complex mixtures and homogenates of soluble and/or membrane extracts with success, aiming to provide answers to the, yet, elusive conundrum of macromolecular recognition: *How and why biomolecules interact?*

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

PK conceived the review and wrote the manuscript with the contributions from FK and JB. FK made figures and collected the data presented in **Figure 2A**. The authors thank all eight reviewers for their highly valuable feedback and the subsequent conceptual improvements that were manifested in our manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Federal Ministry for Education and Research (BMBF, ZIK program) (Grant nos. 03Z22HN23 and 03COV04 to PK), the European Regional Development Funds for Saxony-Anhalt (grant no. EFRE: ZS/2016/04/78115 to PK), funding by Deutsche Forschungsgemeinschaft (DFG) (project number 391498659, RTG 2467), and the Martin-Luther University of Halle-Wittenberg.

REFERENCES

Arimura, Y., Shih, R. M., Froom, R., and Funabiki, H. (2020). Nucleosome structural variations in interphase and metaphase chromosomes. *bioRxiv* [Preprint]. doi: 10.1101/2020.11.12.380386

Barnes, S. J., and Weitzman, P. D. (1986). Organization of citric acid cycle enzymes into a multienzyme cluster. *FEBS Lett.* 201, 267–270. doi: 10.1016/0014-5793(86)80621-4

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., et al. (2011). The quantitative proteome of a human cell line. *Mol. Syst. Biol.* 7:549. doi: 10.1038/msb.2011.82

Bepler, T., Kelley, K., Noble, A. J., and Berger, B. (2020). Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat. Commun.* 11:5208. doi: 10.1038/s41467-020-18952-1

Bepler, T., Morin, A., Noble, A. J., Brasch, J., Shapiro, L., and Berger, B. (2018). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Res. Comput. Mol. Biol.* 10812, 245–247.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Pittsburgh, PA: Association for Computing Machinery). doi: 10.1145/130385.130401

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M., and Forster, F. (2013). Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.* 182, 235–245. doi: 10.1016/j.jsb.2013.03.002

Elias, J. E., and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214. doi: 10.1038/nmeth1019

Fossati, A., Li, C., Sykacek, P., Heusel, M., Frommelt, F., Uliana, F., et al. (2020). Systematic protein complex profiling and differential analysis from co-fractionation mass spectrometry data. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.06.080465

Fursch, J., Kammer, K. M., Kreft, S. G., Beck, M., and Stengel, F. (2020). Proteome-wide structural probing of low-abundant protein interactions by cross-linking mass spectrometry. *Anal. Chem.* 92, 4016–4022. doi: 10.1021/acs.analchem.9b05559

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636. doi: 10.1038/nature04532

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., et al. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* 47, D559–D563. doi: 10.1093/nar/gky973

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., O'Meara, M. J., et al. (2020). A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.22.002386

Gotze, M., Iacobucci, C., Ihling, C. H., and Sinz, A. (2019). A simple cross-linking/mass spectrometry workflow for studying system-wide protein interactions. *Anal. Chem.* 91, 10236–10244. doi: 10.1021/acs.analchem.9b02372

Hammerling, M. J., Fritz, B. R., Yoesep, D. J., Kim, D. S., Carlson, E. D., and Jewett, M. C. (2020). In vitro ribosome synthesis and evolution through ribosome display. *Nat. Commun.* 11:1108. doi: 10.1038/s41467-020-14705-2

Han, B. G., Dong, M., Liu, H., Camp, L., Geller, J., Singer, M., et al. (2009). Survey of large protein complexes in D. vulgaris reveals great structural diversity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16580–16585. doi: 10.1073/pnas.0813068106

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261. doi: 10.1093/nar/gkh036

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., et al. (2012). A census of human soluble protein complexes. *Cell* 150, 1068–1081. doi: 10.1016/j.cell.2012.08.011

Havugimana, P. C., Hu, P., and Emili, A. (2017). Protein complexes, big data, machine learning and integrative proteomics: lessons learned over a decade of systematic analysis of protein interaction networks. *Expert Rev. Proteomics* 14, 845–855. doi: 10.1080/14789450.2017.1374179

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D455. doi: 10.1093/nar/gkh052

Ho, C. M., Li, X., Lai, M., Terwilliger, T. C., Beck, J. R., Wohlschlegel, J., et al. (2020). Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nat. Methods* 17, 79–85. doi: 10.1038/s41592-019-0637-y

Hu, L. Z., Goebels, F., Tan, J. H., Wolf, E., Kuzmanov, U., Wan, C., et al. (2019). EPIC: software toolkit for elution profile-based inference of protein complexes. *Nat. Methods* 16, 737–742. doi: 10.1038/s41592-019-0461-4

Huang, R., Gao, X., Xu, Z., Zhu, W., Wei, D., Jiang, B., et al. (2020). Decision tree searching strategy to boost the identification of cross-linked peptides. *Anal. Chem.* 92, 13702–13710. doi: 10.1021/acs.analchem.0c00452

Huang, Z., and Penczek, P. A. (2004). Application of template matching technique to particle detection in electron micrographs. *J. Struct. Biol.* 145, 29–40. doi: 10.1016/j.jsb.2003.11.004

Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J., and Patwardhan, A. (2016). EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* 13, 387–388. doi: 10.1038/nmeth.3806

Jain, V., and Seung, H. S. (2008). "Natural image denoising with convolutional networks," in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, (Vancouver, BC: Curran Associates Inc).

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., et al. (2003). Assessment of: CAPRI: a critical assessment of PRedicted interactions. *Proteins* 52, 2–9. doi: 10.1002/prot.10381

Jeffery, C. J. (2014). An introduction to protein moonlighting. *Biochem. Soc. Trans.* 42, 1679–1683. doi: 10.1042/BST20140226

Ji, C., Li, S., Reilly, J. P., Radivojac, P., and Tang, H. (2016). XLSearch: a probabilistic database search algorithm for identifying cross-linked peptides. *J. Proteome Res.* 15, 1830–1841. doi: 10.1021/acs.jproteome.6b00004

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., et al. (2020). "High Accuracy Protein Structure Prediction Using Deep Learning," in *Proceedings of the 14th Critical Assessment of Techniques for Protein Structure Prediction*, ed. CASP.

Karim, A. S., and Jewett, M. C. (2016). A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab. Eng.* 36, 116–126. doi: 10.1016/j.ymben.2016.03.002

Kastritis, P. L., and Gavin, A. C. (2018). Enzymatic complexes across scales. *Essays Biochem.* 62, 501–514. doi: 10.1042/EBC20180008

Kastritis, P. L., O'Reilly, F. J., Bock, T., Li, Y., Rogon, M. Z., Buczak, K., et al. (2017). Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol. Syst. Biol.* 13:936. doi: 10.15252/msb.20167412

Kimanius, D., Zickert, G., Nakane, T., Adler, J., Lunz, S., Schonlieb, C.-B., et al. (2021). Exploiting prior knowledge about biological macromolecules in cryo-EM structure determination. *IUCrJ* 8, 60–75. doi: 10.1107/S2052252520014384

Kristensen, A. R., Gsponer, J., and Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* 9, 907–909. doi: 10.1038/nmeth.2131

Kuhlbrandt, W. (2014). Biochemistry. The resolution revolution. *Science* 343, 1443–1444. doi: 10.1126/science.1251652

Kyrilis, F. L., Meister, A., and Kastritis, P. L. (2019). Integrative biology of native cell extracts: a new era for structural characterization of life processes. *Biol. Chem.* 400, 831–846. doi: 10.1515/hsz-2018-0445

Kyrilis, F. L., Semchonok, D. A., Skalidis, I., Tuting, C., Hamdi, F., O'Reilly, F. J., et al. (2021). Integrative structure of a 10-megadalton eukaryotic pyruvate dehydrogenase complex from native cell extracts. *Cell Rep.* 34:108727. doi: 10.1016/j.celrep.2021.108727

Larance, M., Kirkwood, K. J., Tinti, M., Brenes Murillo, A., Ferguson, M. A., and Lamond, A. I. (2016). Global membrane protein interactome analysis using in vivo crosslinking and mass spectrometry-based protein correlation profiling. *Mol. Cell. Proteomics* 15, 2476–2490. doi: 10.1074/mcp.O115.055467

Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., et al. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39, D456–D464. doi: 10.1093/nar/gkq880

Leman, J. K., Weitzner, B. D., Lewis, S. M., dolf-Bryfogle, J. A., Alam, N., Alford, R. F., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* 17, 665–680. doi: 10.1038/s41592-020-0848-2

Liu, F., Rijkers, D. T., Post, H., and Heck, A. J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* 12, 1179–1184. doi: 10.1038/nmeth.3603

Maulik, U., Basu, S., and Ray, S. (2017). Identifying protein complexes in PPI network using non-cooperative sequential game. *Sci. Rep.* 7:8410. doi: 10.1038/s41598-017-08760-x

McCafferty, C. L., Verbeke, E. J., Marcotte, E. M., and Taylor, D. W. (2020). Structural biology in the multi-omics era. *J. Chem. Inf. Model.* 60, 2424–2429. doi: 10.1021/acs.jcim.9b01164

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Mostosi, P., Schindelin, H., Kollmannsberger, P., and Thorn, A. (2020). Haruspex: a neural network for the automatic identification of oligonucleotides and protein secondary structure in cryo-electron microscopy maps. *Angew. Chem. Int. Ed. Engl.* 59, 14788–14795. doi: 10.1002/anie.202000421

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* 23, ii–v. doi: 10.1002/prot.340230303

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi: 10.1038/nmeth.1938

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450. doi: 10.1093/nar/gky1106

Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14, 290–296. doi: 10.1038/nmeth.4169

Ramirez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M., and Sorzano, C. O. S. (2019). DeepRes: a new deep-learning- and aspect-based local resolution method for electron-microscopy maps. *IUCrJ* 6(Pt 6), 1054–1063. doi: 10.1107/S2052252519011692

Rao, V. S., Srinivas, K., Sujini, G. N., and Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics* 2014:147648. doi: 10.1155/2014/147648

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/NECO_a_00990

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10:e1001244. doi: 10.1371/journal.pbio.1001244

Saha, I., Zubek, J., Klingstrom, T., Forsberg, S., Wikander, J., Kierczak, M., et al. (2014). Ensemble learning prediction of protein-protein interactions using proteins functional annotations. *Mol. Biosyst.* 10, 820–830. doi: 10.1039/c3mb70486f

Saha, M., and Morais, M. C. (2012). FOLD-EM: automated fold recognition in medium- and low-resolution (4-15 A) electron density maps. *Bioinformatics* 28, 3265–3273. doi: 10.1093/bioinformatics/bts616

Salas, D., Stacey, R. G., Akinlaja, M., and Foster, L. J. (2020). Next-generation interactomics: considerations for the use of co-elution to measure protein interaction networks. *Mol. Cell. Proteomics* 19, 1–10. doi: 10.1074/mcp.R119.001803

Sanchez-Garcia, R., Gomez-Blanco, J., Cuervo, A., Carazo, J. M., Sorzano, C. O. S., and Vargas, J. (2020a). DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.12.148296

Sanchez-Garcia, R., Segura, J., Maluenda, D., Sorzano, C. O. S., and Carazo, J. M. (2020b). MicrographCleaner: a python package for cryo-EM micrograph cleaning using deep learning. *J. Struct. Biol.* 210:107498. doi: 10.1016/j.jsb.2020.107498

Shatsky, M., Hall, R. J., Brenner, S. E., and Glaeser, R. M. (2009). A method for the alignment of heterogeneous macromolecules from electron microscopy. *J. Struct. Biol.* 166, 67–78. doi: 10.1016/j.jsb.2008.12.008

Silverman, A. D., Karim, A. S., and Jewett, M. C. (2020). Cell-free gene expression: an expanded repertoire of applications. *Nat. Rev. Genet.* 21, 151–170. doi: 10.1038/s41576-019-0186-3

Sindelar, C. V., and Grigorieff, N. (2011). An adaptation of the Wiener filter suitable for analyzing images of isolated single particles. *J. Struct. Biol.* 176, 60–74. doi: 10.1016/j.jsb.2011.06.010

Sinz, A. (2018). Cross-linking/mass spectrometry for studying protein structures and protein-protein interactions: where are we now and where should we go from here? *Angew. Chem. Int. Ed. Engl.* 57, 6390–6396. doi: 10.1002/anie.201709559

Skalidis, I., Tuting, C., and Kastritis, P. L. (2020). Unstructured regions of large enzymatic complexes control the availability of metabolites with signaling functions. *Cell Commun. Signal.* 18:136. doi: 10.1186/s12964-020-00631-9

Stacey, R. G., Skinnider, M. A., Scott, N. E., and Foster, L. J. (2017). A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics* 18:457. doi: 10.1186/s12859-017-1865-8

Su, C. C., Lyu, M., Morgan, C. E., Bolla, J. R., Robinson, C. V., and Yu, E. W. (2021). A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. *Nat. Methods* 18, 69–75. doi: 10.1038/s41592-020-01021-2

Tegunov, D., and Cramer, P. (2019). Real-time cryo-electron microscopy data preprocessing with Warp. *Nat. Methods* 16, 1146–1152. doi: 10.1038/s41592-019-0580-y

Tin Kam, H. (1995). "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC. doi: 10.1109/ICDAR.1995.598994

Titeca, K., Lemmens, I., Tavernier, J., and Eyckerman, S. (2019). Discovering cellular protein-protein interactions: technological strategies and opportunities. *Mass Spectrom. Rev.* 38, 79–111. doi: 10.1002/mas.21574

Torrisi, M., Pollastri, G., and Le, Q. (2020). Deep learning methods in protein structure prediction. *Comput. Struct. Biotechnol. J.* 18, 1301–1310. doi: 10.1016/j.csbj.2019.12.011

Tuting, C., Iacobucci, C., Ihling, C. H., Kastritis, P. L., and Sinz, A. (2020). Structural analysis of 70S ribosomes by cross-linking/mass spectrometry reveals conformational plasticity. *Sci. Rep.* 10:12618. doi: 10.1038/s41598-020-69313-3

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

van Zundert, G. C. P., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., et al. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* 428, 720–725. doi: 10.1016/j.jmb.2015.09.014

Verbeke, E. J., Mallam, A. L., Drew, K., Marcotte, E. M., and Taylor, D. W. (2018). Classification of single particles from human cell extract reveals distinct structures. *Cell Rep.* 24, 259–268.e3. doi: 10.1016/j.celrep.2018.06.022

Verbeke, E. J., Zhou, Y., Horton, A. P., Mallam, A. L., Taylor, D. W., and Marcotte, E. M. (2020). Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections. *J. Struct. Biol.* 209:107416. doi: 10.1016/j.jsb.2019.107416

Vidal, M., Cusick, M. E., and Barabasi, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., et al. (2019). SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol.* 2:218. doi: 10.1038/s42003-019-0437-z

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., et al. (2015). Panorama of ancient metazoan macromolecular complexes. *Nature* 525, 339–344. doi: 10.1038/nature14877

Wang, F., Gong, H., Liu, G., Li, M., Yan, C., Xia, T., et al. (2016). DeepPicker: a deep learning approach for fully automated particle picking in cryo-EM. *J. Struct. Biol.* 195, 325–336. doi: 10.1016/j.jsb.2016.07.006

Wang, L., Wang, H. F., Liu, S. R., Yan, X., and Song, K. J. (2019). Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Sci. Rep.* 9:9848. doi: 10.1038/s41598-019-46369-4

Xie, J., Xu, L., and Chen, E. (2012). "Image denoising and inpainting with deep neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol. 1, (Lake Tahoe, NV: Curran Associates Inc).

Xu, M., Beck, M., and Alber, F. (2011). Template-free detection of macromolecular complexes in cryo electron tomograms. *Bioinformatics* 27, i69–i76. doi: 10.1093/bioinformatics/btr207

Xu, M., Singla, J., Tocheva, E. I., Chang, Y. W., Stevens, R. C., Jensen, G. J., et al. (2019). De novo structural pattern mining in cellular electron cryotomograms. *Structure* 27, 679–691.e14. doi: 10.1016/j.str.2019.01.005

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1496–1503. doi: 10.1073/pnas.1914677117

Zhang, D., and Kabuka, M. (2019). Multimodal deep representation learning for protein interaction identification and protein family classification. *BMC Bioinformatics* 20(Suppl. 16):531. doi: 10.1186/s12859-019-3084-y

Zhang, J., Wang, Z., Chen, Y., Han, R., Liu, Z., Sun, F., et al. (2019). PIXER: an automated particle-selection method based on segmentation using a deep neural network. *BMC Bioinformatics* 20:41. doi: 10.1186/s12859-019-2614-y

Zhao, J., and Lei, X. (2019). Detecting overlapping protein complexes in weighted PPI network based on overlay network chain in quotient space. *BMC Bioinformatics* 20(Suppl. 25):682. doi: 10.1186/s12859-019-3256-9

Zhong, E. D., Bepler, T., Berger, B., and Davis, J. H. (2021). CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* 18, 176–185. doi: 10.1038/s41592-020-01049-4

Zhou, B., Yu, H., Zeng, X., Yang, X., Zhang, J., and Xu, M. (2020). One-shot learning with attention-guided segmentation in cryo-electron tomography. *Front. Mol. Biosci.* 7:613347. doi: 10.3389/fmolb.2020.613347

Zhu, Y., Ouyang, Q., and Mao, Y. (2017). A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics* 18:348. doi: 10.1186/s12859-017-1757-y

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/fmolb.2021.637396

A Deep Graph Network–Enhanced Sampling Approach to Efficiently Explore the Space of Reduced Representations of Proteins

Federico Errica[1†*], *Marco Giulini*[2,3†*], *Davide Bacciu*[1†], *Roberto Menichetti*[2,3†], *Alessio Micheli*[1†] and *Raffaello Potestio*[2,3†]

[1] *Department of Computer Science, University of Pisa, Pisa, Italy,* [2] *Physics Department, University of Trento, Trento, Italy,* [3] *INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, Trento, Italy*

OPEN ACCESS

Edited by:
Elif Ozkirimli,
Roche, Switzerland

Reviewed by:
Ilpo Vattulainen,
University of Helsinki, Finland
Matteo Manica,
IBM Research–Zurich, Switzerland

****Correspondence:***
Federico Errica
federico.errica@phd.unipi.it
Marco Giulini
marco.giulini@unitn.it

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *03 December 2020*
Accepted: *17 February 2021*
Published: *29 April 2021*

Citation:
Errica F, Giulini M, Bacciu D, Menichetti R, Micheli A and Potestio R (2021) A Deep Graph Network–Enhanced Sampling Approach to Efficiently Explore the Space of Reduced Representations of Proteins. Front. Mol. Biosci. 8:637396. doi: 10.3389/fmolb.2021.637396

The limits of molecular dynamics (MD) simulations of macromolecules are steadily pushed forward by the relentless development of computer architectures and algorithms. The consequent explosion in the number and extent of MD trajectories induces the need for automated methods to rationalize the raw data and make quantitative sense of them. Recently, an algorithmic approach was introduced by some of us to identify the subset of a protein's atoms, or mapping, that enables the most informative description of the system. This method relies on the computation, for a given reduced representation, of the associated mapping entropy, that is, a measure of the information loss due to such simplification; albeit relatively straightforward, this calculation can be time-consuming. Here, we describe the implementation of a deep learning approach aimed at accelerating the calculation of the mapping entropy. We rely on Deep Graph Networks, which provide extreme flexibility in handling structured input data and whose predictions prove to be accurate and-remarkably efficient. The trained network produces a speedup factor as large as 10^5 with respect to the algorithmic computation of the mapping entropy, enabling the reconstruction of its landscape by means of the Wang–Landau sampling scheme. Applications of this method reach much further than this, as the proposed pipeline is easily transferable to the computation of arbitrary properties of a molecular structure.

Keywords: molecular dynamics, coarse-grained methods, mapping entropy, deep learning, neural networks for graphs, neural networks

INTRODUCTION

Molecular dynamics (MD) simulations (Alder and Wainwright, 1959; Karplus, 2002) are an essential and extremely powerful tool in the computer-aided investigation of matter. The usage of classical, all-atom simulations has boosted our understanding of a boundless variety of different physical systems, ranging from materials (metals, alloys, fluids, etc.) to biological macromolecules such as proteins. As of today, the latest software and hardware developments have pushed the size of systems that MD simulations can address to the millions of atoms (Singharoy et al., 2019), and the time scales covered by a single run can approach the millisecond for relatively small molecules (Shaw et al., 2009).

In general, a traditional MD-based study proceeds in four steps, here schematically summarized in **Figure 1**. First, the system of interest has to be identified; this apparently obvious problem can

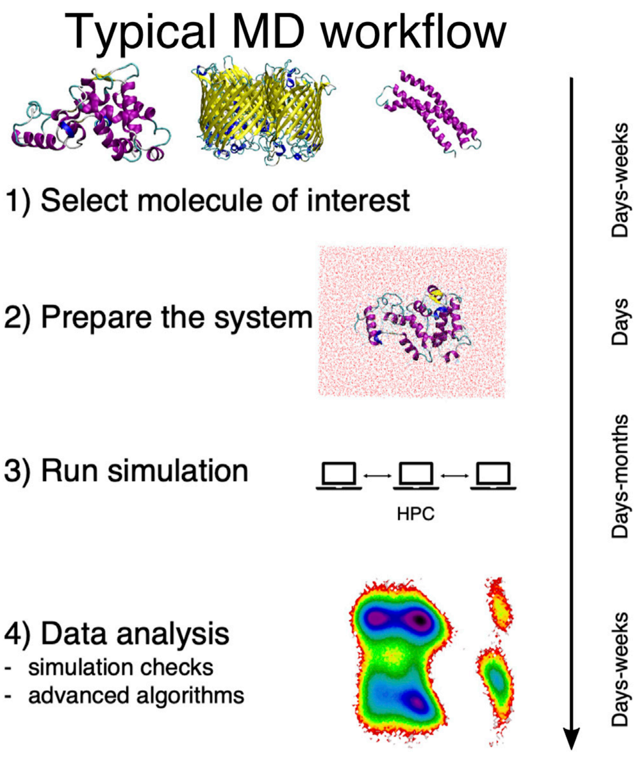


FIGURE 1 | Schematic representation of the typical workflow of a molecular dynamics study. On the right we report the average time scales required for each step of the process.

actually require a substantial effort *per se*, e.g., in the case of dataset-wide investigations. Second, the simulation setup has to be constructed, which is another rather nontrivial step (Kandt et al., 2007). Then the simulation has to be run, typically on a high performance computing infrastructure. Finally, the output has to be analyzed and rationalized *in order to extract information from the data.*

This last step is particularly delicate, and it is acquiring an ever growing prominence as large and long MD simulations can be more and more effortlessly performed. The necessity thus emerges to devise a parameter-free, automated “filtering” procedure to describe the examined system in simpler, intelligible terms and make sense of the immense amount of data we can produce—but not necessarily understand.

In the field of soft and biological matter, coarse-graining (CG) methods represent a notable example of a systematic procedure that aims at extracting, out of a detailed model of a given macromolecular system, the relevant properties of the latter (Marrink et al., 2007; Takada, 2012; Saunders and Voth, 2013; Potestio et al., 2014). This is achieved through the construction of simplified representations of the system that have fewer degrees of freedom with respect to the reference model while retaining key features and properties of interest. In biophysical applications, this amounts to describing a biomolecule, such as a protein, using a number of constituent units, called CG sites, lower than the number of particles composing the original, atomistic system.

The coarse-graining process in soft matter requires two main ingredients, separately addressing two entangled, however conceptually very different, problems (Noid, 2013a). The first ingredient consists of the definition of a *mapping*, that is, the transformation $\mathbf{M}(\mathbf{r}) = \mathbf{R}$ that connects a high-resolution representation $\mathbf{r}$ of the system's configuration to a low-resolution one $\mathbf{R}$. The mapping thus pertains to the *description* of the system's behavior,“filtered” so as to retain only a subset of the original degrees of freedom. The second ingredient is the set of effective interactions introduced among the CG sites; these CG potentials serve the purpose of *reproducing a posteriori* the emergent properties of the system directly from its simplified representation rather than from its higher-resolution model. Both ingredients are highlighted in **Figure 2**, where we display a visual comparison between a high-resolution representation of a protein and one among its possible simplified depictions, as defined by a particular selection of the molecule's retained atoms.

During the past few decades, substantial effort has been invested in the correct parameterization of CG potentials (Noid et al., 2008; Shell, 2008; Noid, 2013b): most of the research focused on accurately reproducing the system's behavior that arises from a model relying on a specific choice of the CG observational filter. Critically, the investigation of the quality of the filter *itself*—that is, the definition of the CG mapping—has received much less attention. Indeed, most methods developed in the field of soft matter do not make use of a system-specific, algorithmic procedure for the selection of the effective sites but rather rely on general criteria, based on physical and chemical intuition, to group together atoms in CG “beads” irrespective of their local environment and global thermodynamics (Kmiecik et al., 2016)—one notable example being the representation of a protein in terms of its α-carbon atoms.

While acceptable in most practical applications, this approach entails substantial limitations: in fact, the CG process implies a loss of information and, through the application of universal mapping strategies, system-specific properties, albeit relevant, might be “lost in translation” from a higher to a lower resolution representation (Foley et al., 2015; Jin et al., 2019; Foley et al., 2020). Hence, a method would be required that enables the automated identification of which subset of retained degrees of freedom of a given system preserves the majority of important detail from the reference, while at the same time reducing the complexity of the problem. In the literature, this task has been addressed through several different techniques, such as graph-theoretical analyses (Webb et al., 2019), geometric criteria (Bereau and Kremer, 2015), and machine learning algorithms (Murtola et al., 2007; Wang and Bombarelli, 2019; Li et al., 2020). These efforts are rooted in the assumption that the optimal CG representation of a system can be determined solely by exploiting a subset of features of the latter. In contrast, taking into account the full information content encoded in the system requires statistical mechanics-based models, where the optimal CG mapping is expected to emerge systematically from the comparison between the CG model and its atomistic counterpart. Within this framework, pioneering works rely on

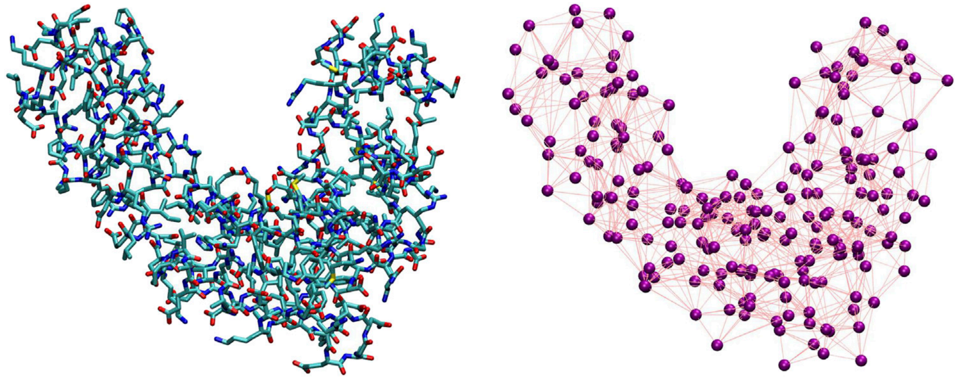


FIGURE 2 | Comparison between an all-atom, detailed description of a protein **(left)** and one of its possible coarse-grained representations **(right)**. The purple spheres on the right plot correspond to CG sites, while the edges connecting them represent the effective interactions.

a simplified description of the system (Koehl et al., 2017; Diggins et al., 2018), e.g., provided by analytically solvable, linearized elastic network models, which cannot faithfully reproduce the complexity of the *true* interaction network.

A recently developed statistical mechanics-based strategy that aims at overcoming such limitations is the one relying on the minimization of the mapping entropy (Giulini et al., 2020), which performs, in an unsupervised manner, the identification of the subset of a molecule's atoms that retains the largest possible amount of information about its behavior. This scheme relies on the calculation of the mapping entropy S_{map} (Shell, 2008; Rudzinski and Noid, 2011; Shell, 2012; Foley et al., 2015), a quantity that provides a measure of the dissimilarity between the probability density of the system configurations in the original, high-resolution description and the one marginalized over the discarded atoms. S_{map} is employed as a cost function and minimized over the possible reduced representations so as to systematically single out the most informative ones.

The method just outlined suffers from two main bottlenecks: on the one hand, the determination of the mapping entropy is *per se* computationally intensive; even though smart workarounds can be conceived and implemented to speed up the calculation, its relative complexity introduces a nontrivial slowdown in the minimization process. On the other hand, the sheer size of the space of possible CG mappings of a biomolecule is so ridiculously large that it makes a random search practically useless and an exhaustive enumeration simply impossible. Hence, an optimization procedure is required to identify the simplified descriptions that entail the largest amount of information about the system. Unfortunately, this procedure nonetheless implies the calculation of S_{map} over a very large number of tentative mappings, making the optimization, albeit possible, computationally intensive and time consuming.

In this work, we present a novel computational protocol that suppresses the computing time of the optimization procedure by several orders of magnitude, while at the same time boosting the sampling accuracy. This strategy relies on the fruitful, and to the best of our knowledge unprecedented combination of two very different techniques: graph-based machine learning models (Micheli, et al., 2009; Bronstein et al., 2017; Hamilton et al., 2017; Battaglia et al., 2018; Zhang et al., 2018; Zhang et al., 2019; Bacciu et al., 2020; Wu et al., 2021) and the Wang–Landau enhanced sampling algorithm (Wang and Landau, 2001a; Wang and Landau, 2001b; Shell et al., 2002; Barash et al., 2017). The first serves the purpose of reducing the computational cost associated with the estimation of the mapping entropy; the second enables the efficient and thorough exploration of the mapping space of a biomolecule.

An essential element of the proposed method is thus a graph-based representation of our object of interest, namely a protein. With their long and successful story both in the field of coarse-graining (Gfeller and Rios, 2007; Webb et al., 2019; Li et al., 2020) and in the prediction of protein properties (Borgwardt et al., 2005; Ralaivola et al., 2005; Micheli et al., 2007; Fout et al., 2017; Gilmer et al., 2017; Torng and Altman, 2019), graph-based learning models represent a rather natural and common choice to encode the (static) features of a molecular structure; here, we show that a graph-based machine learning approach can reproduce the results of mapping entropy estimate obtained by means of a much more time-consuming algorithmic workflow. To this end, we rely on Deep Graph Networks (DGNs) (Bacciu et al., 2020), a family of machine learning models that learn from graph-structured data, where the graph has a variable size and topology; by training the model on a set of tuples (protein, CG mapping, and S_{map}), we can infer the S_{map} values of unseen mappings associated with the same protein making use of a tiny fraction of the extensive amount of information employed in the original method, i.e., the molecular structure viewed as a graph. Compared to the algorithmic workflow presented in Giulini et al. (2020), the trained DGN proves capable of accurately calculating the mapping entropy arising from a particular selection of retained atoms throughout the molecule in a negligible time.

This computational speedup can be leveraged to perform a thorough, quasi-exhaustive characterization of the mapping entropy landscape in the space of possible CG representations of a system, a notable advancement with respect to the relatively limited exploration performed in Giulini et al. (2020). Specifically, by combining inference of the DGNs with the Wang–Landau sampling technique, we here provide an estimate of the density of states associated with the S_{map}, that is, the number of CG representations in the biomolecule mapping space that generate a specific amount of information loss with respect to the all-atom reference. A comparison of the WL results on the DGNs with the exact ones obtained from a random sampling of mappings shows that the machine learning model is able to capture the correct population of CG representations in the S_{map} space. This analysis further highlights the accuracy of the model in predicting a complex observable such as the mapping entropy, which in principle depends on the whole configurational space of the macromolecule, only starting from the sole knowledge of the static structure of the latter.

MATERIALS AND METHODS

In this section, we outline the technical ingredients that lie at the basis of the results obtained in this study. Specifically, in *Mapping entropy* we summarize the mapping entropy protocol for optimizing CG representations presented in Giulini et al. (2020); in *Protein structures and data sets* we briefly describe the two proteins analyzed in this work as well as the data sets fed to the machine learning architecture; in *Data Representation and Machine Learning model* we illustrate our choice for the representation of the input data, together with theoretical and computational details about DGNs; finally, in *Wang–Landau Sampling* we describe our implementation of the Wang–Landau sampling algorithm as applied to the reconstruction of the mapping entropy landscape of a system.

Mapping Entropy

The challenge of identifying maximally informative CG representations for a biomolecular system has been recently tackled by some of us (Giulini et al., 2020); specifically, we developed an algorithmic procedure to find the mappings that minimize the amount of information that is lost when the number of degrees of freedom with which one observes the system is *decimated*, that is, a subset of its atoms is retained while the remainder is integrated out. The quantity that measures this loss is called mapping entropy S_{map} (Shell, 2008; Rudzinski and Noid, 2011; Shell, 2012; Foley et al., 2015), which in the case of decimated CG representations can be expressed as a Kullback–Leibler divergence D_{KL} (Kullback and Leibler, 1951) between two probability distributions (Rudzinski and Noid, 2011),

$$S_{map} = k_B \times D_{KL}\left(p_r(\mathbf{r}) \| \bar{p}_r(\mathbf{r})\right) = k_B \int d\mathbf{r}\, p_r(\mathbf{r}) \ln\left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})}\right]. \quad (1)$$

Here, $p_r(\mathbf{r})$ is the probability of sampling a configuration $\mathbf{r}$ in the high-resolution description, namely, the Boltzmann distribution $p_r(\mathbf{r}) \propto \exp[-\beta u(\mathbf{r})]$, where $u(\mathbf{r})$ is the atomistic potential and $\beta = 1/k_B T$ is the inverse temperature. $\bar{p}_r(\mathbf{r})$, on the other hand, is the distribution obtained by observing the system through the “coarse-graining grid,” i.e., in terms of the selected CG mapping. $\bar{p}_r(\mathbf{r})$ is defined as (Rudzinski and Noid, 2011)

$$\bar{p}_r(\mathbf{r}) = p_R[\mathbf{M}(\mathbf{r})]/\Omega_1[\mathbf{M}(\mathbf{r})], \quad (2)$$

where

$$p_R(\mathbf{R}) = \frac{1}{Z} \int d\mathbf{r} e^{-\beta u(\mathbf{r})} \delta[\mathbf{M}(\mathbf{r}) - \mathbf{R}] \quad (3)$$

is the probability of sampling the configuration $\mathbf{R} = \mathbf{M}(\mathbf{r})$ in the low-resolution description—Z being the canonical partition function of the system—while

$$\Omega_1(\mathbf{R}) = \int d\mathbf{r}\, \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \quad (4)$$

is the number of microstates $\mathbf{r}$ that map onto the CG configuration $\mathbf{R}$.

The mapping entropy quantifies the information loss one experiences by replacing the original, microscopic distribution $p_r(\mathbf{r})$ of the system by an effective one in which the probability of a CG macrostate is equally redistributed to all microstates that map onto it. It follows that different choices of the CG mapping lead to different $\bar{p}_r(\mathbf{r})$ and, consequently, to different amounts of information losses arising from CG’ing.

The definition in **Eq. 1** does not allow, given a CG representation, to directly determine the associated mapping entropy. It is however possible to perform a cumulant expansion of **Eq. 1**; by doing so, Giulini et al. (2020) showed that S_{map} can be approximately calculated as a weighted average over all CG macrostates $\mathbf{R}$ of the variances of the atomistic potential energies of all configurations $\mathbf{r}$ that map onto a specific macrostate. This strategy enables one to measure S_{map} only provided a set of all-atom configurations sampled from $p_r(\mathbf{r})$ and a decimation mapping.

The following, natural step in the analysis is then to identify the reduced representations of a system that are able to preserve the maximum amount of information from the all-atom reference—i.e., which minimize the mapping entropy. However, for a molecule with n atoms, the number of possible decimation mappings is 2^n, an astronomical amount even for the smallest proteins. This number remains huge even narrowing down the exploration to a fixed number of retained atoms N, so that $n!/[N!\,(n-N)!]$ mappings can be constructed. As a complete enumeration of all possible CG representations of a system is unfeasible in practice, Giulini et al. (2020) relied on a stochastic minimization procedure to extract a pool of optimized solutions out of this immense space, namely a simulated annealing approach (Kirkpatrick et al., 1983; Černý, 1985) employing S_{map} as cost function.

Remarkably, the CG mappings singled out by this optimization workflow were discovered to more likely retain atoms directly related to the biological function of the proteins of interest, thus linking the described information-theoretical approach to the properties of biological systems. It follows that this protocol represents not only a practical way to select the most informative mapping in a macromolecular structure, but also a promising paradigm to employ CGing as a controllable filtering procedure that can highlight relevant regions in a system.

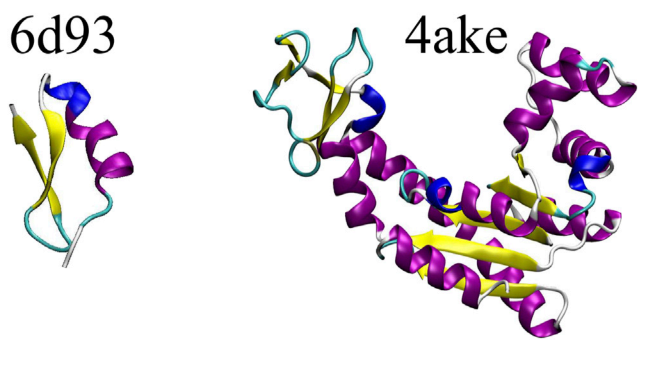


FIGURE 3 | Protein structures employed in this work: the tamapin mutant (PDB code: 6d93) and the open conformation of adenylate kinase (PDB code: 4ake). The former, although small, possesses all the elements of proteins' secondary structures, while the latter is bigger in size and has a much wider structural variability.

The downside of the approach developed in Giulini et al. (2020) is its non-negligible computational cost, which is due to two factors:

1. The protocol requires in input a set of configurations of the high-resolution system that are sampled through an MD simulation, a computationally expensive task.
2. The stochastic exploration of the set of possible CG mappings is limited and time consuming due to the algorithmic complexity associated to S_{map} calculations.

The ultimate aim of this work is, thus, the development and assessment of a protein-specific machine learning model able to swiftly predict the mapping entropy arising from a reduction in the number of degrees of freedom employed to describe the system.

Protein Structures and Data Sets

The DGN-based mapping entropy prediction model developed in this study is applied to two proteins extracted from the set investigated in Giulini et al. (2020), namely *(i) 6d93*, a 31 residues long mutant of *tamapin*—a toxin of the Indian red scorpion (Pedarzani et al., 2002)—whose outstanding selectivity toward the calcium-activated potassium channels SK2 made it an extremely interesting system in the field of pharmacology (Mayorga-Flores et al., 2020); and *(ii) 4ake*, the open conformation of *adenylate kinase* (Müller et al., 1996). This 214-residues enzyme is responsible for the interconversion between adenosine triphosphate (ATP) and adenosine diphosphate + adenosine monophosphate (ADP + AMP) inside the cell.

Figure 3 shows a schematic representation of *6d93* and *4ake*. Both proteins were simulated in explicit solvent for 200 ns in the canonical ensemble by relying on the GROMACS 2018 package (Spoel et al., 2005). For a more detailed discussion of these two molecules and the corresponding MD simulations, please refer to Sec. II.B and II.D of Giulini et al. (2020).

We train the machine learning model of each protein on a data set containing the molecular structure—the first snapshot of the MD trajectory—and many CG representations, the latter being selected with the constraint of having a number of retained sites equal to the number of amino acids composing the molecule. The data sets combine together randomly selected CG mappings (respectively, 4,200 for *6d93* and 1,200 for *4ake*) and optimized ones (768 for both systems). The corresponding mapping entropy values are calculated through the protocol described in Giulini et al. (2020).

Optimized mappings are obtained from independent Simulated Annealing (SA) Monte Carlo runs (Kirkpatrick et al., 1983; Černỳ, 1985): starting from a random selection of retained atoms, S_{map} is minimized for a defined number of steps after which the current mapping is saved and included in the data set. More specifically, at each step of a SA run we randomly swap a retained and a non-retained atom in the CG representation, compute S_{map}, and accept/reject the move based on a Metropolis criterion. The SA effective temperature T decays according to $T(i) = T_0 e^{-i/\nu}$, where i is the SA step and the parameters ν and T_0 are equal to those employed in Giulini et al. (2020). The 768 SA runs of each protein are divided into four groups of 192 elements depending on their length, respectively, 2×10^4 (full optimization, as in Giulini et al. (2020)), 1×10^4, 5×10^3, and 2.5×10^3 steps.

Figure 4 displays the distribution of S_{map} values in the data sets separately for the two systems, discriminating between random (blue) and optimized (red) CG mappings. In both structures the two curves have a negligible overlap, meaning that the set of values spanned by the optimized CG representations cannot be reached by a random exploration of the mapping space, i.e., this region possesses a very low statistical weight. A comparison of the S_{map} distribution of the two proteins, on the other hand, highlights that the mapping entropy increases with the system's size: while the range of values covered has similar width in the two cases, the lower bound in mapping entropy of *4ake* differs of roughly one order magnitude from that of *6d93*.

For each analyzed protein, in **Table 1** we report the computational time required to perform the MD simulation and a single S_{map} estimate. We note that the time associated with the calculation of S_{map} for a single CG mapping through the algorithm discussed in Giulini et al. (2020) grows from 2 to 8 minutes while moving from *6d93* to *4ake*. It is worth stressing that the proteins studied here are small, so that this value would dramatically increase in the case of bigger biomolecules.

Data Representation and Machine Learning Model

We represent each investigated protein structure as a static graph, see **Figure 5**. A graph g can be formally defined as a tuple $(\mathrm{v}_g, \mathcal{E}_g)$, where v_g is the set of vertices (i.e., the entities of interest) and $\mathcal{E}_g = \{\{u, v\} | u, v \in \mathrm{v}_g\}$ is the set of undirected edges (i.e., how entities are related). We define the neighborhood of a vertex v as the set of vertices connected to v by an edge, that is, $\mathcal{N}_v = \{u \in \mathrm{v}_g | \{u, v\} \in \mathcal{E}_g\}$. For the purpose of this work, each

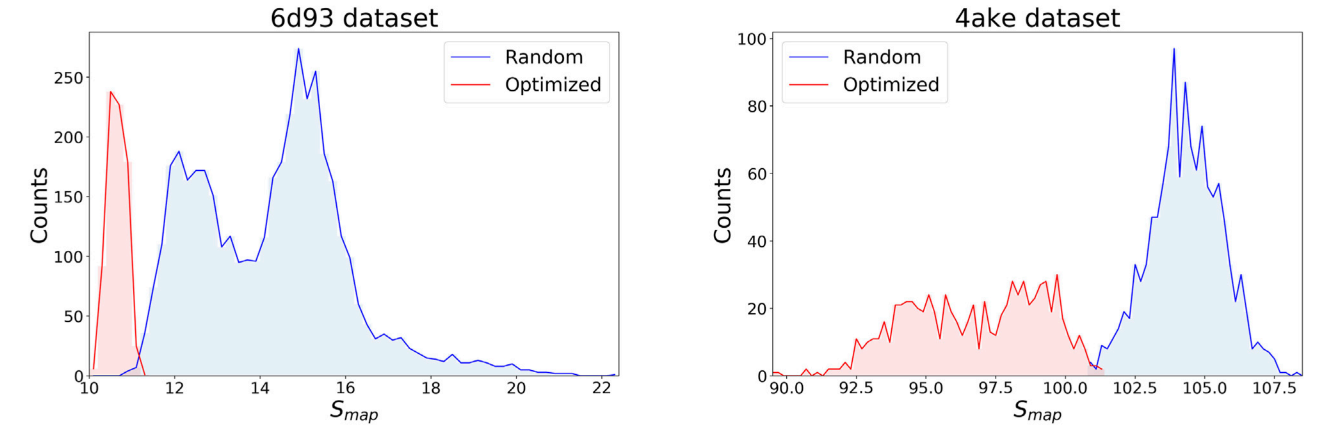


FIGURE 4 | Distributions of target values for both data sets, *6d93* **(left)** and *4ake* **(right)**. For each protein, S_{map} data are displayed in two distinct, non-overlapping histograms depending on their origin: blue curves are filled with random instances, while red histograms represent optimized CG mappings. All values of S_{map} are in kJ/mol/K.

TABLE 1 | Computational cost of all-atom MD simulations and mapping entropy calculations for the two investigated proteins. Specifically, *MD CPU time* (respectively, *MD walltime*) represents the core time (respectively, user time) necessary to simulate the system for 200 ns on the GROMACS 2018 package (Spoel et al., 2005). Both *6d93* and *4ake* runs were performed on Intel Xeon-Gold 5118 processors, respectively, using 16 and 48 cores. *Single measure* is the amount of time that is required to compute, on a single core of the same architecture, the S_{map} of a given CG mapping by relying on the algorithm introduced in Giulini et al. (2020).

Protein	MD CPU time	MD walltime	Single measure
Tamapin (PDB code 6d93)	40.7 days	2.55 days	≃2.1 mins
Adenylate kinase (PDB code 4ake)	153.9 days	3.20 days	≃8.0 mins

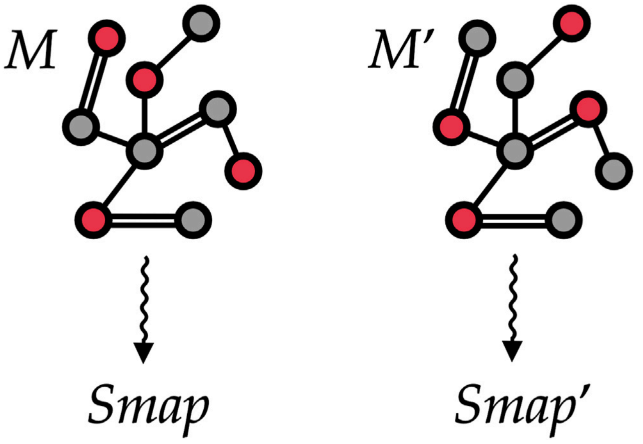


FIGURE 5 | Two different mappings *M* and *M′* associated with the same (schematic) protein structure. To train our machine learning model, we treat each protein as a graph where vertices are atoms and edges are placed among atoms closer than a given threshold. The selected CG sites in each of the two mappings are marked in red and encoded as a vertex feature. Our goal is to automatically learn to associate both mappings to proper values S_{map} and S_{map}' of the mapping entropy.

TABLE 2 | Binary features (0/1) used to describe the physicochemical properties of an atom in the protein, i.e., a vertex in the graph representation of the latter. In this simple model, we only provide the DGN with the chemical nature of the atom and of its residue, together with the flag *Bkb* that specifies if the atom is part of the backbone of the polypeptide chain.

Feature name	Description
C	Carbon atom
N	Nitrogen atom
O	Oxygen atom
S	Sulfur atom
HPhob	Part of a hydrophobic residue
Amph	Part of a amphipathic residue
Pol	Part of a polar residue
Ch	Part of a charged residue
Bkb	Part of the protein backbone
Site	Atom selected as a CG site

heavy atom composing the molecule corresponds to a vertex, and edges connect pairs of atoms that in the reference structure are closer than a selected threshold—in our case, 1 nm. At odds with other definitions of a CG site, the information about the decimation mapping can be directly encoded in the vertices of the protein graph by using a binary feature, with different selections of CG sites—an example being provided in **Figure 5**—corresponding to different values of S_{map}. In addition, we enrich each vertex with 10 features, summarized in **Table 2**, which describe the physicochemical properties of the underlying atom; similarly, we consider the inverse atomic distance e_{uv} between vertices u and v as an edge feature.

TABLE 3 | Basic statistics of the data sets fed to the machine learning model. For each protein, we report the number of vertices (i.e., heavy atoms) in its graph representation, the total number of edges connecting them, and the average number of edges per vertex (Avg. degree). We also report the total number of CG representations of known mapping entropy provided in input to the protocol (Samples), including random and optimized ones.

Protein	Vertices	Edges	Avg. degree	Samples
6d93	230	21,474	93	4,968
4ake	1,656	207,618	125	1,968

Once the protein structure and the CG mapping data sets are converted into this graph-like format (statistics in **Table 3**), we employ DGNs (Bacciu et al., 2020) with the aim of learning the desired property, namely the mapping entropy S_{map}.

The main advantages of DGNs are their efficiency and the ability to learn from graphs of different size and shape. This is possible for two reasons: first, DGNs focus on a local processing of vertex neighbors, so that calculations can be easily distributed; secondly, in a way that is similar to Convolutional Neural Networks for images (LeCun et al., 1995), DGNs stack multiple layers of graph convolutions to let vertices efficiently exchange information. The output of a DGN is a vector for each vertex of the graph, as sketched in **Figure 6**, and these can be aggregated to make predictions about a graph class or property. Again, we remark that the efficiency of the DGN is especially important in our context, where we want to approximate the complex S_{map} computational process in a fraction of the time originally required.

The main building block of a DGN is the "graph convolution" mechanism. At each layer ℓ, the DGN calculates the new state of each vertex v, i.e., a vector $\mathbf{h}_v^{\ell+1} \in \mathbb{R}^K$, as a function of v's neighboring states $\mathbf{h}_{\mathcal{N}_v}^{\ell} = \{\mathbf{h}_u^{\ell} \in \mathbb{R}^K | u \in \mathcal{N}_v\}$, where $K \in \mathbb{N}$ is an hyperparameter of the model.

In general, a graph convolutional layer first applies a permutation-invariant function to the neighbors of each vertex, such as the sum or mean. The resulting aggregated vector is then passed to a multi-layer perceptron (MLP) that performs a nonlinear transformation of the input, thus producing the new vertex state $\mathbf{h}_v^{\ell+1}$.

In this study, we employ an extended version of the GIN model (Xu et al., 2019) or, equivalently, a restricted version of the Gated-GIN model (Errica et al., 2020) to consider edge attributes while keeping the computational burden low. Our graph convolutional layer can be formalized as follows:

$$\mathbf{h}_v^{\ell+1} = MLP^{\ell}\left[\left(1+\epsilon^{\ell}\right) \times \mathbf{h}_v^{\ell} + \sum_{u \in \mathcal{N}_v} \mathbf{h}_u^{\ell} \times e_{uv}\right], \tag{5}$$

where × denotes element-wise scalar multiplication, $\epsilon^{\ell} \in \mathbb{R}$ is an adaptive weight of the model, and e_{uv} is the scalar edge feature holding the inverse atomic distance between two atoms u and v. A pictorial representation of the transition between layer ℓ and layer $\ell + 1$ is presented in **Figure 7**.

A few remarks about **Eq. 5** are in order. First, the initial layer is implemented with a simple nonlinear transformation of the vertex features, that is, $\mathbf{h}_v^1 = MLP^1(\mathbf{x}_v)$, where $\mathbf{x}_v$ is the vector of 10 features associated with each node (see **Table 2**); secondly, at each layer ℓ, we apply the *same* nonlinear transformation MLP^{ℓ} to all the nodes (i.e., a graph traversal), which allows us to treat variable-size graphs. Finally, the MLP weights are not shared across different layers, meaning that we train a different MLP for each layer. It is worth noting that this weight-sharing scheme at each layer resembles the one employed in Convolutional Neural Networks, where the same adaptive filter is applied to all the pixels in an image.

When building a Deep Graph Network, we usually stack L graph convolutional layers, with $L \in \mathbb{N}$ being another hyperparameter, until the model produces a final state for each vertex. We call this state $\mathbf{h}_v$; in addition, we compute a global graph state $\mathbf{h}_g$ by aggregating all vertex states (see **Figure 6**). Being in vectorial form, $\mathbf{h}_g$ can then be fed to standard machine learning models to solve graph regression or classification tasks.

To produce a prediction $\widehat{S}_{map}$, we first need to process and aggregate all node states into a single graph representation. In this work, we take into account the importance of selected (respectively, unselected $\mathcal{V}$) CG sites $\mathrm{v}_g^s \subset \mathrm{v}_g$ (respectively, v_g^n) with a scalar adaptive weight w_s (respectively, w_n). The resulting formula is

$$\widehat{S}_{map} = \mathbf{w}_{out}^T \left\{ \sum_{u \in \mathcal{V}_g^s} \left[\left(\mathbf{h}_u^1, \ldots, \mathbf{h}_u^L\right) \times w_s\right] + \sum_{u \in \mathcal{V}_g^n} \left[\left(\mathbf{h}_u^1, \ldots, \mathbf{h}_u^L\right) \times w_n\right] \right\}, \tag{6}$$

where $\mathbf{w}_{out} \in \mathbb{R}^{K*L}$ is a set of parameters to be learned, while square brackets denote concatenation of the different vertex states computed at different layers.

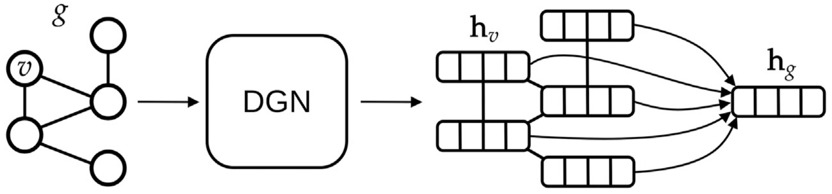


FIGURE 6 | High-level overview of typical deep learning methodologies for graphs. A graph g is given as input to a Deep Graph Network, which outputs one vector, also called embedding or state, for each vertex v of the graph. In this study, we aggregate all vertex states *via* a (differentiable) permutation-invariant operator, i.e., the mean, to obtain a single state that encodes the whole graph structure. Then, the graph embedding is fed into a machine learning regression model (in our case a linear model) to output the S_{map} value associated with g.

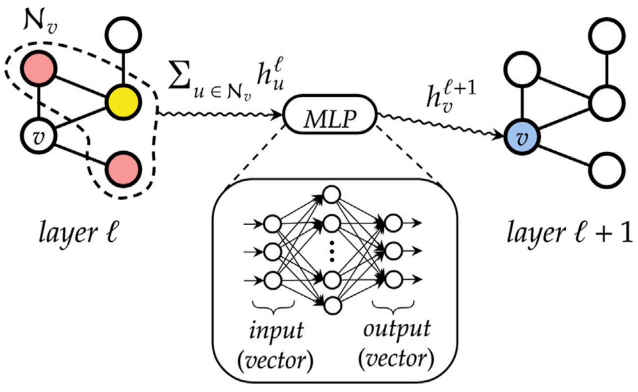


FIGURE 7 | A simplified representation of how a graph convolutional layer works. First, neighboring states of each vertex *v* are aggregated by means of a permutation-invariant function, to abstract from the ordering of the nodes and to deal with variable-sized graphs. Then, the resulting vector is fed into a multi-layer perceptron that outputs the new state for node *v*.

In particular, we use $L = 5$ layers and implement each MLP^ℓ as a one-layer feed-forward network with $K = 64$ hidden units followed by an element-wise rectifier linear unit (ReLU) activation function (Glorot et al., 2011). As the number of weights, without considering the bias, of MLP^ℓ is K^2 ($10 * K$ for MLP^1), the total number of weights in our architecture is $10 * K + K^2 * (L - 1) + (L * K) + (L - 1) + 2 = 17350$.

The loss objective used to train the DGN is the mean absolute error. The optimization algorithm is Adam (Kingma and Ba, 2015) with a learning rate of 0.001 and no regularization. We trained for a maximum of 10,000 epochs with early stopping patience of 1,000 epochs and mini-batch size 8, accelerating the training using a Tesla V100 GPU with 16 GB of memory.

To assess the performance of the model on a single protein, we first split the corresponding data set into training, validation, and test realizations following an 80%/10%/10% hold-out strategy. We trained and assessed the model on each data set separately. We applied early stopping (Prechelt, 1998) to select the training epoch with the best validation score, and the chosen model was evaluated on the unseen test set. The evaluation metric for our regression problem is the coefficient of determination (or R^2 score).

Wang–Landau Sampling

Figure 4 highlights how an attempt of detecting the most informative CG representations of a protein—i.e., those minimizing S_{map}—through a completely unbiased exploration of its mapping space would prove extremely inefficient, if not practically pointless. Indeed, such optimized CG representations live relatively far away in the left tails of the S_{map} distributions obtained from random sampling, thus constituting a region of exponentially vanishing size within the broad mapping space. It would then be desirable to design a sampling strategy in which no specific value of S_{map} is preferred, but rather a *uniform coverage* of the spectra of possible mapping entropies—or at least of a subset of it, vide infra—is achieved.

To obtain this "flattening" of the S_{map} landscape we rely on the algorithm proposed by Wang and Landau (WL) (Wang and Landau, 2001a; Wang and Landau, 2001b; Shell et al., 2002; Barash et al., 2017). In WL sampling, a Markov chain Monte Carlo (MC) simulation is constructed in which a transition between two states M and M' —in our case, two mappings containing N sites but differing in the retainment of one atom—is accepted with probability

$$W(M \to M') = \min\left\{1, \frac{\Omega_N\left[S_{map}(M)\right]}{\Omega_N\left[S_{map}(M')\right]}\right\}. \tag{7}$$

In **Eq. 7**, $\Omega_N(S_{map})$ is the number of CG representations with N retained sites exhibiting a mapping entropy equal to S_{map}, that is, the mapping entropy's density of states,

$$\Omega_N\left(S_{map}\right) = \sum_M \delta[N(M), N]\delta\left[S_{map}(M), S_{map}\right], \tag{8}$$

where the sum is performed over all possible CG representations of the system.

When compounded with a symmetric proposal probability T for the attempted move, $T(M \to M') = T(M' \to M)$, the Markov chain defined in **Eq. 7** generates, at convergence, CG representations distributed according to $P(M) \propto 1/\Omega_N[S_{map}(M)]$ (Wang and Landau, 2001a; Wang and Landau, 2001b). As the equilibrium probability of visiting a mapping is proportional to the inverse of the S_{map} density of states, the WL simulation results in a flat histogram of sampled mapping entropies *over the whole range of possible ones*.

Critically, the density of states $\Omega_N(S_{map})$ is a priori unknown and is itself a byproduct of the WL scheme. $\Omega_N(S_{map})$ is self-consistently constructed by means of a sequence $k = 0,...K$ of nonequilibrium simulations that provide increasingly accurate approximations to the exact result, iterations being stopped when a predefined precision is achieved.

Having divided the range of possible values of the mapping entropy in bins of width δS_{map}, the WL self-consistent protocol is based on three quantities: the overall density of states $\Omega_N(S_{map})$, the histogram of sampled mapping entropies at iteration k, $H_k(S_{map})$, and the modification factor f_k governing convergence–for $k = 0$, one typically initializes $\Omega_N(S_{map}) = 1$ for each value of S_{map} and $f_0 = e$.

At the beginning of WL iteration k, the histogram $H_k(S_{map})$ is reset. Subsequently, a sequence of MC moves among CG mappings driven by the acceptance probability presented in **Eq. 7**, is performed. If a transition between two CG representations M and M'— respectively with mapping entropies S_{map} and $S_{map'}$ predicted by the trained DGNs—is accepted, the entries of the histogram and density of states are updated according to

$$H_k\left(S_{map}{}'\right) = H_k\left(S_{map}{}'\right) + 1, \tag{9}$$

TABLE 4 | Set of parameters employed for the WL exploration of the mapping entropy space for both analyzed proteins. $\ln(f_0)$ and $\ln(f_{end})$ respectively represent the modification factor at the beginning and at the end of the self-consistent scheme in a logarithmic setup, see Sec.*Wang–Landau Sampling*. p_{flat} is the minimal histogram flatness required to halve the modification factor; with $p_{flat} = 0.8$, all bins in the histogram $H(S_{map})$ must have a population between 0.8 and 1.2 times its average $\langle H \rangle$. *range* is the interval of permitted values of the mapping entropy in the WL scheme, while δS_{map} is the bin size employed for its discretization. Both *range* and δS_{map} are expressed in kJ/mol/K.

Parameter	*6d93*	*4ake*
$\ln(f_0)$	1	1
$\ln(f_{end})$	10^{-6}	10^{-6}
p_{flat}	0.8	0.8
range	[10–22.4]	[89.4–108.6]
δS_{map}	0.2	0.2

$$\Omega_N(S_{map}{}') = f_k \times \Omega_N(S_{map}{}'). \qquad (10)$$

In case the move $M \rightarrow M'$ is rejected, one has to replace $S_{map}{}'$ with S_{map} in **Eqs. 9**, **10**.

The sequence of MC moves is stopped—that is, iteration k ends—when $H_k(S_{map})$ is "flat", meaning that each of its entries does not exceed a threshold distance from the average histogram $\langle H_k \rangle$: a typical requirement is $p_{flat} \times \langle H_k \rangle < H_k(S_{map}) < (2 - p_{flat}) \times \langle H_k \rangle$ for every value of S_{map}, p_{flat} being the selected flatness parameter. At this stage, WL iteration $k+1$ begins with a reduced modification factor, where we set $f_{k+1} = \sqrt{f_k}$.

Convergence of the self-consistent scheme is achieved when $f_k \approx 1$ —more precisely, when $\ln(f_k)$ becomes smaller than a predefined value $\ln(f_{end})$. Up to a global multiplicative factor, the resulting density of states $\Omega_N(S_{map})$ reproduces the exact result with an accuracy of order $\ln(f_{end})$ (Landau et al., 2004).

In order to avoid numeric overflow of $\Omega_N(S_{map})$ along the WL simulation, we consider its logarithm $\Sigma_N(S_{map}) = \ln\Omega_N(S_{map})$. Starting from **Eq. 7**, the acceptance probability $W(M \rightarrow M')$ expressed in terms of Σ reads

$$W(M \rightarrow M') = \min\{1, \exp[\Sigma_N(M) - \Sigma_N(M')]\}, \qquad (11)$$

while within iteration k of the self-consistent scheme, the update prescription of Σ after an (accepted) MC move—see **Eq. 10**—becomes

$$\Sigma_N(S'_{map}) = \Sigma_N(S'_{map}) + \ln(f_k). \qquad (12)$$

Finally, in a logarithmic setup, the modification factor $\ln(f_k)$ follows the simple reduction rule $\ln(f_{k+1}) = \ln(f_k)/2$, with $\ln(f_0) = 1$.

The WL algorithm in principle enables the reconstruction of the density of states of an observable over the whole range of possible values of the latter; at the same time, knowledge of the sampling boundaries proves extremely beneficial to the accuracy and rate of convergence of the self-consistent scheme (Wüst and Landau, 2008; Seaton et al., 2009). In our case, for each analyzed protein, such boundaries would correspond to the minimum and maximum achievable mapping entropies S_{map}^{min} and S_{map}^{max} in the space of all CG representations of the system obtained by retaining N of its constituent atoms. As this information is *a priori* unknown, in our implementation of the WL algorithm we limit the range of explorable values of S_{map} by rejecting all MC moves $M \rightarrow M'$ for which $S'_{map} < S_{map}^{min}$ or $S'_{map} > S_{map}^{max}$, in each system setting S_{map}^{min} and S_{map}^{max} as, respectively, the minimum and maximum values of the mapping entropy in the corresponding data set. Note that for each protein S_{map}^{min} is the outcome of a thorough optimization procedure, and can thus be considered a reasonable approximation of the system's *absolute* minimum of the mapping entropy. Imposing an upper bound on S_{map} through S_{map}^{max}, on the other hand, simply amounts at requiring the WL sampling algorithm not to visit uninteresting regions of the mapping space of each biomolecule, that is, CG representations characterized by a huge amount of information loss with respect to the all-atom reference. The values of S_{map}^{min} and S_{map}^{max} employed for the two proteins investigated in this study are presented in **Table 4**, together with the input parameters required by the WL protocol—the bin size δS_{map}, the convergence modification factor $\ln(f_{end})$, and the flatness parameter p_{flat}.

RESULTS AND DISCUSSION

We first analyze the results achieved by DGNs in predicting the mapping entropy associated to a choice of the CG representation of the two investigated proteins; specifically, we employ the R^2 score as the main evaluation metric and the mean average error (MAE) as an additional measure to assess the quality of our model in fitting S_{map} data. The R^2 scores range from $-\infty$ (worst predictor) to 1 (best predictor).

Table 5 reports the R^2 score and MAE in training, validation, and test. We observe that the machine learning model can fit the training set and has excellent performances on the test set. More quantitatively, we achieve extremely low values of MAE for *6d93*, with an R^2 score higher than 0.95 in all cases. The model performs slightly worse in the case of *4ake*: the result of $R^2 = 0.84$ on the test set is still acceptable, although the gap with the training set ($R^2 = 0.92$) is non-negligible.

Figure 8 shows how predicted values for training and test samples differ from the ground truth. Ideally, a perfect result corresponds to the point being on the diagonal dotted line. We can see how close to the true target are both training and test predictions for *6d93*. The deviation from the ideal case becomes

TABLE 5 | Results of the machine learning model in predicting the mapping entropy on the training (TR), validation (VL), and test (TE) sets for the two analyzed proteins. We display both the R^2 score and the mean average error (MAE, kJ/mol/K).

Protein	TR MAE	TR R^2	VL MAE	VL R^2	TE MAE	TE R^2
6d93	0.13	0.99	0.33	0.95	0.33	0.96
4ake	0.91	0.92	1.2	0.85	1.35	0.84

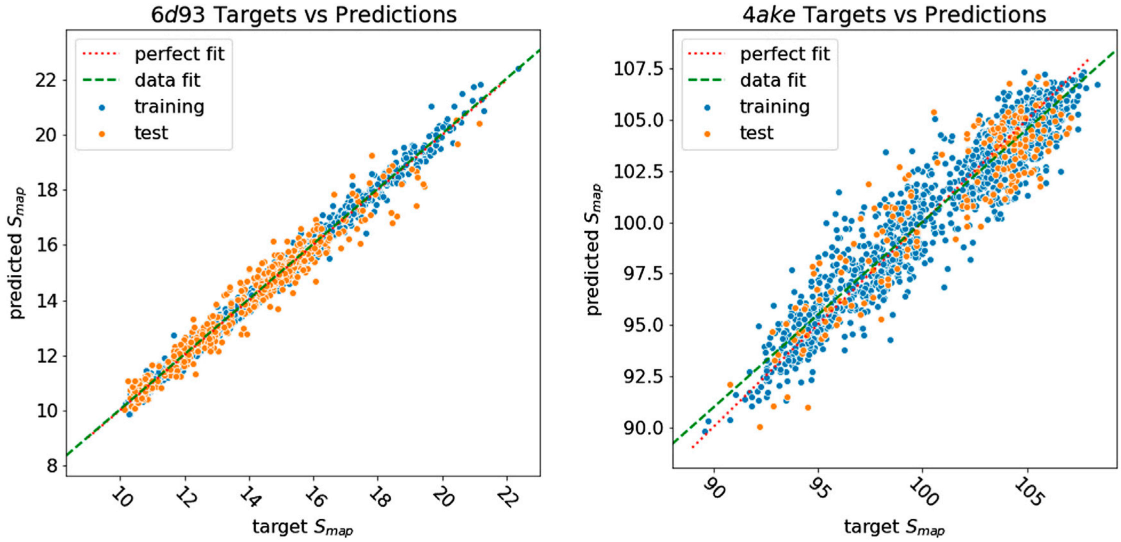


FIGURE 8 | Plot of S_{map} target values against predictions of all samples for *6d93* **(left)** and *4ake* **(right)**. Training samples are in blue, while test samples are in orange. A perfect prediction is represented by points lying on the red dotted diagonal line (perfect fit). To show that in the case of 4*ake*, the model slightly overestimates the S_{map} of optimized mappings and underestimates the rest, we include in the plot the green dashed line obtained by fitting a linear model on the data (data fit). All values of S_{map} are in kJ/mol/K.

TABLE 6 | Comparison between the time required to compute the S_{map} of a single CG mapping through the algorithm presented in Giulini et al. (2020) and the inference time of the model (CPU as well as GPU). For both proteins, CPU calculations were performed on a single core of a Intel Xeon-Gold 5118 processor, while GPU ones were run on a Tesla P100 with 16 GB of memory. The machine learning model generates a drastic speedup, enabling a wider exploration of the S_{map} landscape of each system.

Protein	Single measure	Inference GPU (CPU)	Time ratio GPU (CPU)
6d93	≃2.1 mins	≃0.9(98.7) *ms*	≃140000 × (1276×)
4ake	≃8.0 mins	≃4.8(1103.2) *ms*	≃100000 × (435×)

wider for *4ake*, but no significant outlier is present. A more detailed inspection of the *4ake* scatter plot in **Figure 8**, on the other hand, reveals that the network tends to slightly overestimate the value of S_{map} of optimized CG mappings for $S_{map} \lesssim 100\ kJ/\text{mol}$, whereas the opposite is true for $S_{map} \gtrsim 100\ kJ/\text{mol}$, where random CG mapping values are mildly underestimated.

The dissimilarity in performance between the two data sets is not surprising if one takes a closer look at their nature. In fact, as highlighted in **Figure 3**, adenylate kinase is both larger and more complex than the tamapin mutant, and the CG mapping data set sizes are very different due to the heavy computational requirements associated with the collection of annotated samples for *4ake*. As a consequence, training a model for *4ake* with excellent generalization performance becomes a harder task. What is remarkable, though, is the ability of a completely adaptive machine learning methodology to well approximate, in both structures, the long and computationally intensive algorithm for estimating S_{map} of Giulini et al. (2020). Critically, this is achieved only by relying on a combination of static structural information and few vertex attributes, that is, in absence of a direct knowledge for the DGNs of the complex dynamical behavior of the two systems as obtained by onerous MD simulations.

The computational time required by the machine learning model to perform a single S_{map} calculation is compared to the one of the algorithm presented in Giulini et al. (2020) in **Table 6**. As the protocol of Giulini et al. (2020) relied on a CPU machine, we report results for both CPU and GPU times. Overall, we observe that inference of the model can speed up mapping entropy calculations by a factor of two to five orders of magnitude depending on the hardware used. Noteworthy, these improvements do not come at the cost of a significantly worse performance of the machine learning model. In addition, this methodology is easily applicable to other kinds of molecular structures, as long as a sufficiently large training set is provided as input.

By embedding the trained networks in a Wang–Landau sampling scheme, see Wang–Landau sampling, we are able to retrieve the density of states $\Omega_N(S_{map})$ defined in **Eq. 8** for *6d93* and *4ake*, that is, we can estimate the number of CG representations throughout the mapping space of each protein that exhibits a specific amount of information loss with respect to the all-atom reference. We stress that reaching convergence of the

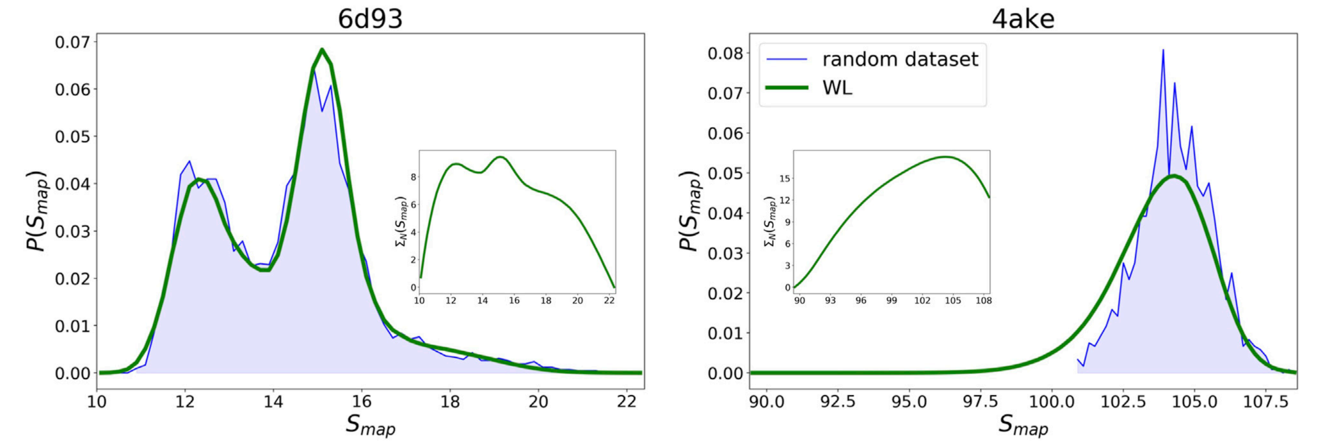


FIGURE 9 | Comparison between the probability densities $P(S_{map})$ for the two systems estimated *via* the Wang–Landau algorithm enhanced by the DGNs (green lines) and the distributions generated by a random sampling of mappings (blue areas). In inset, the logarithm of the WL density of states, $\Sigma(S_{map})$, is reported, after a scaling that assigns to the Σ of the most scarcely populated bin the value of zero. All values of S_{map} are in kJ/mol/K.

self-consistent WL protocol required to probe approximately 4.8×10^6 and 3×10^7 CG representations for *6d93* and *4ake*, respectively: such an extensive sampling is only made feasible by the computational gain provided by the machine learning model.

WL predictions for the logarithm of the density of states $\Sigma_N(S_{map}) = \ln\Omega_N(S_{map})$ of the two proteins are presented in **Figure 9**. As for *6d93*, we observe the presence of a steep increase of Σ starting from low values of the mapping entropy, followed by two main peaks respectively located at $S_{map} \approx 12.5$ and 15 kJ/mol/K. After the second peak Σ decreases, exhibiting a shoulder for high mapping entropies. On the other hand, the Σ of *4ake* displays a relatively gradual growth toward its unique maximum, the latter being located at $S_{map} \approx 105$ kJ/mol/K, before starting to decrease.

Given the WL $\Omega_N(S_{map})$—or equivalently $\Sigma_N(S_{map})$—it is possible to calculate the probability $P(S_{map})$ of observing a particular mapping entropy by performing a completely random exploration of the space of CG representations of a system,

$$P\left(S_{map}\right) = \frac{\Omega_N\left(S_{map}\right)}{\sum_{S_{map}}\Omega_N\left(S_{map}\right)}. \tag{13}$$

Results for the $P(S_{map})$ of *6d93* and *4ake* are shown in **Figure 9**. In the case of *6d93*, we note that the WL sampling scheme produces a probability density that is fully compatible with the (normalized) histograms of **Figure 4**. In particular, the WL graph resembles the histograms in **Figure 4** if we remove the nonrandom, optimized instances whose statistical weight is negligible. This result is highly nontrivial, as it proves that the trained DGN of *6d93* does not overfit the training set and is able to predict the correct population of the true mapping entropy landscape.

As regards *4ake*, the agreement between the two curves presented in **Figure 9** is still remarkable, though not as precise as in the case of *6d93*. More quantitatively, the left tail of the probability density predicted by the WL scheme is shifted of roughly 1 kJ/mol/K toward lower values of S_{map} with respect to the distribution obtained from random sampling. This mismatch can be ascribed to the mild overfitting problem observed in **Figure 8**: the network has the tendency to underestimate (respectively, overestimate) the value of S_{map} associated with random (respectively, optimized) CG representations, resulting in an increase in the predicted population of mappings at the intersection of the two sets.

CONCLUSION AND PERSPECTIVES

Molecular dynamics simulations constitute the core of the majority of research studies in the field of computational biophysics. From protein folding to free energy calculations, an all-atom trajectory of a biomolecule gives access to a vast amount of data, from which relevant information about the system's properties, behavior, and biological function is extracted through an *a posteriori* analysis. This information can be almost immediate to observe (even by naked eye) and quantify in terms of few simple parameters–e.g., the process of ligand binding can be seen in a graphical rendering of the trajectory and made quantitative in terms of the distance between ligand and protein; much more frequently, though, it is a lengthy and nontrivial task, tackled through the introduction of complex "filtering" strategies, the outcomes of which often require additional human intervention to be translated in intuitive terms (Tribello and Gasparotto, 2019; Noé et al., 2020).

A protocol aiming at the unsupervised detection of the relevant features of a biomolecular system was recently proposed (Giulini et al., 2020). The method relies on the concept of mapping entropy S_{map} (Shell, 2008; Rudzinski and Noid, 2011; Shell, 2012; Foley et al., 2015), that is, the information that is lost when the system is observed in terms of a subset of its original degrees of freedom: in Giulini et al. (2020), a minimization of this loss over the space of possible reduced representations, or CG mappings, enabled to single out the most informative ones. By performing a statistical analysis of the properties of such optimized mappings, it was shown that these are more likely to concentrate a finer level of detail—so that more atoms survive the CG'ing procedure—in regions of the system that are directly related to the biological function of the latter. The mapping entropy protocol thus represents a promising filtering tool in an attempt of distilling the relevant information of an overwhelmingly complicated macromolecular structure; furthermore, this information can be immediately visualized and interpreted as it consists of specific subsets of atoms that get singled out from the pool of the constituent ones. Unfortunately, estimating the S_{map} associated with a specific low-resolution representation is a lengthy and computationally burdensome process, thus preventing a thorough exploration of the mapping space to be achieved along the optimization process.

In this work, we have tackled the problem of speeding up the S_{map} calculation procedure by means of deep machine learning models for graphs. In particular, we have shown that Deep Graph Networks are capable of inferring the value of the mapping entropy when provided with a schematic, graph-based representation of the protein and a tentative mapping. The method's accuracy is tested on two proteins of very different size, a tamapin mutant (31 residues) and adenylate kinase (214 residues), with a R^2 test score of 0.96 and 0.84, respectively. These rather promising results have been obtained in a computing time that is up to five orders of magnitude shorter than the algorithm proposed in Giulini et al. (2020).

The presented strategy holds the key for an extensive exploration of the space of possible CG mappings of a biomolecule. In fact, the combination of trained networks and Wang–Landau sampling allows one to characterize the mapping entropy landscape of a system with impressive accuracy.

The natural following step would be to apply the knowledge acquired by the model on different protein structures, so that the network can predict values of S_{map} even in the absence of an MD simulation. As of now, however, it is difficult to assess if the information extracted from the training over a given protein trajectory can be fruitfully employed to determine the mapping entropy of another, by just feeding the structure of the latter as input. More likely one would have to resort to database-wide investigations, training the network over a large variety of different molecular structures before attempting predictions over new data points. In other words, obtaining a transfer effect among different structures by the learning model may not be straightforward, and additional information could be needed to achieve it. Analyses on this topic are on the way and will be the subject of future works.

In conclusion, we point out that the proposed approach is completely general, in that the specific nature and properties of the mapping entropy played no special role in the construction of the deep learning scheme; furthermore, the DGN formalism enables one to input graphs of variable size and shape, relaxing the limitations present in other kinds of deep learning architectures (Giulini and Potestio, 2019). This method can thus be transferred to other problems where different selections of a subset of the molecule's atoms give rise to different values of a given observable (see e.g., Diggins et al., 2018) and pave the way for a drastic speedup in computer-aided computational studies in the fields of molecular biology, soft matter, and material science.

DATA AVAILABILITY STATEMENT

The data sets employed for this study and the code that performs the Wang Landau-based exploration of the mapping space are freely available at https://github.com/CIML-VARIAMOLS/GRAWL.

AUTHOR CONTRIBUTIONS

RP, AM, and RM elaborated the study. FE, AM, and DB designed and realized the DGN model. MG and RM performed the molecular dynamics simulations and the Wang–Landau sampling. MG and FE performed the analysis. All authors contributed to the interpretation of the results and the writing of the manuscript.

FUNDING

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 758588). This work is partially supported by the Università di Pisa under the "PRA-Progetti di Ricerca di Ateneo" (Institutional Research Grants)—Project no. PRA_2020–2021_26 "Metodi Informatici Integrati per la Biomedica".

REFERENCES

Alder, B. J., and Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *J. Chem. Phys.* 31, 459–466. doi:10.1063/1.1730376

Bacciu, D., Errica, F., Micheli, A., and Podda, M. (2020). A gentle introduction to deep learning for graphs. *Neural Netw.* 129, 203–221. doi:10.1016/j.neunet.2020.06.006

Barash, L. Y., Fadeeva, M., and Shchur, L. (2017). Control of accuracy in the Wang-Landau algorithm. *Phys. Rev. E* 96, 043307. doi:10.1103/physreve.96.043307

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint name [Preprint]. Available at: http://arXiv:1806.01261 (Accessed October 4, 2020).

Bereau, T., and Kremer, K. (2015). Automated parametrization of the coarse-grained Martini force field for small organic molecules. *J. Chem. Theor. Comput.* 11, 2783–2791. doi:10.1021/acs.jctc.5b00056

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S. V. N., Smola, A. J., and Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics* 21, i47–i56. doi:10.1093/bioinformatics/bti1007

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Mag.* 34, 2518–2542. doi:10.1109/msp.2017.2693418

Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. optimization Theor. Appl.* 45, 41–51.

Diggins, P., Liu, C., Deserno, M., and Potestio, R. (2018). Optimal coarse-grained site selection in elastic network models of biomolecules. *J. Chem. Theor. Comput.* 15, 648–664. doi:10.1021/acs.jctc.8b00654

Errica, F., Bacciu, D., and Micheli, A. (2020). "Theoretically expressive and edge-aware graph learning," in 28th European symposium on artificial neural networks, computational intelligence and machine learning, Bruges, Belgium, October 2–4, 2020.

Foley, T. T., Kidder, K. M., Shell, M. S., and Noid, W. G. (2020). Exploring the landscape of model representations. *Proc. Natl. Acad. Sci. U.S.A.* 117, 24061–24068. doi:10.1073/pnas.2000098117

Foley, T. T., Shell, M. S., and Noid, W. G. (2015). The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* 143, 243104. doi:10.1063/1.4929836

Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. *Adv. Neural Inf. Process. Syst.* 30, 6530–6539.

Gfeller, D., and Rios, P. D. L. (2007). Spectral coarse graining of complex networks. *Phys. Rev. Lett.* 99, 038701. doi:10.1103/physrevlett.99.038701

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. *Proc. 34th Int. Conf. Machine Learn. (Icml)* 70, 1263–1272.

Giulini, M., Menichetti, R., Shell, M. S., and Potestio, R. (2020). An information-theory-based approach for optimal model reduction of biomolecules. *J. Chem. Theor. Comput.* 16, 6795–6813. doi:10.1021/acs.jctc.0c00676

Giulini, M., and Potestio, R. (2019). A deep learning approach to the structural analysis of proteins. *Interf. Focus.* 9, 20190003. doi:10.1098/rsfs.2019.0003

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proc. Mach. Learn. Res.* 15, 315–323.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Representation learning on graphs: methods and applications. *IEEE Data Eng. Bull.* 40, 52–74.

Jin, J., Pak, A. J., and Voth, G. A. (2019). Understanding missing entropy in coarse-grained systems: addressing issues of representability and transferability. *J. Phys. Chem. Lett.* 10, 4549–4557. doi:10.1021/acs.jpclett.9b01228

Kandt, C., Ash, W. L., and Tieleman, D. P. (2007). Setting up and running molecular dynamics simulations of membrane proteins. *Methods* 41, 475–488. doi:10.1016/j.ymeth.2006.08.006

Karplus, M. (2002). Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* 35, 321–323. doi:10.1021/ar020082r

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in Proceedings of the 3rd international conference on learning representations, Ithaca, United States, December –January 13–22, 2014–2017, (ICLR).

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi:10.1126/science.220.4598.671

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chem. Rev.* 116, 7898–7936. doi:10.1021/acs.chemrev.6b00163

Koehl, P., Poitevin, F., Navaza, R., and Delarue, M. (2017). The renormalization group and its applications to generating coarse-grained models of large biological molecular systems. *J. Chem. Theor. Comput.* 13, 1424–1438. doi:10.1021/acs.jctc.6b01136

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22, 79–86. doi:10.1214/aoms/1177729694

Landau, D. P., Tsai, S.-H., and Exler, M. (2004). A new approach to Monte Carlo simulations in statistical physics: wang-landau sampling. *Am. J. Phys.* 72, 1294–1302. doi:10.1119/1.1707017

LeCun, Y., and Bengio, Y.others (1995). "Convolutional networks for images, speech, and time series," in *The handbook brain theory neural networks*. Cambridge, MA: MIT Press, 3361, 1118.

Li, Z., Wellawatte, G. P., Chakraborty, M., Gandhi, H. A., Xu, C., and White, A. D. (2020). Graph neural network based coarse-grained mapping prediction. *Chem. Sci.* 11, 9524–9531. doi:10.1039/d0sc02458a

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and De Vries, A. H. (2007). The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* 111, 7812–7824. doi:10.1021/jp071097f

Mayorga-Flores, M., Chantôme, A., Melchor-Meneses, C. M., Domingo, I., Titaux-Delgado, G. A., Galindo-Murillo, R., et al. (2020). Novel blocker of onco sk3 channels derived from scorpion toxin tamapin and active against migration of cancer cells. *ACS Med. Chem. Lett.* 11, 1627–1633. doi:10.1021/acsmedchemlett.0c00300

Micheli, A. (2009). Neural network for graphs: a contextual constructive approach. *IEEE Trans. Neural Netw.* 20, 498–511. doi:10.1109/tnn.2008.2010350

Micheli, A., Sperduti, A., and Starita, A. (2007). An introduction to recursive neural networks and kernel methods for cheminformatics. *Curr. Pharm. Des.* 13, 1469–1495. doi:10.2174/138161207780765981

Müller, C., Schlauderer, G., Reinstein, J., and Schulz, G. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147–156. doi:10.1016/s0969-2126(96)00018-4

Murtola, T., Kupiainen, M., Falck, E., and Vattulainen, I. (2007). Conformational analysis of lipid molecules by self-organizing maps. *J. Chem. Phys.* 126, 054707. doi:10.1063/1.2429066

Noé, F., De Fabritiis, G., and Clementi, C. (2020). Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol.* 60, 77–84. doi:10.1016/j.sbi.2019.12.005

Noid, W. G. (2013a). Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.* 139, 090901. doi:10.1063/1.4818908

Noid, W. G. (2013b). Systematic methods for structurally consistent coarse-grained models. *Methods Mol. Biol.* 924, 487–531. doi:10.1007/978-1-62703-017-5_19

Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., et al. (2008). The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* 128, 244114. doi:10.1063/1.2938860

Pedarzani, P., D'hoedt, D., Doorty, K. B., Wadsworth, J. D. F., Joseph, J. S., Jeyaseelan, K., et al. (2002). Tamapin, a venom peptide from the Indian red scorpion (Mesobuthus tamulus) that targets small conductance Ca^{2+}-activated K^{+} channels and after hyperpolarization currents in central neurons. *J. Biol. Chem.* 277, 46101–46109. doi:10.1074/jbc.m206465200

Potestio, R., Peter, C., and Kremer, K. (2014). Computer simulations of soft matter: linking the scales. *Entropy* 16, 4199–4245. doi:10.3390/e16084199

Prechelt, L. (1998). "Early stopping-but when?," in *Neural networks: tricks of the trade*. New York, NY: Springer, 55–69.

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

Rudzinski, J. F., and Noid, W. G. (2011). Coarse-graining entropy, forces, and structures. *J. Chem. Phys.* 135, 214101. doi:10.1063/1.3663709

Saunders, M. G., and Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* 42, 73–93. doi:10.1146/annurev-biophys-083012-130348

Seaton, D. T., Wüst, T., and Landau, D. P. (2009). A wang-landau study of the phase transitions in a flexible homopolymer. *Comput. Phys. Commun.* 180, 587–589. doi:10.1016/j.cpc.2008.11.023

Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J., Mackenzie, K. M., Bank, J. A., et al. (2009). Millisecond-scale molecular dynamics simulations on anton. *Proc. Conf. high Perform. Comput. Netw. Storage Anal.* 65, 1–11. doi:10.1145/1654059.1654126

Shell, M. S., Debenedetti, P. G., and Panagiotopoulos, A. Z. (2002). Generalization of the wang-landau method for off-lattice simulations. *Phys. Rev.* 66, 56703. doi:10.1103/physreve.66.056703

Shell, M. S. (2012). Systematic coarse-graining of potential energy landscapes and dynamics in liquids. *J. Chem. Phys.* 137, 84503. doi:10.1063/1.4746391

Shell, M. S. (2008). The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* 129, 144108. doi:10.1063/1.2992060

Singharoy, A., Maffeo, C., Delgado-Magnero, K. H., Swainsbury, D. J. K., Sener, M., Kleinekathöfer, U., et al. (2019). Atoms to phenotypes: molecular design principles of cellular energy metabolism. *Cell* 179, 1098–1111. doi:10.1016/j.cell.2019.10.021

Spoel, D. V. D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). Gromacs: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi:10.1002/jcc.20291

Takada, S. (2012). Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* 22, 130–137. doi:10.1016/j.sbi.2012.01.010

Torng, W., and Altman, R. B. (2019). Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model.* 59, 4131–4149. doi:10.1021/acs.jcim.9b00628

Tribello, G. A., and Gasparotto, P. (2019). Using dimensionality reduction to analyze protein trajectories. *Front. Mol. biosci.* 6, 46. doi:10.3389/fmolb.2019.00046

Wang, F., and Landau, D. (2001). Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram. *Phys. Rev.* 64, 056101. doi:10.1103/physreve.64.056101

Wang, F., and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86, 2050. doi:10.1103/physrevlett.86.2050

Wang, W., and Bombarelli, R. G. (2019). Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* 5, 1–9. doi:10.1038/s41524-019-0261-5

Webb, M. A., Delannoy, J. Y., and de Pablo, J. J. (2019). Graph-based approach to systematic molecular coarse-graining. *J. Chem. Theor. Comput.* 15, 1199–1208. doi:10.1021/acs.jctc.8b00920

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn Syst.* 32 (1), 4–24. doi:10.1109/TNNLS.2020.297838

Wüst, T., and Landau, D. P. (2008). The HP model of protein folding: a challenging testing ground for Wang-Landau sampling. *Comput. Phys. Commun.* 179, 124–127. doi:10.1016/j.cpc.2008.01.028

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). "How powerful are graph neural networks?," in Proceedings of the 7th international conference on learning representations, Ithaca, United States, October–February 1–22, 2018–2019 (ICLR), 17.

Zhang, S., Tong, H., Xu, J., and Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Comput. Soc. Netw.* 6, 11. doi:10.1186/s40649-019-0069-y

Zhang, Z., Cui, P., and Zhu, W. (2018). Deep learning on graphs: a survey. Preprint repository name [Corr]. Available at: https://arxiv.org/abs/1812.04202 (Accessed December 11, 2018).

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

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 May 2021
doi: 10.3389/fmolb.2021.647424

ML-AdVInfect: A Machine-Learning Based Adenoviral Infection Predictor

Onur Can Karabulut[1†], *Betül Asiye Karpuzcu*[1†], *Erdem Türk*[2†], *Ahmad Hassan Ibrahim*[2†] and *Barış Ethem Süzek*[2,3]*

[1]*Bioinformatics Graduate Program, Graduate School of Natural and Applied Sciences, Muğla Sıtkı Koçman University, Muğla, Turkey,* [2]*Department of Computer Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, Muğla, Turkey,* [3]*Georgetown University Medical Center, Biochemistry and Molecular and Cellular Biology, Washington, DC, United States*

OPEN ACCESS

Edited by:
Tunca Dogan,
Hacettepe University, Turkey

Reviewed by:
Emrah Nikerel,
Yeditepe University, Turkey
A. Ercument Cicek,
Bilkent University, Turkey

Correspondence:
Barış Ethem Süzek
barissuzek@mu.edu.tr

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *29 December 2020*
Accepted: *22 April 2021*
Published: *07 May 2021*

Citation:
Karabulut OC, Karpuzcu BA, Türk E, Ibrahim AH and Süzek BE (2021) ML-AdVInfect: A Machine-Learning Based Adenoviral Infection Predictor. Front. Mol. Biosci. 8:647424. doi: 10.3389/fmolb.2021.647424

Adenoviruses (AdVs) constitute a diverse family with many pathogenic types that infect a broad range of hosts. Understanding the pathogenesis of adenoviral infections is not only clinically relevant but also important to elucidate the potential use of AdVs as vectors in therapeutic applications. For an adenoviral infection to occur, attachment of the viral ligand to a cellular receptor on the host organism is a prerequisite and, in this sense, it is a criterion to decide whether an adenoviral infection can potentially happen. The interaction between any virus and its corresponding host organism is a specific kind of protein-protein interaction (PPI) and several experimental techniques, including high-throughput methods are being used in exploring such interactions. As a result, there has been accumulating data on virus-host interactions including a significant portion reported at publicly available bioinformatics resources. There is not, however, a computational model to integrate and interpret the existing data to draw out concise decisions, such as whether an infection happens or not. In this study, accepting the cellular entry of AdV as a decisive parameter for infectivity, we have developed a machine learning, more precisely support vector machine (SVM), based methodology to predict whether adenoviral infection can take place in a given host. For this purpose, we used the sequence data of the known receptors of AdVs, we identified sets of adenoviral ligands and their respective host species, and eventually, we have constructed a comprehensive adenovirus–host interaction dataset. Then, we committed interaction predictions through publicly available virus-host PPI tools and constructed an AdV infection predictor model using SVM with RBF kernel, with the overall sensitivity, specificity, and AUC of 0.88 ± 0.011, 0.83 ± 0.064, and 0.86 ± 0.030, respectively. ML-AdVInfect is the first of its kind as an effective predictor to screen the infection capacity along with anticipating any cross-species shifts. We anticipate our approach led to ML-AdVInfect can be adapted in making predictions for other viral infections.

Keywords: adenovirus, host susceptibility, host-pathogen interaction, virus-host interaction, PPI prediction, viral infection prediction, virus bioinformatics

INTRODUCTION

Adenoviruses (AdVs) are relatively large, nonenveloped, icosahedral viruses composed of a complex protein capsid surrounding the core proteins and the dsDNA genome. They belong to a diverse family called *Adenoviridae,* with several hundred recognized members capable of infecting a broad variety of cell types across several organisms (Rowe et al., 1953). As of the isolation of the first human AdV from adenoid tissues in 1953, many other novel AdVs were identified, such that, 103 human AdVs genotypes have been classified, to date, into seven "species" named A to G (Author Anonymous, 2020a). The pathogenic human AdVs (HAdV) may lead to serious gastrointestinal, respiratory, urinary, and corneal infections especially in immunosuppressed individuals (Gao et al., 2020). Moreover, recombinant AdVs are the most widely used viral vectors for gene therapy, accounting for 18.6% of vectors used in gene therapy clinical trials. AdVs feature out with their current and potential usage in different fields, including gene therapy, vaccine trials, and cancer treatments as oncolytic viruses (Singh et al., 2019).

Before any further steps leading to the infection may take place, viral pathogenesis requires the viral particle, the virion, to enter into the host cell. For AdVs, the main mechanism of entry is a two-step process, which starts with binding of a viral capsid protein (*i.e.* hexon, penton base, or mostly the fiber) to a primary receptor on the host cell to ensure attachment followed by secondary interactions to enable penetration of virion by clathrin- and dynamin-dependent endocytosis often involving integrins, or by macropinocytosis (Zhang and Bergelson, 2005; Lasswitz et al., 2018).

In explaining the pathogenesis of viral infections, therefore, understanding the viral protein–host receptor interactions plays a pivotal role. Expanding knowledge on AdV interactions, in particular, is essential not only to enhance our understanding of the life cycle, tissue tropism, host specificity/range, and cross-species transmission of the AdVs but also to help researchers in inhibiting adenoviral infections and in constructing efficient adenoviral vectors. Thus, HAdVs serve as a good template to elucidate virus–receptor interactions and as expectedly, identification and characterization of AdV receptors have been performed at varying levels of confirmation through different experimental methodologies by several investigators.

Given their diversity, broad host range, and complex use of receptors, the biological modeling of adenoviral infection poses a challenge to decipher with gaps and controversies in the existing literature. To this end, the use of computational methods on publicly available data about PPIs and the application of machine learning algorithms may accelerate and enrich our exploration of virus–host interactions. The conventional definition of PPI, however, refers to the physical contact with molecular docking between proteins that occur in a cell or in a living organism *in vivo*. As the definition implies, main databases and repositories that include PPIs are not structured from a host–(viral) pathogen point of view (De Las Rivas and Fontanillo, 2010). An exceptional resource which provides interspecies protein interaction data is the pathogen–host interaction search tool (PHISTO) (Durmus Tekir et al., 2013) which has extracted and integrated all PPIs between the human host and a non-human organism from publicly available databases and then manually labeled the respective organisms as pathogenic or not. For collected interactions without a specified method of detection, PHISTO includes a text mining module to predict the experimental method of interaction detection and also houses a user interface allowing visualization of protein networks. The recently launched pathogen–host interactions database (PHI-base), on the other hand, encompasses comprehensive expert-curated molecular and biological information, but does not cover viruses as a pathogen (Urban et al., 2020).

A similar concern also applies for the PPI prediction tools, yet there are several tools developed to predict virus–host interactions, and herein Section *Background*, we provide some background information on the publicly available virus–host PPI prediction tools DeNovo (Eid et al., 2016), HOPITOR (Basit et al., 2018), VHPPI (Alguwaizani et al., 2018), and InterSPPI-HVPPI (Yang et al., 2020) that we have used.

In the presented study, we have first curated the set of primary protein receptors that are essential in the adenoviral entry into the host cell based on the available evidence in the literature; herein Section *Adenoviral Receptors* Background, we provide further details regarding the included receptors. Then, using the public bioinformatics resources, we have identified the host species of adenoviruses, and also found the orthologs for our curated set of protein receptors in identified hosts. Similarly, we also created the set of adenoviral fiber proteins which stand for the ligands occupied in the adenoviral attachment. Next, for each of the fiber protein and adenoviral receptors, we had a dataset of pairs composed of the corresponding host and pathogen pair. Thus, altogether, we have compiled an extensive dataset on AdV–host relations. Next, we calculated the predictions as to whether there is an interaction between this particular virus fiber protein and host receptor as generated by four different existing PPI tools. Although these PPI tools are available individually, to this date, there is no approach that brings predictions of these tools together to make infection predictions. We recognize a virus-host PPI is not sufficient to warrant infection, yet attachment of the virus to a cellular receptor is a necessary condition and the initial step of viral entry which has been used previously as a decisive parameter for AdV infectivity by Hoffman et al. (Hoffmann et al., 2007; Hoffmann et al., 2008). We cannot accurately model, however, whether the viral interaction will cause its internalization or any further viral pathogenesis within the host cell. Taking these constraints into account, we used PPI as a basis for infection prediction. To this end, we applied a machine-learning, more specifically support vector machine (SVM), based methodology to develop the ML-AdVInfect predictor that uses virus-host PPI predictions from several tools in addition to the taxonomy data. This predictor is the first of its kind to carry the interaction prediction forward to anticipate whether adenoviral infection may occur in a given host species. The approach herein referred to yields a versatile and promising method to predict the occurrence of infection, investigate host-specificity, and anticipate cross-species transmissions for viral infections.

BACKGROUND

Adenoviral Receptors

The adenoviral receptors included in the present study contains molecules that were characterized specifically as the primary, proteinaceous, surface receptor for at least one HAdV type according to the available literature, excluding glycan-based interactions, interactions with secretory proteins, as well as any other molecular interactions which are auxiliary in nature. Based on the said criteria, we curated the set of receptors composed of coxsackie and adenovirus receptor (CAR), cluster of differentiation (CD) 46, CD80 and CD86, desmoglein-2 (DSG2), integrin subunit alpha-V (ITAV), macrophage scavenger receptor 1 (MSR1), and lung macrophage scavenger receptor SR-A6 (MARCO) and a brief overview on individual receptors and experimental methodology of receptor identification is given below (Lasswitz et al., 2018; Stasiak and Stehle, 2020).

CAR is a member of the junction adhesion molecule (JAM) family within the immunoglobulin (Ig) superfamily and is present in specialized intracellular junctions. CAR functions as a receptor for all HAdV species, except for the B species and interacts with the knob domain of the viral fiber protein (Tomko et al., 1997). CD46, also known as membrane cofactor protein (MCP), is expressed on all nucleated cells and belongs to the family of regulators of complement activation. For most species B HAdVs, which do not bind CAR, CD46 was shown to function as a cellular receptor (Gaggar et al., 2003). CD80 and CD86 are expressed on the cell surface of human dendritic cells and mature B lymphocytes (Caux et al., 1994). Species B AdVs use CD80 and CD86 as receptors and the fiber knob domain is required for the interaction (Short et al., 2006). DSG2 is a protein that belongs to the cadherin superfamily and was identified as the main receptor for HAdV-3, -7, -11, and -14. Unlike CD46 interactions, high-affinity binding to DSG2 requires both penton base and fiber protein (Wang et al., 2011). Integrins are a family of transmembrane heterodimers combining into 24 proteins in vertebrates which are engaged in a plethora of cellular functions. AdVs employ various integrins via their penton protein to mainly act as co-receptors. However, in a setting with little to no CAR expression, certain integrins from the group of the αv integrins were shown to function as a primary receptor. (Lyle and McCormick, 2010; Nestić et al., 2018). Scavenger receptors constitute a large group of membrane-bound receptors. The interaction with MSR1, also designated as SR-A and CD204, was shown to be responsible for liver uptake of HAdV5 (Haisma et al., 2009). Mutational analysis of AdV capsid proteins and *in vivo* administration in mice revealed that the SR-A interaction is mediated by the hypervariable regions of the AdV hexon protein (Piccolo et al., 2013). Similarly, in murine alveolar macrophage-like MPI cells MARCO was shown to be an entry receptor for HAdV-C5 and hexon protein was suggested to be relevant to the viral ligand (Stichling et al., 2018).

The most commonly used strategies to explore any protein-protein interactions (PPIs) are yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (AP-MS), in addition to other experimental modalities of array-based screening as well as flow cytometry-based binding assays, immunoadhesin/co-immunoprecipitation, luminescence, protease assays, surface plasmon resonance (SPR) and Förster Resonance Energy Transfer (FRET)-based techniques. In order to identify host factors of viral infection, initially, virus overlay protein binding assays (VOPBAs) were employed. For example, VOPBA successfully identified the AdV receptor CD46, among others (Gaggar et al., 2003) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7094377/- bb0405. Likewise, DSG2 was confirmed as a HAdV-3 receptor through binding assays including surface plasmon resonance and gain and loss of function assay. For follow-up analysis and validation of screening hits, genetic and drug-based validation methods including CRISPR/Cas9 and RNA interference are also being utilized. Syrian hamster models have been developed as an animal model for oncolytic species C HAdV vectors; however, AdV receptor studies are otherwise based on cell culture models. From a structural biological point of view, among the primary AdV receptors, only CAR and CD46 have solved structures in complex with their adenoviral ligands according to the entries in Protein Data Bank (PDB) database (Kilcher and Mercer, 2014; Brito and Pinney, 2017; Lasswitz et al., 2018; Hensen et al., 2020; Li et al., 2020; Stasiak and Stehle, 2020).

Machine Learning-Based PPI Prediction Tools

So far, several computational methods have been developed to predict virus-host protein interactions. As the publicly available virus-host PPI data increased, the emphasis on this subject has recently been shifted to machine-learning-based computational techniques to identify virus-host PPIs. PPI prediction tools have been developed based on different machine-learning models such as support vector machines (SVM) (Shen et al., 2007; Cui et al., 2012; Eid et al., 2016), random forest (RF) (Yang et al., 2020) and gradient boosting machine (XGBoost) (Basit et al., 2018; Chen et al., 2020).

An algorithm for predicting PPIs mediated by mimicked short linear motifs (SLiM) between HIV-1 and human has been developed by Becerra and colleagues (Becerra et al., 2017). Also, Eid and colleagues introduced an SVM-based virus-host PPI prediction model, called DeNovo, which uses amino acid sequence similarity-based features (Eid et al., 2016). Based on three PPI sets, containing several bacterial and human protein interactions, DeNovo achieved an average accuracy, sensitivity, and specificity of 97%, 94.5%, and 97.5% respectively. The most important feature that distinguishes DeNovo from other SVM-based prediction tools is that it employs a sequence similarity-based strategy for sampling the negative virus-host PPI data set for SVM training. The DeNovo sampling strategy has inspired other researchers to develop new virus-host PPI methods. HOPITOR, an XGBoost classifier-based host-pathogen predictor, is another method using the DeNovo sampling strategy. However, the sequence similarity between the different virus and host types is rather low. As a consequence, sequence similarity-based prediction methods have some limitations. To cope with this problem, Zhou and colleagues applied Naive Bayes, RF, and SVM models on feature vectors derived from amino acid compositions of interacting host-virus proteins and introduced another SVM-based tool called VirusHostPPI (Zhou et al., 2018). VirusHostPPI has been compared with two different methods, including DeNovo (Eid et al., 2016) and Barman's SVM (Barman et al., 2014), and it achieved an accuracy of 84.47%–79.95%, the sensitivity of 80.00%–76.14% and specificity of 88.94%–83.77% against DeNovo and Barman's SVM, respectively. As a result of the latest efforts in virus-host PPI prediction, Yang and colleagues introduced a doc2vec embedding-based RF classifier called InterSPPI-HVPPI. Using Barman et al.'s dataset, InterSPPI-HVPPI achieved 79.17% accuracy, 81.85% sensitivity, and 76.45% specificity.

In a similar manner to the overall experience in other research fields, the number of machine learning-based approaches to virus-host interaction prediction has been increasing rapidly over time, bringing a gradual decrease in the difference of performances between the developed methods. Besides, considering the host and pathogen diversity, it would be more efficient to develop new PPI prediction methods using ensemble learning techniques instead of highlighting a single method in the literature. Ensemble learning-based approaches use multiple learning algorithms to achieve greater predictive performance than is possible from any single of the constituent learning algorithms alone (Polikar, 2006; Rokach, 2010).

Here, we introduce a machine-learning-based methodology to predict AdV infections based on the utilization of an ensemble of available virus-host PPI prediction tools.

MATERIALS AND METHODS

Identification of Adenovirus Hosts

We constructed a library of AdV hosts using the UniProt knowledgebase (UniProtKB Release 2020_02) (The UniProt, 2017), the Virus-Host DB (Mihara et al., 2016), and the National Center for Biotechnology Information GenBank (Clark et al., 2016). We initially created a list of host organisms using the curated "Virus Hosts" information available in UniProtKB for the "Adenoviridae" family, primary hosts curated in Virus-Host DB, and hosts curated in GenBank records for Adenoviridae complete genomes. Next, we parsed out the hostnames out of the AdV species names (e.g. "Human" for Human Adenovirus). The hostnames from both steps were further curated to obtain a species (or subspecies) level host organism nomenclature, reviewing the related literature and/or sequence submission records (e.g. "*Gallus gallus*" for UniProt: R4N0P7, rather than "fowl"). The list of infecting AdV species is also curated for each host and AdV–host pairs are generated.

Creation of Adenovirus Host Receptor Protein Sets

We identified orthologs of AdV receptors in the hosts using a sequence similarity-based approach. We initially compiled the human protein sequences for the list of receptors we have manually curated, namely CAR/CXAR (UniProt Accession: P78310), CD46 (UniProt Accession: P15529), CD80 (UniProt Accession: P33681), CD86 (UniProt Accession: P42081), ITAV (UniProt Accession: P06756), DSG2 (UniProt Accession: Q14126), MSR1 (UniProt Accession: P21757), and MARCO (UniProt Accession: Q9UEW3). Human receptors are selected as a starting point, as human is the most well-studied AdV host. We ran BLAST (Altschul et al., 1990) searches with human receptor proteins (as query sequences) against locally downloaded protein sequences from UniProtKB for all the hosts with complete proteomes based on the UniProt Proteomes database. Availability of complete proteome was applied as a criterion to make sure that all orthologs are potentially represented in the respective proteomes. We parsed BLAST results to identify orthologs from various hosts using e-value and overlap thresholds. As CAR is the first-identified and most well-studied receptor in mammalian hosts, our aim was to be able to catch all the CAR orthologs in 40 host organisms in the study through BLAST searches. Moreover, we tried to avoid partial CAR orthologs or fragments. Thus, we tried different BLAST e-value and overlap thresholds, and the e-value (<1e-20) and overlap (>66%) thresholds were chosen to maximize the number of full-length orthologs of CAR receptors.

Creation of Adenovirus Fiber Protein Sets

In order to compile a comprehensive set of AdV fiber proteins, we initially curated a fiber protein synonym list using UniProtKB adenoviridae entries to cope with naming inconsistencies. "Fiber," "fibre," "fiber protein," "fiber homolog," "protein fiber," "fibre protein" and "fibre homolog" were among the few terms we identified as possible names assigned for the fiber protein orthologs. We then used UniProt website REST API and our terms to retrieve AdV fiber proteins. Furthermore, to account for uncharacterized fiber proteins (i.e. uncharacterized protein or hypothetical protein, or unknown), we BLAST'ed a local database of AdV sequences using the curated fiber proteins using the same e-value and overlap thresholds described in Section *Creation of Adenovirus Host Receptor Protein Sets*.

Preparing Dataset for Adenovirus Infection Prediction

To apply machine learning classification algorithms to predict adenoviral infection, we created a dataset containing AdV–host pairs. The dataset contained all possible host and AdV pairs, where hosts are from the final list at Section *Creation of Adenovirus Host Receptor Protein Sets* and AdVs are the ones that have the fiber proteins as identified in Section *Creation of Adenovirus Fiber Protein Sets*.

For each AdV and host species pair, we computed a feature vector with two major components and a class label. The first component is predictions of virus–host protein interaction for AdV fiber protein and host receptors; the basic prerequisite for adenoviral infection. The second component is which was incorporated to account for a potential taxonomic preference toward host receptors. Finally, the class label indicates whether the AdV in question is known to infect the respective host based

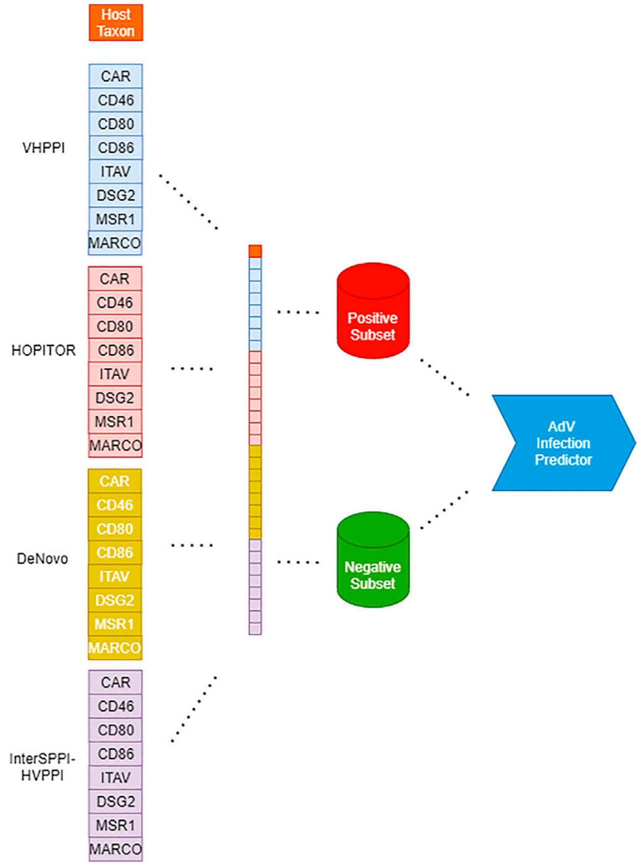


FIGURE 1 | Workflow for creation of adenoviral infection prediction models from left to right: Creation of feature vector based on virus-host PPI predictors, host taxa, and infection class; partitioning of the dataset according to the class. Finally, creation of an infection predictor through various machine-learning algorithms such as RF, SVM, and MLP.

on the known AdV–host pairs generated in Section *Identification of Adenovirus Hosts.*

To serve as the first component of the feature vector, we utilized four virus–host PPI predictors with their respective default parameters DeNovo (run locally, using a dissimilarity threshold of 0.8), HOPITOR (run locally, with default parameters), VHPPI (online version[1] with default parameters), and InterSPPI-HVPPI (run locally, using default specificity threshold of 0.95 as per its web site[2]). In an attempt to factor in the strengths and weaknesses of individual virus-host PPI prediction tools, and their varying prediction performance for different receptors, we applied a stacking-like ensemble technique using DeNovo, HOPITOR, VirusHostPPI, and InterSPPI-HVPPI models. For each one of 10,237 AdV–host pairs, 4 interaction predictions were computed per receptor which resulted in 32 predictions (4 predictors; DeNovo, HOPITOR, VirusHostPPI, and InterSPPI-HVPPI x 8 receptors; CAR, CD46, CD80, CD86, ITAV, DSG2, MSR1, and MARCO). Each feature in this component had a binary value; either 1 (interacting) or 0 (otherwise). For practical purposes, the lack of a specific host receptor is treated as if there were no interaction between that receptor and the fiber protein.

As the second component, we captured host taxa at four taxonomic levels; genus, family, order, and class. National Center for Biotechnology Information (NCBI) Taxonomy Database was used to gather the taxon of each organism (Federhen, 2012).

Finally, the infection class label, for each AdV–host pair, is computed to constitute the ground truth as to whether that particular AdV infects that respective host. For this purpose, we looked at the host portion in the pair to see whether it is identical to the known host of the AdV in that pair (e.g., the known host for human AdV is "homo sapiens"). If these two are identical, class label 1 is assigned as an indication of infection under the assumption that there are no cross-species transmission, while 0 is assigned as an indication of AdV being not infectious for the host in question. Consequently, the feature vectors with class label 1 (one) form the positives (i.e., adenoviral infection happens) while those with class label 0 (zero) form the negatives of our dataset.

An illustration of the creation of a Dataset for Adenovirus Infection Prediction is provided as part of **Figure 1**.

Creation of Adenovirus Infection Prediction Models

We used machine learning classification algorithms RF, SVM, and Multilayer Perceptron (MLP) on the dataset described in Section *Preparing Dataset for Adenovirus Infection Prediction*. The algorithms were chosen based on their use and reported performance on similar problems in bioinformatics such as virus-host protein interaction prediction (See Background). To cope with the class imbalance problem between the number of positives (i.e. adenoviral infection happens) and negatives, we employed random oversampling of minority positives set during the training of the infection prediction model. We experimented using one level of host taxa (genus, family, order, or class) at a time as part of feature vectors. For the classification algorithms requiring numerical values, host taxa which is a categorical feature are encoded using the label encoder in Scikit-Learn. For each machine learning classification algorithm, we first split our dataset into a training set (the 80% portion) to conduct hyperparameter tuning and a test set (the 20% portion) to assess respective performances. During hyperparameter tuning, we used 10-fold cross-validation where we first split the training set into 10 folds and then applied random oversampling on 9 folds which were used for training the classification model and then tested the model performance on the remaining 1 fold. It has been documented that oversampling and undersampling leads to similar performances, provided that the sampling is correctly implemented on the training folds, as we have done, during the cross-validation (Blagus and Lusa, 2015). Following the hyperparameter tuning, the best models trained on the training test (the 80% portion) are used to classify the test set

[1] http://165.246.44.47/VirusHostPPI/Prediction
[2] http://zzdlab.com/hvppi/predict.php

for assessment of the model performances. The following performance metrics to compare our models where TP, FP, TN and FN represent the number of true positives, false positives, true negatives and false negatives, respectively. True positives (TP) contain host proteins which are predicted to correctly interact with a virus protein. True negatives (TN) are non-interactive host proteins that are correctly predicted to be non-interacting with a virus protein. False Positive (FP) is a non-interactive host protein that is wrongly predicted to interact with a virus protein. False negatives (FNs) are host proteins that are wrongly predicted to interact with a virus protein.

Precision measures the ability or quality of a measurement to be consistently reproduced.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity measures the proportion of true positives that are correctly identified.

$$\text{Sensitivity} = \text{Recall} = TPR = \frac{TP}{TP + FN}$$

Specificity measures the proportion of true negatives.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Accuracy is how close a measured value is to the actual (true) value.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

F-Score is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into “positive” or “negative”.

$$F - \text{Score} = \frac{2^{*}\text{Precision}^{*}\text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under Curve (AUC) refers to the area under the receiver operating characteristics curve which is one of the most important evaluation metrics for checking any classification model’s performance. It tells how much the model is capable of distinguishing between classes.

$$\text{Area Under Curve} = \int_{a}^{b} f(x)\mathrm{d}x$$

Matthew’s Correlation Coefficient (MCC) is used in machine learning. It is a measure of the quality of binary (two-class) classifications.

$$\text{MCC} = \frac{TP^{*}TN - FP^{*}FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

All the machine learning models are implemented using Scikit-Learn library (2020b) and random oversampling was implemented using the imbalanced-learn toolbox (Lemaître et al., 2017) for the Python programming language. Unless otherwise is specified, the default parameters of respective implementations of RF, SVM, and MLP were used.

An illustration of the flow of our work steps from the creation of Dataset for Adenovirus Infection Prediction until the creation of Adenovirus Infection Prediction Models is provided in **Figure 1**.

RESULTS

Adenovirus Host/Receptor and Fiber Protein Sets

The identification of AdV hosts resulted in 297 unique species as potential host species. The majority of the hosts (n = 179) are mammalians and primates predominate this class. Once these hosts were sorted out based on the availability of the complete proteomes in UniProt, the remaining 40 host species were included in our final set of hosts. See **Supplementary Table S1** for a full list of identified hosts as well as the information as to whether complete proteome data for the relevant host is available or not. Our results further confirm that the AdVs infect a wide variety of organisms including mammals, lizards, birds, turtles, and frog and toads (See **Figure 2**).

Out of 40 host species, CAR is found in 32 organisms, CD46 in 25, CD80 in 23, CD86 in 33, ITAV in 36, DSG2 in 38, and the scavenger receptors MSR1 and MARCO exist in 25 and 17 of these host species, respectively. For each of the 40 host species, the UniProt accession numbers for existing receptors are provided in **Supplementary Table S2**. As a validation, we have compared the identified receptors’ orthologs against the respective orthologs recorded in the OrthoDB database (Kriventseva et al., 2019). Our orthologs included all those recorded in OrthoDB, and further included some additional uncharacterized orthologs (eg. UniProtKB: M3Y0B3 as a CD86 ortholog in *Mustela furo*).

Our set of AdV fiber proteins is composed of 254 fiber proteins. A full list of these proteins together with the adenoviruses they belong to is provided in **Supplementary Table S3**.

Dataset for Adenovirus Infection Prediction

Our dataset contains a total of 10,237 AdV–host pairs, of which 220 are from the positive class and 10,017 are from the negative class. For each AdV–host pair, each one of the 4 virus–host PPI prediction tools was used separately to make predictions for 8 host receptors. The prediction results where 1 indicates interaction and 0 indicates either no interaction or non-existence of the corresponding receptor are provided in **Supplementary Table S4**.

We compared the prediction results for our dataset using the correlation coefficients between individual virus–host PPI tools which are provided in **Table 1**. The coefficient correlations between the tools range between 0.13 and 0.79. InterSPPI-HVPPI produced a rather low number of positive predictions for the entire set of the receptors which is attributable to its conservative nature. Therefore, it has been excluded from the correlation with the other tools. The longer proteins, which are DSG-2 and ITAV (ca. 1000 amino acids) had the poorest correlation which suggests a size-dependency in the prediction of these tools.

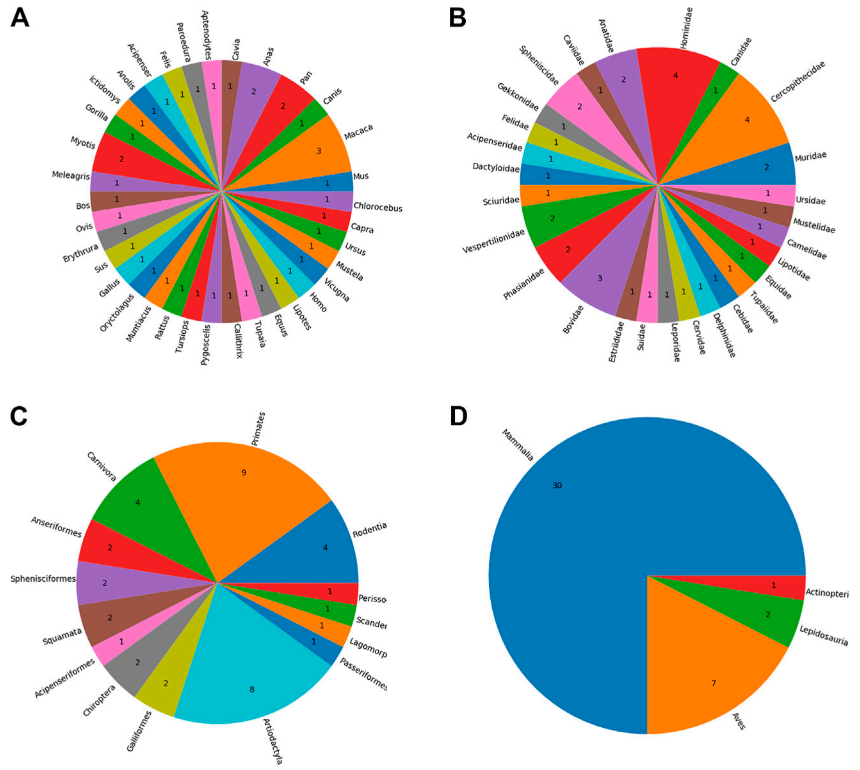


FIGURE 2 | The taxonomic distribution of identified host species by **A**) genus **B**) family **C**) order, and **D**) class.

TABLE 1 | Correlation coefficients of PPI predictors by adenoviral receptor.

		VHPPI	HOPITOR	DeNovo			VHPPI	HOPITOR	DeNovo
CAR	VHPPI	1.00			CD46	VHPPI	1.00		
	HOPITOR	0.52	1.00			HOPITOR	0.41	1.00	
	DeNovo	0.79	0.51	1.00		DeNovo	0.55	0.54	1.00
		VHPPI	HOPITOR	DeNovo			VHPPI	HOPITOR	DeNovo
CD80	VHPPI	1.00			CD86	VHPPI	1.00		
	HOPITOR	0.14	1.00			HOPITOR	0.21	1.00	
	DeNovo	0.38	0.57	1.00		DeNovo	0.31	0.48	1.00
		VHPPI	HOPITOR	DeNovo			VHPPI	HOPITOR	DeNovo
ITAV	VHPPI	1.00			DSG2	VHPPI	1.00		
	HOPITOR	0.13	1.00			HOPITOR	0.13	1.00	
	DeNovo	0.31	0.17	1.00		DeNovo	0.52	0.23	1.00
		VHPPI	HOPITOR	DeNovo			VHPPI	HOPITOR	DeNovo
MSR1	VHPPI	1.00			MARCO	VHPPI	1.00		
	HOPITOR	0.35	1.00			HOPITOR	0.50	1.00	
	DeNovo	0.49	0.44	1.00		DeNovo	0.79	0.66	1.00

The highest correlation, on average, is between DeNovo and VHPPI followed by DeNovo and HOPITOR. The variance of correlations between the tools per receptor reinforces the use of an ensemble of virus-host PPI prediction tools rather than opting for a single one. As none of these tools correlate 100%, we anticipate each will complement each other and boost the overall performance of virus–host PPI prediction. We also validated PPI predictions against public databases. We checked PHISTO and identified that it names only human adenoviral receptor CAR and its interactions with human AdV2 and human AdV12 fiber proteins which we checked against the results from 4 PPI prediction tools and confirmed that all 4 predicted these interactions correctly.

Comparison of Adenovirus Infection Prediction Models

We have used training set (the 80% portion) of our dataset which was generated as described in Section *Creation of Adenovirus*

TABLE 2 | Performance metrics of adenoviral infection prediction models. RBF, Radial Basis Function; AUC, Area Under the Curve; MCC, Matthew's Correlation Coefficient.

Host taxa level	Classifier	Sensitivity	Specificity	Accuracy	F-score	MCC	AUC
Genus	**SVM (kernel = "rbf," gamma = "auto")**	**0.92 ± 0.009**	**0.86 ± 0.047**	**0.92 ± 0.009**	**0.96 ± 0.005**	**0.39 ± 0.035**	**0.89 ± 0.023**
	MLP (activation = "tanh," hidden layer=(16,4))	0.95 ± 0.009	0.73 ± 0.064	0.94 ± 0.009	0.97 ± 0.005	0.40 ± 0.045	0.84 ± 0.031
	Random forest (number of trees = 50, criterion = "entropy," max_depth = 16)	0.96 ± 0.006	0.70 ± 0.071	0.95 ± 0.006	0.98 ± 0.003	0.42 ± 0.043	0.83 ± 0.035
Family	**SVM (kernel = "rbf," gamma = "auto")**	**0.91 ± 0.009**	**0.84 ± 0.057**	**0.91 ± 0.008**	**0.95 ± 0.005**	**0.36 ± 0.031**	**0.88 ± 0.027**
	MLP (activation = "tanh," hidden layer=(16,4))	0.94 ± 0.012	0.72 ± 0.064	0.94 ± 0.011	0.97 ± 0.006	0.37 ± 0.048	0.83 ± 0.031
	Random forest (number of trees = 50, criterion = "entropy", max_depth = 16)	0.95 ± 0.007	0.66 ± 0.068	0.95 ± 0.007	0.97 ± 0.004	0.38 ± 0.043	0.81 ± 0.033
Order	**SVM (kernel = "rbf," gamma = "auto")**	**0.91 ± 0.009**	**0.82 ± 0.057**	**0.90 ± 0.009**	**0.95 ± 0.005**	**0.34 ± 0.028**	**0.86 ± 0.027**
	MLP (activation = "tanh," hidden layer=(16,4))	0.94 ± 0.011	0.70 ± 0.075	0.94 ± 0.011	0.97 ± 0.006	0.37 ± 0.051	0.82 ± 0.038
	Random forest (number of trees = 50, criterion = "entropy," max_depth = 16)	0.95 ± 0.007	0.66 ± 0.070	0.95 ± 0.007	0.97 ± 0.004	0.37 ± 0.040	0.81 ± 0.034
Class	**SVM (kernel = "rbf," gamma = "auto")**	**0.88 ± 0.011**	**0.82 ± 0.061**	**0.88 ± 0.010**	**0.93 ± 0.006**	**0.30 ± 0.028**	**0.85 ± 0.029**
	MLP (activation = "tanh," hidden layer=(16,4))	0.94 ± 0.010	0.68 ± 0.067	0.93 ± 0.010	0.97 ± 0.005	0.35 ± 0.043	0.81 ± 0.032
	Random forest (number of trees = 50, criterion = "entropy," max_depth = 16)	0.95 ± 0.007	0.63 ± 0.071	0.94 ± 0.007	0.97 ± 0.004	0.35 ± 0.043	0.79 ± 0.035
None	**SVM (kernel = "rbf," gamma = "auto")**	**0.88 ± 0.011**	**0.83 ± 0.064**	**0.88 ± 0.010**	**0.93 ± 0.006**	**0.30 ± 0.029**	**0.86 ± 0.030**
	MLP (activation = "tanh," hidden layer=(16,4))	0.94 ± 0.009	0.68 ± 0.079	0.93 ± 0.008	0.96 ± 0.005	0.34 ± 0.043	0.81 ± 0.038
	Random forest (number of trees = 50, criterion = "entropy," max_depth = 16)	0.95 ± 0.007	0.63 ± 0.072	0.94 ± 0.007	0.97 ± 0.004	0.35 ± 0.041	0.79 ± 0.035

Bolded value indicates the implementation of the SVM algorithm yielded the best performance in terms of sensitivity for infection prediction for our particular dataset for all the experiments.

Infection Prediction Models for hyperparameter tuning of SVM-, RF-, and MLP-based models for adenoviral infection prediction. SVM was tested with several kernels (polynomial, radial basis function (RBF), and sigmoid) and gamma values (default = auto, 1, 10). For SVM, the highest sensitivity and AUC scores were consistently achieved with the RBF and otherwise default parameters. We experimented with MLP with activation functions ReLU and tanh along with the different hidden layer configurations. For MLP, tanh yielded the highest sensitivity and AUC scores with a hidden layer configuration of [16, 4]. RF was also experimented with several parameters including depth (50, default = 100, 150), number of trees, and split metrics (gini and entropy) where the best sensitivity and AUC scores were attained with the depth = 16, number of trees = 50, and split metrics = entropy. The results from hyperparameter tuning which were carried out without the host taxa are available in **Supplementary Table S5**.

The performance metrics were computed on the test split (the 20% portion) for the best SVM-, RF-, and MLP-based models which are identified through hyperparameter tuning. The 80%–20% train-test split was repeated 100 times and the mean values and standard deviation are reported in **Table 2**. For our study, we favored higher sensitivity models since our main focus was correctly predicting infection. The implementation of the SVM algorithm yielded the best performance in terms of sensitivity for infection prediction for our particular dataset for all the experiments (bolded in **Table 2**) we conducted with or without the inclusion of the host taxa levels.

According to our findings, the inclusion of the host taxa level led to a slight performance improvement in terms of sensitivity, specificity and MCC. Although it was informative to see the potential benefit of inclusion of host taxa to overall predictor performance, we wanted to avoid any bias introduced by our dataset's limited representation of the real taxonomic diversity of AdV hosts. Hence, we decided to exclude host taxa level in training models at the moment, while deferring the inclusion of host taxa to a later iteration of ML-AdVInfect when more AdV host complete proteomes become available.

For the reasons mentioned above, in this study, we chose SVM with RBF kernel model over alternative models trained without host taxa level. The analysis reported in Section *Discussion* is based on this model. **Supplementary Table S4** also includes the infection predictions of this SVM with RBF kernel-based model.

In order to assess the infection prediction power of a single receptor and a single PPI prediction tool, we used the same set of machine-learning algorithms and parameters as in our hyperparameter tuning experiments described above for the overall AdV infection prediction model. The results for hyperparameter tuning for single receptor/PPI prediction tool experiments, which were carried out without the host taxa, are available in **Supplementary Table S6**. In turn, the performance metrics computed for the test set are available in **Supplementary Table S7**. Based on their performance metrics, we conclude a single-receptor-based or single-PPI-predictor-based infection prediction model is not achievable.

DISCUSSION

AdVs are infectious microorganisms that are particularly harmful to elderly and immunocompromised individuals. Along with their clinical importance, AdVs have further implications as they are promising vectors for gene and vaccine delivery. Therefore, adenoviral interactions with their hosts have been extensively searched. To the best of our knowledge, on the other hand, there is no computational model to estimate whether AdV can cause an infection or not in a given host. The model we propose here encompasses a machine learning-based approach to predict the infection capacity of AdVs.

In our study, we favored models trained without host taxa level as our dataset is not necessarily a representation of a wide diversity of AdV hosts. The highest sensitivity predictor among these models was based on SVM with RBF kernel with performance metrics sensitivity, specificity, and AUC 0.88 ± 0.011, 0.83 ± 0.064, and 0.86 ± 0.030, respectively. Our preference for favoring sensitivity rather than specificity is tailored toward our main goal of correctly predicting infection, but our approach does not preclude favoring higher specificity models such as MLP and RF.

In our analysis, we also identified that a single-receptor-based or single-PPI-predictor-based infection prediction model is not achievable. Yet, the overall performance of ML-AdVInfect demonstrates the utility of a stacking-like ensemble of PPI predictors for infection prediction.

In bioinformatics, several machine learning problems have to handle class-imbalanced data. Ours is not an exception to this. Oversampling techniques to randomly add instances from the minority class or undersampling techniques to randomly drop instances from the majority class are widely used on such imbalanced data (Radivojac et al., 2004; Taft et al., 2009; Kim and Choi, 2014; Li et al., 2014). Yet, as long as the cross-validation is implemented correctly, choice of sampling results in similar model performances (Radivojac et al., 2004). In the light of this, we have opted for oversampling with a correct implementation in the cross-validation process.

According to the documented results in the literature, the available virus–host PPI prediction tools (see Background) have varying performance. The level of agreement between the individual tools was limited based on our correlation analysis (coefficients at a range of 0.13–0.79). This was our main motivation behind using an ensemble of these tools for infection prediction. As our model strictly relies on the performance and use of virus–host PPI prediction tools, improvement in the performance of existing ones and/or the introduction of newly developed ones may help to attain better infection predictions.

We have addressed the main adenoviral entry mechanism into the cells, namely, binding to the primary membrane receptor on the host cell by the viral ligand (namely, CAR, CD46, CD80, CD86, ITAV, DSG2, MSR1, and MARCO) yet it is worth to emphasize that occurrence and spread of adenoviral infection may also make use of interactions between non-proteinaceous portions of molecules, viral binding to soluble host proteins, secondary interactions between the virus and host, as well as the internalization of the virion through caveolin- or clathrin-dependent mechanisms. Similarly, ligand-wize, our dataset comprises merely the AdV fiber proteins which are the most common but indeed are not necessarily the only domain of viral binding. Here, we pursued an approach to ensure the proven determinants of infection are encompassed through manual curation of a set of receptors. This approach can be expanded from both virus and host side to accommodate other interacting proteins if needed.

Although we tried to identify primary human adenoviral receptors and their orthologs to our best effort, we cannot rule out the possibility that there may still be uncharacterized proteins in various hosts or partially sequenced host genomes. Hence, we restricted our dataset to include 40 complete proteomes as curated by UniProt. As a future insight, completed proteomes might be added to this dataset as they become available.

Cross-species transmission of viruses corresponds to the capacity of a virus species to infect other host organism(s) in addition to its original host. In order to assess the capability of our predictor in detecting a potential interspecies shift, we further investigated the false positives of our best predictor, namely the SVM model with RBF kernel, as they might as well catch a cross-species transmission event. Of our false-positive results, 15% accounts for the cases where a HAdV infects another non-human primate which is a well-established zoonotic shift of AdVs (Hoppe et al., 2015). Furthermore, in 26% of the cases a primate AdV was predicted to infect another primate which could potentially be an indication of cross-species transmission. For the primates, we did a literature review and inspected Virus-Host DB. One of our false positive predictions refers to the human infection caused by a titi monkey adenovirus ECC-2011. We have identified that both *Callicebus cupreus* and *Homo sapiens* were reported as host organisms infected by this virus. According to the documented transmission (Chen et al., 2011), a novel adenovirus (TMAdV, titi monkey adenovirus) was identified in a colony of titi monkeys confined in a research center who experienced fulminant pneumonia and hepatitis leading to a devastating outcome; 23 out of 65 monkeys were infected, of whom 18 were lost. Furthermore, the researcher who was in closest contact with these monkeys also developed upper respiratory symptoms and found to be seropositive, and more concerningly, also had a clinically ill family member with no colony contact who was as well tested seropositive. Most likely, this new world monkey colony has acquired the pathogen from an unknown natural reservoir, but this outbreak implies the offending pathogen is capable of breaking the species barrier and may even cause human-to-human transmission. Although remained at a smaller scale on this particular occasion, viruses that can cross the species barrier and infect a broad primate host range may lead to larger epidemics and therefore needs closer attention. Similarly, AdVs may also be transmitted within domestic settings, across humans and domestic animals (Pauly et al., 2015). Out of false positives, 34 could be attributed to the shift of AdVs host from human to domestic animals including dog, goat, and pigs.

The work presented here, namely ML-AdVInfect, is the first of its kind in terms of allowing adenoviral infection prediction. As a step toward this predictor, we have also constructed a comprehensive dataset of AdV–host interactions which may accommodate other studies on AdVs. The proposed approach is an effective predictor to screen the infection capacity along with anticipating any cross-species shifts. It is also versatile as it allows expansion by the addition of novel virus–host PPI predictors, new host organisms, and newly identified AdV species. We anticipate such expansions will make positive contributions to the overall performance of the ML-AdVInfect. Our approach that is composed of identifying hosts, host–virus interacting protein pairs, and creating a machine-learning-based model leveraging individual virus–host PPI prediction tools, can be adapted for making predictions of infection by other viruses. As a prospective work, based on our tool ML-AdVInfect together with its further expansions and/or adaptations, a web platform with a user interface will also be provided.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

BS and ET contributed to the conception and design of the study. OK and AI constructed and tested the models. BK implemented data collection and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. All authors have contributed equally to this work.

FUNDING

This project is supported by The Scientific and Technological Research Council of Turkey (Grant number: 119E664) We would like to acknowledge: Türkiye Bilimsel ve Teknolojik Araştirma Kurumu (Award number(s): 119E664).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.647424/full#supplementary-material

REFERENCES

Alguwaizani, S., Park, B., Zhou, X., Huang, D. S., and Han, K. (2018). Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids. *J. Healthc. Eng.* 2018, 1391265. doi:10.1155/2018/1391265

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/s0022-2836(05)80360-2

Author Anonymous (2020a). *HAdV Working Group* [Online]. Available: http://hadvwg.gmu.edu (Accessed December 24, 2020)

Author Anonymous (2020b). *Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.24.0 Documentation [Online]*. Available: https://scikit-learn.org/stable (Accessed December 24, 2020)

Barman, R. K., Saha, S., and Das, S. (2014). Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods. *PLoS One*. 9, e112034. doi:10.1371/journal.pone.0112034

Basit, A. H., Abbasi, W. A., Asif, A., Gull, S., and Minhas, F. U. A. A. (2018). Training Host-Pathogen Protein-Protein Interaction Predictors. *J. Bioinform. Comput. Biol.* 16, 1850014. doi:10.1142/s0219720018500142

Becerra, A., Bucheli, V. A., and Moreno, P. A. (2017). Prediction of Virus-Host Protein-Protein Interactions Mediated by Short Linear Motifs. *BMC Bioinformatics*. 18, 163. doi:10.1186/s12859-017-1570-7

Blagus, R., and Lusa, L. (2015). Joint Use of over- and Under-sampling Techniques and Cross-Validation for the Development and Assessment of Prediction Models. *BMC Bioinformatics*. 16, 363. doi:10.1186/s12859-015-0784-9

Brito, A. F., and Pinney, J. W. (2017). Protein-Protein Interactions in Virus-Host Systems. *Front. Microbiol.* 8, 1557. doi:10.3389/fmicb.2017.01557

Caux, C., Vanbervliet, B., Massacrier, C., Azuma, M., Okumura, K., Lanier, L. L., et al. (1994). B70/B7-2 Is Identical to CD86 and Is the Major Functional Ligand for CD28 Expressed on Human Dendritic Cells. *J. Exp. Med.* 180, 1841–1847. doi:10.1084/jem.180.5.1841

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., et al. (2020). Improving Protein-Protein Interactions Prediction Accuracy Using XGBoost Feature Selection and Stacked Ensemble Classifier. *Comput. Biol. Med.* 123, 103899. doi:10.1016/j.compbiomed.2020.103899

Chen, E. C., Yagi, S., Kelly, K. R., Mendoza, S. P., Tarara, R. P., Canfield, D. R., et al. (2011). Cross-species Transmission of a Novel Adenovirus Associated with a Fulminant Pneumonia Outbreak in a New World Monkey Colony. *Plos Pathog.* 7, e1002155. doi:10.1371/journal.ppat.1002155

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/nar/gkv1276

Cui, G., Fang, C., and Han, K. (2012). Prediction of Protein-Protein Interactions between Viruses and Human by an SVM Model. *BMC Bioinformatics*. 13 (Suppl. 7), S5. doi:10.1186/1471-2105-13-s7-s5

De Las Rivas, J., and Fontanillo, C. (2010). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *Plos Comput. Biol.* 6, e1000807. doi:10.1371/journal.pcbi.1000807

Durmus Tekir, S., Cakir, T., Ardic, E., Sayilirbas, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: Pathogen-Host Interaction Search Tool. *Bioinformatics*. 29, 1357–1358. doi:10.1093/bioinformatics/btt137

Eid, F.-E., Elhefnawi, M., and Heath, L. S. (2016). DeNovo: Virus-Host Sequence-Based Protein-Protein Interaction Prediction. *Bioinformatics*. 32, 1144–1150. doi:10.1093/bioinformatics/btv737

Federhen, S. (2012). The NCBI Taxonomy Database. *Nucleic Acids Res.* 40, D136–D143. doi:10.1093/nar/gkr1178

Gaggar, A., Shayakhmetov, D. M., and Lieber, A. (2003). CD46 Is a Cellular Receptor for Group B Adenoviruses. *Nat. Med.* 9, 1408–1412. doi:10.1038/nm952

Gao, J., Zhang, W., and Ehrhardt, A. (2020). Expanding the Spectrum of Adenoviral Vectors for Cancer Therapy. *Cancers*. 12, 1139. doi:10.3390/cancers12051139

Haisma, H. J., Boesjes, M., Beerens, A. M., Van Der Strate, B. W. A., Curiel, D. T., Plüddemann, A., et al. (2009). Scavenger Receptor A: A New Route for Adenovirus 5. *Mol. Pharmaceutics*. 6, 366–374. doi:10.1021/mp8000974

Hensen, L. C. M., Hoeben, R. C., and Bots, S. T. F. (2020). Adenovirus Receptor Expression in Cancer and its Multifaceted Role in Oncolytic Adenovirus Therapy. *Int J Mol Sci.* 21, 6828. doi:10.3390/ijms21186828

Hoffmann, D., Bayer, W., Heim, A., Potthoff, A., Nettelbeck, D. M., and Wildner, O. (2008). Evaluation of Twenty-One Human Adenovirus Types and One Infectivity-Enhanced Adenovirus for the Treatment of Malignant Melanoma. *J. Invest. Dermatol.* 128, 988–998. doi:10.1038/sj.jid.5701131

Hoffmann, D., Heim, A., Nettelbeck, D. M., Steinstraesser, L., and Wildner, O. (2007). Evaluation of Twenty Human Adenoviral Types and One Infectivity-Enhanced Adenovirus for the Therapy of Soft Tissue Sarcoma. *Hum. Gene Ther.* 18, 51–62. doi:10.1089/hum.2006.132

Hoppe, E., Pauly, M., Gillespie, T. R., Akoua-Koffi, C., Hohmann, G., Fruth, B., et al. (2015). Multiple Cross-Species Transmission Events of Human Adenoviruses (HAdV) during Hominine Evolution. *Mol. Biol. Evol.* 32, 2072–2084. doi:10.1093/molbev/msv090

Kilcher, S., and Mercer, J. (2014). Next Generation Approaches to Study Virus Entry and Infection. *Curr. Opin. Virol.* 4, 8–14. doi:10.1016/j.coviro.2013.10.002

Kim, S., and Choi, J. (2014). An SVM-Based High-Quality Article Classifier for Systematic Reviews. *J. Biomed. Inform.* 47, 153–159. doi:10.1016/j.jbi.2013.10.005

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., et al. (2019). OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs. *Nucleic Acids Res.* 47, D807–D811. doi:10.1093/nar/gky1053

Lasswitz, L., Chandra, N., Arnberg, N., and Gerold, G. (2018). Glycomics and Proteomics Approaches to Investigate Early Adenovirus-Host Cell Interactions. *J. Mol. Biol.* 430, 1863–1882. doi:10.1016/j.jmb.2018.04.039

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* 18, 559–563.

Li, J., Li, C., Han, J., Zhang, C., Shang, D., Yao, Q., et al. (2014). The Detection of Risk Pathways, Regulated by miRNAs, via the Integration of Sample-Matched miRNA-mRNA Profiles and Pathway Structure. *J. Biomed. Inform.* 49, 187–197. doi:10.1016/j.jbi.2014.02.004

Li, R., Ying, B., Liu, Y., Spencer, J. F., Miao, J., Tollefson, A. E., et al. (2020). Generation and Characterization of an Il2rg Knockout Syrian Hamster Model for XSCID and HAdV-C6 Infection in Immunocompromised Patients. *Dis. Models Mech.* 13. doi:10.1242/dmm.044602

Lyle, C., and Mccormick, F. (2010). Integrin αvβ5 Is a Primary Receptor for Adenovirus in CAR-Negative Cells. *Virol. J.* 7, 1–13. doi:10.1186/1743-422x-7-148

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016). Linking Virus Genomes with Host Taxonomy. *Viruses.* 8, 66. doi:10.3390/v8030066

Nestić, D., Uil, T. G., Ma, J., Roy, S., Vellinga, J., Baker, A. H., Custers, J., and Majhen, D. (2019). αvβ3 Integrin Is Required for Efficient Infection of Epithelial Cells with Human Adenovirus Type 26. *Journal of virology*. 93 1, e01474-18. doi:10.1128/JVI.01474-18

Pauly, M., Akoua-Koffi, C., Buchwald, N., Schubert, G., Weiss, S., Couacy-Hymann, E., et al. (2015). Adenovirus in Rural Côte D`Ivoire: High Diversity and Cross-Species Detection. *Ecohealth.* 12, 441–452. doi:10.1007/s10393-015-1032-5

Piccolo, P., Vetrini, F., Mithbaokar, P., Grove, N. C., Bertin, T., Palmer, D., et al. (2013). SR-A and SREC-I Are Kupffer and Endothelial Cell Receptors for Helper-dependent Adenoviral Vectors. *Mol. Ther.* 21, 767–774. doi:10.1038/mt.2012.287

Polikar, R. (2006). Ensemble Based Systems in Decision Making. *IEEE Circuits Syst. Mag.* 6, 21–45. doi:10.1109/mcas.2006.1688199

Radivojac, P., Chawla, N. V., Dunker, A. K., and Obradovic, Z. (2004). Classification and Knowledge Discovery in Protein Databases. *J. Biomed. Inform.* 37, 224–239. doi:10.1016/j.jbi.2004.07.008

Rokach, L. (2010). Ensemble-based Classifiers. *Artif. Intell. Rev.* 33, 1–39. doi:10.1007/s10462-009-9124-7

Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrott, R. H., and Ward, T. G. (1953). Isolation of a Cytopathogenic Agent from Human Adenoids Undergoing Spontaneous Degeneration in Tissue Culture. *Exp. Biol. Med.* 84, 570–573. doi:10.3181/00379727-84-20714

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting Protein-Protein Interactions Based Only on Sequences Information. *Proc. Natl. Acad. Sci.* 104, 4337–4341. doi:10.1073/pnas.0607879104

Short, J. J., Vasu, C., Holterman, M. J., Curiel, D. T., and Pereboev, A. (2006). Members of Adenovirus Species B Utilize CD80 and CD86 as Cellular Attachment Receptors. *Virus. Res.* 122, 144–153. doi:10.1016/j.virusres.2006.07.009

Singh, S., Kumar, R., and Agrawal, B. (2019). Adenoviral Vector-Based Vaccines and Gene Therapies: Current Status and Future Prospects, in *Adenoviruses*, London: IntechOpen, 53–91.

Stasiak, A. C., and Stehle, T. (2020). Human Adenovirus Binding to Host Cell Receptors: a Structural View. *Med. Microbiol. Immunol.* 209, 325–333. doi:10.1007/s00430-019-00645-2

Stichling, N., Suomalainen, M., Flatt, J. W., Schmid, M., Pacesa, M., Hemmi, S., et al. (2018). Lung Macrophage Scavenger Receptor SR-A6 (MARCO) Is an Adenovirus Type-specific Virus Entry Receptor. *Plos Pathog.* 14, e1006914. doi:10.1371/journal.ppat.1006914

Taft, L. M., Evans, R. S., Shyu, C. R., Egger, M. J., Chawla, N., Mitchell, J. A., et al. (2009). Countering Imbalanced Datasets to Improve Adverse Drug Event Predictive Models in Labor and Delivery. *J. Biomed. Inform.* 42, 356–364. doi:10.1016/j.jbi.2008.09.001

The Uniprot, C. (2017). UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi:10.1093/nar/gkw1099

Tomko, R. P., Xu, R., and Philipson, L. (1997). HCAR and MCAR: the Human and Mouse Cellular Receptors for Subgroup C Adenoviruses and Group B Coxsackieviruses. *Proc. Natl. Acad. Sci.* 94, 3352–3356. doi:10.1073/pnas.94.7.3352

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., et al. (2020). PHI-base: the Pathogen-Host Interactions Database. *Nucleic Acids Res.* 48, D613–D620. doi:10.1093/nar/gkz904

Wang, H., Li, Z.-Y., Liu, Y., Persson, J., Beyer, I., Möller, T., et al. (2011). Desmoglein 2 Is a Receptor for Adenovirus Serotypes 3, 7, 11 and 14. *Nat. Med.* 17, 96–104. doi:10.1038/nm.2270

Yang, X., Yang, S., Li, Q., Wuchty, S., and Zhang, Z. (2020). Prediction of Human-Virus Protein-Protein Interactions through a Sequence Embedding-Based Machine Learning Method. *Comput. Struct. Biotechnol. J.* 18, 153–161. doi:10.1016/j.csbj.2019.12.005

Zhang, Y., and Bergelson, J. M. (2005). Adenovirus Receptors. *Jvi* 79, 12125–12131. doi:10.1128/jvi.79.19.12125-12131.2005

Zhou, X., Park, B., Choi, D., and Han, K. (2018). A Generalized Approach to Predicting Protein-Protein Interactions between Virus and Host. *BMC Genomics*. 19, 568. doi:10.1186/s12864-018-4924-2

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

ORIGINAL RESEARCH
published: 25 May 2021
doi: 10.3389/fmolb.2021.647915

Protein Docking Model Evaluation by Graph Neural Networks

Xiao Wang[1], *Sean T. Flannery*[1] *and Daisuke Kihara*[1,2]*

[1]*Department of Computer Science, Purdue University, West Lafayette, IN, United States,* [2]*Department of Biological Sciences, Purdue University, West Lafayette, IN, United States*

Physical interactions of proteins play key functional roles in many important cellular processes. To understand molecular mechanisms of such functions, it is crucial to determine the structure of protein complexes. To complement experimental approaches, which usually take a considerable amount of time and resources, various computational methods have been developed for predicting the structures of protein complexes. In computational modeling, one of the challenges is to identify near-native structures from a large pool of generated models. Here, we developed a deep learning–based approach named Graph Neural Network–based DOcking decoy eValuation scorE (GNN-DOVE). To evaluate a protein docking model, GNN-DOVE extracts the interface area and represents it as a graph. The chemical properties of atoms and the inter-atom distances are used as features of nodes and edges in the graph, respectively. GNN-DOVE was trained, validated, and tested on docking models in the Dockground database and further tested on a combined dataset of Dockground and ZDOCK benchmark as well as a CAPRI scoring dataset. GNN-DOVE performed better than existing methods, including DOVE, which is our previous development that uses a convolutional neural network on voxelized structure models.

Keywords: protein docking, docking model evaluation, graph neural networks, deep learning, protein structure prediction

OPEN ACCESS

Edited by:
Elif Ozkirimli,
Roche, Switzerland

Reviewed by:
Manon Réau,
Utrecht University, Netherlands
Ilya Vakser,
University of Kansas, United States
Martin Zacharias,
Technical University of Munich,
Germany

****Correspondence:***
Daisuke Kihara
dkihara@purdue.edu

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *30 December 2020*
Accepted: *26 April 2021*
Published: *25 May 2021*

Citation:
Wang X, Flannery ST and Kihara D
(2021) Protein Docking Model
Evaluation by Graph Neural Networks.
Front. Mol. Biosci. 8:647915.
doi: 10.3389/fmolb.2021.647915

INTRODUCTION

Experimentally determined protein structures provide fundamental information about the physicochemical nature of the biological function of protein complexes. With the recent advances in cryo-electron microscopy, the number of experimentally determined protein complex structures has been increasing rapidly. However, experimental methods are costly in terms of money and time. To aid the experimental efforts, computational modeling approaches for protein complex structures, often referred to as protein docking (Aderinwale et al., 2020), have been extensively studied over the past two decades.

Protein docking methods aim to build the overall quaternary structure of a protein complex from the tertiary structure information of individual chains. Similar to other protein structure modeling methods, protein docking can also be divided into two main categories: template-based methods (Tuncbag et al., 2011; Anishchenko et al., 2015), which use a known structure as a scaffold of modeling, and *ab initio* methods, which assemble individual structures and score generated models to choose most plausible ones. In *ab initio* methods, various approaches were used for molecular structure representations (Venkatraman et al., 2009; Pierce et al., 2011). These include docking conformational searches, such as fast Fourier transform (Katchalski-Katzir et al., 1992; Padhorny et al., 2016), geometric hashing (Fischer et al., 1995; Venkatraman et al., 2009), and particle swarm optimization (Moal and Bates, 2010), as well as considering protein flexibility (Gray et al., 2003; Oliwa and Shen, 2015). The development of new methods aims to extend and surpass the capabilities of simple pairwise docking, such as multichain docking (Schneidman-Duhovny et al., 2005; Esquivel-Rodríguez et al., 2012; Ritchie and Grudinin, 2016), peptide–protein docking (Kurcinski et al., 2015; Alam et al., 2017; Kurcinski et al., 2020), docking with disordered proteins (Peterson et al., 2017), docking order prediction (Peterson et al., 2018a; Peterson et al., 2018b), and docking for cryo-EM maps (Esquivel-Rodríguez and Kihara, 2012; van Zundert et al., 2015). Researchers have also applied recent advances in deep learning to further boost docking performance (Akbal-Delibas et al., 2016; Degiacomi, 2019; Gainza et al., 2020).

Although substantial improvements have been made in *ab initio* protein docking, selecting near-native (i.e., correct) models out of a large number of produced models, which are often called decoys, is still challenging. The difficulty is partly due to a substantial imbalance in the number of near-native models and incorrect decoys in a generated decoy pool. The accuracy of scoring decoys certainly determines the overall performance of protein docking, and thus, there is active development of scoring functions (Moal et al., 2013) for docking models. Recognizing the importance of scoring, the Critical Assessment of PRediction of Interactions (CAPRI) (Lensink et al., 2018), which is the community-based protein docking prediction experiment, has arranged a specific category of evaluating scoring methods, where participants are asked to select 10 plausible decoys from thousands of decoys provided by the organizers. Over the last two decades, various approaches have been developed for scoring decoys. The main categories include physics-based potentials (Akbal-Delibas et al., 2016; Degiacomi, 2019; Gainza et al., 2020), scoring based on interface shape (Akbal-Delibas et al., 2016; Kingsley et al., 2016; Degiacomi, 2019; Gainza et al., 2020), knowledge-based statistical potentials (Lu et al., 2003; Huang and Zou, 2008), machine learning methods (Fink et al., 2011), evolutionary profiles of interface residues (Nadaradjane et al., 2018), and deep learning methods using interface structures (Wang et al., 2019).

In our previous work, we developed a model selection method for protein docking, that is, DOVE (Wang et al., 2019), which uses a convolutional deep neural network (CNN) as the core of its architecture. DOVE captures atoms and interaction energies of atoms located at the interface of a docking model using a cube of 20^3 or 40^3 Å^3 and judges if the model is correct or incorrect according to the CAPRI criteria (Janin et al., 2003). We showed that DOVE performed better than existing methods. However, DOVE has a critical limitation—since it captures an interface with a fixed-size cube, only a part of the interface is captured when the interface region is too large. This often caused an erroneous prediction. In addition, a 3D grid representation of an interface often includes voxels of void space where no atoms exist inside, which is not efficient in memory usage and may even be detrimental for accurate prediction. In this work, we address this limitation of DOVE by applying a graph neural network (GNN) (Scarselli et al., 2008; Wu et al., 2020), which has previously been successful in representing molecular properties (Duvenaud et al., 2015; Smith et al., 2017; Lim et al., 2019; Zubatyuk et al., 2019). Using a GNN allows all atoms at an interface of any size to be captured in a more flexible manner. The GNN representation of the interface also is rotationally invariant, meaning arbitrary rotations of a candidate model are accounted for when training and predicting docking scores. To the best of our knowledge, this is the first method that applies GNNs to the protein docking problem. Compared to DOVE and other existing methods, GNN-DOVE demonstrated substantial improvement in a benchmark study.

MATERIALS AND METHODS

We first introduce the datasets used for training and testing GNN-DOVE. Subsequently, we introduce the graph neural network architecture and the training process of GNN-DOVE.

Docking Decoy Datasets

To train and test GNN-DOVE, we first used the Dockground dataset 1.0 (available at http://dockground.compbio.ku.edu/downloads/unbound/decoy/decoys1.0.zip) (Liu et al., 2008). Docking decoys in this dataset were built by Gramm-X (Tovchigrechko and Vakser, 2005). The dataset includes 58 target complexes, each with averages of 9.83 correct and 98.5 incorrect decoys. A decoy was considered as correct following the CAPRI criteria (Lensink et al., 2018), which consider interface root mean square deviation (iRMSD), ligand RMSD (lRMSD), and the fraction of native contacts (fnat). The iRMSD is the Cα RMSD of interface residues with respect to the native structure. Interface residues in a complex are defined as all the residues within 10.0 Å from any residues of the other subunit. lRMSD is the Cα RMSD of ligands when receptors are superimposed, and fnat is the fraction of contacting residue pairs, that is, residue pairs with any heavy atom pairs within 5.0 Å, that exist in the native structure.

To remove redundancy, we grouped the 58 complexes using sequence alignment and TM-align (Zhang and Skolnick, 2004). Two complexes were assigned to the same group if at least one pair of proteins from the two complexes had a TM-score of over 0.5 and sequence identity of 30% or higher. This resulted in 29 groups (**Table 1**). In **Table 1**, complexes (PDB IDs) of the same group are shown in lower case in a parenthesis followed by the PDB ID of the representative. These groups were split into four subgroups to perform four-fold cross-validation, where three subsets were used for training, while one testing subset was used for testing the accuracy of the model. Thus, by cross-validation, we have four models tested on four independent testing sets. Among the training set, we used 80% of the complexes (i.e., unique dimers) for training a model and the remaining 20% of the complexes as a validation set, which was used to determine the best hyper-parameter set for training. In the results, the accuracy of targets when treated in the testing set was reported. To have a fair comparison with DOVE (Wang et al., 2019), DOVE was also newly trained and tested using this protocol.

Subsequently, we further trained and validated the GNN-DOVE network with a combined dataset of Dockground (ver 1.0) and ZDOCK (ver 4.0) (Hwang et al., 2010), which includes 58

TABLE 1 | Dockground dataset splits for training and testing GNN.

Fold	PDB ID
1	1A2K, 1E96 (1he1, 1he8, 1wq1), 1F6M, 1MA9 (2btf), 1G20, 1KU6, 1T6G, 1UGH, 1YVB, 2CKH, 3PRO
2	1AKJ (1p7q, 2bnq), 1DFJ, 1NBF (1r4m, 1xd3, 2bkr), 1GPW, 1HXY, 1U7F, 1UEX, 1ZY8, 2GOO, 1EWY
3	1AVW (1bth, 1bui, 1cho, 1ezu, 1ook, 1oph, 1ppf, 1tx6, 1xx9, 2fi4, 2kai, 1r0r, 2sni, 3sic)
4	1BVN (1tmq), 1F51, 1FM9, 1A2Y (1g6v, 1gpq, 1jps, 1wej, 1l9b, 1s6v), 1W1I, 2A5T, 3FAP

There are in total 29 representative targets shown in the upper case; targets in the lower case in a parenthesis indicate that they belong to the same group.

target complexes from Dockground and 120 target complexes from ZDOCK. ZDOCK has 110 more targets, but they were discarded because either GOAP (Zhou and Skolnick, 2011) or ITScore (Huang and Zou, 2008) failed to process them, or fnat could not be computed due to inconsistency of the sequence in the structures provided in the ZDOCK dataset from the native complex structure in PDB. The same criteria mentioned above were used to group the targets into 71 groups. Among them, we used 45 groups for training, 11 groups for validation, and 15 groups (19 complexes) for testing. Since a decoy set for each target in ZDOCK is much larger (around 54,000) than Dockground, we reduced the number of ZDOCK decoys for a target to 400. Up to 200 correct decoys (i.e., decoys with an acceptable or higher CAPRI quality) were selected if available, including at most 50 high-quality decoys, at most 50 medium-quality decoys, and the rest were selected from acceptable quality decoys. Then, the remaining 400 decoys were filled with negative decoys. One-third of negative decoys were selected from those with an iRMSD less than 7 Å, another third came from those with an iRMSD between 7 and 10 Å, and the rest came from those with ones with an iRMSD over 10 Å.

Finally, we tested GNN-DOVE on decoy sets of 13 targets in the CAPRI Score_set (Lensink and Wodak, 2014), which consists of 13 scoring targets from the CAPRI round 13 to round 26 (Janin, 2010; Janin, 2013). Each decoy set included 500 to 2,000 models generated using different methods by CAPRI participants.

The GNN-DOVE Algorithm

In this section, we describe GNN-DOVE, which uses the graph neural network. The GNN-DOVE algorithm is inspired by a recent work in drug–target interactions (Lim et al., 2019), which designed a two-graph representation for capturing intermolecular interactions for protein–ligand interactions. We will first explain how the 3D structural information of a protein–complex interface is embedded as a graph. Then, we describe how we used a graph attention mechanism to focus on the intermolecular interaction between a receptor and a ligand protein. The overall protocol is illustrated in **Figure 1**. For an input protein docking decoy, the interface region is identified as a set of residues located within 10.0 Å of any residues of the other protein. A residue–residue distance is defined as the shortest distance among any heavy atom pairs across the two residues. Using the extracted interface region, two graphs are built representing two types of interactions: the graph G^1 describes heavy atoms at the interface region, which only considers the covalent bonds between atoms of interface residues within each subunit as edges. Another graph G^2 connects both covalent (thus includes G^1) and non-covalent residue interaction as edges, where a non-covalent atom pair is defined as those which are closer than 10.0 Å of each other. Both graphs will be processed by a graph neural network (GNN) to output a score, which is a probability that the docking decoy has a CAPRI acceptable quality (thus making higher scores better).

Building Graphs

A key feature of this work is the graph representation of an interface region of a complex model. Graph G is defined by $G = (V, E, \text{and } A)$, where V denotes the node set, E is a set of edges, and A is the adjacency matrix, which numerically represents the connectivity of the graph. For a graph G with N nodes, the adjacency matrix A has a dimension of N*N, where $A_{ij} > 0$ if the i-th node and the j-th node are connected, and $A_{ij} = 0$ otherwise. The adjacency matrix A^1 for graph G^1 describes covalent bonds at the interface and thus defined as follows:

$$A^1_{ij} = \begin{cases} 1 & \textit{if atom i and atom j are connected by a covalent bond or if } i = j \\ 0 & \textit{otherwise} \end{cases}. \tag{1}$$

The matrix A^2 for G^2 describes both covalent bonds and non-covalent interactions between atoms within 10.0 Å to each other. It is defined as follows:

$$A^2_{ij} = \begin{cases} A^1_{ij}, & \textit{if } i, j \in \textit{receptor or } i, j \in \textit{ligand} \\ e^{\frac{-\left(d_{ij}-\mu\right)^2}{\sigma}}, & \textit{if } d_{ij} \le 10 \text{ Å and i} \in \text{receptor and j} \in \text{ligand;} \\ & \textit{or if } d_{ij} \le 10 \text{ Å and j} \in \text{receptor and i} \in \text{ligand} \\ 0, & \textit{otherwise} \end{cases} \tag{2}$$

where d_{ij} denotes the distance between the i-th and the j-th atoms. μ and σ are learnable parameters, whose initial values are 0.0 and 1.0, respectively. The formula $e^{-(d_{ij}-\mu)^2/\sigma}$ decays as the distance increases between atoms.

Compared to the previous voxel representation used in DOVE, the graph representation encodes the distance information more flexibly and naturally. Note that the representation is rotationally invariant and any size of interaction regions can be taken into analysis. Also, memory usage is more efficient as void spaces are not represented as is needed for the voxel representation.

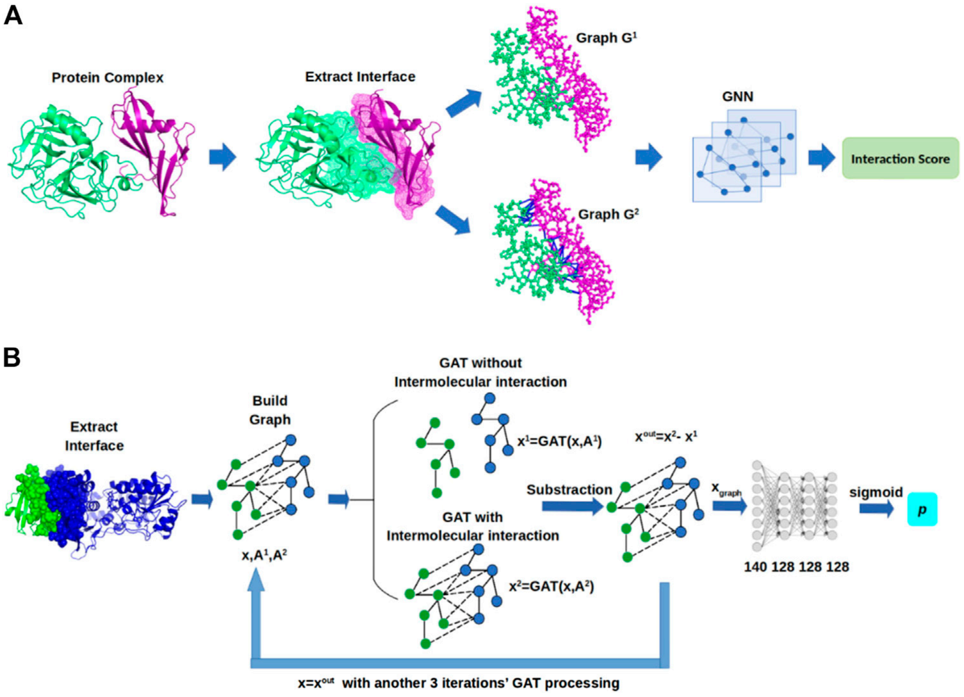


FIGURE 1 | Framework of GNN-DOVE. GNN-DOVE extracts the interface region of protein complex and further reconstructs graph with/without intermolecular interactions as input, then outputs the probability that indicates if the input structure is acceptable or not. **(A)** Overall logical steps of the pipeline. **(B)** Architecture of the GNN network with the gated graph attention mechanism.

TABLE 2 | Atom features.

Features	Representation
Atom type	C, N, O, S, H (one hot)
The degree (connections) of atom	0, 1, 2, 3, 4, 5 (one hot)
The number of connected hydrogen atoms	0, 1, 2, 3, 4 (one hot)
The number of implicit valence electrons	0, 1, 2, 3, 4, 5 (one hot)
Aromatic	0 or 1

As for the node features in the graph, we considered the physicochemical properties of atoms. We used the same features as used in previous works (Lim et al., 2019; Torng and Altman, 2019) as shown in **Table 2**. Thus, the length of a feature vector of a node from **Table 2** was 23 (=5 + 6+5 + 6+1), which was embedded by a one-layer fully connected (FC) network into 140 features.

Attention and Gate-Augmented Mechanism

The constructed graphs are used as the input to the GNN. More formally, graphs are the adjacency matrix A^1 and A^2, and the node features, $x^{in} = \{x_1^{in}, x_2^{in}, \cdots, x_N^{in}\}$ with $x \in \mathbb{R}^F$, where F is the dimension of the node feature.

We first explain the attention mechanism of our GNN. With the input graph of x^{in}, the pure graph attention coefficient is defined in **Eq. 3**, which denotes the relative importance between the i-th and the j-th node:

$$e_{ij} = x_i^{'T} E x_j^{'} + x_j^{'T} E x_i^{'}, \qquad (3)$$

where $x_i^{'}$ and $x_j^{'}$ are the transformed feature representations defined by $x_i^{'} = Wx_i^{in}$ and $x_j^{'} = Wx_j^{in}$. $W, E \in \mathbb{R}^{F\times F}$ are learnable matrices in the GNN. e_{ij} and e_{ji} become identical to satisfy the symmetrical property of the graph by adding $x_i^{'T} E x_j^{'T}$ and $x_i^{'T} E x_i^{'}$. The coefficient will only be computed for i and j where $A_{ij} > 0$.

Attention coefficients will also be computed for elements in the adjacency matrices. They are formulated in the following form for the element (i, j):

$$a_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{j\in N_i} \exp\left(e_{ij}\right)} A_{ij}, \qquad (4)$$

where a_{ij} is the normalized attention coefficient for the i-th and the j-th node pair, e_{ij} is the symmetrical graph attention coefficient computed in **Eq. 3**, and N_i is the set of neighbors of the i-th node that includes interacting nodes j where $A_{ij} > 0$. The purpose of **Eq. 4** is to consider both the physical structure of the interaction, A_{ij}, and the normalized attention coefficient, e_{ij}, to define the attention.

Based on the attention mechanism, the new node feature of each node is updated by considering its neighboring nodes, which is a linear combination of the neighboring node features with the final attention coefficient a_{ij}:

$$x_i'' = \sum_{j\in N_i} a_{ij} x_j'. \tag{5}$$

Furthermore, the gate mechanism is further applied to update the node feature since it is known to significantly boost the performance of GNN (Zhang et al., 2018). The basic idea is similar to that of ResNet (He et al., 2016), where the residual connection from the input helps to avoid information loss, alleviating the gradient collapse problem of the conventional backpropagation. The gated graph attention can be viewed as a linear combination of x_i and x_i'', as defined in **Eq. 6**:

$$x_i^{out} = c_i x_i + (1 - c_i) x_i'', \tag{6}$$

where $c_i = \sigma[D(x_i||x_i'') + b]$, $D \in \mathbb{R}^{2F}$ is a weight vector that is multiplied (dot product) with the vector $x_i||x_i''$, and b is a constant value. Both D and b are learnable parameters and are shared among different nodes. $x_i||x_i''$ denotes the concatenation vector of x_i *and* x_i''.

We refer to attention and gate-augmented mechanism as the gate-augmented graph attention layer (GAT). Then, we can simply denote $x_i^{out} = GAT(x_i^{in}, A)$. The node embedding can be iteratively updated by GAT, which aggregates information from neighboring nodes.

Graph Neural Network Architecture of GNN-DOVE

Using the GAT mechanism described before, we adopted four layers of GAT in GNN-DOVE to process the node embedding information from neighbors and to output the updated node embedding (**Figure 1B**). For the two adjacency matrices A^1 and A^2, we used a shared GAT. The initial input of the network is atom features. With two matrices, A^1 and A^2, we have $x_1 = GAT(x^{in}, A^1)$ and $x_2 = GAT(x^{in}, A^2)$. To focus only on the intermolecular interactions within an input protein complex model, we subtracted the embedding of the two graphs as the final node embedding. By subtracting the updated embedding x_1 from x_2, we can capture the aggregation information that only comes from the intermolecular interactions with other nodes in the protein complex model. Thus, the output node feature is defined as

$$x^{out} = x^2 - x^1. \tag{7}$$

Then, the updated x^{out} will become x^{in} to iteratively augment the information through the three following GAT layers. After the node embeddings were updated by the four GAT layers, the node embedding of the whole graph was summed up as the entire graph representation, which is considered as the overall intermolecular interaction representation of the protein complex model:

$$x_{graph} = \sum_{k\in G} x_k. \tag{8}$$

Finally, FC layers were applied to x_{graph} to classify whether the protein complex model is correct or incorrect. In total, four FC layers were applied. The first layer takes 140 feature values from **Eq. 8**. The three subsequent layers have a dimension of 128. RELU activation functions were used between the FC layers, and a sigmoid function was applied for the last layer to output a probability value.

The source code of GNN-DOVE is available at https://github.com/kiharalab/GNN_DOVE.

Training Networks

Since the dataset was highly imbalanced with more incorrect decoys than acceptable ones, we balanced the training data by sampling the same number of acceptable and incorrect decoys in each batch. We sampled the same number of correct and incorrect decoys. To achieve this, a positive (i.e., correct) decoy may be sampled multiple times in one epoch of training.

For training, cross-entropy loss (Goodfellow et al., 2016) was used as the loss function, and the Adam optimizer (Kingma and Ba, 2015) was used for parameter optimization. To avoid overfitting, a dropout (Srivastava et al., 2014) of 0.3 was applied for every layer, except the last FC layer. Models were trained for 100 epochs with a batch size of 32. Weights of every layer were initialized using the Glorot uniform (Glorot and Bengio, 2010) to have a zero-centered Gaussian distribution, and bias was initialized to 0 for all layers.

First, we performed four-fold cross-validation on the Dockground dataset (**Table 1**). For fold 1, where we used the fold 1 subset as testing and the other three subsets for training and validation, 16 hyper-parameter combinations with learning rates of 0.2, 0.02, 0.002, and 0.0002 and a weight decay in Adam of 0, 1e-1, 1e-2, 1e-3, 1e-4, and 1e-5 were tested. Among these combinations, we found a learning rate of 0.002 with a weight decay of 0 achieved the highest accuracy on the validation set. We used this parameter combination throughout the other three folds in the cross-validation. The training process generally converged after approximately 30 epochs.

Next, we used the combined dataset of Dockground and ZDOCK for further training. We adopted transfer learning on this dataset by starting from the models pretrained on the Dockground dataset. The training was performed in two stages: In the first stage, nine hyper-parameter combinations with learning rates of 0.002, 0.0002, and 0.00002 and weight decay of 1e-4, 1e-5, and 0 were tested on the fold 1 model. We found that a combination of a learning rate of 0.0002 and weight decay of 0 performed the best when evaluated on its validation set. We used this hyper-parameter combination to train the fold 2, 3, and 4 models and selected the fold 1 model for further training because it showed the highest accuracy on the validation set. In the second stage, we used a smaller learning rate of 0.00002 and weight decay 0 to further fine tune the fold 1 model for another 30 epochs. The resulting model was evaluated on the testing set of the combined Dockground and ZDOCK dataset. Further, we applied the model to the dataset of CAPRI scoring targets.

DOVE

We compared the performance of GNN-DOVE with its predecessor, DOVE. Here, we briefly describe the DOVE algorithm. DOVE is a CNN-based method for evaluating protein docking models. It first extracts the interface region of

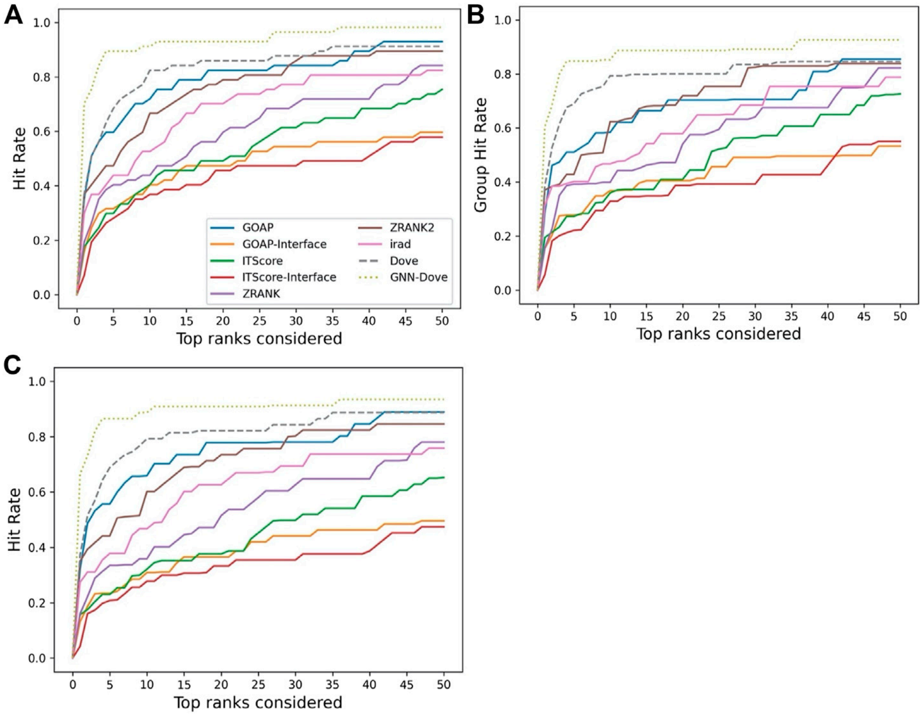


FIGURE 2 | Performance on the Dockground dataset. GNN-DOVE was compared with DOVE and seven other scoring methods. **(A)** The panel shows the fraction of target complexes among the 58 complexes in the benchmark set for which a method selected at least one acceptable model (within top *x* scored models). **(B)** Considering the complexes are grouped into 29 groups, we also compared the hit rate of different methods based on the group classification. The hit rates for complexes in each group were averaged and then re-averaged over the 29 groups. **(C)** Results when 46 complex groups were considered that were formed with interface similarity. The hit rates for complexes in each group were averaged and then re-averaged over the 46 groups.

an input protein complex model, and the region is put into a 40*40*40 Å^3 cube as input. A seven-layer CNN, which consists of three convolutional layers, two pooling layers, and two fully connected layers, was adopted to process the voxel input. The output of DOVE is the probability that indicates whether the input model is acceptable or not. For input features, DOVE took atom types as well as atom-based interaction energy values from GOAP (Zhou and Skolnick, 2011) and ITScore (Huang and Zou, 2008). Since voxelized structure input is not rotationally invariant, DOVE needed to augment training data by rotations.

RESULTS

Performance on the Dockground Dataset

We evaluated the performance of GNN-DOVE on the Dockground dataset. GNN-DOVE was compared with DOVE and five other existing structure model scoring methods, such as GOAP (Zhou and Skolnick, 2011), ITScore (Huang and Zou, 2008), ZRANK (Pierce and Weng, 2007), ZRANK2 (Pierce and Weng, 2008), and IRAD (Vreven et al., 2011). The test set results were reported for GNN-DOVE and DOVE. Both GOAP and ITScore were run in two different ways. First, as originally designed, the entire complex structure model was input. The other way was to input only the interface residues that are within 10 Å of the interacting protein (denoted as GOAP-Interface and ITScore-Interface). Thus, GNN-DOVE was compared with a total of eight methods. As for DOVE, we used a cube size of 40^3 Å^3 and heavy atom distributions as input feature because this setting performed the best among other settings tested on the Dockground dataset in the original paper (Wang et al., 2020) (**Figure 4** in the paper, the setting was named as DOVE-Atom 40). For this work, DOVE was newly retrained using the same four-fold cross-validation as GNN-DOVE.

Figure 2 shows the hit rate of GNN-DOVE in comparison with the other methods. A hit rate of a method is the fraction of target complexes where the method ranked at least one acceptable model based on the CAPRI criteria within each top rank. Targets were evaluated when they were in the heldout testing set from the four-fold cross-testing we performed. In **Figure 2**, we show three panels. Panel A shows the fraction of targets where a method had at least one hit among each rank cutoff. Panel B shows the hit rates for a method were averaged first within each of the 29 groups, and then re-averaged over the groups. Panel C shows the hit rate when targets with similar interface structures were grouped.

Figure 2 shows that GNN-DOVE (dotted line in light green) performed better than the other methods. GNN-DOVE was able to rank correct models within earlier ranks in many target complexes. Within the top 10 rank, GNN-DOVE achieved a hit rate of 89.7%, while the next best method, DOVE, achieved 81.0%, and the third best method, GOAP, obtained 70.7%

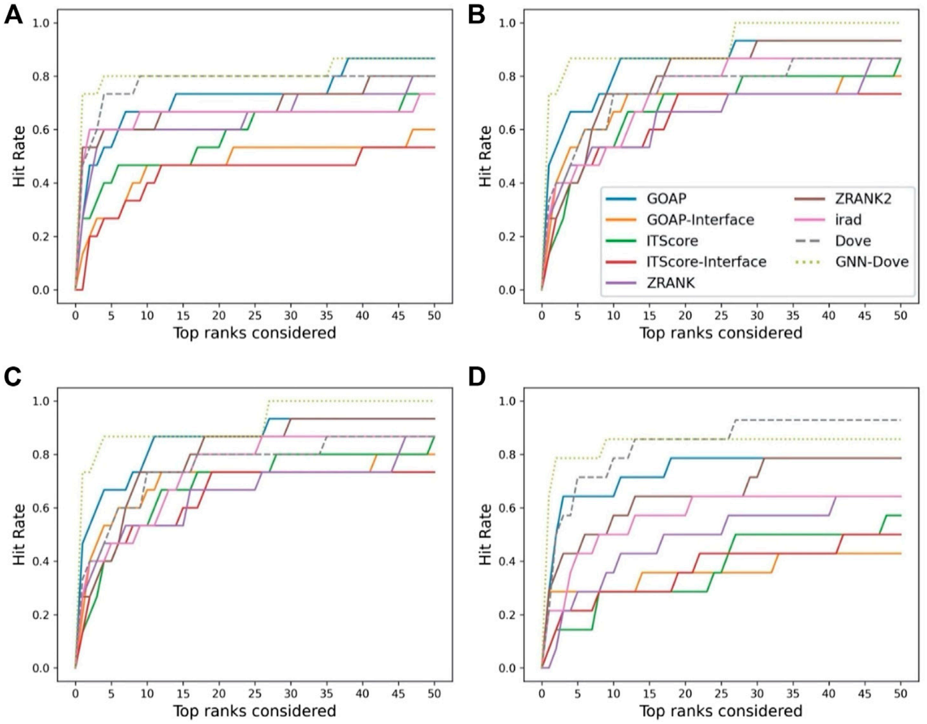


FIGURE 3 | The hit rate is shown for each fold in the cross validation on the Dockground dataset. Protein complexes in the test set of each fold are listed in **Table 1**. In the same way as **Figure 2A**, a hit rate was computed for individual complexes separately and averaged over the complexes. **(A)** The hit rate of the fold 1 test set. The model was trained on the fold 2, 3, and 4 subsets. **(B)** The fold 2 test set. **(C)** The fold 3 test set. **(D)** The fold 4 test set.

(**Figure 2A**). When we further compared the hit rates considering the target groups (**Figure 2B**), GNN-DOVE consistently outperformed other methods. The gap between GNN-DOVE and DOVE against the other existing methods also increased. Among the other seven existing methods, GOAP showed the highest hit rate at 5th rank, followed by ZRANK2 in both panels, while ITScore-Interface had the lowest hit rates on this dataset. In **Figure 2C**, we evaluated the methods' performance when target complexes were grouped considering their docking interface area similarity, which was evaluated by TM-Score. For a complex, an interface was defined as residues that are closer than 10 Å to any residue of the docking partner. To run TM-align to obtain TM-Score for two interfaces, we prepared two versions of PDB files for each interface: one with residues from the receptor first followed by residues from the ligand and the other with the opposite order. Then, we computed TM-Score for four combinations of the files from the two interfaces and selected the largest TM-Score among them. A pair of interfaces was grouped if one of the computed TM-score values of the interface regions was 0.5 or higher. This process formed 46 groups. The hit rate was computed for each complex first, then averaged within each group, and finally re-averaged across 46 groups. GNN-DOVE still showed the highest hit rate among the methods compared when considering top 10 ranks.

In **Figure 3**, we show results on each test set from the four-fold cross validation. GNN-DOVE showed the highest hit rate in early ranks.

In **Figure 4**, we compared iRMSD, lRMSD, and fnat values of the methods. These metrics are used for defining the quality levels in CAPRI. The best value among the top 10 ranked decoys was plotted. For the majority of the cases (49 out of the 58 targets), GNN-DOVE selected a decoy within an iRMSD of 4 Å (one of the criteria for the acceptable quality level in CAPRI). This is in sharp contrast to the other methods (**Figure 4A**), where the iRMSD of many targets they selected were larger (worse) than GNN-DOVE. In terms of iRMSD, the second best method was DOVE, where 44 targets were within an iRMSD of 4 Å. A similar situation was observed for lRMSD. GNN-DOVE selected a decoy within an lRMSD of 10 Å (one of the criteria for the acceptable quality level in CAPRI) for 50 targets, while the second best method, DOVE, selected 45 targets within 10 Å lRMSD. In terms of fnat (larger being more accurate), GNN-DOVE only missed 5 targets in selecting at least one model with an fnat over 0.1 (one of the criteria for acceptable quality level in CAPRI). The plot shows that GNN-DOVE had a larger fnat value than the other existing methods for most of the targets, as indicated by many data points below the diagonal line.

Figure 4B compares GNN-DOVE against DOVE. In terms of iRMSD, lRMSD, and fnat, GNN-DOVE outperformed DOVE for 26 targets (22 ties), 27 targets (20 ties), and 27 targets (17 ties

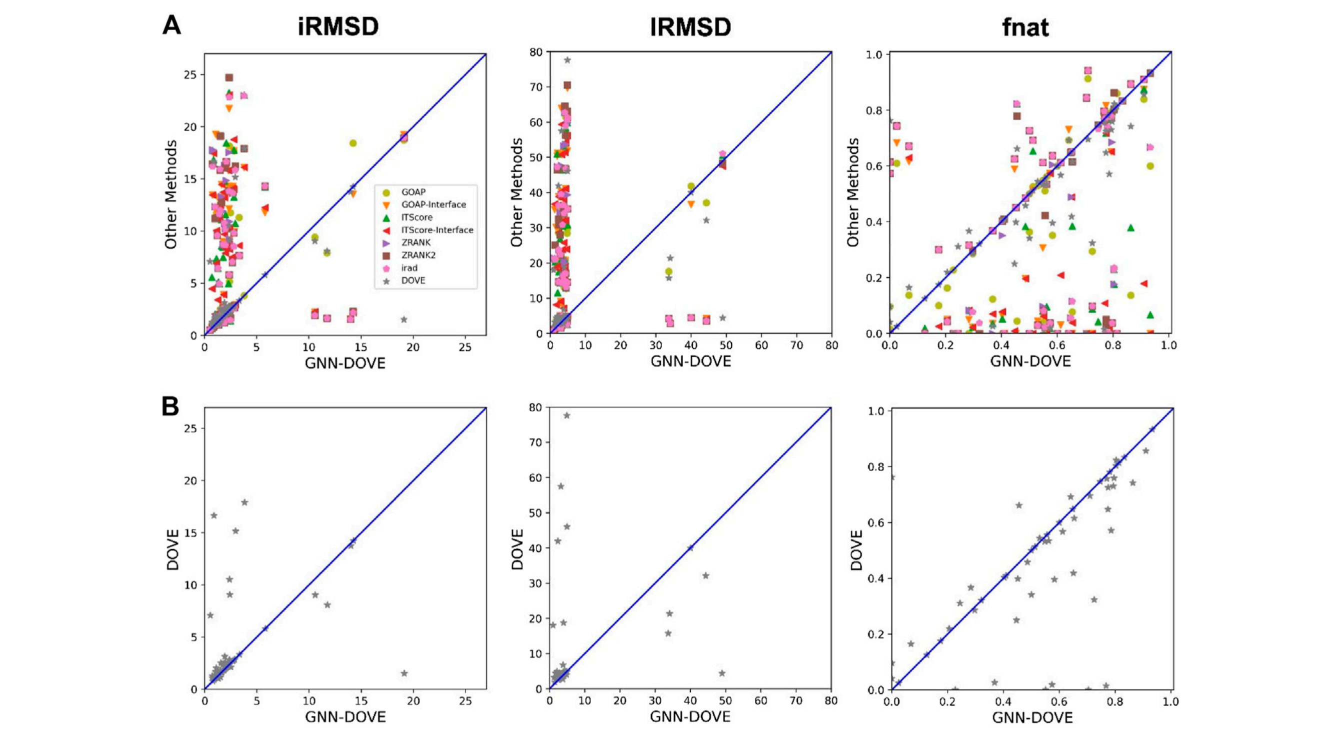


FIGURE 4 | Comparison of iRMSD, IRMSD, and fnat. For each method, the best value among the top 10 scored decoys was plotted. **(A)** Comparison against all eight methods. **(B)** Comparison against DOVE.

targets), respectively. Overall, GNN-DOVE outperformed the eight existing methods for all three metrics.

T-SNE Analysis

To illustrate how GNN-DOVE classified decoys, we used t-SNE (Maaten and Hinton, 2008) to visualize GNN-DOVE's encoding of decoys in **Figure 5**. t-SNE is a dimension-reduction method to visualize similarities of high-dimensional data points. Since we employed a four-fold cross-validation, a plot was provided for each of the four testing sets. In all the plots, particularly in Fold 3 and Fold 4, most of the acceptable decoys (black circles) were distinguished from incorrect ones (gray crosses), which indicates a good representation and generalization ability of the graph neural networks for this problem.

Examples of Decoys for Comparison With DOVE

We mentioned above that a limitation of DOVE is its usage of a fix-sized cube of 40^3 $Å^3$, which cannot capture the entire interface region if the interface is too large to fit in the cube. Here, we show two examples of such cases, which led to misclassification by DOVE but correct classification by GNN-DOVE. In **Figure 6**, the interface region of a decoy is shown in blue and green, and the atoms that did not fit in the cube are shown in a sphere representation in red.

The first example (**Figure 6A**) shows a decoy of a protein complex of plasminogen and staphylokinase (PDB ID: 1bui), which has an acceptable quality by the CAPRI criteria. For this decoy, 59 atoms (in red) out of 1,022 atoms at the interface were not included in the cube. Because of this, it was ranked the 65th out of 110 decoys by DOVE, while it was ranked 15th by GNN-DOVE. For this target, GNN-DOVE ranked five hits within the top 10 scoring decoys and eight hits within the top 20. In contrast, DOVE could not rank any hit within the top 20. The first hit by DOVE was found at the 35th rank.

The second example (**Figure 6B**) is an acceptable model for the nitrogenase complex (PDB ID: 1g20). As shown, many interface atoms, 497 out of 1,843, were outside the cube. DOVE ranked this decoy 28th, while GNN-DOVE ranked this decoy 10th. DOVE had 0 hits within the top 10 and had only one hit within top 20. On the other hand, GNN-DOVE was very successful for this target, where all the top 10 selections were correct models.

Performance on the Combined Dockground and ZDOCK Dataset

Next, we examined the performance of GNN-DOVE on the 19 complexes in the test set of the combined Dockground and ZDOCK dataset. In **Table 3**, we showed the total number of hits among top 10 ranks by GNN-DOVE and the same five other methods, that is, GOAP, ITScore, ZRANK, ZRANK2, and IRAD, as we used in **Figures 2–4**. GNN-DOVE achieved the highest hit rate of 0.842, followed by ZRANK with 0.789. GNN-DOVE ranked at the top among the methods consistently when the

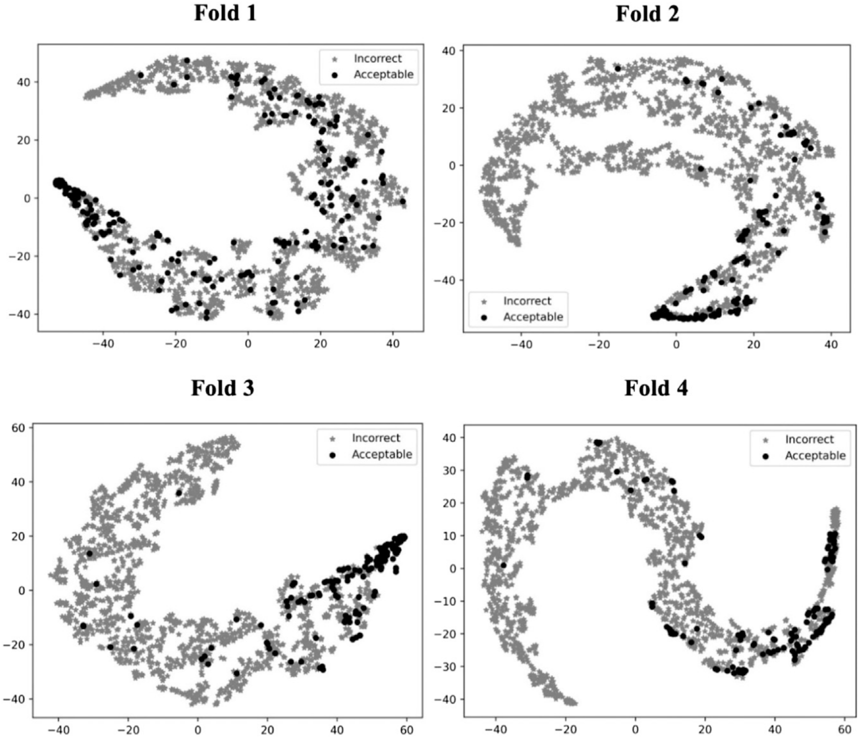


FIGURE 5 | t-SNE plots of decoy selection. Decoys from all the testing target complexes in the four different folds in the cross-testing are plotted, which in total include 580 correct decoys (black circles) and 5,591 incorrect decoys (gray stars). Encoded features of those decoys are taken from the output of the last fully connected layer of GNN, which is a vector of 128 elements. To visualize the different embedding, we use t-SNE to project them into a 2D space. The four panels correspond to the embedding of models on the four-fold testing sets.

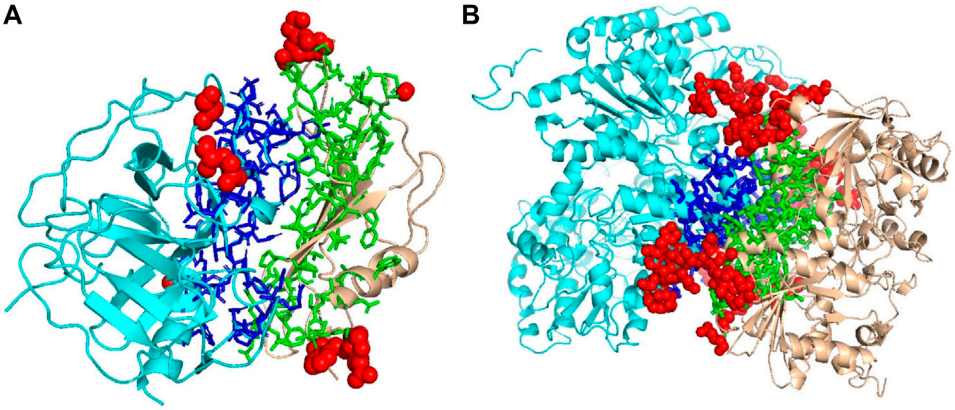


FIGURE 6 | Examples of decoys with an acceptable quality but not selected within the top 10 by DOVE. Two subunits docked are shown in cyan and light brown, and the interface regions of the two subunits are presented in the stick representation and in blue and green, respectively. To highlight the missed atoms from the input cube of DOVE, they are shown in red spheres. **(A)** A medium-quality decoy for 1bui. iRMSD: 2.54 Å, lRMSD: 2.93 Å, fnat: 0.551. **(B)** A medium-quality decoy for 1g20. iRMSD: 2.14 Å, lRMSD: 3.86 Å, fnat: 0.453.

group hit rate was considered. We note that some of the existing methods performed perfectly for specific complexes, choosing 10 hits within the top 10. However, many methods failed to select any top hits for other target complexes. In contrast, GNN-DOVE showed the most stable performance across different complexes.

Performance on the CAPRI Scoring Dataset

Finally, we evaluate GNN-DOVE on another independent dataset, the CAPRI Score_set. This dataset was chosen to be able to compare GNN-DOVE on a larger number of existing methods which participated in the corresponding CAPRI rounds.

TABLE 3 | Performance on the Dockground+ZDOCK testing dataset.

ID	GNN-DOVE	GOAP	ITScore	ZRANK	ZRANK2	IRAD	Total
1AK4	1	10	1	1	7	0	179
1AY7	8	0	3	9	8	8	176
1EER	0	0	0	0	3	0	41
1GLA	5	1	0	8	4	8	165
1HCF	9	0	8	3	3	7	183
1JIW	3	0	2	0	1	2	106
1JTG	8	0	10	10	0	10	177
1KAC	7	0	5	8	2	6	183
1KTZ	0	1	0	1	3	0	77
1MAH	9	0	8	9	0	9	179
2MTA	7	0	4	9	0	9	186
2VDB	9	1	9	7	2	6	173
3D5S	7	0	10	6	1	5	156
1BUH (1)	3	8	9	6	4	9	183
1FQ1 (1)	0	0	0	0	0	0	20
1JWH (1)	6	6	7	6	2	8	171
2OZA (1)	1	0	1	0	0	0	19
1EFN (2)	1	0	0	4	3	4	130
1GCQ (2)	2	9	0	1	8	4	142
Hit rate	0.842	0.368	0.684	0.789	0.737	0.737	—
Group HR	0.867	0.333	0.717	0.833	0.767	0.767	

In the ID column, the number in a parentheses indicates which group the target belongs to. Thus, four complexes belong to the same similarity group, and the other two belong to another group. The rest of the complexes are single entry groups. Group HR indicates the group hit rate. In Group HR, the fraction of complexes within each group that have at least one hit (acceptable model) within the top 10 ranks was first computed, and then averaged across all the groups. The total column indicates the total number of acceptable docking models for a given target.

TABLE 4 | Performance on the CAPRI scoring dataset.

ID	GNN-DOVE	GOAP	ITScore	ZRANK	ZRANK2	IRAD	Total
(T29)	2/0/0	1/0/0	0/0/0	0/0/0	2/2/0	1/1/0	167/78/2
(T30)	1/0/0	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	2/0/0
T32	0/0/0	1/0/0	0/0/0	0/0/0	0/0/0	0/0/0	15/3/0
T35	1/0/0	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	3/0/0
(T37)	0/0/0	1/0/0	3/0/1	1/0/0	4/1/0	4/1/0	99/46/11
T39	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	4/3/0
(T40)	4/4/0	1/0/1	7/3/4	1/1/0	9/8/1	3/3/0	588/206/193
T41	5/0/0	4/2/2	1/1/0	4/0/0	2/0/0	3/0/0	371/120/2
T46	1/0/0	0/0/0	0/0/0	5/0/0	6/0/0	6/0/0	24/0/0
T47	9/4/5	10/0/10	2/1/0	9/5/4	9/3/5	10/2/7	611/307/278
T50	6/0/0	0/0/0	4/1/0	0/0/0	2/0/0	2/0/0	133/36/0
T53	2/2/0	7/6/0	3/0/0	1/0/0	7/3/0	4/2/0	130/17/0
(T54)	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	19/1/0
Hit	9/3/1	7/2/3	6/4/2	6/2/1	8/5/2	8/5/1	13/10/5
Hit-NR	6/2/1	4/2/2	4/3/0	4/1/1	5/2/1	5/2/1	8/7/2

The IDs in parentheses are those which have structure or sequence similarity to one of the complexes used in training. Results for a complex by a method have three numbers separated by /. The first number is the number of decoys selected within the top 10 ranked models, which has an acceptable or better quality. The second and third numbers are the number of models with medium or higher quality, and the number of high-quality models. The numbers in the total column indicate the total number of decoys in the three quality classifications in the decoy set of each target. The last two rows report the summary of the performance. Three numbers are the number of targets where the method identified at least one acceptable or higher-quality models, at least one medium- or higher-quality models, or at least one high-quality model, respectively. The hit row lists the results when all 13 targets were considered. Hit-NR only considers targets that are not in parentheses.

In **Table 4**, we show detailed results of GNN-DOVE and the other five methods for each target. For each method, the number of decoys within the quality categories of acceptable, medium, and high (in this order) of the top 10 models are listed.

GNN-DOVE had hits for the largest number of targets, that is, nine, when decoys of acceptable or higher quality were considered. When decoys in a medium or higher quality were considered, ITScore, ZRANK2, and IRAD had hits for five targets, while GNN-DOVE had hits for three targets. It is worth noting that GNN-DOVE successfully identified correct models in two difficult targets, T30 and T35, which only contained two and three acceptable models in the decoy sets, while all the other methods failed to select any correct decoys among the top 10.

TABLE 5 | Ranking of GNN-DOVE among other scorer groups on the CAPRI scoring dataset.

Group	Performance		# Submitted targets
	All	Nonredundant	
iScore	9/6/2	6/5/1	13 (8)
GNN-DOVE	9/3/1	6/2/1	13 (8)
GraphRank	8/4/1	5/3/1	13 (8)
Bates	8/4/1	5/2/0	10 (5)
Bonvin	8/3/2	5/2/1	9 (5)
Weng	8/2/3	5/2/1	9 (6)
Zou	7/1/4	5/1/2	9 (6)
Wang	6/3/2	4/2/1	6 (4)
Fernandez-Recio	5/3/2	4/4/1	8 (7)
Elber	5/1/1	4/1/0	5 (4)
Wolfson	4/0/1	1/0/0	5 (2)
Camacho	3/1/2	1/1/1	5 (2)

*Results of the existing methods were taken from **Table 2** of the article by Geng et al. (2020). The numbers in the nonredundant column only considered targets in **Table 4** that are not in the parentheses. The last column shows the number of targets that each group has submitted their prediction among the 13 targets listed in **Table 4**. The numbers in parentheses report the number of submitted targets among those which do not have similarity to the training set we used (i.e., discarding the targets in parentheses in **Table 4**).*

In **Table 5**, we further compared GNN-DOVE with the top groups who participated in the model scoring task for the 13 CAPRI scoring targets. The results were taken from **Table 2** of the article by Geng et al. (2020). In total, 37 scoring groups have submitted their scores during this challenge and among them we list here only groups with five or more submitted targets. In addition to the CAPRI participants the table also includes the latest protein docking evaluation approaches, iScore (Geng et al., 2020) and GraphRank (Geng et al., 2020).

GNN-DOVE tied with iScore when decoys of acceptable or higher quality were considered. When medium- or higher-quality decoys were considered, GNN-DOVE performed second to iScore. In this list, except for GNN-DOVE, iScore, and GraphRank, all the other groups were human groups, which may have used manual intervention using expert knowledge. Thus, the results show that GNN-DOVE is also highly competitive against human experts.

DISCUSSION

In this work, we developed GNN-DOVE for protein docking decoy selection, which used a graph neural network (GNN). We used the gate-augmented attention mechanism to capture the atom interaction pattern at the interface region of protein docking models. The benchmark on the Dockground dataset demonstrated that GNN-DOVE outperformed DOVE, along with other existing scoring functions compared. We further trained GNN-DOVE on a larger dataset and evaluated two more datasets, including the CAPRI Score_set, which confirmed superior performance of GNN-DOVE to existing methods.

To assess the quality of structure models, considering multi-body (atom or residue) interactions (Gniewek et al., 2011; Kim and Kihara, 2014; Kim and Kihara, 2016; Olechnovic and Venclovas, 2017) have been proven to be an effective approach. GNNs consider patterns of multiatom interactions by representing the interactions as a graph structure. Since a graph is a natural representation of molecular structures, GNNs may be applied in various problems in structural bioinformatics and cheminformatics.

The performance of GNN-DOVE likely would be improved by considering other physicochemical properties of atoms such as atom-wise binding energies, as well as sequence conservation of residues that can be computed from a multiple sequence alignment of homologous proteins. Application to multichain complexes remains a potential path for future work.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The Dockground docking dataset was downloaded from the Dockground database (http://dockground.compbio.ku.edu) at the link http://dockground.compbio.ku.edu/downloads/unbound/decoy/decoys1.0.zip. The ZDOCK dataset was downloaded from the ZDOCK decoy sets (https://zlab.umassmed.edu/zdock/decoys.shtml) at the link https://zlab.umassmed.edu/zdock/decoys_bm4_zd3.0.2_6deg.tar.gz. The CAPRI score set was downloaded from http://cb.iri.univ-lille1.fr/Users/lensink/Score_set.

AUTHOR CONTRIBUTIONS

XW and STF conceived the initial version of the study. XW and DK designed this work in the current form. XW developed the codes in communication with STF. XW performed the computation, and XW and DK analyzed the results. XW wrote the initial draft of the manuscript, and DK critically edited it. XW, STF, and DK edited the manuscript in the revision.

FUNDING

We declare that all the sources of funding received for this research have been submitted. This work was partly supported by the National Institutes of Health (R01GM133840 and R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, and DBI2003635).

ACKNOWLEDGMENTS

The authors are grateful to Jacob Verburgt for proofreading the manuscript, and Sai Raghavendra Maddhuri Venkata Subramaniya and Aashish Jain for testing the GNN-DOVE code on GitHub.

REFERENCES

Aderinwale, T., Christoffer, C. W., Sarkar, D., Alnabati, E., and Kihara, D. (2020). Computational Structure Modeling for Diverse Categories of Macromolecular Interactions. *Curr. Opin. Struct. Biol.* 64, 1–8. doi:10.1016/j.sbi.2020.05.017

Akbal-Delibas, B., Farhoodi, R., Pomplun, M., and Haspel, N. (2016). Accurate Refinement of Docked Protein Complexes Using Evolutionary Information and Deep Learning. *J. Bioinform. Comput. Biol.* 14, 1642002. doi:10.1142/s0219720016420026

Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D., and Schueler-Furman, O. (2017). High-resolution Global Peptide-Protein Docking Using Fragments-Based PIPER-FlexPepDock. *PLoS Comput. Biol.* 13, e1005905. doi:10.1371/journal.pcbi.1005905

Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., and Vakser, I. A. (2015). Structural Templates for Comparative Protein Docking. *Proteins* 83, 1563–1570. doi:10.1002/prot.24736

Degiacomi, M. T. (2019). Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* 27, 1034–1040. e1033. doi:10.1016/j.str.2019.03.018

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). *Advances in Neural Information Processing Systems*, 2224–2232.

Esquivel-Rodríguez, J., Yang, Y. D., Kihara, D., and Multi-LZerD (2012). Multiple Protein Docking for Asymmetric Complexes. *Proteins: Struct. Funct. Bioinformatics* 80, 1818–1833. doi:10.1002/prot.24079

Esquivel-Rodríguez, J., and Kihara, D. (2012). Fitting Multimeric Protein Complexes into Electron Microscopy Maps Using 3D Zernike Descriptors. *J. Phys. Chem. B.* 116, 6854–6861. doi:10.1021/jp212612t

Fink, F., Hochrein, J., Wolowski, V., Merkl, R., and Gronwald, W. (2011). PROCOS: Computational Analysis of Protein-Protein Complexes. *J. Comput. Chem.* 32, 2575–2586. doi:10.1002/jcc.21837

Fischer, D., Lin, S. L., Wolfson, H. L., and Nussinov, R. (1995). A Geometry-Based Suite of Moleculardocking Processes. *J. Mol. Biol.* 248, 459–477. doi:10.1016/s0022-2836(95)80063-8

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., et al. (2020). Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat. Methods* 17, 184–192. doi:10.1038/s41592-019-0666-6

Geng, C., Jung, Y., Renaud, N., Honavar, V., Bonvin, A. M. J. J., and Xue, L. C. (2020). iScore: a Novel Graph Kernel-Based Function for Scoring Protein-Protein Docking Models. *Bioinformatics* 36, 112–121. doi:10.1093/bioinformatics/btz496

Glorot, X., and Bengio, Y. (2010). *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Gniewek, P., Leelananda, S. P., Kolinski, A., Jernigan, R. L., and Kloczkowski, A. (2011). Multibody Coarse-Grained Potentials for Native Structure Recognition and Quality Assessment of Protein Models. *Proteins* 79, 1923–1929. doi:10.1002/prot.23015

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*. MIT press Cambridge.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., et al. (2003). Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* 331, 281–299. doi:10.1016/s0022-2836(03)00670-3

He, K., Zhang, X., Ren, S., and Sun, J. (2016). *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2016.90

Huang, S.-Y., and Zou, X. (2008). An Iterative Knowledge-Based Scoring Function for Protein-Protein Recognition. *Proteins* 72, 557–579. doi:10.1002/prot.21949

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-protein Docking Benchmark Version 4.0. *Proteins* 78, 3111–3114. doi:10.1002/prot.22830

Janin, J. l., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., et al. (2003). CAPRI: a Critical Assessment of Predicted Interactions. *Proteins* 52, 2–9. doi:10.1002/prot.10381

Janin, J. (2010). The Targets of CAPRI Rounds 13-19. *Proteins* 78, 3067–3072. doi:10.1002/prot.22774

Janin, J. (2013). The Targets of CAPRI Rounds 20-27. *Proteins* 81, 2075–2081. doi:10.1002/prot.24375

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci.* 89, 2195–2199. doi:10.1073/pnas.89.6.2195

Kim, H., and Kihara, D. (2014). Detecting Local Residue Environment Similarity for Recognizing Near-Native Structure Models. *Proteins* 82, 3255–3272. doi:10.1002/prot.24658

Kim, H., and Kihara, D. (2016). Protein Structure Prediction Using Residue- and Fragment-Environment Potentials in CASP11. *Proteins* 84 (Suppl. 1), 105–117. doi:10.1002/prot.24920

Kingma, D. P., and Ba, J. (2015). *Paper Presented at the International Conference on Learning Representations*.

Kingsley, L. J., Esquivel-Rodríguez, J., Yang, Y., Kihara, D., and Lill, M. A. (2016). Ranking Protein-Protein Docking Results Using Steered Molecular Dynamics and Potential of Mean Force Calculations. *J. Comput. Chem.* 37, 1861–1865. doi:10.1002/jcc.24412

Kurcinski, M., Badaczewska-Dawid, A., Kolinski, M., Kolinski, A., and Kmiecik, S. (2020). Flexible Docking of Peptides to Proteins Using CABS-dock. *Protein Sci.* 29, 211–222. doi:10.1002/pro.3771

Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., and Kmiecik, S. (2015). CABS-dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Nucleic Acids Res.* 43, W419–W424. doi:10.1093/nar/gkv456

Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., and Wodak, S. J. (2018). The Challenge of Modeling Protein Assemblies: the CASP12-CAPRI Experiment. *Proteins* 86, 257–273. doi:10.1002/prot.25419

Lensink, M. F., and Wodak, S. J. (2014). Score_set: a CAPRI Benchmark for Scoring Protein Complexes. *Proteins* 82, 3163–3169. doi:10.1002/prot.24678

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. (2019). Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J. Chem. Inf. Model.* 59, 3981–3988. doi:10.1021/acs.jcim.9b00387

Liu, S., Gao, Y., and Vakser, I. A. (2008). DOCKGROUND Protein-Protein Docking Decoy Set. *Bioinformatics* 24, 2634–2635. doi:10.1093/bioinformatics/btn497

Lu, H., Lu, L., and Skolnick, J. (2003). Development of Unified Statistical Potentials Describing Protein-Protein Interactions. *Biophysical J.* 84, 1895–1901. doi:10.1016/s0006-3495(03)74997-2

Maaten, L. v. d., and Hinton, G. (2008). Visualizing Data Using T-SNE. *J. Machine Learn. Res.* 9, 2579–2605.

Moal, I. H., and Bates, P. A. (2010). SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Ijms* 11, 3623–3648. doi:10.3390/ijms11103623

Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013). The Scoring of Poses in Protein-Protein Docking: Current Capabilities and Future Directions. *BMC Bioinformatics* 14, 286. doi:10.1186/1471-2105-14-286

Nadaradjane, A. A., Guerois, R., and Andreani, J. (2018). "Protein-Protein Docking Using Evolutionary Information," in *Protein Complex Assembly* (Springer), 429–447. doi:10.1007/978-1-4939-7759-8_28

Olechnovic, K., and Venclovas, C. (2017). VoroMQA: Assessment of Protein Structure Quality Using Interatomic Contact Areas. *Proteins* 85, 1131–1145. doi:10.1002/prot.25278

Oliwa, T., and Shen, Y. (2015). cNMA: a Framework of Encounter Complex-Based Normal Mode Analysis to Model Conformational Changes in Protein Interactions. *Bioinformatics* 31, i151–i160. doi:10.1093/bioinformatics/btv252

Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., et al. (2016). Protein-protein Docking by Fast Generalized Fourier Transforms on 5D Rotational Manifolds. *Proc. Natl. Acad. Sci. USA* 113, E4286–E4293. doi:10.1073/pnas.1603929113

Peterson, L. X., Shin, W.-H., Kim, H., and Kihara, D. (2018a). Improved Performance in CAPRI Round 37 Using LZerD Docking and Template-Based Modeling with Combined Scoring Functions. *Proteins* 86, 311–320. doi:10.1002/prot.25376

Peterson, L. X., Togawa, Y., Esquivel-Rodriguez, J., Terashi, G., Christoffer, C., Roy, A., et al. (2018b). Modeling the Assembly Order of Multimeric Heteroprotein Complexes. *PLoS Comput. Biol.* 14, e1005937. doi:10.1371/journal.pcbi.1005937

Peterson, L. X., Roy, A., Christoffer, C., Terashi, G., and Kihara, D. (2017). Modeling Disordered Protein Interactions from Biophysical Principles. *PLoS Comput. Biol.* 13, e1005485. doi:10.1371/journal.pcbi.1005485

Pierce, B. G., Hourai, Y., and Weng, Z. (2011). Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PloS one* 6, e24657. doi:10.1371/journal.pone.0024657

Pierce, B., and Weng, Z. (2008). A Combination of Rescoring and Refinement Significantly Improves Protein Docking Performance. *Proteins* 72, 270–279. doi:10.1002/prot.21920

Pierce, B., and Weng, Z. (2007). ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function. *Proteins* 67, 1078–1086. doi:10.1002/prot.21373

Ritchie, D. W., and Grudinin, S. (2016). Spherical Polar Fourier Assembly of Protein Complexes with Arbitrary Point Group Symmetry. *J. Appl. Cryst.* 49, 158–167. doi:10.1107/s1600576715022931

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The Graph Neural Network Model. *IEEE Trans. Neural Netw.* 20, 61–80. doi:10.1109/TNN.2008.2005605

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). Geometry-based Flexible and Symmetric Protein Docking. *Proteins: Struct. Funct. Bioinformatics* 60, 224–231. doi:10.1093/nar/gki481

Smith, J. S., Isayev, O., and Roitberg, A. E. (2017). ANI-1: an Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* 8, 3192–3203. doi:10.1039/c6sc05720a

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *J. Machine Learn. Res.* 15, 1929–1958. doi:10.5555/2627435.2670313

Torng, W., and Altman, R. B. (2019). Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* 59, 4131–4149. doi:10.1021/acs.jcim.9b00628

Tovchigrechko, A., and Vakser, I. A. (2005). Development and Testing of an Automated Approach to Protein Docking. *Proteins* 60, 296–301. doi:10.1002/prot.20573

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting Protein-Protein Interactions on a Proteome Scale by Matching Evolutionary and Structural Similarities at Interfaces Using PRISM. *Nat. Protoc.* 6, 1341–1354. doi:10.1038/nprot.2011.367

van Zundert, G. C. P., Melquiond, A. S. J., and Bonvin, A. M. J. J. (2015). Integrative Modeling of Biomolecular Complexes: HADDOCKing with Cryo-Electron Microscopy Data. *Structure* 23, 949–960. doi:10.1016/j.str.2015.03.014

Venkatraman, V., Yang, Y. D., Sael, L., and Kihara, D. (2009). Protein-protein Docking Using Region-Based 3D Zernike Descriptors. *BMC Bioinformatics* 10, 407. doi:10.1186/1471-2105-10-407

Vreven, T., Hwang, H., and Weng, Z. (2011). Integrating Atom-Based and Residue-Based Scoring Functions for Protein-Protein Docking. *Protein Sci.* 20, 1576–1586. doi:10.1002/pro.687

Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., and Kihara, D. (2019). Protein Docking Model Evaluation by 3D Deep Convolutional Neural Networks. *Bioinformatics* 36, 2113–2118. doi:10.1093/bioinformatics/btz870

Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., and Kihara, D. (2020). Protein Docking Model Evaluation by 3D Deep Convolutional Neural Networks. *Bioinformatics* 36, 2113–2118. doi:10.1093/bioinformatics/btz870

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2020). *A Comprehensive Survey on Graph Neural Networks*. IEEE Transactions on Neural Networks and Learning Systems.

Zhang, J., Shi, X., Xie, J., Ma, H., King, I., and Yeung, D.-y (2018). 34th Conference on Uncertainty in Artificial Intelligence 2018, Monterey, CA. Arlington, VA: AUAI Press.

Zhang, Y., and Skolnick, J. (2004). Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* 57, 702–710. doi:10.1002/prot.20264

Zhou, H., and Skolnick, J. (2011). GOAP: a Generalized Orientation-dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical J.* 101, 2043–2052. doi:10.1016/j.bpj.2011.09.012

Zubatyuk, R., Smith, J. S., Leszczynski, J., and Isayev, O. (2019). Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-In-Molecules Neural Network. *Sci. Adv.* 5, eaav6490. doi:10.1126/sciadv.aav6490

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

frontiers
in Molecular Biosciences

REVIEW
published: 10 June 2021
doi: 10.3389/fmolb.2021.673363

Learning the Regulatory Code of Gene Expression

Jan Zrimec[1], *Filip Buric*[1], *Mariia Kokina*[1,2], *Victor Garcia*[3] *and Aleksej Zelezniak*[1,4]*

[1]*Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden,* [2]*Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark,* [3]*School of Life Sciences and Facility Management, Zurich University of Applied Sciences, Wädenswil, Switzerland,* [4]*Science for Life Laboratory, Stockholm, Sweden*

Data-driven machine learning is the method of choice for predicting molecular phenotypes from nucleotide sequence, modeling gene expression events including protein-DNA binding, chromatin states as well as mRNA and protein levels. Deep neural networks automatically learn informative sequence representations and interpreting them enables us to improve our understanding of the regulatory code governing gene expression. Here, we review the latest developments that apply shallow or deep learning to quantify molecular phenotypes and decode the *cis*-regulatory grammar from prokaryotic and eukaryotic sequencing data. Our approach is to build from the ground up, first focusing on the initiating protein-DNA interactions, then specific coding and non-coding regions, and finally on advances that combine multiple parts of the gene and mRNA regulatory structures, achieving unprecedented performance. We thus provide a quantitative view of gene expression regulation from nucleotide sequence, concluding with an information-centric overview of the central dogma of molecular biology.

Keywords: gene expression prediction, cis-regulatory grammar, gene regulatory structure, mRNA & protein abundance, chromatin accessibility, regulatory genomics, machine learning, deep neural networks

OPEN ACCESS

Edited by:
Artur Yakimovich,
Roche, United Kingdom

Reviewed by:
James J. Cai,
Texas A&M University, United States
Chuan Ku,
Academia Sinica, Taiwan
Luigi Russo,
University of Campania Luigi Vanvitelli, Italy
Kai Song,
Tianjin University, China
Lei Yang,
South China Agricultural University, China

****Correspondence:***
Aleksej Zelezniak
aleksej.zelezniak@chalmers.se

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *27 February 2021*
Accepted: *24 May 2021*
Published: *10 June 2021*

Citation:
Zrimec J, Buric F, Kokina M, Garcia V and Zelezniak A (2021) Learning the Regulatory Code of Gene Expression. Front. Mol. Biosci. 8:673363. doi: 10.3389/fmolb.2021.673363

INTRODUCTION

Genetic information is stored and encoded in genes that produce an organism's phenotype by being expressed through multiple biochemical processes into a variety of functional molecules. The central dogma of molecular biology states that genetic information flows from DNA to the phenotypically relevant proteins of an organism in a unidirectional, two-step process: the transcription of DNA into messenger RNA (mRNA) is followed by translation of mRNA into protein (Watson et al., 2008).

Abbreviations: AA, amino acid; AUC, area under the receiver operating characteristic curve; AUPRC, area under the precision recall curve; biLSTM, bidirectional long short-term memory; BunDLE-seq, binding to designed library, extracting, and sequencing; CAI, codon adaptation index; ChIP-seq, chromatin immunoprecipitation sequencing; CNN, convolutional neural network; CUB, codon usage bias; DBP, DNA-binding protein; DHS, DNase I hypersensitive site; DNase-seq, DNase I hypersensitive sites sequencing; DNN, deep neural network; dsDNA, double-stranded DNA; GM, geometric mean of precision and recall; HTS, high-throughput sequencing (technology); LR, linear regression; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient; ML, machine learning; mRNA, messenger RNA; NN, feedforward neural network; NuSAR, nucleotide sequence activity relationships; OLS, ordinary least squares (regression); ORF, open reading frame; PBM, protein binding microarray; PIC, (eukaryotic) preinitiation complex; PLS, partial least squares (regression); pre-mRNA, precursor mRNA; PTM, post-translational modifications; PWM, position weight matrix; RBP, RNA-binding protein; RBS, ribosome binding site; RF, random forest; RNAP, RNA polymerase; RNN, recurrent neural network; SELEX, systematic evolution of ligands by exponential enrichment; SVM, support vector machine; SNV, single nucleotide variant; ssDNA, single-stranded DNA; TF, transcription factors; TFBS, TF binding site; TIF, transcription initiation frequency; TSS, transcription start site; UTR, untranslated region.

From these molecular phenotypes, further post-translational processing and cellular metabolism shape and define the observable phenotype of the organism (Nielsen, 2017). Some of the most important processes involved in gene expression are regulated at the nucleotide sequence level, spanning the coding and non-coding regulatory regions adjacent to the gene (Watson et al., 2008; Zrimec et al., 2020). For over a decade, a key trend in the field has thus been to develop computational methods that can process nucleotide sequences and interpret the regulatory code within them, to better understand gene expression and improve quantitative predictions (Segal and Widom, 2009; Levo and Segal, 2014; Li et al., 2019a). These developments are not only important for advancing molecular biology, but have practical implications as well: they are crucial for solving problems related to human disease (Lee and Young, 2013; Zhou et al., 2018a) as well as biotechnology applications (de Jongh et al., 2020).

The key interactions that govern gene expression occur among proteins and nucleic acids. Proteins search for their active binding sites by sliding and diffusion, recognizing a particular DNA site *via* physicochemical interactions with the molecule (Tafvizi et al., 2011; Hammar et al., 2012). Typical binding domains of DNA-binding proteins (DBPs), such as transcription factors (TFs) and polymerases, include helix-turn-helix and zinc finger domains (Watson et al., 2008). However, besides direct protein-DNA readout with the major groove of the DNA helix, which offers base-specific hydrogen bond donors, acceptors, and nonpolar groups that can be recognized by complementary groups on the amino acid side chain, the specificities of protein-DNA interactions are defined also by indirect readout (Rohs et al., 2010; Marcovitz and Levy, 2013; Inukai et al., 2017). This comprises "weak" protein-DNA interactions that depend on base pairs that are not directly contacted by the protein and are defined by conformational and physicochemical DNA properties at the specific binding sites or in their vicinity (Rohs et al., 2009; Yang et al., 2017; Zrimec and Lapanje, 2018). On the other hand, RNA is a single stranded molecule with a softer backbone than DNA and thus has more extensive secondary and tertiary structure. RNA-binding proteins (RBPs) recognize single or double stranded RNA, three-dimensional structural features of folded RNAs, or even bind RNA non-specifically (Re et al., 2014). In regulating translation, however, multiple conserved RNA sequence motifs have been uncovered that play a key role typically *via* single strand or secondary structure-recognition mechanisms (Watson et al., 2008; Leppek et al., 2018). Therefore, despite the apparent monomeric simplicity of nucleic acid sequences, the problem of extracting information from them is quite complex, as they encode a rich grammar of motif occurrences, combinations and sequential properties that needs to be correctly interpreted (Siggers and Gordân, 2014; Slattery et al., 2014; Li et al., 2019a; Nagy and Nagy, 2020).

In this regard, machine learning (ML) comprises a set of algorithms that are capable of mapping complex relationships between input and target variables in a supervised fashion. The resulting predictive/descriptive models can perform classification of discrete target variables or regression of continuous ones. Classical algorithms, which include (multiple) linear regression (LR), support vector machines (SVMs), tree-based methods such as random forests (RFs), and feedforward neural networks (NNs) (Hastie et al., 2013; Géron, 2019), commonly referred to as "shallow" methods, have in recent years been superseded by deep neural networks (DNNs) (LeCun et al., 2015). DNNs resolve many problems inherent to the shallow methods, such as the reliance on feature engineering and selection, but come at the cost of requiring orders of magnitude more training data and computational resources (Angermueller et al., 2016; Eraslan et al., 2019a). In the current big data era, however, this is a diminishing problem. The result is that the information in nucleotide sequences can now be deciphered at unprecedented scale and quality, elucidating the regulatory grammar and greatly expanding our understanding of the underlying processes and capacity to accurately predict the outcomes of gene expression (Zhou et al., 2018a; Eraslan et al., 2019a; Zrimec et al., 2020).

In the present review, we provide an overview of the latest published developments that apply ML to nucleotide sequence data in order to understand gene expression in the most well studied model organisms, including bacteria (*Escherichia coli*), unicellular eukaryotes (yeast, *Saccharomyces cerevisiae*) and multicellular eukaryotes (human, *Homo sapiens*). Since these organisms represent the whole spectrum of genetic regulatory complexity, with gene densities ranging from 892 (bacteria) to six (human) genes per Mbp (Zrimec et al., 2020), the knowledge and principles presented here are generally applicable to all other organisms including insects and plants (Haberle and Stark, 2018; Wang H. et al., 2020). We specifically focus on the latest developments with deep learning and compare them to the state of the art solutions with shallow methods. By reasoning from first principles, the problem of predicting gene expression levels from nucleotide sequence data is explained from the ground up by deconstructing it into the basic regulatory processes and grammatical elements. We first focus on modeling the protein-DNA interactions important for initiating transcription, which include TF binding and nucleosome positioning. We then detail the current understanding of the regulatory grammar carried within the specific coding and non-coding regulatory regions, and its involvement in defining transcript and protein abundance. Based on these principles, we review advanced modeling approaches that use multiple different parts of the gene regulatory structure or whole nucleotide sequences, demonstrating how this increases their predictive power. Finally, by considering all the results, we provide an information-centric overview of the field, and discuss the applicative potential and future outlook of the presented modeling approaches.

LEARNING THE PROTEIN-DNA INTERACTIONS INITIATING GENE EXPRESSION

One of the key regulation strategies of gene expression is at the level of transcription initiation (Watson et al., 2008), which is also the most studied and modeled regulatory mechanism (Segal and Widom, 2009; Levo and Segal, 2014). Transcription initiation is

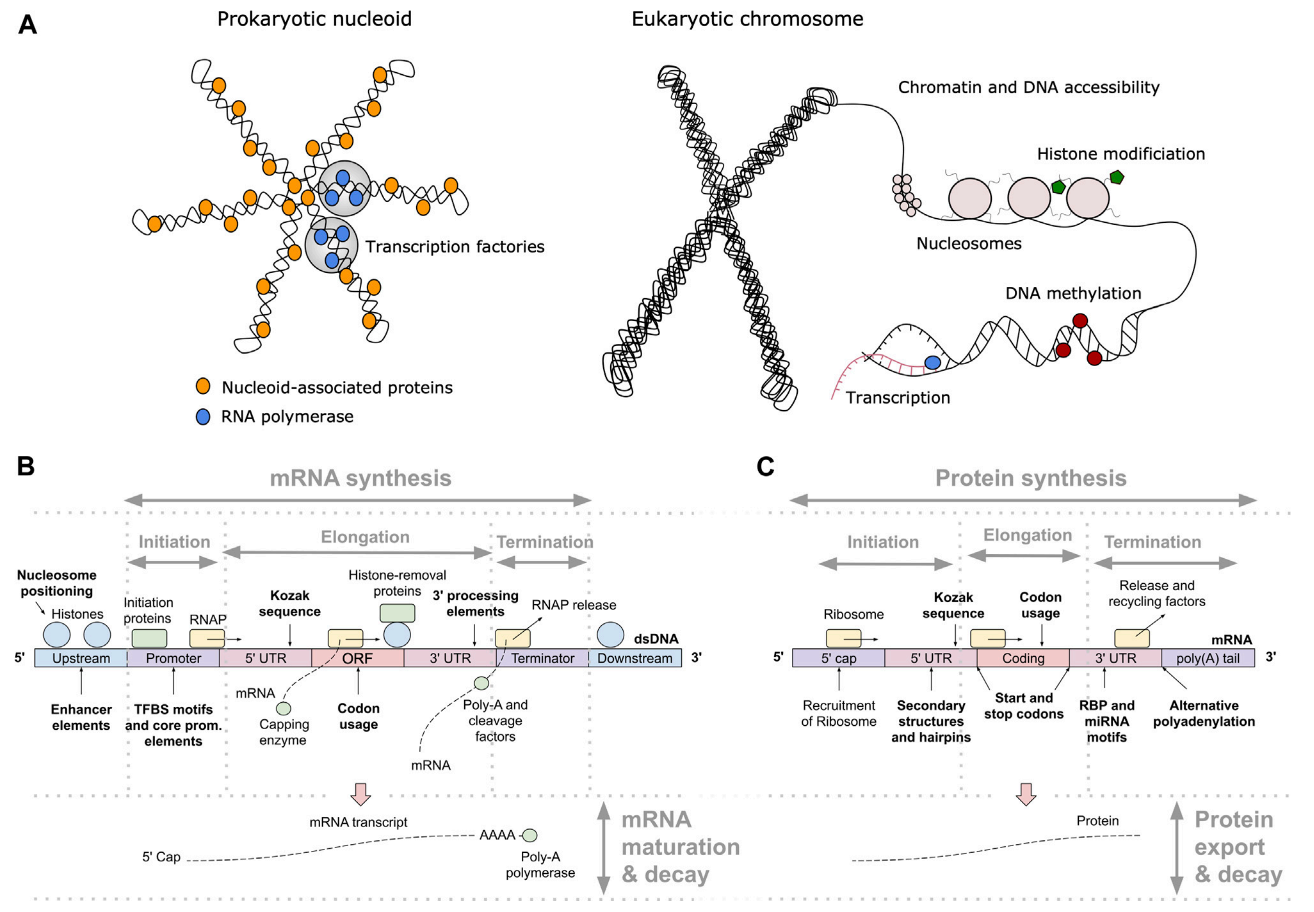


FIGURE 1 | Principles of gene expression. **(A)** Protein-DNA interactions in prokaryotic nucleoid and eukaryotic chromosome structure, epigenetics and transcription initiation. The basic repeating structural unit of chromatin is the nucleosome, which contains eight histone proteins. Bacterial nucleoid-associated proteins are the main regulators of nucleoid structure, where the circular genome is supercoiled and uncoiled by these proteins. In cells, genes are switched on and off based on the need for product in response to cellular and environmental signals. This is regulated predominantly at the level of transcription initiation, where chromatin and nucleoid structure open and close, controlling the accessibility of DNA and defining areas with high amounts of transcription (factories) upon demand. **(B)** Depiction of eukaryotic transcription across the gene regulatory structure that includes coding and non-coding regulatory regions. The open reading frame (ORF) carries the coding sequence, constructed in the process of splicing by joining introns and removing exons. Each region carries specific regulatory signals, including transcription factor binding sites (TFBS) in enhancers, core promoter elements in promoters, Kozak sequence in 5′ untranslated regions (UTRs), codon usage bias of coding regions and multiple termination signals in 3′ UTRs and terminators, which are common predictive features in ML (highlighted bold). RNAP denotes RNA polymerase, mRNA messenger RNA. **(C)** Depiction of eukaryotic translation across the mRNA regulatory structure, where initiation involves the 5′ cap, Kozak sequence and secondary structures in the 5′ UTR. Codon usage bias affects elongation, whereas RNA-binding protein (RBP) sites, microRNA (miRNA) response elements and alternative polyadenylation in the 3′ UTR affect post-translational processing and final expression levels. These regulatory elements are common predictive features in ML (highlighted bold).

a complex process involving many different interacting DNA and protein components, including: 1) activating or repressing TFs that bind 6–12 bp long TF binding sites (TFBS) in enhancer and promoter regions (Watson et al., 2008) with different binding affinities and specificities (Levo and Segal, 2014), 2) nucleosomes that form around 147 bp long DNA stretches and define chromatin accessibility, acting as general transcriptional repressors by competing with TFs for DNA binding (Segal and Widom, 2009; Struhl and Segal, 2013), 3) other components of the transcription initiation enzymatic machinery including sigma factors (σ) in prokaryotes and components (TFIID/SAGA, mediator) of the preinitiation complex (PIC) in eukaryotes (Feklístov et al., 2014; Haberle and Stark, 2018), and 4) physicochemical and thermodynamic properties related to protein binding (Rohs et al., 2010; Inukai et al., 2017) and transcription initiation (Chen et al., 2010; Zrimec and Lapanje, 2015), such as strand dissociation around the transcription start site (TSS), giving enzymatic access to the DNA (**Figure 1A**). The DNA sequence preferences of nucleosomes define nucleosome organization *in vivo* and have been shown to account for the general depletion of nucleosomes around the starts and ends of genes as well as around TFBS, which might assist in directing TFs to their appropriate genomic sites (Segal and Widom, 2009). Apart from the DNA-guided nucleosome positioning, other epigenetic mechanisms (where functionally relevant changes to the genome do not involve a change in the nucleotide sequence), such as histone modification and DNA methylation, also play a vital part in transcriptional regulation (Gibney and Nolan, 2010; Miller and Grant, 2013). Together, they control the accessibility of DNA for protein binding and enzymatic processing (Watson et al., 2008) (**Figure 1A**). The epigenome is established and maintained by the site-specific recruitment of chromatin-modifying enzymes and their cofactors. Identifying the *cis* elements that regulate transcription initiation and epigenomic modification is critical for understanding the regulatory mechanisms that control gene expression patterns.

Machine learning is used to predict the locations of TFBS and their TF binding specificities, other *cis*-regulatory elements and binding sites, larger DNA non-coding regions such as enhancers and promoters, as well as nucleosome binding landscapes and epigenetic states. The computational tasks for inferring TFBS from DNA sequence or modeling TFBS specificity based on TF activity measurements can be framed as binary/multiclass classification and regression problems, respectively. TFBS can be predicted from the genome *de novo* (Jayaram et al., 2016), or analyzed based on separate measurements (Kim et al., 2007; Visel et al., 2009; Ghandi et al., 2014) or massively parallel reporter assays using high-throughput quantitative sequencing technologies (HTS), giving peak calls for various regulatory (epigenetic and transcriptional) activities across tissues and isolated cell types (Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015). These include: 1) ChIP-seq (Chromatin immunoprecipitation sequencing) (Johnson et al., 2007) and ChIP-nexus (addition of exonuclease digestion step) (He et al., 2015) to map TF binding sites and histone modification presence, 2) DNase-seq (DNase I hypersensitive sites sequencing) (Song and Crawford, 2010) and ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing) (Buenrostro et al., 2013) to measure DNA chromatin accessibility, which typically mark nucleosomes and TF-bound sites, and 3) other methods, such as PBMs (protein binding microarrays) (Berger et al., 2006), SELEX (Systematic evolution of ligands by exponential enrichment) (Blackwell and Weintraub, 1990) and BunDLE-seq [Binding to Designed Library, Extracting, and sequencing) (Levo et al., 2015) that can provide quantitative measurements of TF binding to thousands sequences within a single experiment (further details can be found in the following publication (Barshai et al., 2020)].

Common measures for evaluating the performance of ML classifiers, typically on unseen data, include: 1) precision and recall, 2) the area under the receiver operating characteristic curve (AUC) that measures the tradeoff between the true positive rate (recall) and false positive rate for different thresholds, as well as 3) the area under the precision recall curve (AUPRC) that measures the tradeoff between precision and recall for different thresholds [for technical details we refer the reader to a recent review (Jiao and Du, 2016)]. Regression models are frequently evaluated using a correlation coefficient or the coefficient of variation (R^2) (de Boer et al., 2020; Zrimec et al., 2020).

Classical Machine Learning Relies on Engineered Features

The goal of supervised ML is to learn a response function y (target variable) from the set of features x (explanatory variables) present in the training dataset, where y describes some property related to gene expression, such as TF binding, ChIP-seq signal or mRNA abundance. With shallow learning, the DNA sequence that generally serves as the explanatory variable must be described with numerical features, such as position weight matrices (PWMs) (Stormo, 2000; Jayaram et al., 2016; Lu and Rogan, 2018), ungapped or gapped k-mer frequencies (Fletez-Brant et al., 2013; Ghandi et al., 2014; Zrimec et al., 2020), pseudo k-tuple nucleotide composition (Lin et al., 2014; Chen et al., 2015) or physicochemical and conformational (structural) properties (Rohs et al., 2009; Meysman et al., 2012; Zrimec, 2020a). Shallow methods thus require some features and methods that can describe or interpret the DNA regulatory motifs, and then use these features or motifs to build predictors. Due to their dependence on feature engineering, the shallow model training and evaluation methodology also commonly includes feature selection on all variables, retaining only the feature sets most informative for predicting the target variable. Afterward, ML models are trained on the engineered and selected feature subsets and finally, validation is performed on a held out portion of the data to assess the model performance (Ghandi et al., 2014; Zelezniak et al., 2018; Zrimec and Lapanje, 2018) (**Figure 2A**).

Comparison of 26 different approaches to model and learn a protein's DNA-binding specificity based on PBMs for various mouse TFs (Weirauch et al., 2013) showed that, for most TFs examined, simple models based on mononucleotide PWMs can perform similarly to more complex models, falling short only in specific cases that represented less than 10% of the examined TFs. The best-performing motifs typically have relatively low information content, consistent with widespread degeneracy in eukaryotic TF sequence preferences. Out of multiple *de novo* motif discovery tools that can be used locally for creating PWMs from HTS data and for scanning them against DNA, FIMO (Grant et al., 2011) and MCast (Grant et al., 2016) were found to have the best performance in their respective classes of methods that predict individual TFBSs or identify clusters, respectively (**Table 1**) (Jayaram et al., 2016). In an approach termed "Catchitt" for predicting cell type-specific TFBS using ensemble classifiers (Keilwagen et al., 2019), standard PWM motifs from databases were expanded with motifs learned by *de novo* motif discovery from ChIP-seq and DNase-seq data using sparse local inhomogeneous mixture (Slim) models (Keilwagen and Grau, 2015), which capture short to mid-range intra-motif dependencies. Catchitt earned a shared first rank in the 2017 ENCODE-DREAM *in vivo* TFBS prediction challenge, achieving a median AUPRC of 0.41 on test data. Despite the success of PWM-based methods, ML approaches have been shown to achieve similar or even better results. For instance, the method "QBiC-Pred" was developed to quantitatively predict TF binding changes due to sequence variants (Martin et al., 2019), using ordinary least squares (OLS) regression and HTS data containing single nucleotide variants (SNVs). The OLS models of TF binding specificity were accurate in predicting mutational effects on TF binding *in vitro* and *in vivo* (R^2 up to 0.95), outperforming widely used PWM models as well as recently developed DNNs (Alipanahi et al., 2015) on the tested data. The problem with any ML approach using k-mers as features is that it becomes susceptible to noisy training k-mer frequencies once k becomes large. This was

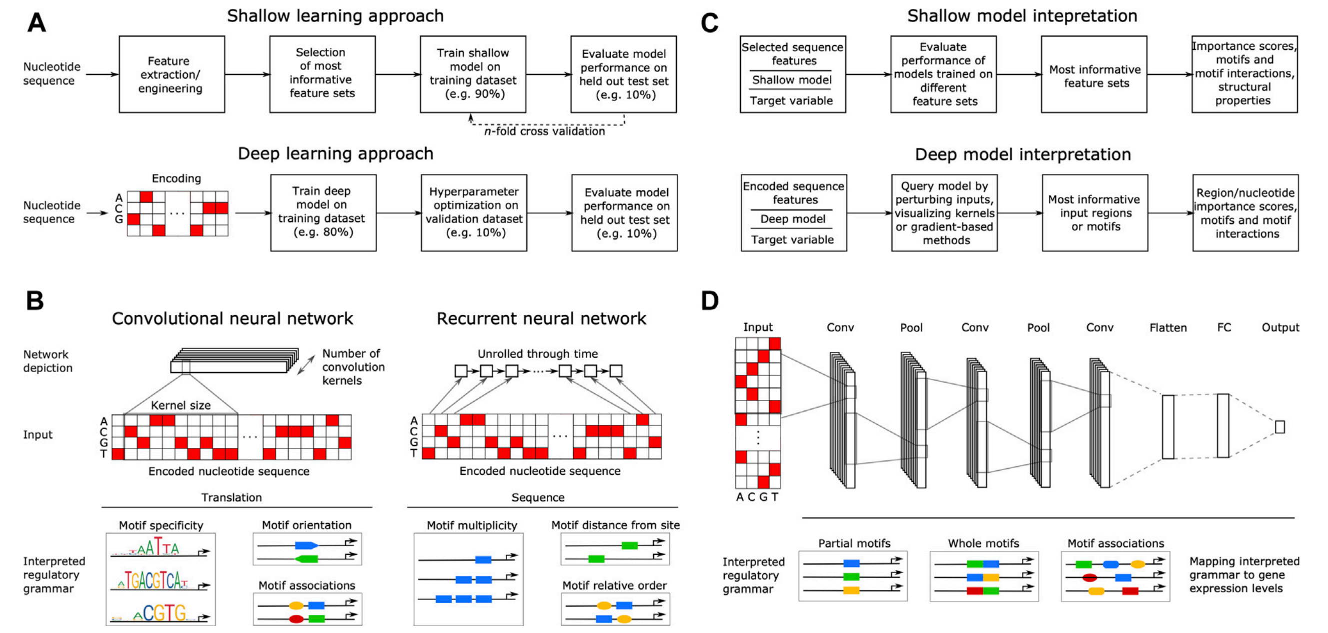

A
Shallow learning approach
Nucleotide sequence
Feature extraction/ engineering
Selection of most informative feature sets
Train shallow model on training dataset (e.g. 90%)
Evaluate model performance on held out test set (e.g. 10%)
n-fold cross validation
Deep learning approach
Nucleotide sequence
Encoding
A C G
Train deep model on training dataset (e.g. 80%)
Hyperparameter optimization on validation dataset (e.g. 10%)
Evaluate model performance on held out test set (e.g. 10%)
C
Shallow model intepretation
Selected sequence features
Shallow model
Target variable
Evaluate performance of models trained on different feature sets
Most informative feature sets
Importance scores, motifs and motif interactions, structural properties
Deep model interpretation
Encoded sequence features
Deep model
Target variable
Query model by perturbing inputs, visualizing kernels or gradient-based methods
Most informative input regions or motifs
Region/nucleotide importance scores, motifs and motif interactions
B
Convolutional neural network
Recurrent neural network
Network depiction
Number of convolution kernels
Unrolled through time
Kernel size
Input
A C G T
Encoded nucleotide sequence
Encoded nucleotide sequence
Translation
Sequence
Motif specificity
Motif orientation
Motif multiplicity
Motif distance from site
Interpreted regulatory grammar
AATTA
TGACGTCA
ACGTG
Motif associations
Motif relative order
D
Input
Conv
Pool
Conv
Pool
Conv
Flatten
FC
Output
A C G T
Interpreted regulatory grammar
Partial motifs
Whole motifs
Motif associations
Mapping interpreted grammar to gene expression levels

FIGURE 2 | Principles of machine learning from nucleotide sequence. **(A)** Flowcharts of a typical supervised shallow modeling approach **(top)** and a typical supervised deep modeling approach **(bottom)**, depicting a one-hot encoding that equals k-mer embedding with $k = 1$. **(B)** Overview of convolutional (CNN) and recurrent neural networks (RNN) in interpreting DNA regulatory grammar. A CNN scans a set of motif detectors (kernels) of a specified size across an encoded input sequence, learning motif properties such as specificity, orientation and co-association. An RNN scans the encoded sequence one nucleotide at a time, learning sequential motif properties such as multiplicity, distance from e.g. transcription start site and the relative order of motifs. **(C)** Interpreting shallow models **(top)** by evaluating their performance when trained on different feature sets can yield feature importance scores, motifs and motif interactions, as well as compositional and structural properties. Similarly, interpreting the regulatory grammar learned by deep models **(bottom)**, by e.g. perturbing the input, visualizing kernels or using gradient-based methods, can yield feature importance scores spanning nucleotides up to whole regions, as well as motifs and motif interactions. **(D)** Example of a typical deep neural network (DNN) comprising three separate convolutional layers (Conv) connected *via* pooling layers (Pool) and a final fully connected network (FC) producing the output gene expression levels. Pool stages compute the maximum or average of each motif detector's rectified response across the sequence, where maximizing helps to identify the presence of longer motifs and averaging helps to identify cumulative effects of short motifs. The DNN learns distributed motif representations in the initial Conv layers and motif associations that have a joint effect on predicting the target in the final Conv layer, representing DNA regulatory grammar that is mapped to gene expression levels.

solved with methods for robust estimation of k-mer frequencies based on alternative feature sets, where gapped k-mers were introduced as a followup to the initial k-mer method "kmer-SVM" (Lee et al., 2011). The new classifier termed "gkm-SVM" predicted functional genomic regulatory elements with significantly improved accuracy compared to the original kmer-SVM, increasing the precision by up to 2-fold and achieving an AUC of 0.97 for TFBS prediction, compared to 0.91 with kmer-SVM (Ghandi et al., 2014). In this case however, the PWM-based classifier still outperformed both methods (AUC = 0.98).

In the case of epigenetic states that underlie DNA accessibility, it was shown that histone modifications can be predicted with remarkable accuracy from TF-binding profiles using LR classifiers (avg. AUC ~0.86 to 0.95 on different DNA regions in H1 cells), recapitulating known interactions between TFs and chromatin-modifying enzymes (Benveniste et al., 2014). This demonstrated that associations between gene expression and histone modifications do not necessarily imply a direct regulatory role for these modifications, but can be explained equally well as an indirect effect of interactions between TFs and chromatin-modifying enzymes. Similarly, a pipeline termed "Epigram" (Whitaker et al., 2015) was developed to predict histone modification and DNA methylation patterns from DNA motifs. The authors also cataloged novel *cis* elements by *de novo* motif finding, showing that numerous motifs that have location preference and represented interactions with the site-specific DNA-binding factors that establish and maintain epigenomic modifications. Using their method gkm-SVM (Ghandi et al., 2014) to encode cell type–specific regulatory sequence vocabularies, Lee and colleagues (Lee et al., 2015) devised a sequence-based computational method to predict the effect of regulatory variation. The effect of sequence variants was quantified by the induced change in the gkm-SVM score, "deltaSVM," which accurately predicted the impact of SNVs on DNase I hypersensitivity in their native genomes and could identify risk-conferring functional variants in validated data including autoimmune diseases, demonstrating the usefulness of this approach.

Apart from the base DNA sequence properties, structural properties have been found to improve model performance in certain cases, such as when predicting: 1) TFBS and their specificities (Abe et al., 2015; Tsai et al., 2015; Mathelier et al., 2016; Yang et al., 2017), 2) promoters and TSS sites (Meysman et al., 2012; Bansal et al., 2014; Kumar and Bansal, 2017), and 3) σ factor binding sites (Zrimec, 2020a). These properties are directly related to protein-DNA recognition and binding (Rohs et al.,

TABLE 1 | Overview of studies modeling protein-DNA interactions that govern the initiation of gene expression from nucleotide sequence properties. Highest achieved or average scores are reported, on test sets where applicable, and include precision (*prec*) and recall (*rec*), area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), the coefficient of variation (R^2), Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and Matthews correlation coefficient (MCC).

Ref.	Strategy	Target var.	Explan. vars.	Method	Score	Organism
(Jayaram et al., 2016)	Shallow	TFBS prediction	PWMs	PWM alignment algorithms	*prec* = 0.73, *rec*. = 0.82	Human
(Keilwagen et al., 2019)	Shallow	TFBS prediction	DNA motif and chromatin-based features	Classifier ensembles	AUPRC = 0.81	Human
(Ghandi et al., 2014)	Shallow	TFBS prediction	PWMs, gapped k-mers	SVM classification	AUC = 0.98	Human
(Levo et al., 2015)	Shallow	TF binding specificity	k-mers, DNA structural variables	L1-regularized LR	R^2 = 0.90	Yeast
(Yang et al., 2017)	Shallow	TF binding specificity	k-mers, DNA structural variables	L2-regularized multiple LR	R^2 = 0.90	Human
(Martin et al., 2019)	Shallow	TF binding specificity	k-mers	OLS regression	R^2 = 0.95	Human
(Lin et al., 2014)	Shallow	σ54 promoter prediction	Pseudo k-tuple nucleotide composition	SVM classification	MCC = 0.88	*E. coli*
(He et al., 2018)	Shallow	σ70 promoter prediction	Trinucleotide-based features	SVM classification	MCC = 0.92	*E. coli*
(Benveniste et al., 2014)	Shallow	Histone modifications	k-mers, TF CHIP-seq data	LR classification	AUC = 0.95	Human
(Whitaker et al., 2015)	Shallow	Histone modifications, DNA methylation	DNA motifs	RF classification	AUC = 0.96	Human
(Lee et al., 2015)	Shallow	DNA chromatin accessibility	PWMs, gapped k-mers	SVM classification	AUC = 0.75	Human
(Trabelsi et al., 2019)	Deep	TFBS prediction	k-mers	CNN + biLSTM classification	AUC = 0.93	Human
(Zeng et al., 2016)	Deep	TFBS prediction	DNA sequence	CNN classification	AUC = 0.88	Human
(Kelley, 2020)	Deep	TFBS prediction	DNA sequence	CNN classification	AUC = 0.82	Human, mouse
(Chen et al., 2021)	Deep	TFBS prediction	DNA sequence	CNN + biLSTM + attention classification	AUC = 0.99	Human
(Alipanahi et al., 2015)	Deep	TF binding specificity	DNA sequence	CNN classification	AUC = 0.90	Human
(Wang et al., 2018)	Deep	TF binding specificity	DNA sequence	CNN regression	ρ = 0.81	Human
(Avsec et al., 2021)	Deep	TF binding specificity	DNA sequence	CNN regression	ρ = 0.62	Human
(Van Brempt et al., 2020)	Deep	Transcription initiation frequency	DNA sequence	CNN ordinal regression	R^2 = 0.88	*E. coli*
(Zhou and Troyanskaya, 2015)	Deep	Multitask chromatin profiling data	DNA sequence	CNN classification	AUC = 0.96	Human
(Quang and Xie, 2016)	Deep	Multitask chromatin profiling data	DNA sequence	CNN + biLSTM classification	AUC = 0.97	Human
(Park et al., 2020)	Deep	Multitask chromatin profiling data	DNA sequence	CNN + biLSTM + attention classification	AUC = 0.95	Human
(Singh et al., 2016)	Deep	Histone modifications	DNA sequence	CNN classification	AUC = 0.80	Human
(Singh et al., 2017)	Deep	Histone modifications	DNA sequence	LSTM + attention classification	AUC = 0.81	Human
(Kelley et al., 2016)	Deep	DNA chromatin accessibility	DNA sequence	CNN classification	AUC = 0.90	Human
(Kelley et al., 2018)	Deep	DNA chromatin accessibility	DNA sequence	CNN regression	r = 0.86	Human
(Angus and Eyuboglu, 2018)	Deep	DNA chromatin accessibility	DNA sequence	CNN + attention regression	ρ = 0.59	Human
(Angermueller et al., 2017)	Deep	DNA methylation	DNA sequence and features	CNN classification	AUC = 0.83	Human
(Tian et al., 2019)	Deep	DNA methylation	DNA sequence	CNN regression	AUC = 0.97	Human

2009; Bishop et al., 2011; Zrimec, 2020b) and include DNA shape (Mathelier et al., 2016), thermodynamic stability (SantaLucia, 1998) and propensity for duplex destabilization (Zrimec and Lapanje, 2015), as well as flexibility and curvature related properties (Brukner et al., 1995; Geggier and Vologodskii, 2010). For instance, the dependence of TF binding specificity on the TFBS core and flanking sequence was studied using LR and BunDLE-seq data on thousands of designed sequences with single or multiple Gcn4 or Gal4 binding sites (Levo et al., 2015). By supplanting k-mer frequencies at each position with DNA structural properties, 15 bp flanking sequences (15 bp) of core binding sites were shown to affect the binding of TFs, as models based on combined core and flanking regions explained the highest amount of variance in the measurements (R^2 up to 0.9 for Gal4). The contribution of DNA shape readout and its importance in core motif-flanking regions was further demonstrated using LR and HT-SELEX data across a diverse set of 215 mammalian TFs from 27 families (Yang et al., 2017), as regression models that used k-mer and shape features generally outperformed k-mer models by ~10% (R^2 up to 0.90). Using feature selection techniques, positions in the TFBSs could be pinpointed where DNA shape readout is most likely to occur, and accordingly, novel DNA shape logos were proposed to visualize the DNA shape preferences of TFs. Similarly, SVM regression models of TF binding specificity based on PBM data for 68 mammalian TFs showed that shape-augmented models compared favorably to sequence-based models (Zhou et al., 2015), as DNA shape features reduced the dimensionality of the feature space. The authors from Rohs lab also provide an updated database of TFBS shape logos in 2020 (Chiu et al., 2020). Moreover, derivatives of DNA structural properties, such as pseudo k-tuple nucleotide compositions (Lin et al., 2014) and trinucleotide features including position-specific propensity and electron-ion potential (He et al., 2018), were applied to the problem of predicting bacterial σ54 and σ70 promoters in *E. coli*, which transcribe carbon and nitrogen-related genes or regulate the transcription of most genes, respectively. The respective ML classifiers termed "iPro54-PseKNC" (Lin et al., 2014) and "70ProPred" (He et al., 2018) could accurately distinguish the specific promoters from negative examples (AUC = 0.98 and 0.99, respectively).

Deep Neural Networks can Learn Regulatory Grammar Automatically

In contrast to shallow architectures that are limited in their applications even when large datasets are available, deep architectures are abstracted by multiple hidden layers between *x* and *y*. Each layer learns a new representation of the data before passing it on to the successive layers, finding hidden data structures to make accurate predictions (Mhaskar et al., 2017). The most common DNN architectures in genomics include convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as bidirectional long short-term memory (biLSTM) networks. CNNs are regularized fully connected networks that progressively scan a DNA molecule within a receptive field, where they learn to recognize the occurrence of DNA motifs (e.g. specificity, orientation and co-association) (Eraslan et al., 2019a) (**Figure 2B**). Despite the capability of RNNs to learn sequential information (e.g. multiplicity, relative order), they are computationally expensive to train and certain improvements to CNNs, such as dilation (Yu and Koltun, 2015) and self-attention (Wang et al., 2017; Bello et al., 2019; Repecka et al., 2021), enable them to outperform RNNs (Gupta and Rush, 2017; Strubell et al., 2017; Trabelsi et al., 2019). Dilated convolution uses kernels with gaps to allow each kernel to capture information across a larger stretch of the input sequence, without incurring the increased cost of using RNNs (Gupta and Rush, 2017; Strubell et al., 2017). Similarly, self-attention is a special case of attention mechanism that allows kernels to focus on specific parts of the input when producing the output, allowing positions across the entire input sequence to interact and contribute to the result with different attention weights (Vaswani et al., 2017).

Deep learning does not require feature engineering or selection, since this is an inherent feature of the DNN learning process (Webb, 2018). However, it does require representing the categorical nucleotide sequence data numerically using an encoding scheme, such as one-hot, which transforms the sequence into a binary matrix with columns corresponding to each category. DNNs have thus been applied mostly on one-hot encoded nucleotide sequences as input (Eraslan et al., 2019a; Alipanahi et al., 2015), with recent reports showing that the use of k-mer embedding to represent the input sequences can improve model performance compared to one-hot encoding (itself a special case of k-mer embedding where $k = 1$) (Trabelsi et al., 2019). These inputs are well suited for comprehending the base DNA motif information as well as higher order interactions that describe the DNA regulatory grammar of gene expression (Eraslan et al., 2019a; Zrimec et al., 2020). Thus, DNNs achieve high predictive accuracies often surpassing those of models based on engineered features and, in our experience, using structural DNA properties does not lead to improved predictive performance with DNNs (Zrimec et al., 2020). Due to the large amount of model *hyper*parameters, such as network structure (e.g. number and size of kernels, **Figure 2B**) and training algorithm (e.g. learning rate), a special step termed hyperparameter optimization (Bergstra et al., 2015) is required for finding the best combinations of these hyperparameters and is an integral part of DNN training. To train DNNs, the data is typically split into training, validation, and testing datasets, where: 1) the model is trained on the training set by minimizing a loss function commonly MSE for regression and cross entropy for classification (Géron, 2019), 2) hyperparameter tuning is performed on the validation set and the best performing model on the validation set is chosen, and 3) the performance of the final model is evaluated on the testing set, also verifying if it overfits the data (Eraslan et al., 2019a; Zrimec et al., 2020) (**Figure 2A**). With DNN testing, cross-validation is rarely performed due to the large dataset sizes and issues with algorithmic efficiency. Commonly, 10% test splits are used for testing the models trained on 80% of the data, whereas another 10% of the training data is used for the internal validation of hyperparameter selection (Géron, 2019). For further technical details we refer the reader to excellent recent reviews (Eraslan et al., 2019a; Barshai et al., 2020).

Deep methods are frequently trained on HTS peak profiles, either converted to binary scores or left continuous as a regression problem, and the underlying TFBS and specificities are interpreted by the network itself. The first such method to showcase the efficiency of DNNs for analysis of TF binding specificities was DeepBind (Alipanahi et al., 2015), where a single CNN layer was trained on sequence specificities of DNA and RNA-binding proteins as measured by several types of HTS assays (including PBM, HT-SELEX, and ChIP-seq), in a combined 12 terabases of mouse and human data. DeepBind captured binding specificities from raw sequence data by jointly discovering new motifs of hundreds of TFs along with the rules for combining them into a predictive binding score. The resulting DeepBind models could then be used to identify binding sites in test sequences and to score the effects of novel mutations, uncovering the regulatory role of disease-associated genetic variants that can affect TF binding and gene expression. Importantly, the method outperformed 14 other methods (Weirauch et al., 2013) and achieved the highest score when applied to the *in vivo* ChIP-seq data (avg. AUC = 0.90), suggesting that it can generalize from HT-SELEX (Jolma et al., 2013) to other data acquisition technologies despite being based on a general-purpose ML framework.

The basic approach of DeepBind was further explored and expanded upon in subsequent studies with different network layers. For instance, Zeng and co. (Zeng et al., 2016). performed a systematic exploration of CNN architectures for predicting DNA sequence binding using a similarly large set of TF data. To control potentially confounding effects, like positional or motif strength bias, they chose to explore two specific classification tasks of motif discovery (bound vs. dinucleotide shuffles per TF and cell type) and motif occupancy (bound vs. non-bound). In both tasks, classification performance increased with the number of convolution kernels (AUC up to 0.88), and the use of local pooling or additional layers had little effect on the performance. CNN architectures that took advantage of these insights exceeded the classification performance of DeepBind, emphasizing the need to use sufficient kernels to capture motif variants. With deepRAM, a tool providing an implementation of a wide selection of architectures (Trabelsi et al., 2019), it was shown that deeper, more complex architectures provide a clear advantage with sufficient training data, with hybrid CNN + RNN architectures outperforming other methods in terms of accuracy (AUC = 0.93 with 1xCNN + biLSTM). However, although RNNs improve model accuracy, this comes at the expense of a loss in the interpretability of the features learned by the model. Kelley (Kelley, 2020) developed a strategy to train deep CNNs simultaneously on human and mouse genomes, which improved gene expression prediction accuracy on held out and variant sequences. Applying mouse regulatory models to analyze human genetic variants associated with molecular phenotypes and disease improved model performance (AUROC increased from 0.80 to 0.82), showing that the thousands of available non-human transcriptional and epigenetic profiles can be leveraged for more effective investigation of how gene regulation affects human disease. Moreover, the performance of assessing the functional impact of non-coding variants (e.g. SNVs) was further improved with DeFine (Wang et al., 2018), a regression model based on large-scale TF ChIP-seq data and capable of accurately predicting real-valued TF binding intensities (Spearman's ρ up to 0.81). Here, the predicted changes in the TF binding intensities between the altered sequence and the reference sequence reflected the degree of functional impact for the variant, and could accurately identify the causal functional variants from measured disease-associated variants. Similar networks have also been used in bacteria, where the online promoter design tool (ProD) (Van Brempt et al., 2020) is based on forward engineering of promoter transcription initiation frequency (TIF). By training a CNN with high-throughput DNA sequencing data from fluorescence-activated cell sorted promoter libraries of *E. coli* σ70 and *Bacillus subtilis* σB-, σF- and σW-dependent promoters, prediction models were capable of predicting both TIF and orthogonality of the σ-specific promoters, which facilitated development of tailored promoters, where predictions explained ~88% of the variance of experimental observations.

With prediction of epigenetic states, the "DeepSEA" method (Zhou and Troyanskaya, 2015) was the first to utilize three CNN layers trained for multi-task predictions of large-scale chromatin-profiling data, including transcription factor (TF) binding, DNase I hypersensitivity sites (DHSs) and histone-mark profiles across multiple cell types. The method significantly outperformed gkm-SVM (avg. AUC of 0.96 vs. 0.90) and enabled high-performance sequence-based prediction of both DHSs (avg. AUC = 0.92) and histone modifications (avg. AUC = 0.86). In the "DanQ" model (Quang and Xie, 2016) trained on similar data as DeepSEA, a hybrid CNN + RNN architecture was used in order to enhance its perception of regulatory grammar, where the CNN captured regulatory motifs and the RNN captured long-term dependencies between the motifs. The model achieved improved performance compared to DeepSEA (avg. AUC = 0.97) as well as compared to a LR baseline model, which despite its simplicity was an effective predictor (AUROC >0.70). Similarly, with histone modifications, the CNN "DeepChrome" (Singh et al., 2016) was shown to consistently outperform both SVM and RF classifiers (avg. AUC of 0.80 vs. 0.66 and 0.59, respectively). Kelley and co. (Kelley et al., 2016) introduced the open source package "Basset" that trains CNNs on a set of accessible genomic sites mapped in 164 cell types by DNase-seq, achieving improved predictive accuracy compared to previous methods, such as gkm-SVM (avg. AUC = 0.90 vs. 0.78), and good overlap of SNV predictions with previous observations. Furthermore, Kelley and co. (Kelley et al., 2018) developed another CNN, "Basenji," to predict mammalian cell-type specific epigenetic and transcriptional profiles, where an unprecedented input sequence size of 131 kbp around TSS was used, spanning distal as well as proximal regulatory elements. Indeed, model predictions regarding the influence of SNVs on gene expression were shown to align well to known variants in human populations related to disease loci (avg. Pearson's r = 0.86).

To map associations between DNA sequence patterns and methylation levels at CpG-site resolution, Angermuller and co. developed "DeepCpG" (Angermueller et al., 2017). The method was evaluated on single-cell methylation data across different cell types and HTS protocols, and yielded more accurate predictions than shallow methods, such as RF (avg. AUC = 0.83 vs. 0.80). The authors also showed that interpretation of the model parameters could provide insights into how sequence composition affects methylation variability. A more recent alternative approach termed "MRCNN" (Tian et al., 2019) outperformed DeepCpG (AUC up to 0.97), and *de novo* discovered motifs from the trained CNN kernels were shown to match known motifs.

Finally, by expanding DNN architectures with attention mechanisms to model complex dependencies among input signals, favourable results can be achieved compared to the non-attentive DNN counterparts. This was shown with multiple prediction tasks, including: 1) TFBS prediction, where "DeepGRN" (Chen et al., 2021) achieved higher unified scores in 6 of 13 targets than any of the top four methods in the 2016 ENCODE-DREAM challenge including Catchitt (Keilwagen et al., 2019), 2) histone modification, where "AttentiveChrome" (Singh et al., 2017) outperformed DeepChrome (Singh et al., 2016) in 50 out of 56 human cell types (avg. AUC of 0.81 vs. 0.80), 3) DNA chromatin accessibility, where the attention-based model (Angus and Eyuboglu, 2018) outperformed standard CNNs (ρ = 0.59 vs. 0.54) as well as dilated convolutions on specific experiments, and 4) multitask chromatin profiling data, where "TBiNet" (Park et al., 2020) outperformed DeepSea (Zhou and

Troyanskaya, 2015) and DanQ (Quang and Xie, 2016) in the TF-DNA binding prediction task (avg. AUC of 0.95 vs. 0.90 and 0.93, respectively). This suggests that attention is an effective strategy to incorporate long-range sequence context into local predictions and particularly effective for gene-expression prediction.

Interpreting Models to Retrieve the Learned Regulatory Grammar

With shallow models, the most informative feature sets are interpreted by evaluating the performance of models trained on different feature sets (Ghandi et al., 2014; Zrimec and Lapanje, 2018; de Boer et al., 2020) (**Figure 2C**). This can yield feature importance scores, motifs (k-mers or PWMs, depending on the provided input features, **Figure 2A**) and motif interactions (Ghandi et al., 2014; Keilwagen and Grau, 2015), as well as compositional and structural properties (Lin et al., 2014; Yang et al., 2017), all of which comprise a compendium of regulatory grammar, informative for understanding the regulation of gene expression. Due to the inherent capability of DNNs to learn predictive motif representations, rules for cooperative TF binding interactions (Avsec et al., 2021) and higher-order sequence features, such as secondary motifs and local sequence context (Zeng et al., 2016), as well as genotypic variation effects (Zhou and Troyanskaya, 2015), they represent a powerful approach to uncover the detailed *cis*-regulatory grammar of genomic sequences (**Figure 2C**) (Koo and Ploenzke, 2020a; He et al., 2020). This is achieved by interpreting the models using approaches that include: 1) CNN kernel visualization, where typically motifs in the initial layers are visualized, 2) input perturbation-based (sensitivity) analysis, which highlights the parts of a given input sequence that are most influential for the model prediction by occluding or mutating them (Alipanahi et al., 2015; Ancona et al., 2017), 3) gradient-based methods that estimate feature importance with iterative backward and forward propagations through the network (Shrikumar et al., 2017; Montavon et al., 2018; Shrikumar et al., 2018), yielding e.g. saliency maps (Simonyan et al., 2013) and 4) higher-order interactions among sequence elements, which can be assessed e.g. by using association rule analysis (Naulaerts et al., 2015; Zrimec et al., 2020), second-order perturbations (Koo et al., 2018), self-attention networks (Ullah and Ben-Hur, 2020) or by visualizing kernels in deeper layers (Maslova et al., 2020) [interested readers are referred to (Eraslan et al., 2019a; Koo and Ploenzke, 2020a)]. Moreover, attention mechanisms were recently shown to be more effective in discovering known TF-binding motifs compared to non-attentive DNNs (Park et al., 2020), as the learned attention weights correlate with informative inputs, such as DNase-Seq coverage and DNA motifs (Chen et al., 2021), and they can provide better interpretation than other established feature visualization methods, such as saliency maps (Lanchantin et al., 2016; Singh et al., 2017).

Since these are computational approaches, they extract statistical patterns that may not immediately reflect physical properties of the variables and should be treated as hypotheses that need to be further examined (Koo and Eddy, 2019). For instance, a method can point out certain motifs or associations that are important for the model in predicting the target, but how this reflects actual physicochemical interactions can be rather hard to interpret from the model alone. Nevertheless, this is an active area of research and new solutions are frequently developed (Lundberg and Lee, 2017; Chen and Capra, 2020; Koo and Ploenzke, 2020b), where rigorous testing as well as experimentally verifying predictions will highlight the most promising approaches (Ancona et al., 2017). On the other hand, an alternative trend that is arguably more appropriate than interpreting black box models is the development of inherently interpretable models (Rudin, 2019), where prior knowledge of gene expression can be built into the deep network structure itself (Ma et al., 2018; Tareen and Kinney, 2019; Liu et al., 2020). We refer interested readers to the excellent recent review by Azodi and co. (Azodi et al., 2020).

REGULATORY MECHANISMS IN SPECIFIC CODING AND NON-CODING REGIONS

Both transcription and translation comprise multiple steps that include initiation, elongation and termination (Watson et al., 2008). Transcription of protein coding genes is controlled *via* the gene regulatory structure, comprised of coding and *cis*-regulatory regions that include promoters, untranslated regions (UTRs) and terminators, and generally proceeds in the direction from the upstream 5′ to downstream 3′ end (**Figure 1B**). Initiation is regulated by enhancers, promoters and 5′ UTRs, where the transcriptional machinery including RNA polymerase (RNAP) is guided to the correct sites on the DNA. In the elongation phase, mRNA is synthesized (transcribed) from the coding sequence, and this process terminates toward the 3′ UTR and terminator regions carrying termination signals. Afterward, the process of mRNA decay is triggered, which occurs in eukaryotes after the mRNA strand is matured by 5′ capping and 3′ poly(A) tail extension, and precursor mRNA (pre-mRNA) transcripts are processed by the spliceosome, removing introns (non-coding regions) and joining exons (coding regions) together (Watson et al., 2008; Wilkinson et al., 2020). The rates of mRNA synthesis and decay define the actual mRNA levels in the cell that are commonly measured with RNA-Seq (Wang et al., 2009). The DNA regions involved in mRNA synthesis carry multiple regulatory motifs, with codon usage in coding regions detailing which nucleotide triplets encoding an amino acid (AA) are used at each position, contributing to the base regulatory grammar of transcription (Plotkin and Kudla, 2011; Cheng et al., 2017). As described above, the general genomic architecture, defined by binding of histones in eukaryotes (Struhl and Segal, 2013) and nucleoid-associated proteins (NAPs) in prokaryotes (Dillon and Dorman, 2010), acts as a master regulator of transcription by controlling the accessibility of DNA to proteins (Curran et al., 2014; Morse et al., 2017).

Translation also proceeds in the direction from the 5′ to the 3′ end of an mRNA (**Figure 1C**) and, in bacteria, occurs simultaneously with transcription in the cytoplasm of the cell, whereas in eukaryotes transcription occurs in the nucleus and translation occurs in the cytoplasm (Watson et al., 2008).

TABLE 2 | Overview of studies modeling gene expression-related properties from separate regulatory or coding regions. Highest achieved or average scores are reported, on test sets where applicable, and include accuracy (*acc*), area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), the coefficient of variation (R^2) and Pearson's correlation coefficient (*r*).

Ref.	Strategy	Region	Target var.	Explan. vars.	Method	Score	Organism
(Leman et al., 2018)	Shallow	Coding	Splice site prediction	Sequence and PWM features	Logistic regression	*acc* = 0.96%	Human
(Signal et al., 2018)	Shallow	Coding	Branch point prediction	Sequence features	Gradient boosting classification	AUC = 0.94	Human
(Zhang et al., 2017a)	Shallow	Coding	Branch point prediction	Sequence features	Mixture models classification	AUC = 0.82	Human
(Trösemeier et al., 2019)	Shallow	Coding	Protein abundance	Codon usage features	COSEM mathematical model	R^2 = 0.45, 0.51, 0.37, respectively	*E. coli*, yeast, human
(Ferreira et al., 2020)	Shallow	Coding	Protein abundance	Codon usage	AdaBoost regression	R^2 = 0.95	Yeast
(Tunney et al., 2018)	Shallow	Coding	Ribosome density at each codon	Codon usage	NN regression	*r* = 0.57	Yeast
(Zuallaert et al., 2018)	Deep	Coding	Splice site prediction	DNA sequence	CNN classification	AUPRC = 0.61	Human, *A. thaliana*
(Wang et al., 2019)	Deep	Coding	Splice site prediction	DNA sequence	CNN classification	AUC = 0.98	Human
(Jaganathan et al., 2019)	Deep	Coding	Splice site prediction	DNA sequence	CNN classification	AUPRC = 0.98	Human
(Paggi and Bejerano, 2018)	Deep	Coding	Branch point prediction	DNA sequence	biLSTM classification	AUC = 0.71	Human
(Nazari et al., 2019)	Deep	Coding	Branch point prediction	DNA sequence	biLSTM + CNN classification	AUC = 0.81	Human
(Xu et al., 2017)	Deep	Coding	Alternative splicing prediction	Sequence and epigenetic features	Dense DNN classification	AUPRC = 0.89	Human
(Lee et al., 2020)	Deep	Coding	Alternative splicing prediction	Sequence and epigenetic features	RNN classification	AUPRC = 0.8	Human
(Zhang et al., 2019)	Deep	Coding	Alternative splicing prediction	RNA-seq data	Dense DNN + bayesian hypothesis testing	AUC = 0.87	Human
(Fu et al., 2020)	Deep	Coding	Protein abundance	DNA sequence	Multilayer biLSTM regression	R^2 = 0.52	*E. coli*
(Fujimoto et al., 2017)	Deep	Coding	Optimal codon usage	DNA sequence	biLSTM encoder-decoder	*acc* = 0.97	*E. coli*
(Yang et al., 2019)	Deep	Coding	Transcript abundance	DNA sequence	biLSTM transducer	*acc* = 0.67	*E. coli*, human
(Grossman et al., 2017)	Shallow	Enhancer	Transcript abundance	Motifs and pairwise motif interactions	L1-regularized LR	R^2 = 0.38 (natural), 0.52 (synthetic)	Human
(Lee et al., 2011)	Shallow	Enhancer	Enhancer prediction	k-mers	SVM classification	AUC = 0.93	Human
(Min et al., 2017)	Deep	Enhancer	Enhancer prediction	DNA sequence	CNN classification	AUPRC = 0.92	Human
(Cohn et al., 2018)	Deep	Enhancer	Enhancer prediction	DNA sequence	CNN classification	AUC = 0.92	17 mammalian species including human
(Niu et al., 2019)	Deep	Enhancer	Transcript abundance	DNA sequence	CNN regression	AUC = 0.92	Human
(Chen and Capra, 2020)	Deep	Enhancer	Multitask regulatory properties	DNA sequence	Deep residual NN classification	AUPRC = 0.98	Human
(Lubliner et al., 2015)	Shallow	Promoter	Core promoter activity via reporter fluorescence	k-mers	LR	R^2 = 0.72	Yeast
(Urtecho et al., 2019)	Shallow	Promoter	mRNA abundance	σ factor binding sites	NN regression	R^2 = 0.96	*E. coli*
(Einav and Phillips, 2019)	Shallow	Promoter	mRNA abundance	σ factor binding sites	Biophysical model	R^2 = 0.91	*E. coli*
(de Boer et al., 2020)	Shallow	Promoter	Protein abundance	TF binding and sequence features	L2-regularized multiple LR	R^2 = 89 (natural), 94 (synthetic)	Yeast
(Hossain et al., 2020)	Shallow	Promoter	mRNA abundance	TF binding and sequence features	L1-regularized multiple LR	R^2 = 0.49	*E. coli*, yeast
(Leiby et al., 2020)	Deep	Promoter	Transcription initiation rate	DNA sequence	CNN regression	R^2 = 0.90	*E. coli*
(Kotopka and Smolke, 2020)	Deep	Promoter	Protein abundance	DNA sequence	CNN regression	R^2 = 0.79	Yeast
(Dvir et al., 2013)	Shallow	5′ UTR	Protein levels	DNA sequence features + k-mers	LR	R^2 = 0.52	Yeast
(Bonde et al., 2016)	Shallow	5′ UTR	Protein abundance	RBS features	RF regression	R^2 = 0.89	*E. coli*

(Continued on following page)

TABLE 2 | (*Continued*) Overview of studies modeling gene expression-related properties from separate regulatory or coding regions. Highest achieved or average scores are reported, on test sets where applicable, and include accuracy (*acc*), area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), the coefficient of variation (R^2) and Pearson's correlation coefficient (*r*).

Ref.	Strategy	Region	Target var.	Explan. vars.	Method	Score	Organism
(Salis et al., 2009; Salis, 2011)	Shallow	5′ UTR	Protein abundance	RBS features	Thermodynamic model, LR	R^2 = 0.54 (natural), 0.84 (synthetic)	*E. coli*
(Espah Borujeni et al., 2017)	Shallow	5′ UTR	Translation initiation rate	N-terminal mRNA structures	Biophysical model, LR	R^2 = 0.78	*E. coli*
(Ding et al., 2018)	Shallow	5′ UTR	Protein abundance	DNA sequence activity relationships	Partial least-squares (PLS) regression	R^2 = 0.60 (natural), 0.71 (synthetic)	Yeast
(Decoene et al., 2018)	Shallow	5′ UTR	Translation initiation rate	DNA sequence features	PLS regression	R^2 = 0.73	Yeast
(Cuperus et al., 2017)	Deep	5′ UTR	Protein abundance	DNA sequence	CNN regression	R^2 = 0.62	Yeast
(Sample et al., 2019)	Deep	5′ UTR	Mean ribosome load	DNA sequence	CNN regression	R^2 = 0.82	Human
(Morse et al., 2017)	Shallow	3′ UTR, terminator	Protein abundance	Nucleosome occupancy score	Weighted LR	R^2 = 0.84	Yeast
(Cambray et al., 2013)	Shallow	Terminator	Termination efficiency	DNA sequence features (12)	Multiple LR	*r* = 0.9	*E. coli*
(Vogel et al., 2010)	Shallow	3′ UTR, terminator	mRNA abundance	k-mers	L1-regularized logistic regression	*r* = 0.41	Yeast
(Bogard et al., 2019)	Deep	3′ UTR	Alternative polyadenylation signals	DNA sequence	CNN regression	R^2 = 0.88	Human

Prokaryotic mRNAs have a ribosome binding site (RBS) located in the 5′ UTR that aids recruitment of the translation machinery (Omotajo et al., 2015). In eukaryotes, mRNAs are modified at their 5′ and 3′ ends to facilitate translation by 5′ capping, which recruits the ribosome to the mRNA, and addition of a 3′ poly(A) tail, promoting higher translation by efficient recycling of ribosomes (Mayr, 2017). The key factors for initiation are ribosome recruitment to the mRNA and correct positioning over the start codon, where the presence of a Kozak sequence in the 5′ UTR also increases the efficiency of translation (Nakagawa et al., 2008; Hinnebusch et al., 2016). Elongation is mostly driven by codon usage, where ribosomes synthesize proteins by concatenating one AA per codon according to the genetic code (Saier, 2019). In the termination phase, release factors terminate translation in response to stop codons and the ribosomes are recycled.

Open Reading Frame and Coding Region

Alternative splicing plays a crucial role for protein diversity in eukaryotic cells and produces several mRNA molecules from a single pre-mRNA molecule with ~95% of human genes (Wilkinson et al., 2020). Conversely, in yeast, ~6% of genes carry introns and very few alternative splice forms exist. RNA splicing requires a mandatory set of splicing signals including: 1) the splice donor site (5'ss) and splice acceptor site (3'ss) that define the exon/intron junction of each intron at the 5′ and 3′ ends, respectively, and are characterized by highly conserved dinucleotides (mainly GT and AG, respectively), and 2) the branch point site, a short and degenerate motif usually located between 18 and 44 bp upstream of 3'ss and as far as 400 bp upstream (Mercer et al., 2015). Alterations of these signals were found to be the most frequent cause of hereditary disease (Anna and Monika, 2018). Since 5'ss and 3'ss sequences are well characterized, reliable tools dedicated to splice site predictions have emerged, such as the logistic regression-based "SPiCE" (Leman et al., 2018), trained on 395 splice-site variants of 11 human genes, which achieved an accuracy of 95.6% and correctly predicted the impact on splicing for 98.8% of variants (**Table 2**). To predict the position of splice sites on long genomic sequences, "SpliceRover" (Zuallaert et al., 2018) and "SpliceFinder" (Wang et al., 2019) were developed using CNNs, both outperforming existing splice site prediction tools. SpliceRover achieved ~10% improvement over an existing SVM-based model (Sonnenburg et al., 2007) (AUPRC = 0.61 vs. 0.54) and SpliceFinder compared favourably to both LSTM and SVM-based approaches (AUC of 0.98 vs. 0.95 and 0.93, respectively). A deeper, 32-layer CNN termed "SpliceAI" that accurately predicts splice junctions in pre-mRNAs was developed by Jaganathan and co. (Jaganathan et al., 2019), enabling precise prediction of noncoding genetic variants that cause cryptic splicing and outperforming shallow methods (AUPRC = 0.98 vs. 0.95). The study also found that splice-altering mutations are significantly enriched in patients with rare genetic disorders, causing an estimated 9–11% of pathogenic mutations. For identification of relevant branch points, the method "Branchpointer" (Signal et al., 2018) based on gradient boosting machines showed the best performance to detect the branch points upstream of constitutive and alternative 3'ss (accuracy of 99.48 and 65.84%, respectively). Alternatively, for variants occurring in a branch point area, the mixture-model based "BPP" (Zhang et al., 2017a) emerged as having the best performance to predict effects on mRNA splicing, with an accuracy of 89.17%. Interestingly, two deep learning methods based on bidirectional LSTMs, "LaBranchoR" (Paggi and Bejerano, 2018) and "RNABPS" (Nazari et al., 2019), both performed worse than the above shallow methods when assessed on large scale datasets (AUC of 0.71 and 0.81, respectively, vs. 0.82 with BPP using constitutive 3'ss) (Leman et al., 2020).

Further deep learning studies on alternative splicing prediction have shown that a comprehensive splicing code should include not only genomic sequence features but also epigenetic properties. For instance, 16 histone modifications were used with a multi-label DNN for human embryonic stem cell differentiation in an approach termed "DeepCode" (Xu et al., 2017), achieving an AUPRC up to 0.89. Lee and co. (Lee et al., 2020) built an interpretable RNN that mimics the physical layout of splicing regulation, where the chromatin context progressively changes as the RNAP moves along the guide DNA, achieving an AUPRC of over 0.8 and showing that adjacent epigenetic signals carry useful information in addition to the actual nucleotide sequence of the guide DNA strand. Finally, to enable the characterization of differential alternative splicing between biological samples based on RNA-seq datasets even with modest coverage, the approach DARTS (Zhang et al., 2019) was developed based on a DNN and a Bayesian statistical framework used to determine the statistical significance of differential splicing events in RNA-seq data across biological conditions.

The genetic code is degenerate as most AAs are coded by multiple codons, and these codons would appear in equal frequencies if use of specific codons would not amount to any change in cellular fitness. However, the unequal use of codons that decode the same AA, termed codon usage bias (CUB), cannot be explained by mutation bias alone and is generally believed to arise from selection for improved translational efficiency (Plotkin and Kudla, 2011). Due to variations in transfer RNA (tRNA) abundances, favoring the usage of codons that correspond to more abundant tRNA can lead to faster translation. Such codons are preferred or "optimal" for translation speed up (termed codon optimality) (Hershberg and Petrov, 2008). This is supported by multiple findings in both prokaryotes and eukaryotes, showing that CUB correlates with translation efficiency (protein numbers per mRNA) (Tuller et al., 2010), certain protein structural motifs and tRNA levels (Hanson and Coller, 2018), and affects mRNA translation initiation rates and elongation rates. Furthermore, CUB indices of genes, such as the codon adaptation index (CAI) (Sharp and Li, 1987; Carbone et al., 2003), tend to correlate with the genes' expression (Ghaemmaghami et al., 2003). The role of the coding region extends beyond codon usage, however. mRNA structure was found to regulate translation (Yu et al., 2019) and mRNA hairpins can obstruct translation and override the effect of codon usage bias on translation (Cambray et al., 2018).

The strong association of mRNA levels with protein expression in a variety of organisms (Schwanhäusser et al., 2011; Csárdi et al., 2015; Liu et al., 2016) indicates a more complex background process. The selection pressure for increased protein expression can manifest in changes of DNA that optimize both translation and transcription, improving protein expression and mRNA levels, respectively. Multiple lines of recent evidence corroborate this dual role of synonymous codon changes in transcription and translation, suggesting that selection is shaping codon usage not only to optimize translational efficiency, but in response to conditions imposed by the transcription machinery as well as the physical properties of mRNA (Zhou et al., 2016; Zhou et al., 2018b). For instance, in fungi, codon optimization was found to increase mRNA and protein levels in a promoter-independent manner (Zhou et al., 2016), with CUB shown to be predictive of mRNA and protein levels, affect mRNA stability (Presnyak et al., 2015) and toxicity (Mittal et al., 2018), coevolve with transcription termination (Zhou et al., 2018b) as well as be influenced by mRNA local secondary structure (Trotta, 2013). Similarly, in *E. coli*, CUB was found to affect mRNA stability by defining mRNA folding at the ribosomal site (Kudla et al., 2009).

Multiple modeling studies have been performed to analyze the causes and effects of CUB as well as to find ways to optimize codon usage in order to boost gene expression levels. Codon optimization is a mature field with tools readily available on most biotechnology and DNA synthesis companies' websites (e.g. www.thermofisher.com, www.genewiz.com, www.twistbioscience.com) as well as in standalone solutions (Puigbò et al., 2007; Gould et al., 2014; Rehbein et al., 2019). Most existing optimization strategies are based on biological indices, such as CAI (Sharp and Li, 1987; Puigbò et al., 2007), and use the host's preferred codons to replace less frequently occurring ones, while also adjusting the new sequences to match the natural codon distribution in order to preserve the slow translation regions that are important for protein folding (Richardson et al., 2006; Angov et al., 2008; Hershberg and Petrov, 2009; Gaspar et al., 2012). Standard codon usage metrics were shown to be highly predictive of protein abundance. For instance, an AdaBoost model trained on a number of codon usage metrics in *S. cerevisiae* genes coding for high-abundance proteins (top 10%) and low-abundance proteins (lowest 10%) was highly predictive of these extremes of protein abundance (R^2 = 0.95) (Ferreira et al., 2020).

However, while explicitly modeling existing frequency-based indices has helped to engineer high-yield proteins, it is unclear what other biological features (e.g. RNA secondary structure) should be considered during codon selection for protein synthesis maximization. To address this issue, inspired by natural language processing, deep learning was recently also applied to model CUB. Fujimoto and co. (Fujimoto et al., 2017) showed that their biLSTM-based deep language model that "translates" from DNA to optimal codon sequences, is more robust than existing frequency-based methods due to its reliance on contextual information and long-range dependencies. Similarly, a biLSTM-Transducer model of codon distribution in highly expressed bacterial and human transcripts was able to predict the next codon in a genetic sequence with improved accuracy and lower perplexity on a held out set of transcripts, outperforming previous state-of-the-art frequency-based approaches (accuracy of 0.67 vs. 0.64) (Yang et al., 2019). Another deep learning-based codon optimization approach introduced the concept of *codon boxes*, enabling DNA sequences to be transformed into codon box sequences, while ignoring the order of bases, and thus converting the problem of codon optimization to sequence annotation of corresponding AAs with codon boxes (Fu et al., 2020). Sequences optimized by these biLSTM codon optimization models with ones optimized by Genewiz and ThermoFisher were compared using protein expression experiments in *E. coli*, demonstrating that the method is efficient and competitive.

Alternatively, an algorithmic approach to replacing codons by the target organism's preferred codons was developed by Trösemeier and co. (Trösemeier et al., 2019), termed "COSEM," which simulates ribosome dynamics during mRNA translation and informs about protein synthesis rates per mRNA in an organism and context-dependent way. Protein synthesis rates from COSEM were integrated with further relevant covariates such as translation accuracy into a protein expression score that was used for codon optimization, with further algorithmic fine-tuning implemented in their software "OCTOPOS." The protein expression score produced competitive predictions on proteomic data from prokaryotic and eukaryotic expression systems and was shown to be superior to standard methods, achieving 3-fold increases in protein yield compared to wildtype and commercially optimized sequences (Trösemeier et al., 2019). Moreover, since ribosomes do not move uniformly along mRNAs, Tunney and co. (Tunney et al., 2018) modeled the variation in translation elongation by using a shallow NN to predict the ribosome density at each codon as a function of its sequence neighborhood. This enabled them to study sequence features affecting translation elongation and to design synonymous variants of a protein coding sequence in budding yeast that closely tracked the predicted translation speeds across their full range *in vivo*, demonstrating that control of translation elongation alone is sufficient to produce large quantitative differences in protein output.

Enhancer and Promoter

Transcriptional enhancers are located upstream of the transcription start site (TSS) and regulate spatiotemporal tissue-specific gene expression patterns over long genomic distances, which is achieved through the binding of TFs to cognate motifs (Shlyueva et al., 2014). They can typically be found farther away from the TSS with increasing genomic complexity of the organism (Mora et al., 2016; Clément et al., 2018; Zicola et al., 2019), as far as a million bps in mammals (Pennacchio et al., 2013). Enhancer function and TF binding are influenced by various features, such as the chromatin state of the genomic locus, binding site affinities, activity of bound TFs as well as interactions among TFs (Shlyueva et al., 2014; Chen and Capra, 2020). The nature of how TF interactions influence enhancer function was explored in a recent systematic analysis using *in vivo* binding assays with 32,115 natural and synthetic enhancers (Grossman et al., 2017). The activity of enhancers that contain motifs for PPARγ, a TF that serves as a key regulator of adipogenesis, were shown to depend on varying contributions from dozens of TFs in their immediate vicinity. Importantly, different pairs of motifs followed different interaction rules, including subadditive, additive, and superadditive interactions among specific classes of TFs, with both spatially constrained and flexible grammars.

One of the key ML tasks shedding new light on DNA features affecting enhancer function is identification of enhancer regions in genomic sequences. For instance, a k-mer based SVM framework was able to accurately identify specific types of enhancers (EP300-bound) using only genomic sequence features (Lee et al., 2011), outperforming PWM-based classifiers (AUC = 0.93 vs. 0.87). The predictive sequence features identified by the SVM classifier revealed both enriched and depleted DNA sequence elements in the enhancers, many of which were found to play a role in specifying tissue-specific or developmental-stage-specific enhancer activity, and others that operate in a general or tissue-independent manner. The first deep learning approach to facilitate the identification of enhancers, termed "DeepEnhancer" (Min et al., 2017), relied purely on DNA sequences to predict enhancers using CNNs and transfer learning to fine-tune the model on cell line-specific enhancers. The method was superior to gkm-SVM by ~7% in both AUC and AUPRC scores, and visualizing CNN kernels as sequence logos identified motifs similar to those in the JASPAR database (Khan et al., 2018). Similarly, Cohn and co. (Cohn et al., 2018). trained deep CNNs to identify enhancer sequences in 17 mammalian species using simulated sequences, *in vivo* binding data of single TFs and genome-wide chromatin maps of active enhancers. High classification accuracy was obtained by combining two training strategies that identified both short (1–4 bp) low-complexity motifs and TFBS motifs unique to enhancers. The performance improved when combining positive data from all species together, demonstrating how transfer of learned parameters between networks trained on different species can improve the overall performance and supporting the existence of a shared mammalian regulatory architecture. Although identification of enhancer locations across the whole genome is necessary, it can be more important to predict in which specific tissue types they will be activated and functional. The existing DNNs, though achieving great successes in the former, cannot be directly employed in tissue-specific enhancer predictions because a specific cell or tissue type only has a limited number of available enhancer samples for training. To solve this problem, Niu and co. (Niu et al., 2019) employed a transfer learning strategy, where models trained for general enhancer predictions were retrained on tissue-specific enhancer data and achieved a significantly higher performance (geometric mean of precision and recall, GM = 0.81 vs. 0.70), also surpassing gkm-SVM (GM = 0.53). Interestingly, a very small amount of retraining epochs (~20) were required to complete the retraining process, giving insight into the tissue-specific regulatory rewiring and suggesting that tissue specific responses are mediated by precise changes on a small subset of binding features.

Promoters are adjacent regions directly upstream, as well as a short distance downstream, of the TSS typically spanning from 50 to a couple of 100 bp (Sharon et al., 2012; Redden and Alper, 2015). Besides TFBS and enhancers, they contain core promoters (Lubliner et al., 2015; Haberle and Stark, 2018) in eukaryotes and σ factor binding sites (Feklístov et al., 2014) in prokaryotes, to which the RNAP is recruited and where it acts to initiate transcription. The core promoter contains several motifs with fixed positioning relative to the TSS (Haberle and Stark, 2018), including: 1) the TATA-box motif (consensus 5′-TATAWAW-3′), located ~30 bp upstream of TSS and conserved from yeast to humans but found only in a minority of core promoters, 2) the initiator (Inr) motif, which directly overlaps the TSS and is more abundant than the TATA-box but not universal, with differing consensus sequence among organisms, 3) the downstream promoter element (DPE) that can accompany Inr in promoters that lack a TATA-box and is positioned downstream of the TSS, and 4) other motifs with defined positions relative to the TSS, including TFIIB recognition elements (BREs) and downstream core elements (DCEs) in humans (Watson et al., 2008; Haberle and Stark, 2018). A comprehensive study of yeast core promoter activity and TSS locations in thousands of native and designed sequences (Lubliner et al., 2015) showed that core promoter activity is highly correlated to that of the entire promoter and is in fact predictable from the sequence variation in core promoters (R^2 up to 0.72). Interestingly, orthologous core promoters across yeast species have conserved activities, with transcription initiation in highly active core promoters focused within a narrow region and location, orientation, and flanking bases critically affecting motif function. De Boer and co. (de Boer et al., 2020) recently transcended the limitations of using native and engineered sequences with insufficient scale, instead measuring the expression output of >100 million fully random synthetic promoter sequences in yeast. Using shallow ML they built interpretable models of transcriptional regulation that predicted 94 and 89% of the expression driven from independent test promoters and native yeast promoter fragments, respectively, with a deep model mentioned to have achieved 96%. These models allowed them to characterize each TF's specificity, activity and interactions with chromatin, showing that expression level is influenced by weak regulatory interactions, which confound designed-sequence studies, further supporting that interactions between elements in regulatory regions play an important role in orchestrating gene expression. Moreover, based on promoter libraries comprising >1,000,000 constitutive and inducible promoters and using deep learning, Kotopka and Smolke (Kotopka and Smolke, 2020) developed accurate predictors of promoter activity (R^2 = 0.79) that were used for model-guided design of large, sequence-diverse promoter sets, confirmed to be highly active *in vivo*.

Prokaryotic promoters are marked by σ factor binding sites with five distinct motifs controlling transcription initiation rates by mediating RNAP recruitment: the −35, extended −10, −10, and discriminator motifs recognized by σ; and the UP element recognized by other RNAP domains (Browning and Busby, 2004; Feklístov et al., 2014). The −35 (consensus 5′-TTGACA-3′) and −10 motifs (consensus 5′-TATAAT-3′) are the most abundant, though the extended −10 motif can supplant −35 for initiation, both of which are recognized as dsDNA, with the remaining motifs recognized as ssDNA (Feklístov et al., 2014). By building and testing a library of 10,898 σ70 promoter variants consisting of combinations of −35, −10 and UP elements, spacers, and backgrounds in *E. coli* (Urtecho et al., 2019), the −35 and −10 sequence elements were shown to explain over 95% of the variance in promoter strength using a shallow NN. This was an improvement over using a simple log-linear statistical model, which explained ~74% of the variance, likely due to capturing nonlinear interactions with the spacer, background, and UP elements. Based on the same data from Urtecho and co. (Urtecho et al., 2019), the central claim in energy matrix models of gene expression, stating that each promoter element contributes independently and additively to gene expression and contradicting experimental measurements, was tested using biophysical models (Einav and Phillips, 2019). A "multivalent" modeling framework incorporated the effect of avidity between the −35 and −10 RNAP binding sites and could successfully characterize the full suite of gene expression data (R^2 = 0.91), suggesting that avidity represents a key physical principle governing RNAP-promoter interaction, with overly tight binding inhibiting gene expression. Another use of the data by Urtecho and co. (Urtecho et al., 2019) was with deep learning, where CNN models were trained to predict a promoter's transcription initiation rate directly from its DNA sequence without requiring expert-labeled sequence elements (Leiby et al., 2020). The model performed comparably to the above shallow models (R^2 = 0.90) and corroborated the consensus −35, −10 and discriminator motifs as key contributors to σ70 promoter strength. Similarly, using a "Nonrepetitive Parts Calculator" to rapidly generate and experimentally characterize thousands of bacterial promoters with transcription rates that varied across an almost 1e6-fold range, a ML model was built to explain how specific interactions controlled the promoters' transcription rates, supporting that the number of −35 and −10 motif hexamer mismatches is a potent sequence determinant (Hossain et al., 2020).

5′ Untranslated Region

The key known sequence elements affecting gene expression in 5′ UTRs are the RBS, known as the Shine-Dalgarno sequence, in prokaryotes (Omotajo et al., 2015) and the Kozak sequence in eukaryotes (Nakagawa et al., 2008). The Shine-Dalgarno sequence is a ~6 bp highly conserved sequence (consensus 5′-AGGAGG-3′) (Shine and Dalgarno, 1975) located 3–9 bp from the start codon, which aids recruitment of the ribosome to the mRNA and has a strong effect on the translation initiation rate, thus being highly predictive of expression (Bonde et al., 2016). In order to design synthetic RBS and enable rational control over protein expression levels, the "RBS calculator" was developed a decade ago (Salis et al., 2009; Salis, 2011). Experimental validations in *E. coli* showed that the method is accurate to within a factor of 2.3 over a range of 100,000-fold (R^2 = 0.54 on natural sequences and 0.84 on synthetic ones), correctly predicting the large effects of genetic context on identical RBS sequences that result in different protein levels. The tool was further expanded in a subsequent study (Espah Borujeni et al., 2017), where the N-terminal mRNA structures that need to be unfolded by the ribosome during translation initiation were precisely determined by designing and measuring expression levels of 27 mRNAs with N-terminal coding structures with varying positioning and energetics. The folding energetics of the N-terminal mRNA structures were determined to control translation rates only when the N-terminal mRNA structure overlaps with the ribosomal footprint, which extends 13 nucleotides past the start codon. By utilizing this improved quantification of the ribosomal footprint length, their biophysical model could more accurately predict the translation rates of 495 characterized mRNAs with diverse sequences and structures (R^2 = 0.78). The contribution of the Shine-Dalgarno sequence to protein expression was further comprehensively assessed and used to develop the tool

"EMOPEC," which can modulate the expression level of any *E. coli* gene by changing only a few bases (Bonde et al., 2016). Measured protein levels for 91% of the designed sequences were within twofold of the desired target levels, and predictions of these levels with RF regressors wastly outperformed RBS calculator with an R^2 of 0.89 compared to 0.44.

In eukaryotes, the nucleotide composition of the 5′ UTR changes across genes and species, with highly expressed genes in *S. cerevisiae* preferring A-rich and G-poor 5′ UTRs. The Kozak sequence, which helps to initiate translation in most mRNAs and occupies the first 6–9 nucleotides upstream of the START codon AUG, thus has the consensus 5′-WAMAMAA-3′ in yeast (Li et al., 2017a), whereas in humans this is 5′-GCCGCCRMC-3′ (Nakagawa et al., 2008). Measurement of protein abundance in 2,041 5′-UTR sequence variants, differing only in positions −10 to −1, showed that in yeast, key regulatory elements, including AUG sequence context, mRNA secondary structure, nucleosome occupancy and out-of-frame upstream AUGs conjointly modulate protein levels (Dvir et al., 2013). Based on these features, a predictive model could be developed that explains two-thirds of the expression variation. Recently, however, it was shown that also nucleotides upstream of the Kozak sequence are highly important (Li et al., 2017a). Ding and co. (Ding et al., 2018) synthesized libraries of random 5′ UTRs of 24 nucleotides and used a mathematical model accounting for strong epistatic interactions among bases to predict protein abundance. Then, by stepwise engineering the 5′ UTRs according to nucleotide sequence activity relationships (NuSAR), through repeated cycles of backbone design, directed screening, and model reconstruction, the predictive accuracy of the model was improved (R^2 = 0.71 vs. initial 0.60), resulting in strong 5′ UTRs with 5-fold higher protein abundance than the initial sequences. Similarly, a computational approach for predicting translation initiation rates, termed "yUTR calculator," was developed using partial least-squares (PLS) regression and multiple predictive features, including presence of upstream AUGs (Decoene et al., 2018). This enabled the *de novo* design of 5′ UTRs with a diverse range of desired translation efficiencies, which were confirmed *in vivo*. Moreover, the importance of mRNA secondary structures in 5′ UTRs (Leppek et al., 2018) was also confirmed by inserting hairpin RNA structures into mRNA 5′ UTRs, which tuned expression levels by 100-fold by inhibiting translation (Weenink et al., 2018). This enables generating libraries with predicted expression outputs.

To facilitate deep learning of 5′ UTR function in yeast, a library of half a million 50 bp random 5′ UTRs was constructed and their activity assayed with growth selection experiments (Cuperus et al., 2017). A CNN model was generated that could accurately predict protein levels of both random and native sequences (R^2 = 0.62), and was used to evolve highly active 5′ UTRs that were experimentally confirmed to lead to higher protein expression rates than the starting sequences. Similarly, in human cells, polysome profiling of a library of 280,000 randomized 5′ UTRs was used to develop a CNN, termed "Optimus 5-Prime," that could quantitatively capture the relationship between 5′ UTR sequences and their associated mean ribosome load (R^2 = 0.93 vs. 0.66 with k-mer based LR) (Sample et al., 2019). Combined with a genetic algorithm, the model was used to engineer new 5′ UTRs that accurately directed specified levels of ribosome loading, and also enabled finding disease-associated SNVs that affect ribosome loading and may represent a molecular basis for disease.

3′ Untranslated Region and Terminator

Regulatory motifs within the 3′ UTR and terminator region influence transcription termination, with 3′ UTR regulating polyadenylation, localization and stability (decay) of mRNA as well as translation efficiency (Barrett et al., 2012; Ren et al., 2017). The 3′ UTR contains both binding sites for regulatory proteins and microRNAs that can decrease gene expression by either inhibiting translation or directly causing mRNA degradation. It carries the A-rich 'positioning' element (consensus 5′-AAWAAA-3′ in yeast and 5′-AATAAA-3′ in humans) that directs addition of several hundred adenine residues called the poly(A) tail to the end of the mRNA transcript - the poly(A) site 5′-Y(A)n-3′, the TA-rich 'efficiency' element (most frequently 5′-TATWTA-3′) upstream of the positioning element and multiple T-rich sites (Guo and Sherman, 1996; Zhao et al., 1999; Curran et al., 2015). Based on these motifs, Curran and co. (Curran et al., 2015) developed a panel of short 35–70 bp synthetic terminators for modulating gene expression in yeast, the best of which resulted in a 3.7-fold increase in protein expression compared to that of the common CYC1 terminator. Further investigation of the effects of 13,000 synthetic 3′ end sequences on constitutive expression levels in yeast showed that the vast majority (~90%) of strongly affecting mutations localized to a single positive TA-rich element, similar to the efficiency element (Vogel et al., 2010). Based on the strength of this element, dependent also on the GC content of the surrounding sequence, their classification model could explain a significant amount of measured expression variability in native 3′ end sequences (r = 0.41). Moreover, similarly as with promoters (Curran et al., 2014), Morse and co. (Morse et al., 2017) showed that terminator function can be modulated on the basis of predictions of nucleosome occupancy, with LR models highly predictive of protein output based on nucleosome occupancy scores (R^2 = 0.84). Designed terminators depleted of nucleosomes achieved an almost 4-fold higher net protein output than their original counterparts, with the main mode of action through increased termination efficiency, rather than half-life increases, suggesting a role in improved mRNA maturation.

Most genes express mRNAs with alternative polyadenylation sites at their 3′ ends (Tian and Manley, 2017), which were found to be remarkably heterogeneous across different yeast species. The polyadenylation pattern is determined by a broad degenerate sequence as well as local sequence reliant on poly(A) residues that can adopt secondary structures to recruit the polyadenylation machinery (Moqtaderi et al., 2013). In humans, alternative polyadenylation leads to multiple RNA isoforms derived from a single gene, and a CNN termed 'APARENT' was trained on isoform expression data from over three million reporters to infer alternative polyadenylation in synthetic and human 3′UTRs (Bogard et al., 2019). APARENT was shown to recognize known sequence motifs for polyadenylation, such as the positioning element, and also discover new ones, enabling the authors to engineer precisely defined polyadenylation signals and study disease-related genetic variants.

Bacterial transcription termination is known to occur *via* two distinct mechanisms: factor-dependent or factor-independent termination. The former relies on a regulatory protein Rho at Rho-dependent terminator sequences and is responsible for ~20% of termination events in *E. coli* (Peters et al., 2009), whereas factor-independent termination accounts for the remaining ~80% of transcription termination events and occurs at defined sequence regions known as "intrinsic terminators" that contain GC-rich regions (Roberts, 2019). Cambray and co. (Cambray et al., 2013) assembled a collection of 61 natural and synthetic intrinsic terminators that encode termination efficiencies across an 800-fold dynamic range in *E. coli* and, by simulating RNA folding, they found that secondary structures extending beyond the core terminator stem are likely to increase terminator activity. They developed linear sequence-function models that can accurately predict termination efficiencies ($r = 0.67$), further improving their performance by excluding terminators encoding the context-confounding structural elements ($r = 0.9$).

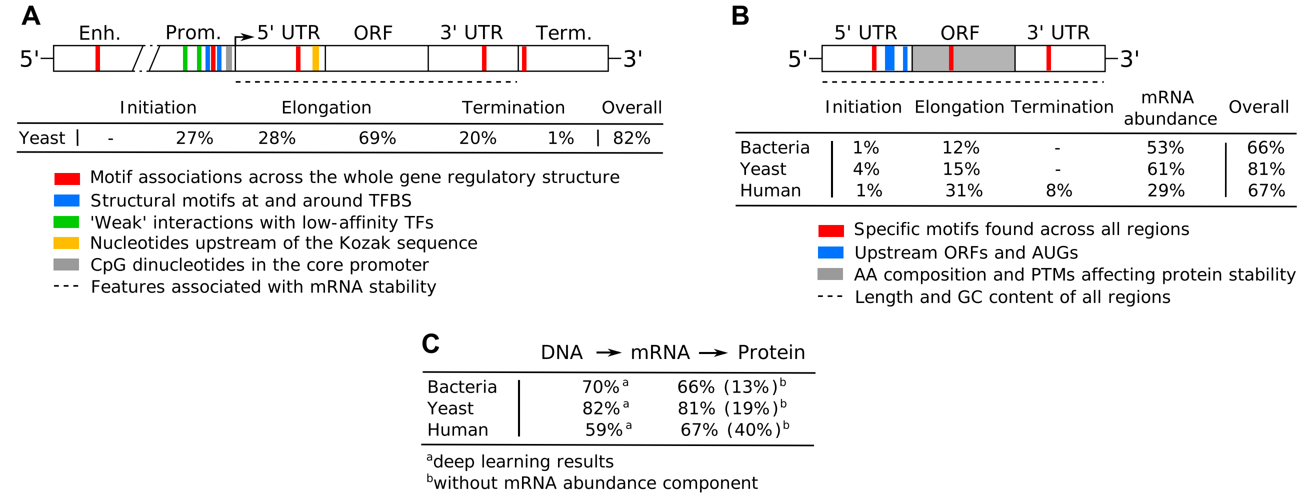
PREDICTING TRANSCRIPT AND PROTEIN LEVELS FROM MULTIPLE REGULATORY PARTS

The whole nucleotide sequence is involved in gene expression. When predicting the outcomes of transcription and translation, e.g. transcript and protein abundance, it is important to consider that many of the underlying steps in these processes are dependent on the outcome of the previous steps and some can occur in tandem (Watson et al., 2008) (**Figures 1B,C**). Each region of the gene and mRNA regulatory structures carries distinct regulatory signals that control the specific enzymatic interactions and thus encodes a significant amount of information related to mRNA (Shalem et al., 2015; Cheng et al., 2017; Cuperus et al., 2017; Zrimec et al., 2020) and protein levels (Vogel et al., 2010; Guimaraes et al., 2014; Lahtvee et al., 2017). Moreover, multiple lines of evidence support that the gene regulatory structure is a coevolving unit in both multicellular (Castillo-Davis et al., 2004; Wittkopp et al., 2004; Hahn, 2007; Wittkopp and Kalay, 2011; Arbiza et al., 2013; Naidoo et al., 2018; Washburn et al., 2019) and unicellular eukaryotes (Tirosh et al., 2009; Park et al., 2012; Chen et al., 2016; Zrimec et al., 2020), as genes display a coupling of coding and regulatory sequence evolution (Wittkopp et al., 2004; Tirosh et al., 2009; Zrimec et al., 2020) with approximately half of all functional variation found in non-coding regions (Hahn, 2007). However, although data from multiple regions was already used in prediction of mRNA and protein levels with shallow models (Vogel et al., 2010; Guimaraes et al., 2014; Lahtvee et al., 2017), predictions based on whole gene regulatory structures spanning multiple kilobases have started to emerge only recently, with the support of deep learning (Washburn et al., 2019; Zrimec et al., 2020). Accounting for multiple regions in ML models can lead to important observations, such as differentiating and quantifying the effects of separate vs. combined regions, and determining the DNA variables across the regions as well as their interactions, which affect predictions (**Figure 3A**).

DNNs are highly useful in learning the regulatory code of gene expression across regulatory structures. Despite hybrid CNN + RNN architectures outperforming them in terms of accuracy, CNNs work sufficiently well for this task (Yu and Koltun, 2015; Gupta and Rush, 2017; Strubell et al., 2017) and excel in learning rich higher-order sequence features that define the *cis*-regulatory grammar (Siggers and Gordân, 2014; Zeng et al., 2016). Systematic analyses of network properties, such as CNN kernel size, number of kernels and number of layers as well as pooling designs (pooling layers between connected CNNs), have exemplified how DNNs decode the regulatory grammar in sequence-based learning tasks (Trabelsi et al., 2019; Zeng et al., 2016; Koo and Eddy, 2019) (**Figure 2D**). In a multilayer DNN, the initial one to two layers capture information on single motif occurrence, with the first layer potentially learning partial motif representations. This can be useful in complicated tasks, such as learning DNA regulatory grammar, because a wider array of representations can be combinatorially constructed from partial representations to capture the rich array of biologically important sequence patterns *in vivo* (Siggers and Gordân, 2014; Zrimec et al., 2020). Successive layers (e.g. Third layer) then learn to recognize motif interactions (i.e. associations in predicting the target variable) across the regulatory structure (Zeng et al., 2016; Zrimec et al., 2020). The extent to which sequence motif representations are learned by first layer kernels is influenced by kernel size and pooling, which enforce a spatial information bottleneck either from the sequence to the CNN or between successive CNN layers, respectively. For instance, large max-pooling (≥10) was shown to force kernels to learn whole motifs, whereas CNNs that employ a low max-pool size (≤4) capture partial motifs (Koo and Eddy, 2019). Similarly, the size of successive convolutional kernels can also affect the ability to assemble whole motifs in deeper layers. Moreover, the number of kernels in the first layer sets a hard constraint on the number of different sequence patterns that can be detected (Koo and Eddy, 2019). Since the scope of initial characterized sequence features limits the range and complexity of grammar representations that can be built downstream, this parameter was generally found to have the greatest impact on CNN performance (Zeng et al., 2016). Therefore, in contrast to learning tasks where the main features are simple, such as occurrence of a PWM-like motif in TFBS prediction, using multiple parts of the regulatory structures requires deeper and more complex architectures that learn distributed motif representations to address the more complex sequence patterns (Koo and Eddy, 2019; Trabelsi et al., 2019) (**Figure 2D**).

Predicting messenger RNA Levels From Nucleotide Sequence

Despite the importance of the whole gene regulatory structure in gene expression, very few combinations of regulatory elements



	Initiation		Elongation		Termination		Overall
Yeast	-	27%	28%	69%	20%	1%	82%

🟥 Motif associations across the whole gene regulatory structure
🟦 Structural motifs at and around TFBS
🟩 'Weak' interactions with low-affinity TFs
🟨 Nucleotides upstream of the Kozak sequence
▨ CpG dinucleotides in the core promoter
--- Features associated with mRNA stability

	Initiation	Elongation	Termination	mRNA abundance	Overall
Bacteria	1%	12%	-	53%	66%
Yeast	4%	15%	-	61%	81%
Human	1%	31%	8%	29%	67%

🟥 Specific motifs found across all regions
🟦 Upstream ORFs and AUGs
▨ AA composition and PTMs affecting protein stability
--- Length and GC content of all regions

	DNA → mRNA	→ Protein
Bacteria	70%[a]	66% (13%)[b]
Yeast	82%[a]	81% (19%)[b]
Human	59%[a]	67% (40%)[b]

[a]deep learning results
[b]without mRNA abundance component

FIGURE 3 | Quantifying gene expression and interpreting its regulatory grammar with machine learning. **(A)** Recently identified DNA regulatory elements predictive of mRNA abundance that expand the base knowledge depicted in **Figure 1B**. These include motif associations (Zrimec et al., 2020) (red), structural motifs (e.g. DNA shape, blue) (Zhou et al., 2015; Yang et al., 2017), weak interactions (de Boer et al., 2020) (green), nucleotides upstream of the Kozak sequence (Li et al., 2017a) (yellow), CpG dinucleotides (Agarwal and Shendure, 2020) (gray) and mRNA stability features (Neymotin et al., 2016; Cheng et al., 2017; Agarwal and Shendure, 2020; Zrimec et al., 2020) (dashed line, see text for details) identified in specific regions or across the whole gene regulatory structure. The table specifies the variation of mRNA abundance explained by DNA sequence and features using deep learning (Zrimec et al., 2020). Note that with alternative approaches, higher predictive values were obtained for certain regions in **Table 2**. **(B)** mRNA regulatory elements recently found to be predictive of protein abundance apart from features depicted in **Figure 1C**. These include specific motifs found across all regions (Li et al., 2019a; Eraslan et al., 2019b) (red), upstream ORFs (Vogel et al., 2010; Li et al., 2019a) and AUGs (Neymotin et al., 2016; Li et al., 2019a) (blue), AA composition (Vogel et al., 2010; Guimaraes et al., 2014) and post-translational modifications (PTMs) (Eraslan et al., 2019b) (gray) as well as lengths and GC content of all regions (Neymotin et al., 2016; Cheng et al., 2017; Li et al., 2019a) (dashed line). The table specifies the variation of protein abundance explained by mRNA levels and translational elements, using comparable shallow approaches in *E. coli* (Guimaraes et al., 2014), *S. cerevisiae* (Lahtvee et al., 2017) and *H. sapiens* (Vogel et al., 2010). Note that with alternative approaches, higher values were obtained for certain regions in **Table 2**. **(C)** Quantifying the central dogma of molecular biology with variance explained by mapping DNA to mRNA levels (Agarwal and Shendure, 2020; Zrimec et al., 2020) and mRNA levels to protein abundance (Vogel et al., 2010; Guimaraes et al., 2014; Lahtvee et al., 2017), using deep and shallow learning, respectively. Note that highly different modeling approaches were used.

have been tested and their functional interactions remain poorly explored. To estimate the contribution of individual regulatory parts in gene expression, a combinatorial library of regulatory elements including different enhancers, core promoters, 5′ UTRs and transcription terminators was constructed in *S. cerevisiae* (Dhillon et al., 2020). A strong interaction was found between enhancers and promoters, showing that, while enhancers initiate gene expression, core promoters modulate the levels of enhancer-mediated expression and can positively or negatively affect expression from even the strongest enhancers. Interestingly, principal component analysis indicated that enhancer and promoter function can be explained by a single principal component. Espinar and co. (Espinar et al., 2018) tested if promoters and coding regions can be understood in isolation, or if they interact, by measuring mRNA levels for 10,000 constructs. The strength of cotranslational regulation on mRNA levels from either inducible or constitutive promoter architecture was explored using LR, where a novel mechanism for co-regulation with inducible promoters was identified (RNA helicase Dbp2), whereas with constitutive promoters, most of the information on mRNA levels was found in the coding region and not in the promoter (**Table 3**). Neymotin and co. (Neymotin et al., 2016) analyzed both coding regions and their interactions with other *cis*-regulatory variables in mRNA transcripts that affect mRNA degradations rates (which in turn affect overall mRNA abundance) using multiple LR. Multiple transcript properties were significantly associated with variation in mRNA degradation rates, including transcript length, ribosome density, CUB and GC content at the third codon position, and a model incorporating these properties explained ~50% of the genome-wide variance. A similar quantitative model based on functional mRNA sequence features explained 59% of the half-life variation between genes, predicting half-life at a median relative error of 30% (Cheng et al., 2017). mRNA sequence features found to most strongly affect mRNA stability included CUB (R^2 = 0.55), destabilizing 3′ UTR motifs, upstream AUG codons, UTR lengths and GC content.

Recently, deep learning was applied on over 20,000 mRNA datasets in seven model organisms that included bacteria, yeast and human, to examine how individual coding and non-coding regions of the gene regulatory structure interact and contribute to mRNA abundance (Zrimec et al., 2020). The CNN-based approach, termed "DeepExpression," could predict the variation of transcript levels directly from DNA sequence in all organisms, with up to 82 and 70% achieved in *S. cerevisiae* and *E. coli*, respectively, outperforming shallow methods by over 13%. Apart from the DNA sequence, CUB and features associated with mRNA stability, including lengths of UTRs and open reading

TABLE 3 | Overview of studies modeling transcript and protein-abundance related properties from combined regulatory and coding regions. Highest achieved or average scores are reported, on test sets where applicable, and include area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), the coefficient of variation (R^2) and Spearman's correlation coefficient (ρ).

Ref.	Strategy	Region	Target var.	Explan. vars.	Method	Score	Organism
(Espinar et al., 2018)	Shallow	Promoter, coding	mRNA abundance	DNA sequence features	LR	R^2 = 0.64	Yeast
(Neymotin et al., 2016)	Shallow	mRNA transcript	mRNA stability (degradation rates)	mRNA features	Multiple LR	R^2 = 0.50	Yeast
(Cheng et al., 2017)	Shallow	mRNA transcript	mRNA stability (half-life)	mRNA features	Multivariate LR	R^2 = 0.59	Yeast
(Zhou et al., 2018a)	Deep	Whole gene regulatory structure	mRNA abundance	DNA sequence	CNN + L2-regularized LR	AUC = 0.82	Human
(Zrimec et al., 2020)	Deep	Whole gene regulatory structure	mRNA abundance	DNA sequence and features	CNN regression	R^2 = 0.82, 0.70, 0.42, respectively	Yeast, *E. coli*, human
(Agarwal and Shendure, 2020)	Deep	Promoter, coding	mRNA abundance	DNA sequence and features	CNN regression	R^2 = 0.59	Human
(Zhang et al., 2020)	Deep	Whole gene regulatory structure	mRNA abundance	DNA sequence	ResNet regression	ρ = 0.80	Human
(Guimaraes et al., 2014)	Shallow	mRNA transcript	Protein abundance	mRNA features	PLS regression	R^2 = 0.66	*E. coli*
(Lahtvee et al., 2017)	Shallow	mRNA transcript	Protein abundance	mRNA features	MARS nonlinear regression	R^2 = 0.81	Yeast
(Vogel et al., 2010)	Shallow	mRNA transcript	Protein abundance	mRNA features	MARS nonlinear regression	R^2 = 0.67	Human
(Li et al., 2019a)	Shallow	mRNA transcript	Translation rates	mRNA features	Multivariate LR	R^2 = 0.81, 0.42, respectively	Yeast, human
(Terai and Asai, 2020)	Shallow	mRNA transcript	Protein abundance	mRNA features of translation initiation	RF regression	ρ = 0.76	*E. coli*
(Li et al., 2017b)	Shallow	mRNA transcript	Translation rates	mRNA features	Bayesian model	R^2 = 0.20 (TR_{mD},; 0.80 (TR_{mIND})	Yeast
(Eraslan et al., 2019b)	Shallow	mRNA transcript	Protein-to-RNA ratio	mRNA sequence and features	Multivariate LR	R^2 = 0.62	Human
(Zhang et al., 2017b)	Deep	mRNA transcript	Translation initiation sites	mRNA sequence	CNN + RNN classification	AUPRC = 0.62	Human
(Zhang et al., 2017c)	Deep	mRNA transcript	Translation elongation dynamics	mRNA sequence	CNN classification	AUC = 0.88, 0.83, respectively	Yeast, human

frames (ORFs), UTR GC content and GC content at each codon position (Neymotin et al., 2016; Cheng et al., 2017), were found to increase the predictive power of the models. Compared to single interpreted DNA motifs, motif associations could explain a much larger portion of the dynamic range of mRNA levels (84 vs. 57%), suggesting that instead of single motifs and regions, the entire gene regulatory structure with specific combinations of regulatory elements defines gene expression levels (**Figure 3A**). This was also supported by observations of co-evolution among coding and non-coding regions across 14 related yeast species. With similar objectives, Agarwal and Schendure (Agarwal and Shendure, 2020) developed "Xpresso," which could explain 59 and 71% of variation in steady-state mRNA levels in human and mouse, respectively, based only on promoter sequences and explanatory features associated with mRNA stability. They showed that Xpresso more than doubles the accuracy of alternative sequence-based models and model interpretation revealed that promoter-proximal CpG dinucleotides strongly predict transcriptional activity.

To predict the tissue-specific transcriptional effects of genome variation, including rare or unseen mutations, Zhou and co. (Zhou et al., 2018a) developed a DNN–based framework termed "ExPecto." Using ExPecto to profile over 140 million promoter-proximal mutations, the authors characterized the regulatory mutation space for human RNAP II–transcribed genes, which enables probing of evolutionary constraints on gene expression and *ab initio* prediction of mutation disease effects. A similar model was constructed using residual networks (ResNets), which are multilayer CNNs that utilize *skip connections* to jump over some layers (He et al., 2016), termed "ExpResNet" (Zhang et al., 2020). By utilizing almost 100 kb of sequence around each gene"s TSS, ExpResNet outperformed existing models, including ExPecto (ρ = 0.80 vs. 0.75), across four tested tissues. Interestingly, by comparing the performance achieved with different input sequence sizes, we can observe that the majority of regulatory information in humans is constrained to ~10 kb of regulatory structure around the TSS (ρ = 0.77, 0.79, 0.80 with 10, 40 and 95 kb, respectively), likely since this is sufficient for the majority of genes, whereas enhancers outside of this region are gene-specific and positioned highly variably.

Predicting Protein Abundance From mRNA Sequence

In multiple organisms, protein levels at steady state are primarily determined by mRNA levels, where up to ~85% of the variation of protein expression can be attributed to mRNA transcription rather than protein translation (Schwanhäusser et al., 2011; Csárdi et al., 2015; Liu et al., 2016). Nevertheless, the spatial and temporal variations of mRNAs and the local availability of resources for protein biosynthesis strongly influence the relationship between protein levels and their transcripts (Liu et al., 2016). Thus, in many scenarios, transcript levels by themselves are not sufficient to predict protein levels and multiple other mRNA-related properties and processes affect translation and define the final gene expression levels. It was also shown that, due to translation rates per mRNA molecule being positively correlated with mRNA abundance, protein levels do not scale linearly with mRNA levels, but instead scale with the abundance of mRNA raised to the power of an "amplification exponent" (Csárdi et al., 2015). Li and co. (Li et al., 2017b) proposed that, to quantify translational control, the translation rate must be decomposed mathematically into two components: one that is dependent on mRNA abundance (TR_{mD}), defining also the amplification exponent, and one that is not (TR_{mIND}). In yeast, TR_{mD} represented ~20% of the variance in translation, whereas TR_{mIND} constituted the remaining ~80% of the variance in translation. The components were also preferentially determined by different mRNA sequence features: TR_{mIND} by the length of the ORF and TR_{mD} by a ~60 nt element spanining the initiating AUG and by CUB, implying that these components are under different evolutionary selective pressures.

Quantification of absolute protein and mRNA abundances for over 1,025 genes from the human Daoy medulloblastoma cell line showed that the combined contribution of mRNA levels and sequence features can explain ⅔ of protein abundance variation at steady state (Vogel et al., 2010) (**Figure 3B**). Using multivariate adaptive regression splines (MARS), a nonlinear regression technique, the variation in protein abundance was primarily explained by translation elongation factors (31%), with an impact similar to that of mRNA abundance (29%). The strongest individual correlates of protein levels were translation and degradation-related features including mRNA sequence length, AA properties, upstream ORFs and 5′ UTR secondary structures. Interestingly, characteristics of the 3′ UTR explained a larger proportion of protein abundance variation (8%) than characteristics of the 5′ UTR (1%). A similar analysis performed with 824 genes in *E. coli*, which used PLS regression and over 100 mRNA sequence features, also derived a model that explained ⅔ of the total variation of protein abundance (Guimaraes et al., 2014). The model suggests that protein abundance is primarily determined by the transcript level (53%) and by effectors of translation elongation (12%), which included both CUB and specific AA composition, whereas only a small fraction of the variation is explained by translation initiation (1%). Lahtvee and co. (Lahtvee et al., 2017). measured absolute abundances of 5,354 mRNAs and 2,198 proteins in yeast under different environmental conditions, showing that the overall correlation between mRNA and protein abundances across all conditions is much higher for a subset of 202 differentially expressed proteins than all of them (avg. $r = 0.88$ vs. 046). On a subset of 1,117 proteins, for which translation efficiencies were calculated, MARS detected that mRNA abundance and translation elongation were the dominant factors controlling protein synthesis, explaining 61 and 15% of its variance, with only a small fraction (4%) explained by translation initiation (**Figure 3B**).

On the other hand, multiple recent studies show that general mRNA features control a much larger fraction of the variance in translation rates or protein abundance than previously realized. For instance, Li and co. (Li et al., 2019a) quantified the contributions of mRNA sequence features to predicting translation rates using LR across multiple organisms, including yeast and human, where they specified 81 and 42% of the variance in translation rates, respectively. The identified informative mRNA features included similar ones as found in previous studies: 5′ UTR secondary structures, nucleotides flanking AUG, upstream ORFs, ORF length and CUB (Vogel et al., 2010; Neymotin et al., 2016; Cheng et al., 2017).

Eraslan and co. (Eraslan et al., 2019b) also showed that a large fraction of protein abundance variation can be predicted from mRNA sequence in humans, by analyzing 11,575 proteins across 29 human tissues using matched transcriptomes and proteomes. Their initial LR model explained on average 22% of the variance from sequence alone, and by including additional experimentally characterized interactions and modifications, including mRNA methylation (Zhao et al., 2017), miRNA and RBP binding sites (Mayr, 2017) and post-translational modifications (Millar et al., 2019), the explained variance increased to 62%. Their findings support much of the previously identified mRNA regulatory elements and also uncover new sequence motifs across the whole transcript. Importantly, they also developed a new metric of codon optimality, termed "Protein-to-mRNA adaptation index" that captures the effects of codon frequency on protein synthesis and degradation. Terai and Asai (Terai and Asai, 2020) evaluated six types of structural features in *E. coli*, including mRNA accessibility, which is the probability that a given region around the start codon has no base-paired nucleotides. When calculated by a log-linear model, accessibility showed the highest correlation with protein abundance. This was significantly higher than the widely used minimum free energy ($\rho = 0.71$ vs. 0.55), and combining it with activity of the Shine-Dalgarno sequence yielded a highly accurate method for predicting protein abundance ($\rho = 0.76$). Moreover, similarly as in eukaryotes, secondary structures in bacterial mRNAs were shown to be highly important for protein production and to generally limit translation initiation in a large-scale assay involving 244,000 designed sequences with varying features (Cambray et al., 2018).

Deep learning was recently applied to prediction of translation initiation sites in a method termed "TITER" (Zhang et al., 2017b), using HTS data quantitatively profiling initiating ribosomes (QTI-seq) at single-nucleotide resolution (Gao et al., 2015). Using a hybrid CNN + RNN approach, TITER integrates the prior preference of TIS codon composition with translation initiation features extracted from the surrounding sequence to greatly outperform other state-of-the-art methods in predicting the initiation sites. The method captures the sequence motifs of different start codons, including a Kozak sequence-like motif for AUG, and quantifies mutational effects on translation initiation.

TABLE 4 | The current advantages, disadvantages and further challenges of machine learning methods in genetics and genomics.

	Deep methods	Shallow methods
Advantages	Lower entry barrier to develop new models and save research time by abstracting mathematical details (Eraslan et al., 2019a)	
	Scale effectively with data and support use of latest computational and technological advances, including large genomic datasets and results of HTS technologies (Barshai et al., 2020)	Classic statistical models are better characterized mathematically and some ML algorithms are easier to understand and explain (Hastie et al., 2013)
	Ability to automatically learn features from raw input data and unlock an additional level of information from it (Barshai et al., 2020; Zrimec et al., 2020)	Less computationally expensive and faster to train leading to more iterations and testing of different techniques in a shorter period of time
	Ability to learn and approximate complex functions without prior assumptions, frequently achieving improved predictive power (Barshai et al., 2020)	Possibility to train on much smaller datasets (e.g. hundreds of examples vs. thousands or more with deep learning) (Playe and Stoven, 2020; Zrimec et al., 2020)
	Capability to integrate multiple pre-processing steps into a single end-to-end model (Eraslan et al., 2019a)	Can be easier to interpret due to inherently interpretable structure and direct feature engineering/selection (**Figures 2A,C**) (Azodi et al., 2020)
	Ability to effectively model multimodal data (Eraslan et al., 2019a)	Usually a small number of hyperparameters (Hastie et al., 2013)
	Highly useful as experiment simulators due to the ability to generalize over an experimental dataset (Barshai et al., 2020)	Useful for proof-of-principle and initial model or parameter testing using only numerical variables
	Easily adaptable to different domains and applications, with transfer learning on pre-trained deep networks accelerating training and improving performance	—
Disadvantages	Dependence on accurately labeled data: cannot achieve higher accuracy than that allowed by the noise inherent to the given experimental target labels (Li et al., 2019b; Barshai et al., 2020)	
	Data driven instead of hypothesis driven modeling (Barshai et al., 2020)	
	Dependence on large amounts of data (at least thousands of training examples) and specialized computational resources (e.g. GPUs)	Dependence on feature engineering
	Potential problems with generalizability, as can be overfit to the experiment rather than biological function (Barshai et al., 2020)	Many different algorithms each with its own advantages and disadvantages can be daunting and require extensive specialized study (Hastie et al., 2013)
	Potential lack of model interpretability (Zou et al., 2019; Barshai et al., 2020)	Cannot unlock information directly from nucleotide sequence (Azodi et al., 2020; Zrimec et al., 2020)
Challenges	Methods to interpret heterogeneous multi-omic and highly dimensional data (Azodi et al., 2020)	
	Methods and high quality datasets to benchmark existing and new interpretation strategies (Azodi et al., 2020)	
	Methods to join findings from multiple interpretation strategies into more complete and coherent interpretations of both models (Azodi et al., 2020) and the studied molecular phenomena	
	Making interpretable ML more accessible to biologists by further lowering the entry barriers and requirements of computational knowledge (Azodi et al., 2020)	

Another DNN framework, termed "ROSE" was used to analyze translation elongation dynamics in both human and yeast *via* ribosome stalling, which is manifested by the local accumulation of ribosomes at specific codon positions of mRNA (Zhang et al., 2017c). ROSE estimates the probability of a ribosome stalling event occurring at each genomic location, achieving higher prediction accuracy than conventional prediction models such as gkm-SVM with AUC increases by up to 18.4%.

DISCUSSION

As can be surmised from the presented ML results (**Table1**, **Table 2**, and **Table 3**), deep methods frequently outperform shallow ones, and we outline the main advantages, disadvantages and challenges of these approaches in **Table 4**. The capability of DNNs to more accurately recapitulate experimental data stems mainly from their ability to extract information directly from the raw input nucleotide sequences, automatically learning regulatory grammar (**Figures 2B,D**), which boosts predictive accuracy (Zrimec et al., 2020). However, although multiple different methods exist for interpreting deep methods, many are a work in progress and no explicit solutions currently exist to benchmark these methods or to combine the findings into more complete and coherent interpretations (Azodi et al., 2020). Nevertheless, ML in general lowers the entry barrier to development of new models and saves research time by abstracting mathematical details (Eraslan et al., 2019a), though this has also been used to criticize such approaches, as they rely on data driven instead of hypothesis driven modeling (Barshai et al., 2020). An important limitation of all ML methods is their dependence on accurately labeled data, since they cannot achieve higher accuracy than that allowed by the noise inherent to the given target labels (Li et al., 2019b; Barshai et al., 2020). For instance, *in vivo* measurements, such as those produced by ChIP-seq, ATAC-seq and DNAse-seq, are prone to experimental noise and technological artifacts and subject to the complexity of the cellular environment, affected by chromatin structure and nucleosome positioning, thus concealing the full picture of DBP-DNA interactions. Alternatively, *in vitro* methods, such as PBMs, HT-SELEX and BunDLE-seq, can capture purely direct protein-nucleic acid interactions or cooperative binding of specific factors and allow sampling of the full spectrum of binding sites (Barshai et al., 2020). Fortunately, novel computational methods allow researchers to easily estimate the noise-constrained upper bound of ML regression model performance (Li et al., 2019b).

Despite the knowledge that whole regulatory structures are involved in gene expression, the majority of approaches still focus only on single regulatory or coding regions. For instance, with mRNA abundance prediction, the contribution of the separate parts of the gene regulatory structure has been quantified only in yeast (Zrimec et al., 2020) (**Figure 3A**). The results across the remaining studies are highly variable, likely due to using very different methods and protocols (**Table 2** and **Table 3**). The trend of using whole regulatory structures is however more common with protein abundance prediction, where, apart from mRNA abundance, also the parts involved in translational processing have been quantified across all three major model organisms (**Figure 3B**). Nevertheless, both the fact that these studies were performed using classical shallow models as well as results from other studies suggest that there is potential for improvement. For instance, results from multiple studies focusing on individual regions show that a much higher amount of information can be extracted from these regions [**Table 2**: e.g. 62% of protein abundance variation explained from yeast 5′ UTRs with DNNs (Decoene et al., 2018)] than was achieved with shallow learning on whole mRNAs in **Figure 3B**. Based on other results, we can also presume that it is possible to not only further boost predictive performance but also uncover new mRNA regulatory grammar.

Pooling the highest-scoring results across organisms in an information-centric view of the central dogma of molecular biology (**Figure 3C**) suggests that about ⅔ of the variation of mRNA and protein levels can be explained from DNA sequence. Unequal approaches were employed however, with deep learning used only with mRNA abundance modeling. Here, the lower results with *H. sapiens* might be a result of accounting for only promoter regions and mRNA stability-associated features in the model (Agarwal and Shendure, 2020), though our own analysis had shown that these stability features alone can explain 38% of the mRNA abundance variation in yeast (Zrimec et al., 2020). Interestingly, by omitting the mRNA abundance component from protein abundance predictions, we can observe the possibility of an increasing trend of explained variance with increasing organism complexity (**Figure 3C**: 13–40% from bacteria to human). This would indicate that mRNAs of multicellular eukaryotes carry more regulatory information involved in translation than those of unicellular eukaryotes and prokaryotes. It might also reflect the fact that gene expression regulation is more intricate in multicellular organisms due to the multiple additional regulatory processes that control expression of a much more complex set of biomolecules and phenotypes than in unicellular organisms (Benelli et al., 2016).

Regulatory information seems to be localized *around* the gene, as multiple studies show that the region spanning <10 kb around the TSS has the largest measurable effect on gene expression, likely as the majority of regulatory signals are clustered in this region in most genes and organisms (Agarwal and Shendure, 2020; Ansariola et al., 2020; Zrimec et al., 2020). Enhancers on the other hand are highly variably spaced and act in a gene-specific manner, which makes them much harder to recognize, and also requires processing enormous sizes of input sequences (e.g. >100 kb upstream of genes in human data) that require more training resources. Therefore, the true effect of such regions is still hard to decipher. Procedures handling larger input sequence sizes or whole genomes will likely lead to improved analysis and quantification of the contributions of enhancers to gene expression control, in relation to other parts of the regulatory structure (Singh et al., 2019; Tang et al., 2020). Another potential trend is building DNNs using biophysical (Tareen and Kinney, 2019) or physicochemical properties (Yang et al., 2017; Liu et al., 2020), as deep models trained on these features might uncover novel patterns in data and lead to improved understanding of the physicochemical principles of protein-nucleic acid regulatory interactions, as well as aid model interpretability. Other novel approaches include: 1) modifying DNN properties to improve recovery of biologically meaningful motif representations (Koo and Ploenzke, 2021), 2) transformer networks (Devlin et al., 2018) and attention mechanisms (Vaswani et al., 2017), widely used in protein sequence modeling (Jurtz et al., 2017; Rao et al., 2019; Vig et al., 2020; Repecka et al., 2021), 3) graph convolutional neural networks, a class of DNNs that can work directly on graphs and take advantage of their structural information, with the potential to give us great insights if we can reframe genomics problems as graphs (Cranmer et al., 2020; Strokach et al., 2020), and 4) generative modeling (Foster, 2019), which may help exploit current knowledge in designing synthetic sequences with desired properties (Killoran et al., 2017; Wang Y. et al., 2020). With the latter, unsupervized training is used with approaches including: 1) autoencoders, which learn efficient representations of the training data, typically for dimensionality reduction (Way and Greene, 2018) or feature selection (Xie et al., 2017), 2) generative adversarial networks, which learn to generate new data with the same statistics as the training set (Wang Y. et al., 2020; Repecka et al., 2021), and 3) deep belief networks, which learn to probabilistically reconstruct their inputs, acting as feature detectors, and can be further trained with supervision to build efficient classifiers (Bu et al., 2017). Moreover, the advent of single-cell HTS technologies such as single-cell RNA-seq will offer many novel research opportunities, including modeling of cell-type or cell-state specific enhancer or TFBS activations and chromatin changes (Angermueller et al., 2017; Gustafsson et al., 2020; Kawaguchi et al., 2021).

To conclude, the application of ML in genomics has augmented experimental methods and facilitated accumulating a vast amount of knowledge on gene expression regulation. DNNs, due to their ability to learn biologically relevant information directly from sequence, while performing similarly to or better than classical approaches, are the method of choice for quantifying gene expression and interpreting the predictive features hidden in nucleotide sequence data. DNN-isolated features can be as predictive as models relying on experimental ChIP-seq data (Agarwal and Shendure, 2020), suggesting that current computational approaches are achieving a level of accuracy that might soon allow substituting wet-lab HTS experiments with fully computational pipelines (Keilwagen et al., 2019). Such pipelines can also become indispensable for analysis of human disease-associated regulatory mutations, identifying clinically relevant noncoding variants and expression perturbations, grouping patients in drug treatment trials, disease subtyping as well as personalized treatment (Zhou et al., 2018a; Dagogo-Jack and Shaw, 2018). Since controlling the expression of genes is also one of the key challenges of synthetic biology, the computational models represent excellent starting points in procedures to predictably design regulatory sequences, control protein expression and fine-tune biosynthetic pathways in both prokaryotic and eukaryotic systems (Nielsen and Keasling, 2016; de Jongh et al., 2020; Wang H. et al., 2020).

For readers willing to learn and apply some of the discussed ML approaches, many excellent resources exist, including: 1) specialized packages for model development and interpretation, such as "DragoNN" (https://kundajelab.github.io/dragonn/) (Movva et al., 2019) "Janggu" (https://github.com/BIMSBbioinfo/janggu) (Kopp et al., 2020) and "Pysster" (https://github.com/budach/pysster) (Budach and Marsico, 2018), 2) repositories of trained models, such as "Kipoi" (https://kipoi.org/) (Avsec et al., 2019), 3) other genomics tutorials and code examples (https://github.com/vanessajurtz/lasagne4bio) (Jurtz et al., 2017), as well as 4) resources with a much broader scope than mere genomics, including online courses (https://www.coursera.org/specializations/deep-learning) and books (https://github.com/ageron/handson-ml2) (Géron, 2019).

AUTHOR CONTRIBUTIONS

JZ, FB, and AZ conceptualized the project; JZ performed the literature research and wrote the paper; FB contributed to the literature research on translation; MK contributed to the introduction and discussion; VG contributed to the literature research and text on codon usage; FB, AZ and VG contributed to writing the final manuscript.

FUNDING

The study was supported by SciLifeLab funding and Swedish Research council (Vetenskapsrådet) starting Grant No. 2019-05356.

ACKNOWLEDGMENTS

We thank Sandra Viknander for insightful discussions on deep neural networks and their applications.

REFERENCES

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., et al. (2015). Deconvolving the Recognition of DNA Shape from Sequence. *Cell* 161, 307–318. doi:10.1016/j.cell.2015.02.008

Agarwal, V., and Shendure, J. (2020). Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep* 31, 107663. doi:10.1016/j.celrep.2020.107663

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* 33, 831–838. doi:10.1038/nbt.3300

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). *Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks*. Ithaca, NY: arXiv [cs.LG].

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning. *Genome Biol.* 18, 67. doi:10.1186/s13059-017-1189-z

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep Learning for Computational Biology. *Mol. Syst. Biol.* 12, 878. doi:10.15252/msb.20156651

Angov, E., Hillier, C. J., Kincaid, R. L., and Lyon, J. A. (2008). Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with Those of the Expression Host. *PLoS One* 3, e2189. doi:10.1371/journal.pone.0002189

Angus, G., and Eyuboglu, S. (2018). "Regulatory Activity Prediction with Attention-Based Models," in 32nd Conference on Neural Information Processing Systems (. NIPS 2018).

Anna, A., and Monika, G. (2018). Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation. *J. Appl. Genet.* 59, 253–268. doi:10.1007/s13353-018-0444-7

Ansariola, M., Fraser, Valerie. N., Filichkin, Sergei. A., Ivanchenko, Maria. G., Bright, Zachary. A., Gould, Russell. A., et al. (2020). *Accurate Transcription Start Sites Enable Mining for the Cis-Regulatory Determinants of Tissue Specific Gene Expression*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. doi:10.1101/2020.09.01.278424

Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., et al. (2013). Genome-wide Inference of Natural Selection on Human Transcription Factor Binding Sites. *Nat. Genet.* 45, 723–729. doi:10.1038/ng.2658

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., et al. (2021). Base-resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax. *Nat. Genet.* 53, 354–366. doi:10.1038/s41588-021-00782-6

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., et al. (2019). The Kipoi Repository Accelerates Community Exchange and Reuse of Predictive Models for Genomics. *Nat. Biotechnol.* 37, 592–600. doi:10.1038/s41587-019-0140-0

Azodi, C. B., Tang, J., and Shiu, S.-H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends Genet.* 36, 442–455. doi:10.1016/j.tig.2020.03.005

Bansal, M., Kumar, A., and Yella, V. R. (2014). Role of DNA Sequence Based Structural Features of Promoters in Transcription Initiation and Gene Expression. *Curr. Opin. Struct. Biol.* 25, 77–85. doi:10.1016/j.sbi.2014.01.007

Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012). Regulation of Eukaryotic Gene Expression by the Untranslated Gene Regions and Other Non-coding Elements. *Cell. Mol. Life Sci.* 69, 3613–3634. doi:10.1007/s00018-012-0990-9

Barshai, M., Tripto, E., and Orenstein, Y. (2020). Identifying Regulatory Elements *via* Deep Learning. *Annu. Rev. Biomed. Data Sci.* 3, 315–338. doi:10.1146/annurev-biodatasci-022020-021940

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). *Attention Augmented Convolutional Networks*. Ithaca, NY: arXiv [cs.CV].

Benelli, D., La Teana, A., and Londei, P. (2016). "Evolution of Translational Initiation: From Archaea to Eukarya," in *Evolution of the Protein Synthesis Machinery and its Regulation*. (Berlin, Germany: Springer), 61–79.

Benveniste, D., Sonntag, H.-J., Sanguinetti, G., and Sproul, D. (2014). Transcription Factor Binding Predicts Histone Modifications in Human Cell Lines. *Proc. Natl. Acad. Sci. U. S. A.* 111, 13367–13372. doi:10.1073/pnas.1412081111

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, Universal DNA Microarrays to Comprehensively Determine Transcription-Factor Binding Site Specificities. *Nat. Biotechnol.* 24, 1429–1435. doi:10.1038/nbt1246

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015). Hyperopt: a Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discov.* 8, 014008. doi:10.1088/1749-4699/8/1/014008

Bishop, E. P., Rohs, R., Parker, S. C. J., West, S. M., Liu, P., Mann, R. S., et al. (2011). A Map of Minor Groove Shape and Electrostatic Potential from Hydroxyl Radical Cleavage Patterns of DNA. *ACS Chem. Biol.* 6, 1314–1320. doi:10.1021/cb200155t

Blackwell, T. K., and Weintraub, H. (1990). Differences and Similarities in DNA-Binding Preferences of MyoD and E2A Protein Complexes Revealed by Binding Site Selection. *Science* 250, 1104–1110. doi:10.1126/science.2174572

Bogard, N., Linder, J., Rosenberg, A. B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* 178, 91–106. doi:10.1016/j.cell.2019.04.046

Bonde, M. T., Pedersen, M., Klausen, M. S., Jensen, S. I., Wulff, T., Harrison, S., et al. (2016). Predictable Tuning of Protein Expression in Bacteria. *Nat. Methods* 13, 233–236. doi:10.1038/nmeth.3727

Browning, D. F., and Busby, S. J. (2004). The Regulation of Bacterial Transcription Initiation. *Nat. Rev. Microbiol.* 2, 57–65. doi:10.1038/nrmicro787

Brukner, I., Sanchez, R., Suck, D., and Pongor, S. (1995). Sequence-dependent Bending Propensity of DNA as Revealed by DNase I: Parameters for Trinucleotides. *EMBO J.* 14, 1812–1818. doi:10.1002/j.1460-2075.1995.tb07169.x

Bu, H., Gan, Y., Wang, Y., Zhou, S., and Guan, J. (2017). A New Method for Enhancer Prediction Based on Deep Belief Network. *BMC Bioinformatics* 18, 418. doi:10.1186/s12859-017-1828-0

Budach, S., and Marsico, A. (2018). Pysster: Classification of Biological Sequences by Learning Sequence and Structure Motifs with Convolutional Neural Networks. *Bioinformatics* 34, 3035–3037. doi:10.1093/bioinformatics/bty222

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nat. Methods* 10, 1213–1218. doi:10.1038/nmeth.2688

Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018). Evaluation of 244,000 Synthetic Sequences Reveals Design Principles to Optimize Translation in *Escherichia coli*. *Nat. Biotechnol.* 36, 1005–1015. doi:10.1038/nbt.4238

Cambray, G., Guimaraes, J. C., Mutalik, V. K., Lam, C., Mai, Q.-A., Thimmaiah, T., et al. (2013). Measurement and Modeling of Intrinsic Transcription Terminators. *Nucleic Acids Res.* 41, 5139–5148. doi:10.1093/nar/gkt163

Carbone, A., Zinovyev, A., and Képès, F. (2003). Codon Adaptation index as a Measure of Dominating Codon Bias. *Bioinformatics* 19, 2005–2015. doi:10.1093/bioinformatics/btg272

Castillo-Davis, C. I., Hartl, D. L., and Achaz, G. (2004). cis-Regulatory and Protein Evolution in Orthologous and Duplicate Genes. *Genome Res.* 14, 1530–1536. doi:10.1101/gr.2662504

Chen, C., Hou, J., Shi, X., Yang, H., Birchler, J. A., and Cheng, J. (2021). DeepGRN: Prediction of Transcription Factor Binding Site across Cell-Types Using Attention-Based Deep Neural Networks. *BMC Bioinformatics* 22, 38. doi:10.1186/s12859-020-03952-1

Chen, J., Darst, S. A., and Thirumalai, D. (2010). Promoter Melting Triggered by Bacterial RNA Polymerase Occurs in Three Steps. *Proc. Natl. Acad. Sci. U. S. A.* 107, 12523–12528. doi:10.1073/pnas.1003533107

Chen, L., and Capra, J. A. (2020). Learning and Interpreting the Gene Regulatory Grammar in a Deep Learning Framework. *Plos Comput. Biol.* 16, e1008334. doi:10.1371/journal.pcbi.1008334

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K. C. (2015). PseKNC-General: a Cross-Platform Package for Generating Various Modes of Pseudo Nucleotide Compositions. *Bioinformatics* 31, 119–120. doi:10.1093/bioinformatics/btu602

Chen, Y., Pai, A. A., Herudek, J., Lubas, M., Meola, N., Järvelin, A. I., et al. (2016). Principles for RNA Metabolism and Alternative Transcription Initiation within Closely Spaced Promoters. *Nat. Genet.* 48, 984–994. doi:10.1038/ng.3616

Cheng, J., Maier, K. C., Avsec, Ž., Rus, P., and Gagneur, J. (2017). Cis-regulatory Elements Explain Most of the mRNA Stability Variation across Genes in Yeast. *RNA* 23, 1648–1659. doi:10.1261/rna.062224.117

Chiu, T.-P., Xin, B., Markarian, N., Wang, Y., and Rohs, R. (2020). TFBSshape: an Expanded Motif Database for DNA Shape Features of Transcription Factor Binding Sites. *Nucleic Acids Res.* 48, D246–D255. doi:10.1093/nar/gkz970

Clément, Y., Torbey, P., and Gilardi-Hebenstreit, P. (2018). *Genome-wide Enhancer-Gene Regulatory Maps in Two Vertebrate Genomes*. Cold Spring Harbor, NY: bioRxiv.

Cohn, D., Zuk, O., and Kaplan, T. (2018). Enhancer Identification Using Transfer and Adversarial Deep Learning of DNA Sequences. *Cold Spring Harbor Lab.* 264200. doi:10.1101/264200

Cranmer, M., Sanchez-Gonzalez, Alvaro., Battaglia, Peter., Xu, Rui., Cranmer, Kyle., Spergel, David., et al. (2020). *Discovering Symbolic Models from Deep Learning with Inductive Biases*. Ithaca, NY: arXiv [cs.LG].

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., and Drummond, D. A. (2015). Accounting for Experimental Noise Reveals that mRNA Levels, Amplified by post-transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *Plos Genet.* 11, e1005206. doi:10.1371/journal.pgen.1005206

Cuperus, J. T., Groves, B., and Kuchina, A. (2017). Deep Learning of the Regulatory Grammar of Yeast 5′ Untranslated Regions from 500,000 Random Sequences. *Genome Res.* 27, 1–10. doi:10.1101/gr.224964.117

Curran, K. A., Crook, N. C., Karim, A. S., Gupta, A., Wagman, A. M., and Alper, H. S. (2014). Design of Synthetic Yeast Promoters *via* Tuning of Nucleosome Architecture. *Nat. Commun.* 5, 4002. doi:10.1038/ncomms5002

Curran, K. A., Morse, N. J., Markham, K. A., Wagman, A. M., Gupta, A., and Alper, H. S. (2015). Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. *ACS Synth. Biol.* 4, 824–832. doi:10.1021/sb5003357

Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour Heterogeneity and Resistance to Cancer Therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi:10.1038/nrclinonc.2017.166

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., and Regev, A. (2020). Deciphering Eukaryotic Gene-Regulatory Logic with 100 Million Random Promoters. *Nat. Biotechnol.* 38, 56–65. doi:10.1038/s41587-019-0315-8

de Jongh, R. P. H., van Dijk, A. D. J., Julsing, M. K., Schaap, P. J., and de Ridder, D. (2020). Designing Eukaryotic Gene Expression Regulation Using Machine Learning. *Trends Biotechnol.* 38, 191–201. doi:10.1016/j.tibtech.2019.07.007

Decoene, T., Peters, G., De Maeseneire, S. L., and De Mey, M. (2018). Toward Predictable 5′UTRs in *Saccharomyces cerevisiae*: Development of a yUTR Calculator. *ACS Synth. Biol.* 7, 622–634. doi:10.1021/acssynbio.7b00366

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Ithaca, NY: arXiv [cs.CL].

Dhillon, N., Shelansky, R., Townshend, B., Jain, M., Boeger, H., Endy, D., et al. (2020). Permutational Analysis of *Saccharomyces cerevisiae* Regulatory Elements. *Synth. Biol.* 5, ysaa007. doi:10.1093/synbio/ysaa007

Dillon, S. C., and Dorman, C. J. (2010). Bacterial Nucleoid-Associated Proteins, Nucleoid Structure and Gene Expression. *Nat. Rev. Microbiol.* 8, 185–195. doi:10.1038/nrmicro2261

Ding, W., Cheng, J., Guo, D., Mao, L., Li, J., Lu, L., et al. (2018). Engineering the 5’ UTR-Mediated Regulation of Protein Abundance in Yeast Using Nucleotide Sequence Activity Relationships. *ACS Synth. Biol.* 7, 2709–2714. doi:10.1021/acssynbio.8b00127

Dvir, S., Velten, L., Sharon, E., and Zeevi, D. (2013). Deciphering the Rules by Which 5′-UTR Sequences Affect Protein Expression in Yeast. *Proc. Natl. Acad. Sci.* 110, E2792–E2801. doi:10.1073/pnas.1222534110

Einav, T., and Phillips, R. (2019). How the Avidity of Polymerase Binding to the -35/-10 Promoter Sites Affects Gene Expression. *Proc. Natl. Acad. Sci. U. S. A.* 116, 13340–13345. doi:10.1073/pnas.1905615116

ENCODE Project Consortium (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B. M., Uhlén, M., et al. (2019). Quantification and Discovery of Sequence Determinants of Protein-per-mRNA Amount in 29 Human Tissues. *Mol. Syst. Biol.* 15. doi:10.15252/msb.20188513

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep Learning: New Computational Modelling Techniques for Genomics. *Nat. Rev. Genet.* 20, 389–403. doi:10.1038/s41576-019-0122-6

Espah Borujeni, A., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., and Salis, H. M. (2017). Precise Quantification of Translation Inhibition by mRNA Structures that Overlap with the Ribosomal Footprint in N-Terminal Coding Sequences. *Nucleic Acids Res.* 45, 5437–5448. doi:10.1093/nar/gkx061

Espinar, L., Schikora Tamarit, M. À., Domingo, J., and Carey, L. B. (2018). Promoter Architecture Determines Cotranslational Regulation of mRNA. *Genome Res.* 28, 509–518. doi:10.1101/gr.230458.117

Roadmap Epigenomics Consortium (2015). Integrative Analysis of 111 Reference Human Epigenomes. *Nature* 518, 317–330. doi:10.1038/nature14248

Feklístov, A., Sharon, B. D., Darst, S. A., and Gross, C. A. (2014). Bacterial Sigma Factors: a Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.* 68, 357–376. doi:10.1146/annurev-micro-092412-155737

Ferreira, M., Ventorim, R., Almeida, E., Silveira, S., and Silveira, W. (2020). *Protein Abundance Prediction through Machine Learning Methods*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. doi:10.1101/2020.09.17.302182

Fletez-Brant, C., Lee, D., McCallion, A. S., and Beerkmer, M. A. S. V. M. (2013). A Web Server for Identifying Predictive Regulatory Sequence Features in Genomic Data Sets. *Nucleic Acids Res.* 41, W544–W556. doi:10.1093/nar/gkt519

Foster, D. (2019). *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. Sebastopol, CA: 'O'Reilly Media, Inc.'.

Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., Zhang, H., et al. (2020). Codon Optimization with Deep Learning to Enhance Protein Expression. *Sci. Rep.* 10, 17617. doi:10.1038/s41598-020-74091-z

Fujimoto, M. S., Bodily, Paul. M., Lyman, Cole. A., Jacobsen, Andrew. J., Snell, Quinn., and Clement, Mark. J. (2017). "Modeling Global and Local Codon Bias with Deep Language Models," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 151–156.

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., and Qian, S.-B. (2015). Quantitative Profiling of Initiating Ribosomes *In Vivo*. *Nat. Methods* 12, 147–153. doi:10.1038/nmeth.3208

Gaspar, P., Oliveira, J. L., Frommlet, J., Santos, M. A. S., and Moura, G. (2012). EuGene: Maximizing Synthetic Gene Design for Heterologous Expression. *Bioinformatics* 28, 2683–2684. doi:10.1093/bioinformatics/bts465

Geggier, S., and Vologodskii, A. (2010). Sequence Dependence of DNA Bending Rigidity. *Proc. Natl. Acad. Sci. U. S. A.* 107, 15421–15426. doi:10.1073/pnas.1004809107

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: 'O'Reilly Media, Inc.'.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., et al. (2003). Global Analysis of Protein Expression in Yeast. *Nature* 425, 737–741. doi:10.1038/nature02046

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features. *Plos Comput. Biol.* 10, e1003711. doi:10.1371/journal.pcbi.1003711

Gibney, E. R., and Nolan, C. M. (2010). Epigenetics and Gene Expression. *Heredity* 105, 4–13. doi:10.1038/hdy.2010.54

Gould, N., Hendy, O., and Papamichail, D. (2014). Computational Tools and Algorithms for Designing Customized Synthetic Genes. *Front. Bioeng. Biotechnol.* 2, 41. doi:10.3389/fbioe.2014.00041

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics* 27, 1017–1018. doi:10.1093/bioinformatics/btr064

Grant, C. E., Johnson, J., Bailey, T. L., and Noble, W. S. (2016). MCAST: Scanning for Cis-Regulatory Motif Clusters. *Bioinformatics* 32, 1217–1219. doi:10.1093/bioinformatics/btv750

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., et al. (2017). Systematic Dissection of Genomic Features Determining Transcription Factor Binding and Enhancer Function. *Proc. Natl. Acad. Sci. U. S. A.* 114, E1291–E1300. doi:10.1073/pnas.1621150114

Guimaraes, J. C., Rocha, M., and Arkin, A. P. (2014). Transcript Level and Sequence Determinants of Protein Abundance and Noise in *Escherichia coli*. *Nucleic Acids Res.* 42, 4791–4799. doi:10.1093/nar/gku126

Guo, Z., and Sherman, F. (1996). 3'-end-forming Signals of Yeast mRNA. *Trends Biochem. Sci.* 21, 477–481. doi:10.1016/s0968-0004(96)10057-8

Gupta, A., and Rush, A. M. (2017). *Dilated Convolutions for Modeling Long-Distance Genomic Dependencies*. Ithaca, NY: arXiv [q-bio.GN].

Gustafsson, J., Held, F., Robinson, J. L., Björnson, E., Jörnsten, R., and Nielsen, J. (2020). Sources of Variation in Cell-type RNA-Seq Profiles. *PLoS One* 15, e0239495. doi:10.1371/journal.pone.0239495

Haberle, V., and Stark, A. (2018). Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation. *Nat. Rev. Mol. Cel Biol.* 19, 621–637. doi:10.1038/s41580-018-0028-8

Hahn, M. W. (2007). Detecting Natural Selection on Cis-Regulatory DNA. *Genetica* 129, 7–18. doi:10.1007/s10709-006-0029-y

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The Lac Repressor Displays Facilitated Diffusion in Living Cells. *Science* 336, 1595–1598. doi:10.1126/science.1221648

Hanson, G., and Coller, J. (2018). Codon Optimality, Bias and Usage in Translation and mRNA Decay. *Nat. Rev. Mol. Cel Biol.* 19, 20–30. doi:10.1038/nrm.2017.91

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer Science & Business Media.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus Enables Improved Detection of *In Vivo* Transcription Factor Binding Footprints. *Nat. Biotechnol.* 33, 395–401. doi:10.1038/nbt.3121

He, W., Jia, C., Duan, Y., and Zou, Q. 70Pro. Pred. (2018). A Predictor for Discovering Sigma70 Promoters Based on Combining Multiple Features. *BMC Syst. Biol.* 12, 44. doi:10.1186/s12918-018-0570-1

He, Y., Shen, Z., Zhang, Q., Wang, S., and Huang, D.-S. (2020). A Survey on Deep Learning in DNA/RNA Motif Mining. *Brief. Bioinform.*, 1–10. doi:10.1093/bib/bbaa229

Hershberg, R., and Petrov, D. A. (2009). General Rules for Optimal Codon Choice. *Plos Genet.* 5, e1000556. doi:10.1371/journal.pgen.1000556

Hershberg, R., and Petrov, D. A. (2008). Selection on Codon Bias. *Annu. Rev. Genet.* 42, 287–299. doi:10.1146/annurev.genet.42.110807.091442

Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. (2016). Translational Control by 5'-untranslated Regions of Eukaryotic mRNAs. *Science* 352, 1413–1416. doi:10.1126/science.aad9868

Hossain, A., Lopez, E., Halper, S. M., Cetnar, D. P., Reis, A. C., Strickland, D., et al. (2020). Automated Design of Thousands of Nonrepetitive Parts for Engineering Stable Genetic Systems. *Nat. Biotechnol.* 38, 1466–1475. doi:10.1038/s41587-020-0584-2

Inukai, S., Kock, K. H., and Bulyk, M. L. (2017). Transcription Factor-DNA Binding: beyond Binding Site Motifs. *Curr. Opin. Genet. Dev.* 43, 110–119. doi:10.1016/j.gde.2017.02.007

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176 (3), 535–548. doi:10.1016/j.cell.2018.12.015

Jayaram, N., Usvyat, D., and R Martin, A. C. (2016). Evaluating Tools for Transcription Factor Binding Site Prediction. *BMC Bioinformatics* 17, 547. doi:10.1186/s12859-016-1298-9

Jiao, Y., and Du, P. (2016). Performance Measures in Evaluating Machine Learning Based Bioinformatics Predictors for Classifications. *Quantitative Biol.* 4, 320–330. doi:10.1007/s40484-016-0081-2

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide Mapping of *In Vivo* Protein-DNA Interactions. *Science* 316, 1497–1502. doi:10.1126/science.1141319

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-binding Specificities of Human Transcription Factors. *Cell* 152, 327–339. doi:10.1016/j.cell.2012.12.009

Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., et al. (2017). An Introduction to Deep Learning on Biological Sequence Data: Examples and Solutions. *Bioinformatics* 33, 3685–3690. doi:10.1093/bioinformatics/btx531

Kawaguchi, R. K., Tang, Z., Fischer, S., Tripathy, R., Koo, P. K., and Gillis, J. (2021). *Exploiting Marker Genes for Robust Classification and Characterization of Single-Cell Chromatin Accessibility*. Cold Spring Harbor, NY: bioRxiv. doi:10.1101/2021.04.01.438068

Keilwagen, J., and Grau, J. (2015). Varying Levels of Complexity in Transcription Factor Binding Motifs. *Nucleic Acids Res.* 43, e119. doi:10.1093/nar/gkv577

Keilwagen, J., Posch, S., and Grau, J. (2019). Accurate Prediction of Cell Type-specific Transcription Factor Binding. *Genome Biol.* 20, 9. doi:10.1186/s13059-018-1614-y

Kelley, D. R. (2020). Cross-species Regulatory Sequence Activity Prediction. *Plos Comput. Biol.* 16, e1008050. doi:10.1371/journal.pcbi.1008050

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018). Sequential Regulatory Activity Prediction across Chromosomes with Convolutional Neural Networks. *Genome Res.* 28, 739–750. doi:10.1101/gr.227819.117

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset. (2016). Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks. *Genome Res.* 26, 990–999. doi:10.1101/gr.200535.115

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and its Web Framework. *Nucleic Acids Res.* 46, D1284. doi:10.1093/nar/gkx1188

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. (2017). *Generating and Designing DNA with Deep Generative Models*. Ithaca, NY: arXiv [cs.LG].

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., et al. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* 128, 1231–1245. doi:10.1016/j.cell.2006.12.048

Koo, P. K., Anand, P., Paul, S. B., and Eddy, S. R. (2018). *Inferring Sequence-Structure Preferences of Rna-Binding Proteins with Convolutional Residual Networks*. Cold Spring Harbor, NY: BioRxiv. doi:10.1101/418459

Koo, P. K., and Eddy, S. R. (2019). Representation Learning of Genomic Sequence Motifs with Convolutional Neural Networks. *Plos Comput. Biol.* 15, e1007560. doi:10.1371/journal.pcbi.1007560

Koo, P. K., and Ploenzke, M. (2020). Deep Learning for Inferring Transcription Factor Binding Sites. *Curr. Opin. Syst. Biol.* 19, 16–23. doi:10.1016/j.coisb.2020.04.001

Koo, P. K., and Ploenzke, M. (2021). Improving Representations of Genomic Sequence Motifs in Convolutional Networks with Exponential Activations. *Nat. Machine Intelligence* 3, 258–266. doi:10.1038/s42256-020-00291-x

Koo, P. K., and Ploenzke, M. (2020). *Interpreting Deep Neural Networks beyond Attribution Methods: Quantifying Global Importance of Genomic Features*. Cold Spring Harbor, NY: bioRxiv.

Kopp, W., Monti, R., Tamburrini, A., Ohler, U., and Akalin, A. (2020). Deep Learning for Genomics Using Janggu. *Nat. Commun.* 11, 3488. doi:10.1038/s41467-020-17155-y

Kotopka, B. J., and Smolke, C. D. (2020). Model-driven Generation of Artificial Yeast Promoters. *Nat. Commun.* 11, 2113. doi:10.1038/s41467-020-15977-4

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence Determinants of Gene Expression in *Escherichia coli*. *Science* 324, 255–258. doi:10.1126/science.1170160

Kumar, A., and Bansal, M. (2017). Unveiling DNA Structural Features of Promoters Associated with Various Types of TSSs in Prokaryotic Transcriptomes and Their Role in Gene Expression. *DNA Res.* 24, 25–35. doi:10.1093/dnares/dsw045

Lahtvee, P.-J., Sánchez, B. J., Smialowska, A., Kasvandik, S., Elsemman, I. E., Gatto, F., et al. (2017). Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-specific Translation Efficiency in Yeast. *Cell Syst* 4, 495e5–504. doi:10.1016/j.cels.2017.03.003

Lanchantin, J., Singh, R., Wang, B., and Qi, Y. (2016). “DEEP MOTIF DASHBOARD: VISUALIZING AND UNDERSTANDING GENOMIC SEQUENCES USING DEEP NEURAL NETWORKS,” in *Biocomputing 2017* (World Scientific), 254–265.

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

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., et al. (2015). A Method to Predict the Impact of Regulatory Variants from DNA Sequence. *Nat. Genet.* 47, 955–961. doi:10.1038/ng.3331

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative Prediction of Mammalian Enhancers from DNA Sequence. *Genome Res.* 21, 2167–2180. doi:10.1101/gr.121905.111

Lee, D., Zhang, J., Liu, J., and Gerstein, M. (2020). Epigenome-based Splicing Prediction Using a Recurrent Neural Network. *Plos Comput. Biol.* 16, e1008006. doi:10.1371/journal.pcbi.1008006

Lee, T. I., and Young, R. A. (2013). Transcriptional Regulation and its Misregulation in Disease. *Cell* 152, 1237–1251. doi:10.1016/j.cell.2013.02.014

Leiby, N., Hossain, A., and Salis, H. M. (2020).*Convolutional Neural Net Learns Promoter Sequence Features Driving Transcription Strength*. Manchester, United Kingdom: EasyChair. doi:10.29007/8fmw

Leman, R., Gaildrat, P., Le Gac, G., Ka, C., Fichou, Y., Audrezet, M.-P., et al. (2018). Novel Diagnostic Tool for Prediction of Variant Spliceogenicity Derived from a Set of 395 Combined In Silico/*In Vitro* Studies: an International Collaborative Effort. *Nucleic Acids Res.* 46, 7913–7923. doi:10.1093/nar/gky372

Leman, R., Tubeuf, H., Raad, S., Tournier, I., Derambure, C., Lanos, R., et al. (2020). Assessment of branch point Prediction Tools to Predict Physiological branch Points and Their Alteration by Variants. *BMC Genomics* 21, 86. doi:10.1186/s12864-020-6484-5

Leppek, K., Das, R., and Barna, M. (2018). Functional 5’ UTR mRNA Structures in Eukaryotic Translation Regulation and How to Find Them. *Nat. Rev. Mol. Cel Biol.* 19, 158–174. doi:10.1038/nrm.2017.103

Levo, M., and Segal, E. (2014). In Pursuit of Design Principles of Regulatory Sequences. *Nat. Rev. Genet.* 15, 453–468. doi:10.1038/nrg3684

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A. C., Kalma, Y., Lotam-Pompan, M., et al. (2015). Unraveling Determinants of Transcription Factor Binding outside the Core Binding Site. *Genome Res.* 25, 1018–1029. doi:10.1101/gr.185033.114

Li, G., Zrimec, Jan., Ji, Boyang., Geng, Jun., Larsbrink, Johan., Zelezniak, Aleksej., et al. (2019). *Performance of Regression Models as a Function of experiment Noise*. arXiv [q-bio.BM].

Li, J. J., Chew, G.-L., and Biggin, M. D. (2017). Quantitating Translational Control: mRNA Abundance-dependent and Independent Contributions and the mRNA Sequences that Specify Them. *Nucleic Acids Res.* 45, 11821–11836. doi:10.1093/nar/gkx898

Li, J. J., Chew, G.-L., and Biggin, M. D. (2019). Quantitative Principles of Cis-Translational Control by General mRNA Sequence Features in Eukaryotes. *Genome Biol.* 20, 162. doi:10.1186/s13059-019-1761-9

Li, J., Liang, Q., Song, W., and Marchisio, M. A. (2017). Nucleotides Upstream of the Kozak Sequence Strongly Influence Gene Expression in the Yeast *S. cerevisiae*. *J. Biol. Eng.* 11, 25. doi:10.1186/s13036-017-0068-1

Lin, H., Deng, E. Z., Ding, H., Chen, W., and Chou, K. C. (2014). iPro54-PseKNC: a Sequence-Based Predictor for Identifying Sigma-54 Promoters in Prokaryote with Pseudo K-Tuple Nucleotide Composition. *Nucleic Acids Res.* 42, 12961–12972. doi:10.1093/nar/gku1019

Liu, Y., Barr, K., and Reinitz, J. (2020). Fully Interpretable Deep Learning Model of Transcriptional Control. *Bioinformatics* 36, i499–i507. doi:10.1093/bioinformatics/btaa506

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550. doi:10.1016/j.cell.2016.03.014

Lu, R., and Rogan, P. K. (2018). Transcription Factor Binding Site Clusters Identify Target Genes with Similar Tissue-wide Expression and Buffer against Mutations. *F1000Res* 7, 1933. doi:10.12688/f1000research.17363.1

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., and Segal, E. (2015). Core Promoter Sequence in Yeast Is a Major Determinant of Expression Level. *Genome Res.* 25, 1008–1017. doi:10.1101/gr.188193.114

Lundberg, S., and Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. Ithaca, NY: arXiv [cs.AI].

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using Deep Learning to Model the Hierarchical Structure and Function of a Cell. *Nat. Methods* 15, 290–298. doi:10.1038/nmeth.4627

Marcovitz, A., and Levy, Y. (2013). Weak Frustration Regulates Sliding and Binding Kinetics on Rugged Protein-DNA Landscapes. *J. Phys. Chem. B* 117, 13005–13014. doi:10.1021/jp402296d

Martin, V., Zhao, J., Afek, A., Mielko, Z., and Gordân, R. (2019). QBiC-Pred: Quantitative Predictions of Transcription Factor Binding Changes Due to Sequence Variants. *Nucleic Acids Res.* 47, W127–W135. doi:10.1093/nar/gkz363

Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., et al. (2020). Deep Learning of Immune Cell Differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 117, 25655–25666. doi:10.1073/pnas.2011795117

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions *In Vivo*. *Cel Syst* 3, 278–286. doi:10.1016/j.cels.2016.07.001e4

Mayr, C. (2017). Regulation by 3’-Untranslated Regions. *Annu. Rev. Genet.* 51, 171–194. doi:10.1146/annurev-genet-120116-024704

Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., et al. (2015). Genome-wide Discovery of Human Splicing Branchpoints. *Genome Res.* 25, 290–303. doi:10.1101/gr.182899.114

Meysman, P., Marchal, K., and Engelen, K. (2012). DNA Structural Properties in the Classification of Genomic Transcription Regulation Elements. *Bioinform. Biol. Insights* 6, 155–168. doi:10.4137/BBI.S9426

Mhaskar, H., Liao, Q., and Poggio, T. (2017). "When and Why Are Deep Networks Better Than Shallow Ones?," in AAAI, 31.

Millar, A. H., Heazlewood, J. L., Giglione, C., Holdsworth, M. J., Bachmair, A., and Schulze, W. X. (2019). The Scope, Functions, and Dynamics of Posttranslational Protein Modifications. *Annu. Rev. Plant Biol.* 70, 119–151. doi:10.1146/annurev-arplant-050718-100211

Miller, J. L., and Grant, P. A. (2013). The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans. *Subcell. Biochem.* 61, 289–317. doi:10.1007/978-94-007-4525-4_13

Min, X., Zeng, W., Chen, S., Chen, N., Chen, T., and Jiang, R. (2017). Predicting Enhancers with Deep Convolutional Neural Networks. *BMC Bioinformatics* 18, 478. doi:10.1186/s12859-017-1878-3

Mittal, P., Brindle, J., Stephen, J., Plotkin, J. B., and Kudla, G. (2018). Codon Usage Influences Fitness through RNA Toxicity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 8639–8644. doi:10.1073/pnas.1810022115

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal. Process.* 73, 1–15. doi:10.1016/j.dsp.2017.10.011

Moqtaderi, Z., Geisberg, J. V., Jin, Y., Fan, X., and Struhl, K. (2013). Species-specific Factors Mediate Extensive Heterogeneity of mRNA 3' Ends in Yeasts. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11073–11078. doi:10.1073/pnas.1309384110

Mora, A., Sandve, G. K., Gabrielsen, O. S., and Eskeland, R. (2016). The Loop: Promoter-Enhancer Interactions and Bioinformatics. *Brief. Bioinform.* 17, 980–995. doi:10.1093/bib/bbv097

Morse, N. J., Gopal, M. R., Wagner, J. M., and Alper, H. S. (2017). Yeast Terminator Function Can Be Modulated and Designed on the Basis of Predictions of Nucleosome Occupancy. *ACS Synth. Biol.* 6, 2086–2095. doi:10.1021/acssynbio.7b00138

Movva, R., Greenside, P., Marinov, G. K., Nair, S., Shrikumar, A., and Kundaje, A. (2019). Deciphering Regulatory DNA Sequences and Noncoding Genetic Variants Using Neural Network Models of Massively Parallel Reporter Assays. *PLoS One* 14, e0218073. doi:10.1371/journal.pone.0218073

Nagy, G., and Nagy, L. (2020). Motif Grammar: The Basis of the Language of Gene Expression. *Comput. Struct. Biotechnol. J.* 18, 2026–2032. doi:10.1016/j.csbj.2020.07.007

Naidoo, T., Sjödin, P., Schlebusch, C., and Jakobsson, M. (2018). Patterns of Variation in Cis-Regulatory Regions: Examining Evidence of Purifying Selection. *BMC Genomics* 19, 95. doi:10.1186/s12864-017-4422-y

Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H., and Miura, K.-I. (2008). Diversity of Preferred Nucleotide Sequences Around the Translation Initiation Codon in Eukaryote Genomes. *Nucleic Acids Res.* 36, 861–871. doi:10.1093/nar/gkm1102

Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Vanden Berghe, W., Goethals, B., et al. (2015). A Primer to Frequent Itemset Mining for Bioinformatics. *Brief. Bioinform.* 16, 216–231. doi:10.1093/bib/bbt074

Nazari, I., Tayara, H., and Chong, K. T. (2019). Branch Point Selection in RNA Splicing Using Deep Learning. *IEEE Access* 7, 1800–1807. doi:10.1109/access.2018.2886569

Neymotin, B., Ettorre, V., and Gresham, D. (2016). Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in *Saccharomyces cerevisiae*. *G* 6, 3475–3483. doi:10.1534/g3.116.032276

Nielsen, J., and Keasling, J. D. (2016). Engineering Cellular Metabolism. *Cell* 164, 1185–1197. doi:10.1016/j.cell.2016.02.004

Nielsen, J. (2017). Systems Biology of Metabolism. *Annu. Rev. Biochem.* 86, 245–275. doi:10.1146/annurev-biochem-061516-044757

Niu, X., Yang, K., Zhang, G., Yang, Z., and Hu, X. (2019). A Pretraining-Retraining Strategy of Deep Learning Improves Cell-specific Enhancer Predictions. *Front. Genet.* 10, 1305. doi:10.3389/fgene.2019.01305

Omotajo, D., Tate, T., Cho, H., and Choudhary, M. (2015). Distribution and Diversity of Ribosome Binding Sites in Prokaryotic Genomes. *BMC Genomics* 16, 604. doi:10.1186/s12864-015-1808-6

Paggi, J. M., and Bejerano, G. (2018). A Sequence-Based, Deep Learning Model Accurately Predicts RNA Splicing Branchpoints. *RNA* 24, 1647–1658. doi:10.1261/rna.066290.118

Park, C., Qian, W., and Zhang, J. (2012). Genomic Evidence for Elevated Mutation Rates in Highly Expressed Genes. *EMBO Rep.* 13, 1123–1129. doi:10.1038/embor.2012.165

Park, S., Koh, Y., Jeon, H., Kim, H., Yeo, Y., and Kang, J. (2020). Enhancing the Interpretability of Transcription Factor Binding Site Prediction Using Attention Mechanism. *Sci. Rep.* 10, 13413. doi:10.1038/s41598-020-70218-4

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: Five Essential Questions. *Nat. Rev. Genet.* 14, 288–295. doi:10.1038/nrg3458

Peters, J. M., Mooney, R. A., Kuan, P. F., Rowland, J. L., Keles, S., and Landick, R. (2009). Rho Directs Widespread Termination of Intragenic and Stable RNA Transcription. *Proc. Natl. Acad. Sci. U. S. A.* 106, 15406–15411. doi:10.1073/pnas.0903846106

Playe, B., and Stoven, V. (2020). Evaluation of Deep and Shallow Learning Methods in Chemogenomics for the Prediction of Drugs Specificity. *J. Cheminform.* 12, 11. doi:10.1186/s13321-020-0413-0

Plotkin, J. B., and Kudla, G. (2011). Synonymous but Not the Same: the Causes and Consequences of Codon Bias. *Nat. Rev. Genet.* 12, 32–42. doi:10.1038/nrg2899

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., et al. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* 160, 1111–1124. doi:10.1016/j.cell.2015.02.029

Puigbò, P., Guzmán, E., Romeu, A., and Garcia-Vallvé, S. (2007). OPTIMIZER: a Web Server for Optimizing the Codon Usage of DNA Sequences. *Nucleic Acids Res.* 35, W126–W131. doi:10.1093/nar/gkm219

Quang, D., and Xie, X. Dan. Q. (2016). A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of DNA Sequences. *Nucleic Acids Res.* 44, e107. doi:10.1093/nar/gkw226

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., et al. (2019). Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* 32, 9689–9701.

Re, A., Joshi, T., Kulberkyte, E., Morris, Q., and Workman, C. T. (2014). RNA-protein Interactions: an Overview. *Methods Mol. Biol.* 1097, 491–521. doi:10.1007/978-1-62703-709-9_23

Redden, H., and Alper, H. S. (2015). The Development and Characterization of Synthetic Minimal Yeast Promoters. *Nat. Commun.* 6, 7810. doi:10.1038/ncomms8810

Rehbein, P., Berz, J., Kreisel, P., and Schwalbe, H. (2019). 'CodonWizard' - an Intuitive Software Tool with Graphical User Interface for Customizable Codon Optimization in Protein Expression Efforts. *Protein Expr. Purif.* 160, 84–93. doi:10.1016/j.pep.2019.03.018

Ren, G.-X., Guo, X.-P., and Sun, Y.-C. (2017). Regulatory 3' Untranslated Regions of Bacterial mRNAs. *Front. Microbiol.* 8, 1276. doi:10.3389/fmicb.2017.01276

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., et al. (2021). Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks. *Nat. Machine Intelligence* 3, 324–333. doi:10.1038/s42256-021-00310-5

Richardson, S. M., Wheelan, S. J., Yarrington, R. M., and Boeke, J. D. (2006). GeneDesign: Rapid, Automated Design of Multikilobase Synthetic Genes. *Genome Res.* 16, 550–556. doi:10.1101/gr.4431306

Roberts, J. W. (2019). Mechanisms of Bacterial Transcription Termination. *J. Mol. Biol.* 431, 4030–4039. doi:10.1016/j.jmb.2019.04.003

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.* 79, 233–269. doi:10.1146/annurev-biochem-060408-091030

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The Role of DNA Shape in Protein–DNA Recognition. *Nature* 461, 1248–1253. doi:10.1038/nature08473

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Machine Intelligence* 1, 206–215. doi:10.1038/s42256-019-0048-x

Saier, M. H., Jr. (2019). Understanding the Genetic Code. *J. Bacteriol.* 201. doi:10.1128/JB.00091-19

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression. *Nat. Biotechnol.* 27, 946–950. doi:10.1038/nbt.1568

Salis, H. M. (2011). The Ribosome Binding Site Calculator. *Methods Enzymol.* 498, 19–42. doi:10.1016/b978-0-12-385120-8.00002-4

Sample, P. J., Wang, B., Reid, D. W., Presnyak, V., McFadyen, I. J., Morris, D. R., et al. (2019). Human 5' UTR Design and Variant Effect Prediction from a Massively Parallel Translation Assay. *Nat. Biotechnol.* 37, 803–809. doi:10.1038/s41587-019-0164-5

SantaLucia, J., Jr. (1998). A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1460–1465. doi:10.1073/pnas.95.4.1460

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global Quantification of Mammalian Gene Expression Control. *Nature* 473, 337–342. doi:10.1038/nature10098

Segal, E., and Widom, J. (2009). From DNA Sequence to Transcriptional Behaviour: a Quantitative Approach. *Nat. Rev. Genet.* 10, 443–456. doi:10.1038/nrg2591

Shalem, O., Sharon, E., Lubliner, S., Regev, I., Lotan-Pompan, M., Yakhini, Z., et al. (2015). Systematic Dissection of the Sequence Determinants of Gene 3'end Mediated Expression Control. *Plos Genet.* 11, e1005147. doi:10.1371/journal.pgen.1005147

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., et al. (2012). Inferring Gene Regulatory Logic from High-Throughput Measurements of Thousands of Systematically Designed Promoters. *Nat. Biotechnol.* 30, 521–530. doi:10.1038/nbt.2205

Sharp, P. M., and Li, W. H. (1987). The Codon Adaptation Index--a Measure of Directional Synonymous Codon Usage Bias, and its Potential Applications. *Nucleic Acids Res.* 15, 1281–1295. doi:10.1093/nar/15.3.1281

Shine, J., and Dalgarno, L. (1975). Determinant of Cistron Specificity in Bacterial Ribosomes. *Nature* 254, 34–38. doi:10.1038/254034a0

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional Enhancers: from Properties to Genome-wide Predictions. *Nat. Rev. Genet.* 15, 272–286. doi:10.1038/nrg3682

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). *Learning Important Features through Propagating Activation Differences*. Ithaca, NY: arXiv [cs.CV].

Shrikumar, A., Tian, Katherine., Avsec, Žiga., Shcherbina, Anna., Banerjee, Abhimanyu., Sharmin, Mahfuza., et al. (2018). *Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) Version 0.5.6.5*. Ithaca, NY: arXiv [cs.LG].

Siggers, T., and Gordân, R. (2014). Protein-DNA Binding: Complexities and Multi-Protein Codes. *Nucleic Acids Res.* 42, 2099–2111. doi:10.1093/nar/gkt1112

Signal, B., Gloss, B. S., Dinger, M. E., and Mercer, T. R. (2018). Machine Learning Annotation of Human Branchpoints. *Bioinformatics* 34, 920–927. doi:10.1093/bioinformatics/btx688

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). *Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. Ithaca, NY: arXiv [cs.CV].

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. Deep. Chrome. (2016). Deep-learning for Predicting Gene Expression from Histone Modifications. *Bioinformatics* 32, i639–i648. doi:10.1093/bioinformatics/btw427

Singh, R., Lanchantin, J., Sekhon, A., and Qi, Y. (2017). Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin. *Adv. Neural Inf. Process. Syst.* 30, 6785–6795.

Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. *Quantitative Biol.* 7, 122–137. doi:10.1007/s40484-019-0154-0

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a Simple Code: How Transcription Factors Read the Genome. *Trends Biochem. Sci.* 39, 381–399. doi:10.1016/j.tibs.2014.07.002

Song, L., and Crawford, G. E. (2010). *DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells*. Cold Spring Harbor, NY: Cold Spring Harb. Protoc. doi:10.1101/pdb.prot5384

Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Rätsch, G. (2007). Accurate Splice Site Prediction Using Support Vector Machines. *BMC Bioinformatics* 8 (Suppl. 10), S7. doi:10.1186/1471-2105-8-S10-S7

Stormo, G. D. (2000). DNA Binding Sites: Representation and Discovery. *Bioinformatics* 16, 16–23. doi:10.1093/bioinformatics/16.1.16

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. (2020). Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cel Syst* 11, 402e4–411. doi:10.1016/j.cels.2020.08.016

Strubell, E., Verga, P., Belanger, D., and McCallum, A. (2017). *Fast and Accurate Sequence Labeling with Iterated Dilated Convolutions*.

Struhl, K., and Segal, E. (2013). Determinants of Nucleosome Positioning. *Nat. Struct. Mol. Biol.* 20, 267–273. doi:10.1038/nsmb.2506

Tafvizi, A., Mirny, L. A., and van Oijen, A. M. (2011). Dancing on DNA: Kinetic Aspects of Search Processes on DNA. *Chemphyschem* 12, 1481–1489. doi:10.1002/cphc.201100112

Tang, L., Hill, M. C., Wang, J., Wang, J., Martin, J. F., and Li, M. (2020). Predicting Unrecognized Enhancer-Mediated Genome Topology by an Ensemble Machine Learning Model. *Genome Res.* 30, 1835–1845. doi:10.1101/gr.264606.120

Tareen, A., and Kinney, J. B. (2019). *Biophysical Models of Cis-Regulation as Interpretable Neural Networks*. Ithaca, NY: arXiv [q-bio.MN].

Terai, G., and Asai, K. (2020). Improving the Prediction Accuracy of Protein Abundance in *Escherichia coli* Using mRNA Accessibility. *Nucleic Acids Res.* 48, e81. doi:10.1093/nar/gkaa481

Tian, B., and Manley, J. L. (2017). Alternative Polyadenylation of mRNA Precursors. *Nat. Rev. Mol. Cel Biol.* 18, 18–30. doi:10.1038/nrm.2016.116

Tian, Q., Zou, J., Tang, J., Fang, Y., Yu, Z., and Fan, S. (2019). MRCNN: a Deep Learning Model for Regression of Genome-wide DNA Methylation. *BMC Genomics* 20, 192. doi:10.1186/s12864-019-5488-5

Tirosh, I., Reikhav, S., Levy, A. A., and Barkai, N. (2009). A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* 324, 659–662. doi:10.1126/science.1169766

Trabelsi, A., Chaabane, M., and Ben-Hur, A. (2019). Comprehensive Evaluation of Deep Learning Architectures for Prediction of DNA/RNA Sequence Binding Specificities. *Bioinformatics* 35, i269–i277. doi:10.1093/bioinformatics/btz339

Trösemeier, J.-H., Rudorf, S., Loessner, H., Hofner, B., Reuter, A., Schulenborg, T., et al. (2019). Optimizing the Dynamics of Protein Expression. *Sci. Rep.* 9, 7511. doi:10.1038/s41598-019-43857-5

Trotta, E. (2013). Selection on Codon Bias in Yeast: a Transcriptional Hypothesis. *Nucleic Acids Res.* 41, 9382–9395. doi:10.1093/nar/gkt740

Tsai, Z. T.-Y., Shiu, S.-H., and Tsai, H.-K. (2015). Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast. *Plos Comput. Biol.* 11, e1004418. doi:10.1371/journal.pcbi.1004418

Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010). Translation Efficiency Is Determined by Both Codon Bias and Folding Energy. *Proc. Natl. Acad. Sci. U. S. A.* 107, 3645–3650. doi:10.1073/pnas.0909910107

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., and Lareau, L. F. (2018). Accurate Design of Translational Output by a Neural Network Model of Ribosome Distribution. *Nat. Struct. Mol. Biol.* 25, 577–582. doi:10.1038/s41594-018-0080-2

Ullah, F., and Ben-Hur, A. (2020). *A Self-Attention Model for Inferring Cooperativity between Regulatory Features*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. doi:10.1101/2020.01.31.927996

Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H., and Kosuri, S. (2019). Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. *Biochemistry* 58, 1539–1551. doi:10.1021/acs.biochem.7b01069

Van Brempt, M., Clauwaert, J., Mey, F., Stock, M., Maertens, J., Waegeman, W., et al. (2020). Predictive Design of Sigma Factor-specific Promoters. *Nat. Commun.* 11, 5822. doi:10.1038/s41467-020-19446-w

Vaswani, A., Shazeer, Noam., Parmar, Niki., Uszkoreit, Jakob., Jones, Llion., Gomez, Aidan. N., et al. (2017). *Attention Is All You Need*. arXiv [cs.CL].

Vig, J., Madani, Ali., Varshney, Lav. R., Xiong, Caiming., Socher, Richard., and Rajani, Nazneen. Fatema. (2020). *BERTology Meets Biology: Interpreting Attention in Protein Language Models*. Ithaca, NY: arXiv [cs.CL].

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., et al. (2009). ChIP-seq Accurately Predicts Tissue-specific Activity of Enhancers. *Nature* 457, 854–858. doi:10.1038/nature07730

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., et al. (2010). Sequence Signatures and mRNA Concentration Can Explain Two-Thirds of Protein Abundance Variation in a Human Cell Line. *Mol. Syst. Biol.* 6, 400. doi:10.1038/msb.2010.59

Wang, M., Tai, C., E, W., and Wei, L. De. Fine. (2018). Deep Convolutional Neural Networks Accurately Quantify Intensities of Transcription Factor-DNA Binding and Facilitate Evaluation of Functional Non-coding Variants. *Nucleic Acids Res.* 46, e69. doi:10.1093/nar/gky215

Wang, R., Wang, Z., Wang, J., and Li, S. (2019). SpliceFinder: Ab Initio Prediction of Splice Sites Using Convolutional Neural Network. *BMC Bioinformatics* 20, 652. doi:10.1186/s12859-019-3306-3

Wang, X., Girshick, R., Gupta, A., and He, K. (2017). *Non-local Neural Networks*. arXiv [cs.CV].

Wang, H., Cimen, E., Singh, N., and Buckler, E. (2020). Deep Learning for Plant Genomics and Crop Improvement. *Curr. Opin. Plant Biol.* 54, 34–41. doi:10.1016/j.pbi.2019.12.010

Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., and Wang, X. (2020). Synthetic Promoter Design in *Escherichia coli* Based on a Deep Generative Network. *Nucleic Acids Res.* 48, 6403–6412. doi:10.1093/nar/gkaa325

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., et al. (2019). Evolutionarily Informed Deep Learning Methods for Predicting Relative Transcript Abundance from DNA Sequence. *Proc. Natl. Acad. Sci. U. S. A.* 116, 5542–5549. doi:10.1073/pnas.1814551116

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Losick, R. (2008). *Molecular Biology of the Gene*. 6th. ed.. San Francisco, CA: Pearson/Benjamin Cummings.

Way, G. P., and Greene, C. S. (2018). Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. *Pac. Symp. Biocomput.* 23, 80–91.

Webb, S. (2018). Deep Learning for Biology. *Nature* 554, 555–557. doi:10.1038/d41586-018-02174-z

Weenink, T., van der Hilst, J., McKiernan, R. M., and Ellis, T. (2018). Design of RNA Hairpin Modules that Predictably Tune Translation in Yeast. *Synth. Biol.* 3, ysy019. doi:10.1093/synbio/ysy019

Weirauch, M. T., au, fnm., Cote, A., Norel, R., Annala, M., Zhao, Y., et al. (2013). Evaluation of Methods for Modeling Transcription Factor Sequence Specificity. *Nat. Biotechnol.* 31, 126–134. doi:10.1038/nbt.2486

Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the Human Epigenome from DNA Motifs. *Nat. Methods* 12 (), 265–272. 7 p following 272. doi:10.1038/nmeth.3065

Wilkinson, M. E., Charenton, C., and Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* 89, 359–388. doi:10.1146/annurev-biochem-091719-064225

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary Changes in Cis and Trans Gene Regulation. *Nature* 430, 85–88. doi:10.1038/nature02698

Wittkopp, P. J., and Kalay, G. (2011). Cis-regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence. *Nat. Rev. Genet.* 13, 59–69. doi:10.1038/nrg3095

Xie, R., Wen, J., Quitadamo, A., Cheng, J., and Shi, X. (2017). A Deep Auto-Encoder Model for Gene Expression Prediction. *BMC Genomics* 18, 845. doi:10.1186/s12864-017-4226-0

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep Learning of the Splicing (Epi)genetic Code Reveals a Novel Candidate Mechanism Linking Histone Modifications to ESC Fate Decision. *Nucleic Acids Res.* 45, 12100–12112. doi:10.1093/nar/gkx870

Yang, D. K., Goldman, S. L., Weinstein, E., and Marks, D. (2019). "Generative Models for Codon Prediction and Optimization," in *Machine Learning in Computational Biology*.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., et al. (2017). Transcription Factor Family-specific DNA Shape Readout Revealed by Quantitative Specificity Models. *Mol. Syst. Biol.* 13, 910. doi:10.15252/msb.20167238

Yu, F., and Koltun, V. (2015). *Multi-Scale Context Aggregation by Dilated Convolutions*. arXiv [cs.CV].

Yu, M., Guo, W., Wang, Q., and Chen, J. Q. (2019). *Widespread Positive Selection for mRNA Secondary Structure at Synonymous Sites in Domesticated Yeast*. bioRxiv.

Zelezniak, A., Vowinckel, J., Capuano, F., Messner, C. B., Demichev, V., Polowsky, N., et al. (2018). Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cel Syst* 7, 269–283. e6. doi:10.1016/j.cels.2018.08.001

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional Neural Network Architectures for Predicting DNA-Protein Binding. *Bioinformatics* 32, i121–i127. doi:10.1093/bioinformatics/btw255

Zhang, Q., Fan, X., Wang, Y., Sun, M.-a., Shao, J., and Guo, D. (2017). BPP: a Sequence-Based Algorithm for branch point Prediction. *Bioinformatics* 33, 3166–3172. doi:10.1093/bioinformatics/btx401

Zhang, S., Hu, H., Jiang, T., Zhang, L., and Zeng, J. T. I. T. E. R. (2017). Predicting Translation Initiation Sites by Deep Learning. *Bioinformatics* 33, i234–i242. doi:10.1093/bioinformatics/btx247

Zhang, S., Hu, H., Zhou, J., He, X., Jiang, T., and Zeng, J. (2017). Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cel Syst* 5, 212–220. e6. doi:10.1016/j.cels.2017.08.004

Zhang, Y., Zhou, X., and Cai, X. (2020). *Predicting Gene Expression from DNA Sequence Using Residual Neural Network*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. doi:10.1101/2020.06.21.163956

Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., et al. (2019). Deep-learning Augmented RNA-Seq Analysis of Transcript Splicing. *Nat. Methods* 16, 307–310. doi:10.1038/s41592-019-0351-9

Zhao, B. S., Roundtree, I. A., and He, C. (2017). Post-transcriptional Gene Regulation by mRNA Modifications. *Nat. Rev. Mol. Cel Biol.* 18, 31–42. doi:10.1038/nrm.2016.132

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3′ Ends in Eukaryotes: Mechanism, Regulation, and Interrelationships with Other Steps in mRNA Synthesis. *Microbiol. Mol. Biol. Rev.* 63 (2), 405. doi:10.1128/MMBR.63.2.405-445.1999

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk. *Nat. Genet.* 50, 1171–1179. doi:10.1038/s41588-018-0160-6

Zhou, J., and Troyanskaya, O. G. (2015). Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nat. Methods* 12, 931–934. doi:10.1038/nmeth.3547

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., et al. (2015). Quantitative Modeling of Transcription Factor Binding Specificities Using DNA Shape. *Proc. Natl. Acad. Sci. U. S. A.* 112, 4654–4659. doi:10.1073/pnas.1422023112

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.-h., Fu, J., et al. (2016). Codon Usage Is an Important Determinant of Gene Expression Levels Largely through its Effects on Transcription. *Proc. Natl. Acad. Sci. U. S. A.* 113, E6117–E6125. doi:10.1073/pnas.1606724113

Zhou, Z., Dang, Y., Zhou, M., Yuan, H., and Liu, Y. (2018). Codon Usage Biases Co-evolve with Transcription Termination Machinery to Suppress Premature Cleavage and Polyadenylation. *Elife* 7, e33569. doi:10.7554/eLife.33569

Zicola, J., Liu, L., Tänzler, P., and Turck, F. (2019). Targeted DNA Methylation Represses Two Enhancers of FLOWERING LOCUS T in *Arabidopsis thaliana*. *Nat. Plants* 5, 300–307. doi:10.1038/s41477-019-0375-2

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A Primer on Deep Learning in Genomics. *Nat. Genet.* 51, 12–18. doi:10.1038/s41588-018-0295-5

Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., et al. (2020). Deep Learning Suggests that Gene Expression Is Encoded in All Parts of a Co-evolving Interacting Gene Regulatory Structure. *Nat. Commun.* 11, 6141. doi:10.1038/s41467-020-19921-4

Zrimec, J., and Lapanje, A. (2018). DNA Structure at the Plasmid Origin-Of-Transfer Indicates its Potential Transfer Range. *Sci. Rep.* 8, 1820. doi:10.1038/s41598-018-20157-y

Zrimec, J., and Lapanje, A. (2015). Fast Prediction of DNA Melting Bubbles Using DNA Thermodynamic Stability. *Ieee/acm Trans. Comput. Biol. Bioinform.* 12, 1137–1145. doi:10.1109/tcbb.2015.2396057

Zrimec, J. (2020). Multiple Plasmid Origin-Of-Transfer Regions Might Aid the Spread of Antimicrobial Resistance to Human Pathogens. *Microbiologyopen* 9, e1129. doi:10.1002/mbo3.1129

Zrimec, J. (2020). "Structural Representations of DNA Regulatory Substrates Can Enhance Sequence-Based Algorithms by Associating Functional Sequence Variants," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Association for Computing Machinery), 1–6.

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., and De Neve, W. (2018). SpliceRover: Interpretable Convolutional Neural Networks for Improved Splice Site Prediction. *Bioinformatics* 34, 4180–4188. doi:10.1093/bioinformatics/bty497

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 June 2021
doi: 10.3389/fmolb.2021.657087

N6-Methyladenosine RNA Methylation Regulator-Related Alternative Splicing (AS) Gene Signature Predicts Non–Small Cell Lung Cancer Prognosis

Zhenyu Zhao[1,2], *Qidong Cai*[1,2], *Pengfei Zhang*[1,2], *Boxue He*[1,2], *Xiong Peng*[1,2], *Guangxu Tu*[1,2], *Weilin Peng*[1,2], *Li Wang*[1,2], *Fenglei Yu*[1,2] *and Xiang Wang*[1,2]*

[1]*Department of Thoracic Surgery, The Second Xiangya Hospital of Central South University, Changsha, China,* [2]*Hunan Key Laboratory of Early Diagnosis and Precise Treatment of Lung Cancer, The Second Xiangya Hospital of Central South University, Changsha, China*

OPEN ACCESS

Edited by:
Arzucan Ozgur,
Boğaziçi University, Turkey

Reviewed by:
Junguk Hur,
University of North Dakota,
United States
Pınar Pir,
Gebze Technical University, Turkey

***Correspondence:**
Xiang Wang
wangxiang@csu.edu.cn

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 January 2021*
Accepted: *12 May 2021*
Published: *11 June 2021*

Citation:
Zhao Z, Cai Q, Zhang P, He B, Peng X, Tu G, Peng W, Wang L, Yu F and Wang X (2021) N6-Methyladenosine RNA Methylation Regulator-Related Alternative Splicing (AS) Gene Signature Predicts Non–Small Cell Lung Cancer Prognosis. Front. Mol. Biosci. 8:657087. doi: 10.3389/fmolb.2021.657087

Aberrant N6-methyladenosine (m6A) RNA methylation regulatory genes and related gene alternative splicing (AS) could be used to predict the prognosis of non–small cell lung carcinoma. This study focused on 13 m6A regulatory genes (METTL3, METTL14, WTAP, KIAA1429, RBM15, ZC3H13, YTHDC1, YTHDC2, YTHDF1, YTHDF2, HNRNPC, FTO, and ALKBH5) and expression profiles in TCGA-LUAD ($n = 504$) and TCGA-LUSC ($n = 479$) datasets from the Cancer Genome Atlas database. The data were downloaded and bioinformatically and statistically analyzed, including the gene ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses. There were 43,948 mRNA splicing events in lung adenocarcinoma (LUAD) and 46,020 in lung squamous cell carcinoma (LUSC), and the data suggested that m6A regulators could regulate mRNA splicing. Differential HNRNPC and RBM15 expression was associated with overall survival (OS) of LUAD and HNRNPC and METTL3 expression with the OS of LUSC patients. Furthermore, the non–small cell lung cancer prognosis-related AS events signature was constructed and divided patients into high- *vs.* low-risk groups using seven and 14 AS genes in LUAD and LUSC, respectively. The LUAD risk signature was associated with gender and T, N, and TNM stages, but the LUSC risk signature was not associated with any clinical features. In addition, the risk signature and TNM stage were independent prognostic predictors in LUAD and the risk signature and T stage were independent prognostic predictors in LUSC after the multivariate Cox regression and receiver operating characteristic analyses. In conclusion, this study revealed the AS prognostic signature in the prediction of LUAD and LUSC prognosis.

Keywords: non–small cell lung cancer, m6A, alternative splicing, The Cancer Genome Atlas, prognostic signature

INTRODUCTION

Lung cancer is still the most significant health burden in the world, accounting for 14% of all newly diagnosed cancer cases as the second most common cancer and 18% of all cancer-related deaths as the leading cause of cancer death globally in 2018 and 2020 (de Martel et al., 2020; Sung et al., 2021). Lung cancer is also prevalent and the leading cause of cancer death in men (Sung et al., 2021). Histologically, lung cancer can be divided into small cell lung cancer and non–small cell lung cancer (NSCLC), and the latter accounts for 85% of all lung cancer cases, and the overall 5-year survival rate of lung cancer remains to be approximately 15% (Balata et al., 2019). NSCLC can be further classified as lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), and larger cell carcinoma; however, LUAD and LUSC are the main histological subtypes of NSCLC (Tanoue et al., 2015) and major contributors to NSCLC morbidity and mortality (Hirsch et al., 2017). The outcome data were from our most recent advancement and improvement in early detection, prevention, improved surgical procedures, neoadjuvant therapy, immunotherapy, and targeted therapy. To date, treatment of NSCLC is dependent on the stage of disease at diagnosis, and early-staged NSCLC could be surgically cured, whereas the advanced staged diseases can only be subjected to chemotherapy, radiation therapy, immunotherapy, and/or targeted therapy (Maconachie et al., 2019; Planchard et al., 2018) and their prognosis is, therefore, still poor, approximately less than 5–7% at the best according to the American Cancer Society data (https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html) or after advanced therapy (Zhang et al., 2021). Thus, the search and development of biomarkers for early detection and prediction of prognosis and treatment outcome are urgently needed to effectively conquer this now deadly disease clinically.

Newly transcribed RNA could undergo different chemical modifications and N6-methyladenosine (m6A) is the most prevailing one in polyadenylated RNAs (Bokar et al., 1997). Methylation of the adenosine is directed in cells by a large m6A methyltransferase complex containing METTL3 as the SAM-binding subunit (Bokar et al., 1997). The biological functions of m6A are through a group of RNA-binding proteins that can specifically recognize the methylated adenosine on RNA molecules to regulate cell activities (Ji et al., 2018), for example, N6-methyladenosine (m6A) RNA modification could regulate RNA splicing, stability, translocation, and translation and therefore, to influence gene expression and functions in cells (Deng, et al., 2018). These binding proteins to m6A are regarded as the m6A readers, and m6A methyltransferases are considered as the writers, whereas demethylases are considered as the erasers. Altogether, these proteins form a complex mechanism of m6A regulation in which writers and erasers determine the distributions of m6A on RNA, whereas readers mediate m6A-dependent functions (Liu et al., 2014; Wang et al., 2014). Deregulation of the m6A on an RNA molecule has been implicated in the development of various human cancers (Liu et al., 2014; Wang et al., 2014). According to the recent studies, there were 13 m6A regulator genes confirmed to affect cancer progression, including the "writer" (KIAA1429, METTL3, METTL14, RBM15, WTAP, and ZC3H13), the "readers" (HNRNPC, YTHDC1, YTHDC2, YTHDF1, and YTHDF2), and the "erasers" (ALKBH5 and FTO) (Zhang et al., 2020c; Zhang et al., 2020b). Further studies of the m6A regulator genes showed that the m6A regulator genes were also the mRNA splicing factors for gene alternative splicing (GAS) and the m6A regulator genes could interact with the AS events (Kasowitz et al., 2018; Yoshimi et al., 2019). Human cancer cells frequently showed the GAS events, which were regulated by the m6A regulators (Dai et al., 2018). For example, METTL3 was able to regulate the mRNA alternative splicing by the p53 pathway (Alarcón et al., 2015). YTHDC1 could recruit SRSF10 to its target mRNA regions and modulate their exon skipping (Xiao et al., 2016). Abnormal splicing factor expression in normal cells could lead to the formation of the specific pro-oncogenic splicing subtypes and carcinogenesis (Kasowitz et al., 2018).

Indeed, gene alternative splicing (GAS), a posttranscriptional process, subjects a single pre-mRNA molecule to splice into different exons for coding and expression of various protein isoforms (David and Manley, 2008). A molecular structure called a spliceosome is assembled on the pre-mRNA to join the exons together at the splicing site to form a particular mRNA molecule, while the introns are discarded (Papasaikas and Valcárcel, 2016). The assembly of spliceosomes on pre-mRNA is usually affected by the SF and some exons (alternative exons) are variably incorporated into mRNA; thus, under different alternative splicing patterns (including exon skip, retained intron, alternate donor site, alternate acceptor site, alternate promoter, alternate terminator, and mutually exclusive exons), the whole exons of a gene could be spliced into mRNA or excluded (Hanahan and Weinberg, 2011). The different GAS events could lead to the diversity of protein functions and normal GAS will maintain normal cell functions, which is mediated by the production of the diverse and multifunctional proteome to ensure "normal" RNA molecules to maintain normal cell functions; however, abnormal GAS will promote tumorigenesis and cancer development (Bonnal et al., 2020), which could be mainly due to the up or downregulation of the related splicing factors, for example, alterations in the upstream signaling pathways or mutations in the splicing site sequences all lead to abnormal mRNA splicing (Li et al., 2019). Accumulating evidence suggests the contribution of abnormal GAS to cancer phenotypes, like increases in cell proliferation, angiogenesis, but inhibition of apoptosis and drug resistance, and the GAS events form a novel and separate hallmark in cancer (Bonnal et al., 2020; Urbanski et al., 2018). For example, in gastric cancer, abnormal GAS could lead to activation of tumor cell invasion and metastasis (Pio and Montuenga, 2009; Sun and Ma, 2019), while in breast cancer, abnormal GAS results in drug resistance (Yang et al., 2019). In lung cancer, the GAS events could be used as biomarkers for tumor diagnosis (Sholl, 2017). Aberrant *BCL2L1*, *MDM2*, *MDM4*, *NUMB*, and *MET* mRNA splicing occurred in lung cancer and altered cell apoptosis, proliferation, and cohesion (Coomer, Black). Thus, further investigation related to the abnormal GAS events to lung tumorigenesis (Coomer et al., 2019) and novel strategy for cancer targeting therapy (Frankiw et al., 2019) as well as biomarkers for various human cancers, including NSCLC (Li et al., 2017; Paschalis et al., 2018; Yang et al., 2019; Liu et al., 2020; Zhang et al., 2020a).

In this study, we focused on the m6a-related splicing factors for aberrant expression and AS events to associate them with NSCLC clinicopathological and prognostic data from patients using the online The Cancer Genome Atlas (TCGA) data. We then explored the role of the abnormally expressed m6a-related splicing factors in the regulation of the GAS events and constructed the risk signature of these factors to predict NSCLC prognosis after the gene ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. This study could provide a novel insight into the discovery of biomarkers in the prediction of NSCLC prognosis and possibly the underlying molecular mechanisms of NSCLC oncogenesis and development.

MATERIALS AND METHODS

Data Download and Analysis

In this study, we first searched and downloaded differential gene expression profiles in LUAD and LUSC tissue specimens from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The corresponding clinicopathological data were subsequently downloaded from the University of California Santa Cruz database (https://xena.ucsc.edu/), which included 514 LUAD and 488 LUSC tissue samples. However, patients with incomplete clinical information and follow-up duration less than 30 days were excluded from our data analysis, resulting in 504 LUAD and 479 LUSC samples in this study. Moreover, the gene alternative splicing (GAS) events in LUAD and LUSC were download from TCGA Splice Seq (https://bioinformatics.mdanderson.org/TCGASpliceSeq/PSIdownload) and then calculated for the percent spliced in index (PSI) value, a quantifiable GAS indicator after the comparison of single and multiple samples between subgroups, that is, calculation of the percentage of GAS value for each GAS event, which was typically used to quantify GAS events according to a previous study (Lin and Krainer, 2019). We downloaded the contents that included seven main GAS types, that is, the exon skip (ES), retained intron (RI), alternate donor site (AD), alternate acceptor site (AA), alternate promoter (AP), alternate terminator (AT), and mutually exclusive exons (ME).

Selection and Analysis of N6-Methyladenosine RNA Methylation Regulatory Genes

In this study, we selected 13 m6A RNA methylation regulatory genes, that is, N6-adenosine-methyltransferase 70-kDa subunit (METTL3), methyltransferase-like 14 (METTL14), Wilms' tumor-1 associated protein (WTAP), KIAA1429, RNA-binding protein 15 (RBM15), zinc finger CCCH domain-containing protein 13 (ZC3H13), YTH domain-containing protein 1 (YTHDC1), YTHDC2, YTH domain family, member 1 (YTHDF1), YTHDF2, heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC), fat mass and obesity-associated protein (FTO), and m6A demethylase alkB homolog 5 (ALKBH5). We assessed their role in diagnosis, progression, and prognosis of LUAD and LUSC, that is, we first imported data on these m6A regulators into Cytoscape software [version 3.8.2 (Shannon et al., 2003)] and analyzed the data using the ClueGO plugin. After that, we performed the gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and network analysis, and the GO terms included the cellular component (CC), molecular function (MF), biological process (BP), an adjusted $p < 0.05$ as a statistically significant value. Afterward, we utilized the "LIMMA" package (Ritchie et al., 2015) to select and analyze the differentially expressed m6A RNA methylation regulatory genes in LUAD and LUSC tissue specimens using the cutoff value of |log2 fold change (FC)| ≥ 1 and adjusted p-value < 0.05. The "euclidean" and "ward.D2" methods were utilized to cluster the tumor samples, and the "p heat map" R package was used to plot the differential expression analysis results and cluster analysis results. Spearman's correlation analysis was performed to analyze the correlation between the clusters and clinical traits. We also performed the univariate and multivariate Cox regression analyses to predict the association of these m6A RNA methylated regulatory genes with the overall survival of patients using the "survival" R package (Rizvi et al., 2019).

Association of Gene Alternative Splicing Events With Overall Survival of Patients

We utilized the "WGCNA" (weighted gene co-expression network analysis) package (Langfelder and Horvath, 2008) to associate the GAS events with the overall survival of LUAD and LUSC patients. The WGCNA package is able to analyze thousands of the most varied genetic information to identify the sets of genes to associate with tumor phenotypes (Meng et al., 2019); thus, this tool allowed us to analyze the information regarding the GAS event data in association with clinical traits data and profile of m6A regulatory gene expressions, but avoided any unnecessary procedures for multiple hypothesis testing and corrections. In this analysis, we first estimated the standard scale-free network according to formula A to generate the adequate β value (appropriate soft threshold power). We then constructed the weighted adjacency matrix using the formula B and converted the data into a topological overlap matrix (TOM). After that, we utilized the dynamic tree cutting method according to the hierarchical clustering to identify the modules that highly correlated with GAS events. The GAS event then used 1-TOM as distance measurement with the depth (the cutoff value of 2) and the minimum size (the cutoff value of 60). After that, the highly similar modules were fused by clustering and height truncation of 0.3 according to previous studies (Niemira et al., 2019; Wan et al., 2018; Xie and Xie, 2019). Last, we performed the Spearman's correlation and module eigengenes analysis of m6A regulator genes expression for association with the clinical traits and prognosis of 487 LUAD patients. Significant data on the association of this m6A regulator genes expression with clinicopathological data were further analyzed according to a previous study using the |correlation coefficients| between m6A regulators and GAS events module more than 0.4 and adjusted $p < 0.05$ (Langfelder and Horvath, 2008).

In the LUSC cohort of patient's data, we selected and further analyzed the modules of the most significant correlation between clinical features and m6A regulator genes, that is, the |correlation coefficients| between m6A regulators and GAS events module more than 0.4 and adjusted $p < 0.05$. We then performed univariate and multivariate Cox Regression analyses to screen associate the GAS events with the overall survival of LUAD and LUSC patients.

The formula A: $\mathrm{Aij} = \mathrm{power}\,(\mathrm{Sij}, \beta) = |\mathrm{Smn}|^{\beta}$ (i and j represent the GAS event of i and j, respectively, while m and n were the numbers of node connections, and β was the appropriate soft threshold power). The formula B: $TOMij = \frac{\sum_{u} AiuAju + Aij}{\min(Ki,Kj) + 1 - Aij}$ (i and j represent the GAS event of i and j, respectively, while the letter u represents clinical traits and prognostic information).

The Gene Ontology Term and Kyoto Encyclopedia of Genes and Genomes Pathway Analyses of Overall Survival-Related Gene Alternative Splicing Genes

After screening the m6A-related GAS events, we performed the GO terms and KEGG pathways analysis of these overall survival-related GAS genes. Specifically, we imported data on the m6A-related GAS genes and m6A regulator genes into Cytoscape software (3.8.2) and analyzed them using the ClueGO plugin. After associating the overall survival of patients, we performed the GO terms and KEGG pathway analyses of these GAS events-related genes and the GO terms included the cellular component (CC), molecular function (MF), and biological process (BP) using an adjusted $p < 0.05$ as a statistically significant value according to a previous study (Amado et al., 2014). After that, we constructed the functional network of these corresponding genes using the Cytoscape software.

Risk Model Construction

We constructed the risk model using the LASSO Cox regression analysis that could prevent any overfitting of the overall survival-related genes according to a previous study (Tang et al., 2017), and then performed the multivariate Cox regression analysis to predict the usefulness of these overall survival-related genes using the following formula:

$J = 1n \sum i = 1n\,(f\,(xi) - yi)2 + \lambda \|w\| 1$ (the greater the value of J, the better the prediction value; the letter w indicates a globally optimal value of lost J).

$Risk\ \ score = \sum_{n=x}^{n} \mathrm{coef}\,(\mathrm{X}) * \mathrm{PSI}\,(\mathrm{X})$ [Coef(X) is the coefficient of each GAS gene and PSI(X) is the PSI value of the AS genes].

According to the hazard ratio (HR) values after the multivariate Cox regression analysis, we classified the m6A-related prognostic AS events into protective/risky AS events (HR > 1 as a risk factor; HR < 1 as a protective factor), and showed the Sankey diagram that plotted is by "ggalluvial, dplyr, and ggplot2" R packages (Graedel, 2019; Soh et al., 2019). According to the median value of the risk score of the signature of each cohort, we divided the LUAD and LUSC cohorts into two subgroups, that is, the high- and low-risk groups. We then utilized the "survival" and "survminer" package to calculate the survival significance of the high-/low-risk group in these NSCLC patients. We performed the Kaplan–Meier survival analysis and receiver operating characteristic (ROC) curves to further verify the predictive ability of the risk signature using the "survivalROC" package and "Survival" package in R (Park et al., 2004). The concordance index (C-index) was used to validate and quantify the discrimination ability of the risk signature. At last, we performed the univariate and multivariate Cox regression analyses to assess whether these risk models and clinicopathological features were independent predictors for the survival of LUAD and LUSC patients.

Predictive Nomogram Construction

After the univariate and multivariate Cox regression analyses, we used the "RMS" package to construct the nomogram of the independent risk factors (Zhang et al., 2019). We then performed the Wilcoxon rank-sum test to verify the association of the risk model with clinical characteristics (using an adjusted p-value < 0.05 as the statistical significance cutoff). For the construction of the predictive nomogram, we utilized the calibration curves to evaluate and validate the application ability of the nomogram performance.

Statistical Analysis

The "Limma" R package was utilized to analyze the difference in gene expression profiles, while the "WGCNA" package was used to select m6A-related GAS events. Moreover, the univariate, LASSO and multivariate Cox regression analyses were performed to construct the risk signature, while the "Survival," "survivalROC," and "survminer" packages were utilized to verify the predictive efficacy of the risk signature in patients, while the area under the curve (AUC) value (ranged between 0.5 and 0.9) was used to assess the diagnostic ability of the risk signature (larger AUC value, better diagnostic value) (Verbakel et al, 2020). The Wilcoxon rank-sum and Spearman's correlation tests were used to analyze the subgroup differences, while the "RMS" R package was utilized to plot the nomogram and calibrate the analytic data. The GO terms and KEGG pathways enrichment analysis were performed using Cytoscape software (3.8.2) and the "ggplot" R package was to plot the resulting data. All statistical analyses were performed using R software (version 3.6.1). A two-sided p value < 0.05 was considered statistically significant, while an adjusted p-value < 0.05 was applied as the threshold to avoid missing any significant changes.

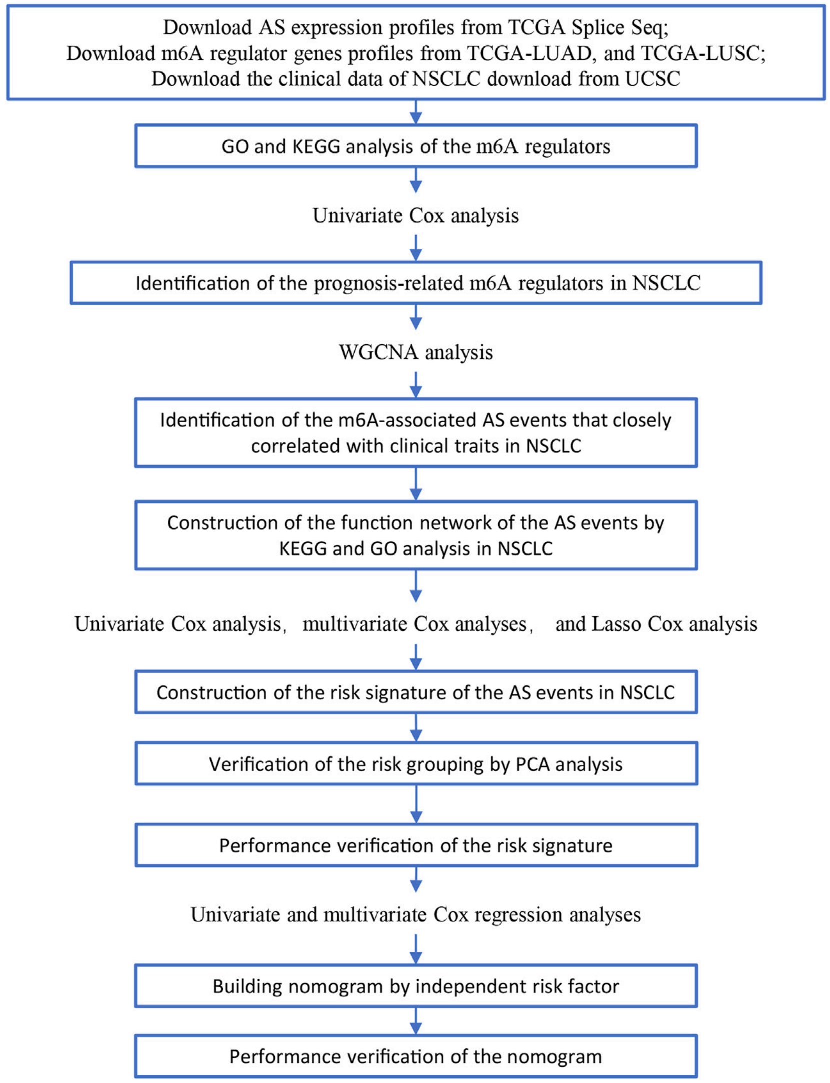

Download AS expression profiles from TCGA Splice Seq;
Download m6A regulator genes profiles from TCGA-LUAD, and TCGA-LUSC;
Download the clinical data of NSCLC download from UCSC
GO and KEGG analysis of the m6A regulators
Univariate Cox analysis
Identification of the prognosis-related m6A regulators in NSCLC
WGCNA analysis
Identification of the m6A-associated AS events that closely correlated with clinical traits in NSCLC
Construction of the function network of the AS events by KEGG and GO analysis in NSCLC
Univariate Cox analysis, multivariate Cox analyses, and Lasso Cox analysis
Construction of the risk signature of the AS events in NSCLC
Verification of the risk grouping by PCA analysis
Performance verification of the risk signature
Univariate and multivariate Cox regression analyses
Building nomogram by independent risk factor
Performance verification of the nomogram

FIGURE 1 | Illustration of the workflow in this study.

RESULTS

Characteristics of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma Data and mRNA Splicing Events in The Cancer Genome Atlas Datasets

In this study, we searched and downloaded expression profiles of TCGA-LUAD and TCGA-LUSC from TCGA database (https://portal.gdc.cancer.gov/) and clinicopathological data from the University of California Santa Cruz database (https://xena.ucsc.edu/). We obtained 486 LUAD and 479 LUSC cases for our data analysis (**Figure 1**). In LUAD samples, there were 224 men and 262 women with a median age of 64.94 years old (arranged between 33 and 88 years). The patients were at the TNM stage of I/II (n = 381) and III/IV (n = 105) and 167 cases had lymph node tumor metastasis (**Table 1**). In LUSC cases, there were 353 men and 126 women with a median age of 64.294 years old (arranged between 39 and 90 years). The patients were at the TNM stage of I/II (n = 391) and III/IV (n = 88) and 169 cases had lymph node tumor metastasis (**Table 1**). Our data analyses identified a total of 43,948 mRNA splicing events in LUAD tissue samples and 46,020 mRNA splicing events in LUSC samples. The exon skip (ES) events were the most GAS events in both LUAD and LUSC groups of samples (**Supplementary Figure S1**).

Association of N6-Methyladenosine RNA Methylation Regulatory Gene Expressions With Lung Adenocarcinoma and Lung Squamous Cell Carcinoma Prognosis

We focused on 13 m6A RNA methylation regulatory genes, including METTL3, METTL14, WTAP, KIAA1429, RBM15, ZC3H13, YTHDC1, YTHDC2, YTHDF1, YTHDF2, HNRNPC, FTO, and ALKBH5. We performed the GO terms and KEGG pathway analyses and found that these m6A regulators were

TABLE 1 | The univariate and multivariate cox regression analysis of clinicopathological data from TCGA-LUAD and LUSC.

Variables	Univariate analysis				Multivariate analysis			
	HR	**HR.95L**	**HR.95H**	***Adj-p***	**HR**	**HR.95L**	**HR.95H**	***Adj-p***
LUAD								
Age (≥65/<65)	1.111	0.823	1.500	4.92E-01	1.204	0.884	1.640	2.38E-01
Gender (female/male)	1.111	0.826	1.494	4.86E-01	1.077	0.791	1.467	6.37E-01
Smoking (yes/no)	0.864	0.440	1.697	6.72E-01	1.863	0.878	3.955	1.05E-01
TNM stage (I + II/III + IV)	1.635	1.422	1.881	***	1.564	1.187	2.062	***
T	1.548	1.296	1.848	***	1.066	0.865	1.315	5.48E-01
N	1.962	1.461	2.634	***	1.108	0.722	1.701	6.39E-01
M	2.181	1.302	3.653	**	0.740	0.355	1.543	4.21E-01
Risk score	1.441	1.346	1.543	***	1.390	1.281	1.508	***
LUSC								
Age (≥ 65/<65)	1.017	0.999	1.035	6.40E-02	1.028	1.007	1.050	*
Gender (female/male)	1.095	0.782	1.534	5.97E-01	1.225	0.834	1.800	3.01E-01
TNM stage (I + II/III + IV)	1.283	1.079	1.526	**	1.132	0.726	1.765	5.83E-01
T	1.352	1.121	1.630	**	1.434	1.083	1.897	**
N	1.170	0.952	1.438	1.35E-01	1.092	0.735	1.622	6.62E-01
M	2.455	0.905	6.659	7.77E-02	1.111	0.279	4.420	8.82E-01
Risk score	1.054	1.040	1.069	***	1.884	1.473	2.386	***

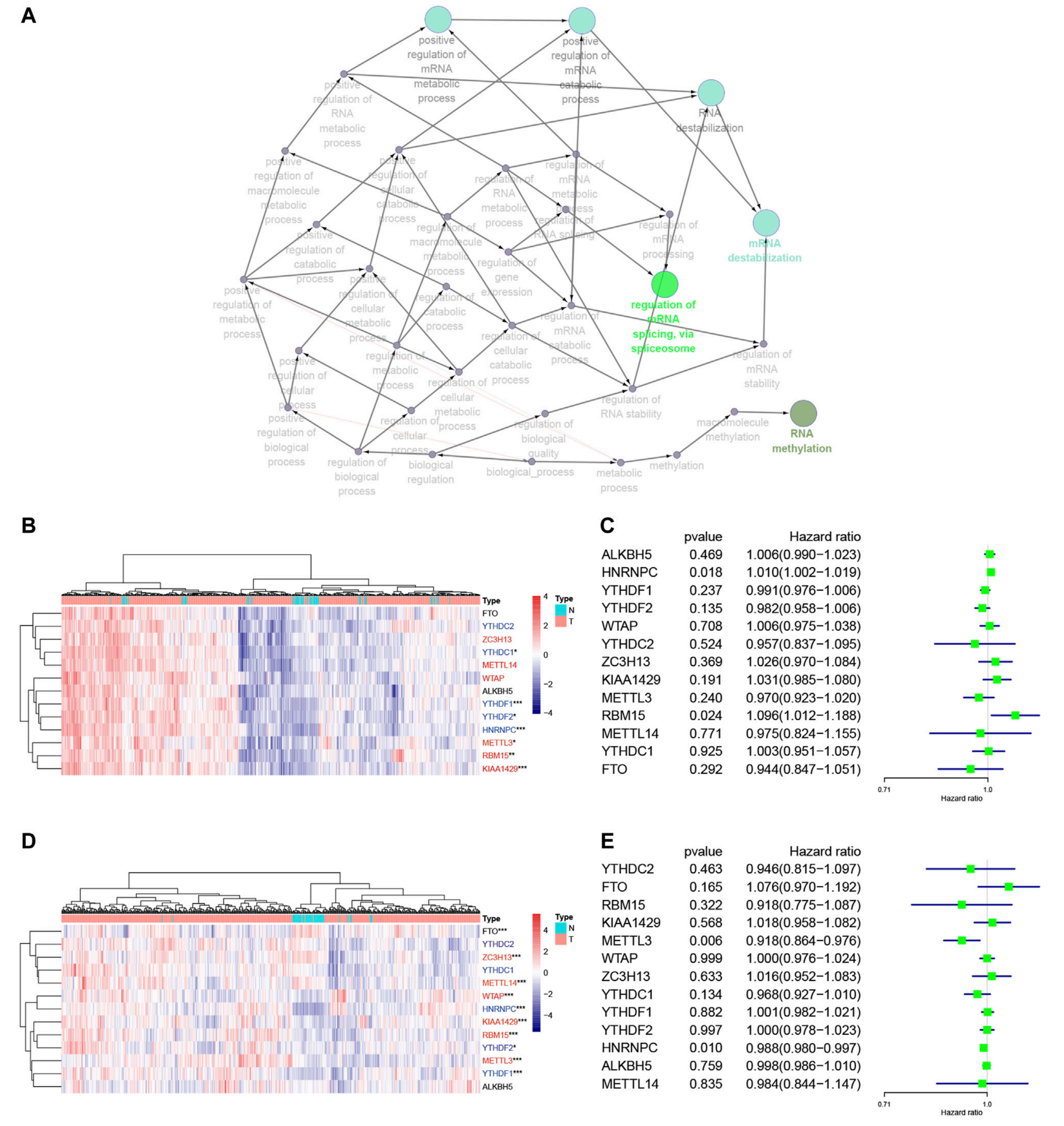
LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; TNM: tumor-node metastasis; HR: hazard ratio. represents* p < 0.05*; ** represents* p < 0.05*; *** represents* p < 0.001*.*

significantly enriched in the mRNA splicing spliceosome biology process, RNA methylation biology process, and RNA destabilization biology process (**Figure 2A**, **Supplementary Table S1**). We then performed the Wilcoxon signed-rank test and found that KIAA1429, HNRNPC, RBM15, METTL3, YTHDF1, YTHDF2, and YTHDC1 were differentially expressed between normal and LUAD tissues (**Figure 2B**). YTHDF1, YTHDF2, WTAP, KIAA1429, RBM15, METTL3, METTL14, FTO, HNRNPC, and ZC3H13 were differentially expressed between normal tissues and LUSC tissues (**Figure 2D**). Furthermore, KIAA1429, HNRNPC, RBM15, METTL3, YTHDF1, and YTHDF2 were all highly expressed in both LUAD and LUSC tissues (**Figures 2B,D**). The LUAD and LUSC patients were clustered into four groups according to the expression of m6A regulators in the heat map (**Supplementary Figure S2A,B**). The correlation analysis suggested that the differentially expressed m6A regulators were associated with status, smoking, TNM stage, and N stage in LUAD samples, and the differentially expressed m6A regulators were associated with status and age in LUSC samples (**Table 2**). These results suggested that m6A regulators play an important role in NSCLC development. The univariate Cox regression data revealed that HNRNPC and RBM15 expression were able to predict overall survival (OS) of LUAD patients (**Figure 2C**), while HNRNPC and METTL3 expression were associated with the OS of LUSC patients (**Figure 2E**); thus, these three m6A regulators genes were subjected to the subsequent analysis for association with OS of NSCLC patients as the splicing factors (**Figures 2C,E**).

Association of N6-Methyladenosine RNA Methylation Regulatory Genes With Lung Adenocarcinoma and Lung Squamous Cell Carcinoma Clinical Features

After that, we correlated the GAS events with the weighted gene co-expression network, and they were consistent with the scale-free network (**Supplementary Figure S3**). The hierarchical clustering analysis of the samples using the Euclidean distance showed log10-transformed RNA-seq fractional counts (**Supplementary Figure S4**), while the dynamic tree cutting method identified the modules with a similar expression spectrum and combine similar modules (**Figures 3A,C**, **Supplementary Figure S5**). We then utilized the "WGCNA" package to analyze the GAS events and Spearman's correlation test to associate the expression of m6A regulator genes with clinical traits. The data showed that the MEbrown module was significantly associated with expression of the m6A regulator genes (RBM15, p = 3e-24, the |correlation coefficient| = −0.44), gender (p = 0.03, the coefficient correlation = −0.1), and tobacco smoking (p = 0.006, the coefficient correlation = −0.12) of LUAD patients (**Figure 3B**, **Supplementary Figure S6A**). Furthermore, the MEred, MEblue, and MEroyalblue modules were significantly associated with expression of the m6A regulator genes (MEred, HNRNPC with an adj p = 1e-24 and the coefficient correlation = −0.44; MEblue, HNRNPC with an adj p = 2e-27 and the coefficient correlation = −0.47; MEroyalblue, HNRNPC with an adj p = 3e-33 and the coefficient correlation = −0.51). These three modules were also associated with the age of patients (MEred, adj p = 0.007 and the coefficient correlation = 0.12), the TNM stage (MEred, adj p = 0.03 and the coefficient correlation = −0.1), and the N stage (MEblue, adj p = 0.04 and the coefficient correlation = −0.092; MEroyalblue, adj p = 0.03 and the coefficient correlation = 0.098; in **Figure 3D**, **Supplementary Figure S6B**). These results suggested that the m6A-related AS events in the MEred, MEblue, and MEroyalblue modules could predict NSCLC development and lymph node metastasis, while the age of patients might also affect the m6A-related AS events.

Furthermore, the MEbrown module included 1.102 GAS events, and the MEred, MEblue, and MEroyalblue modules included 1.5150 GAS events. The most significant enrichment



C

	pvalue	Hazard ratio
ALKBH5	0.469	1.006(0.990−1.023)
HNRNPC	0.018	1.010(1.002−1.019)
YTHDF1	0.237	0.991(0.976−1.006)
YTHDF2	0.135	0.982(0.958−1.006)
WTAP	0.708	1.006(0.975−1.038)
YTHDC2	0.524	0.957(0.837−1.095)
ZC3H13	0.369	1.026(0.970−1.084)
KIAA1429	0.191	1.031(0.985−1.080)
METTL3	0.240	0.970(0.923−1.020)
RBM15	0.024	1.096(1.012−1.188)
METTL14	0.771	0.975(0.824−1.155)
YTHDC1	0.925	1.003(0.951−1.057)
FTO	0.292	0.944(0.847−1.051)

E

	pvalue	Hazard ratio
YTHDC2	0.463	0.946(0.815−1.097)
FTO	0.165	1.076(0.970−1.192)
RBM15	0.322	0.918(0.775−1.087)
KIAA1429	0.568	1.018(0.958−1.082)
METTL3	0.006	0.918(0.864−0.976)
WTAP	0.999	1.000(0.976−1.024)
ZC3H13	0.633	1.016(0.952−1.083)
YTHDC1	0.134	0.968(0.927−1.010)
YTHDF1	0.882	1.001(0.982−1.021)
YTHDF2	0.997	1.000(0.978−1.023)
HNRNPC	0.010	0.988(0.980−0.997)
ALKBH5	0.759	0.998(0.986−1.010)
METTL14	0.835	0.984(0.844−1.147)

FIGURE 2 | Association of m6A RNA methylation regulatory genes with NSCLC prognosis. **(A)** The GO terms and KEGG enrichment pathway analysis of the m6A regulator gene, the different colors represent the different pathways. **(B)** Differential expression of these 13 m6A RNA methylated regulator genes in LUAD (red: "writer"; blue: "readers"; black: "erasers"). **(C)** Forest plot of the univariate Cox regression analytic data. The 13 m6A RNA methylation regulators in LUAD were analyzed using the univariate Cox regression and the data are plotted using the forest plot. **(D)** Differential expression of these 13 m6A RNA methylated regulators in LUSC (red: "writer"; blue: "readers"; black: "erasers"). **(E)** Forest plot of the univariate Cox regression analysis. The 13 m6A RNA methylation regulators in LUAD were analyzed using the univariate Cox regression and the data are plotted using the forest plot. ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$. The KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; m6A, N6-methyladenosine; N, Normal; T, Tumor.

TABLE 2 | The correlation analysis of clinical traits and m6A clusters from TCGA-LUAD and LUSC.

		Overall	Cluster 1	Cluster 3	Cluster 4	Cluster 4	*p*
LUAD							
n		486	148	86	129	123	
Status (%)	Alive	366 (76.0)	123 (81.8)	61 (70.9)	76 (64.9)	106 (86.2)	***
	Dead	120 (24.0)	25 (18.2)	25 (29.1)	53 (35.1)	17 (13.8)	
Gender (%)	Female	262 (53.7)	74 (50.0)	52 (60.5)	69 (52.0)	67 (54.5)	0.483
	Male	224 (46.3)	74 (50.0)	34 (39.5)	60 (48.0)	56 (45.5)	
Age (%)	<65	215 (45.5)	49 (43.9)	40 (46.5)	65 (43.0)	61 (49.6)	0.71
	≥65	271 (54.5)	99 (56.1)	46 (53.5)	64 (57.0)	62 (50.4)	
Smoking (%)	No	13 (2.6)	1 (0.7)	7 (8.1)	1 (0.7)	4 (3.3)	**
	Yes	495 (97.4)	147 (99.3)	79 (91.9)	150 (99.3)	119 (96.7)	
Stage (%)	Stage I	263 (55.7)	85 (56.1)	38 (44.2)	60 (54.3)	80 (65.0)	*
	Stage II	118 (23.4)	38 (26.4)	25 (29.1)	33 (21.9)	22 (17.9)	
	Stage III	80 (15.7)	14 (9.5)	20 (23.3)	30 (19.9)	16 (13.0)	
	Stage IV	25 (5.1)	11 (8.1)	3 (3.5)	6 (4.0)	5 (4.1)	
T (%)	T1	165 (34.1)	59 (30.4)	26 (30.2)	30 (34.4)	50 (40.7)	0.104
	T2	256 (52.8)	74 (58.1)	47 (54.7)	73 (48.3)	62 (50.4)	
	T3	44 (9.1)	7 (5.4)	10 (11.6)	16 (10.6)	11 (8.9)	
	T4	21 (4.1)	8 (6.1)	3 (3.5)	10 (6.6)	0 (0.0)	
N (%)	N0	319 (67.3)	102 (69.6)	49 (57.0)	77 (65.6)	91 (74.0)	*
	N1	91 (17.7)	29 (18.9)	19 (22.1)	25 (16.6)	18 (14.6)	
	N2	70 (13.8)	12 (8.1)	18 (20.9)	26 (17.2)	14 (11.4)	
	N3	6 (1.2)	5 (3.4)	0 (0.0)	1 (0.7)	0 (0.0)	
M (%)	M0	460 (94.9)	136 (91.9)	83 (96.5)	123 (96.0)	118 (95.9)	0.274
	M1	26 (5.1)	12 (8.1)	3 (3.5)	6 (4.0)	5 (4.1)	
LUSC							
n		479	84	169	125	101	
Status (%)	Alive	293 (61.3)	60 (71.4)	108 (64.1)	63 (50.4)	62 (61.4)	*
	Dead	186 (38.8)	24 (28.6)	61 (35.9)	62 (49.6)	39 (38.6)	
Gender (%)	FEMALE	126 (26.5)	24 (28.6)	37 (22.4)	36 (28.8)	29 (28.7)	0.516
	MALE	353 (73.5)	60 (71.4)	132 (77.6)	89 (71.2)	72 (71.3)	
Age (%)	<65	166 (34.8)	37 (44.0)	61 (36.5)	30 (24.0)	38 (37.6)	*
	≥65	313 (65.2)	47 (56.0)	108 (63.5)	95 (76.0)	63 (62.4)	
Stage (%)	Stage I	235 (49.2)	35 (41.7)	83 (49.4)	60 (48.0)	57 (56.4)	0.176
	Stage II	156 (32.5)	33 (39.3)	54 (31.8)	46 (36.8)	23 (22.8)	
	Stage III	81 (16.9)	14 (16.7)	32 (18.8)	16 (12.8)	19 (18.8)	
	Stage IV	7 (1.5)	2 (2.4)	0 (0.0)	3 (2.4)	2 (2.0)	
T (%)	T1	109 (22.9)	16 (19.0)	34 (20.6)	32 (25.6)	27 (26.7)	0.567
	T2	281 (58.5)	51 (60.7)	103 (60.6)	67 (53.6)	60 (59.4)	
	T3	68 (14.2)	13 (15.5)	25 (14.7)	22 (17.6)	8 (7.9)	
	T4	21 (4.4)	4 (4.8)	7 (4.1)	4 (3.2)	6 (5.9)	
N (%)	N0	310 (64.8)	54 (64.3)	106 (62.9)	82 (65.6)	68 (67.3)	0.263
	N1	125 (26.0)	23 (27.4)	49 (28.8)	35 (28.0)	18 (17.8)	
	N2	39 (8.1)	6 (7.1)	14 (8.2)	6 (4.8)	13 (12.9)	
	N3	5 (1.0)	1 (1.2)	0 (0.0)	2 (1.6)	2 (2.0)	
M (%)	M0	472 (98.5)	82 (97.6)	169 (100.0)	122 (97.6)	99 (98.0)	0.264
	M1	7 (1.5)	2 (2.4)	0 (0.0)	3 (2.4)	2 (2.0)	

*LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; T: tumor; N: node; M: metastasis; * represents* $p < 0.05$; *** represents* $p < 0.05$; **** represents* $p < 0.001$.

of the GO terms and KEGG pathway analysis of LUAD cohort revealed the m6A-related AS events were significantly enriched in the GPCR signaling pathway, DNA metabolic process, DNA repair, cellular response to DNA damage stimulus, carbon–oxygen lyase activity, and cell adhesion molecule binding pathways, while the m6A-related AS events in LUSC cohort were significantly enriched in the TGF-beta signaling pathway, peptidyl-serine phosphorylation, peptidyl-serine modification, regulation of actin cytoskeleton, intracellular signaling by second messengers, and early endosome pathway (adj $p < 0.001$; **Figure 4**, **Supplementary Table S2**).

Association of N6-Methyladenosine-Related Alternative Splicing Events With Lung Adenocarcinoma and Lung Squamous Cell Carcinoma Prognosis

We first performed the univariate Cox regression analysis to identify m6A-related AS events for association with LUAD and LUSC prognosis. We found 292 prognostic AS events in LUAD and 922 prognostics AS events in LUSC ($p < 0.05$; **Figures 5A,D**, **Supplementary Table S3**). The LASSO Cox regression analysis confirmed 13 of the prognostic AS events in LUAD (**Figures 5B,C**) and 15 in LUSC (**Figures 5E,F**), while the multivariate

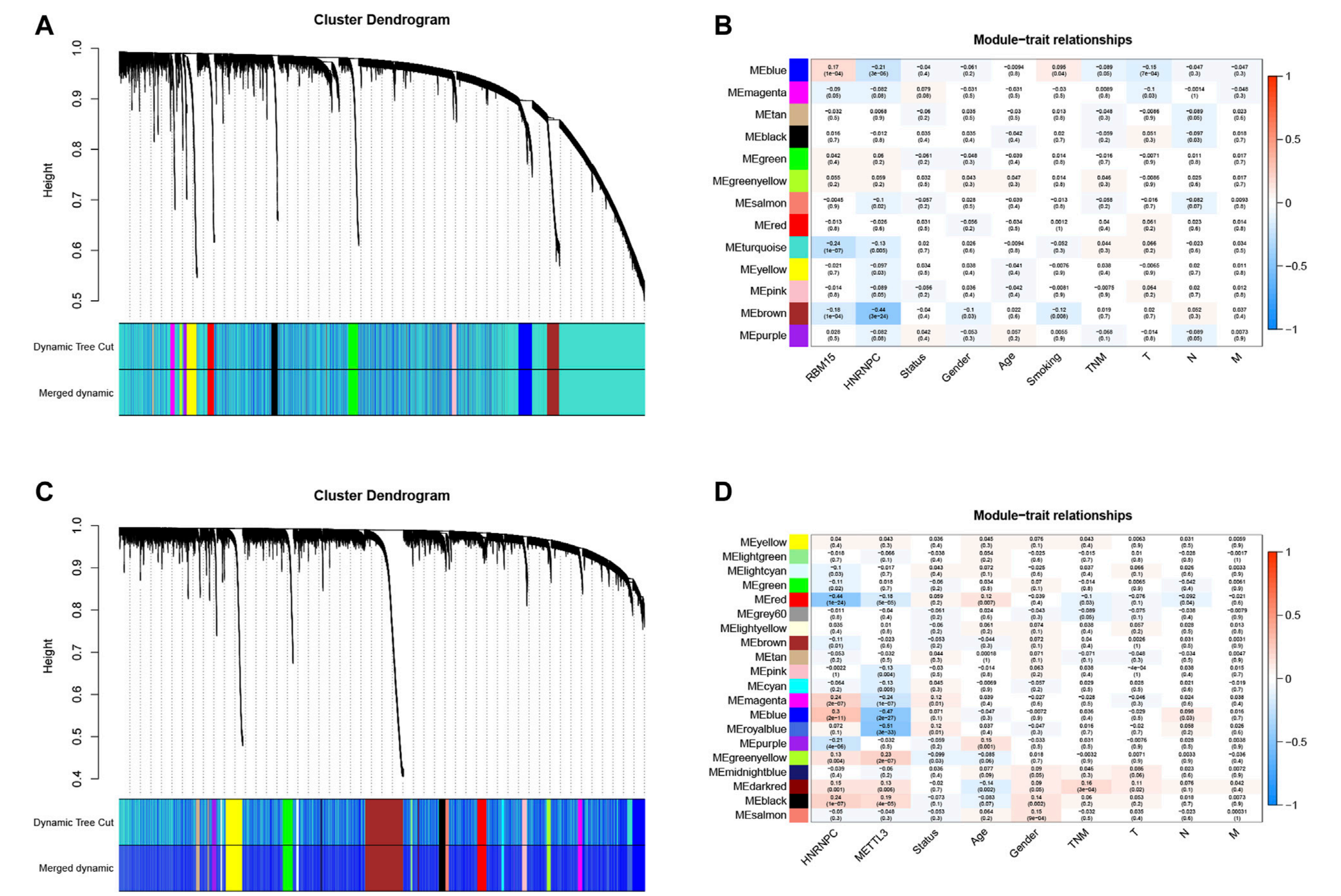


FIGURE 3 | Identification of the AS events that are associated with clinical traits and the expression of m6A regulators. **(A)** Hierarchical cluster tree of the AS events in TCGA-LUAD samples. The tree indicates a unique TCGA-LUAD name and experiment identifier. **(B)** Association of the AS events with gender, age, LUAD lymphatic infiltration, tumor status, and pathological stages. The number represents the Pearson correlation values between the module and the features. The number in parenthesis is the *p*-value, while the numbers in the color column represent the intensity of the correlation in the table (red = 1, blue = – 1). **(C)** Hierarchical clustering tree of the AS events in TCGA-LUSC samples. The tree indicates a unique name and experiment identifier marked TCGA-LUSC. **(D)** Association of the AS events with gender, age, LUSC lymphatic vessel invasion, tumor status, and pathological stages. The number represents the Pearson correlation values between the module and the features The number in parenthesis is the *p*-value, while the numbers in the color column represent the intensity of the correlation in the table (red = 1, blue = 1). T, tumor scope; N, degree of diffusion to lymph nodes; M, there is a transfer; m6A, N6-methyladenosine, LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; AS, alternative splicing.

Cox regression analysis further confirmed seven of the prognostic AS events in LUAD and 14 in LUSC (**Supplementary Table S4**).

Non–Small Cell Lung Cancer Prognosis-Related Gene Alternative Splicing Events Signature

We utilized these seven and 14 AS genes in LUAD and LUSC, respectively, to further construct the LUAD and LUSC risk signature (**Supplementary Table S5**). The Sankey diagram shows that DGKZ|15540|AP and PMP22|39340|AP were the risky m6A-related AS events in LUAD (HR > 1), whereas ABCC6|34219|AT, KIAA0586|27718|ES, LDB1|12935|AP, RPS25|19054|ES, and S100A14|7729|AP were the protective m6A-related AS events in LUAD (HR < 1) (**Figure 6A, Supplementary Table S5**). Furthermore, AKR1E2|10639|ES and SSH1|24258|ES were the risky m6A-related AS events in LUSC (HR > 1), whereas ALPK1|70369|ES, FAM63A|7531|AP, CHMP1A|38102|ES, TSTD2|87013|AT, KIAA1598|13239|AP, ASXL3|45046|AT, VPS37A|82796|ES, TOX2|59455|ES, ZNF544|52429|ES, NOL8|86863|ES, FAM124B|57772|AT, and PTCHD4|76446|AT were the m6A-related protective AS events in LUSC (HR < 1; **Figure 6D, Supplementary Table S5**). We then divided LUAD and LUSC patients into high- and low-risk groups according to their risk scores (high-risk LUAD group, *n* = 240; high-risk LUSC group, *n* = 239; low-risk LUAD group, *n* = 246; and low-risk LUSC group, *n* = 240; **Supplementary Table S6**). The Kaplan–Meier curve analyses showed that the high-risk group had a poorer OS than the low-risk group ($p < 0.001$; **Figures 6B,E**). The ROC analysis revealed that the AUC values were 0.868, 0.834, and 0.801 for 1-, 3-, and 5-year OS of LUAD, respectively, while the AUC values were 0.893, 0.824, and 0.849

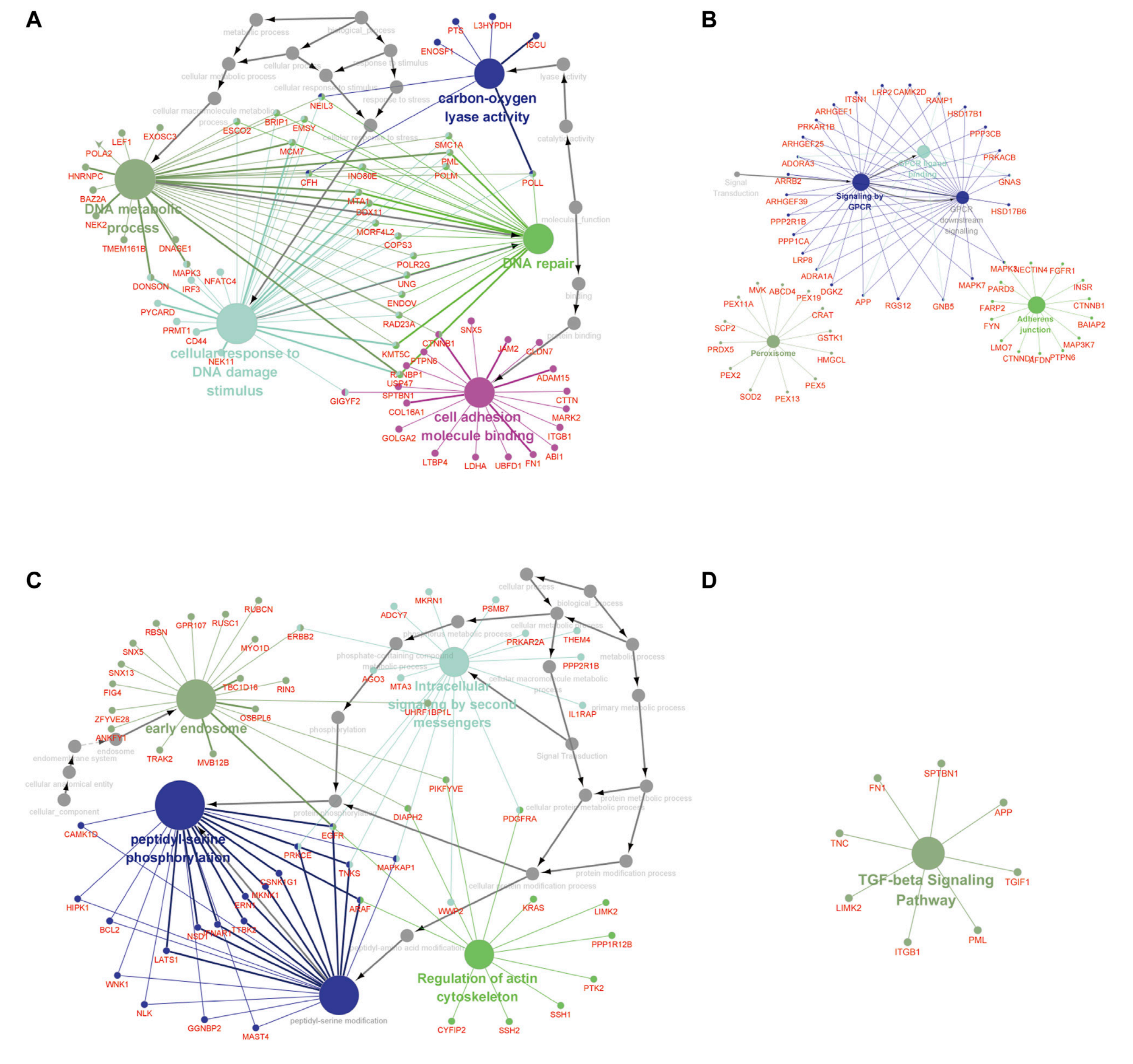

A
B
C
D
carbon-oxygen lyase activity
DNA metabolic process
DNA repair
cellular response to DNA damage stimulus
cell adhesion molecule binding
Signaling by GPCR
GPCR downstream signalling
GPCR ligand binding
Signal Transduction
Peroxisome
Adherens junction
early endosome
Intracellular signaling by second messengers
peptidyl-serine phosphorylation
peptidyl-serine modification
Regulation of actin cytoskeleton
TGF-beta Signaling Pathway
SPTBN1
FN1
APP
TNC
TGIF1
LIMK2
ITGB1
PML

FIGURE 4 | The GO terms and KEGG pathway analyses of genes from m6A-related AS events in NCSLC. The different colors represent the different pathways **(A, B)** in the regulation network. The GO terms **(A)** and KEGG pathways enrichment **(B)** analysis of the m6A-related prognostic AS genes in LUAD. **(C, D)** The regulation network. The GO terms **(C)** and KEGG pathways enrichment **(D)** analysis of the m6A-related prognostic AS genes in LUSC. The KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; AS, alternative splicing.

for 1-, 3-, and 5-year LUSC OS, respectively (**Figures 6C,F**). The concordance index (C-index) of OS was 0.847 [95% confidence interval (CI): 0.788–0.847] in LUAD and 0.832 in LUSC (95% CI: 0.788–0.877; **Supplementary Table S7**).

Furthermore, we found that the risk signature (univariate Cox analysis, HR: 1.543, 95% CI: 1.441–1.346; $p < 0.001$; multivariate Cox analysis, HR: 1.390, 95% CI: 1.281–1.508; $p < 0.001$) and the TNM stage (univariate Cox analysis, HR: 1.635, and 95% CI: 1.422–1.881; $p < 0.001$; multivariate Cox analysis, HR: 1.564, 95% CI: 1.187–2.062; $p = 0.001$) were independent prognostic factors of LUAD, while the risk signature (univariate Cox analysis, HR: 1.054, 95% CI: 1.040–1.069; $p < 0.001$; multivariate Cox analysis, HR: 1.884, 95% CI: 1.472–2.386; $p < 0.001$) and T stage (univariate Cox analysis, HR: 1.352, 95% CI: 1.121–1.630; $p = 0.002$; multivariate Cox analysis, HR: 1.434, 95% CI: 1.083–1.897; $p = 0.012$) were independent prognostic factors in LUSC (**Figures 7A,B**, **Table 2**). The risk curve and scatterplot of the risk score and survival of each NSCLC sample and the heat map of these AS genes in NSCLC samples are shown in **Figures 7Aiii–v, iii–v**.

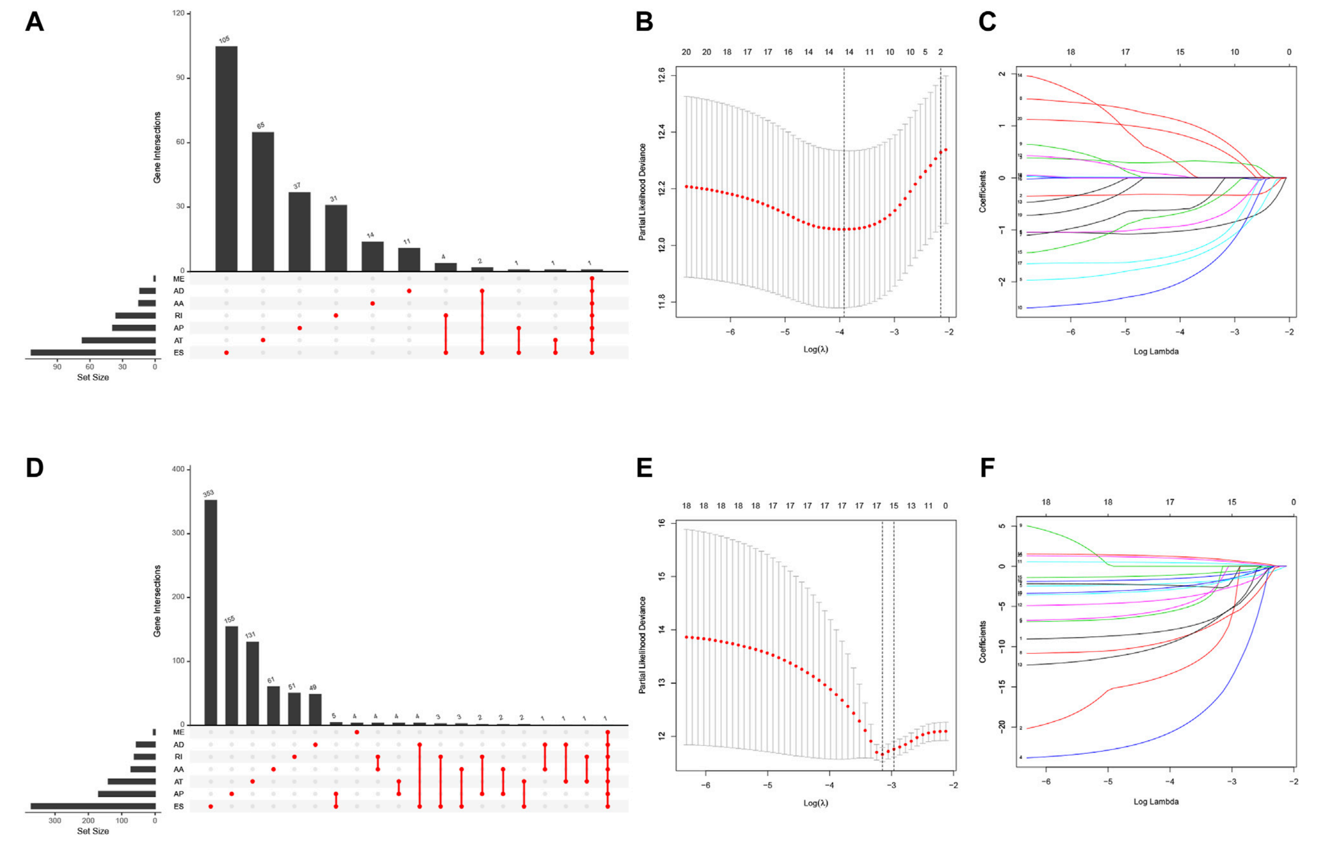


FIGURE 5 | Identification of NSCLC prognosis-related AS events. **(A)** The prognostic upset plot. The data exhibit the m6A-related prognostic AS events in LUAD. **(B, C)** The LASSO Cox analysis. These 13 m6A-related AS events associated with LUAD prognostics and the optimal values of the penalty parameter were assessed using the 10-round cross-validation. **(D)** The prognostic upset plot. The data exhibit the m6A-related prognostic AS events in LUSC. **(E, F)** The LASSO Cox analysis. These 15 m6A-related AS events associated with LUSC prognostics and the optimal values of the penalty parameter were determined by the 10-round cross-validation. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; m6A, N6-methyladenosine; OS, overall survival: AS, alternative splicing.

Association of These Prognostic Signatures With Non–Small Cell Lung Cancer Clinicopathologies

After that, we associated these prognostic signatures with NSCLC clinicopathologies and found that the LUAD risk signature was associated with the gender of patients and tumor T, N, and TNM stages (adj $p < 0.05$; **Figures 8A,C**, **Table 3**), although there was no association occurred between the LUSC risk signature and clinical features (**Table 3**). Furthermore, male ($n = 224$), TNM stage III–IV ($n = 105$), N stage 1–3 ($n = 167$), and T stage 3–4 ($n = 65$) of LUAD patients in had significantly higher risk scores than female ($n = 262$), TNM stage I–II ($n = 381$), N stage 0 ($n = 319$), and T stage 1–2 ($n = 421$; all adj $p < 0.05$; **Figure 8**, **Table 3**). Older age ($n = 271$) and M1 stage ($n = 26$) of LUAD patients also had the higher risk scores (all adj $p > 0.05$; **Figure 8B**, **Table 3**). Notably, male ($n = 353$), TNM stage III–IV ($n = 88$), N stage 1–3 ($n = 169$), M1 stage ($n = 7$), older age ($n = 313$), and T stage 3–4 ($n = 89$) of LUSC patients also had higher risk scores than those of the corresponding subgroups, but the differences did not appear statistically significant (**Figure 8D**).

Usefulness of the Predictive Nomogram in Non–Small Cell Lung Cancer

So far, we showed the risk signature and TNM stage as independent prognostic predictors in LUAD and the risk signature and T stage as independent prognostic predictors in LUSC. We thus, constructed the nomogram using these parameters to assess and apply this risk model for NSCLC (**Figures 9A,C**) and verified the calibration curves of the nomogram (**Figures 9B,D**). We were able to use the numerous values of this risk model to predict the 1-, 3-, and 5-year survival of NSCLC patients.

DISCUSSION

In the current study, we analyzed the aberrant expression of 13 m6A regulatory genes and related GAS events to construct a risk gene signature to predict the overall survival of NSCLC patients. We found that a number of them were highly expressed in LUAD or LUSC tissues vs. their normal ones, which could be used to predict the survival of patients. Furthermore, we found 43,948

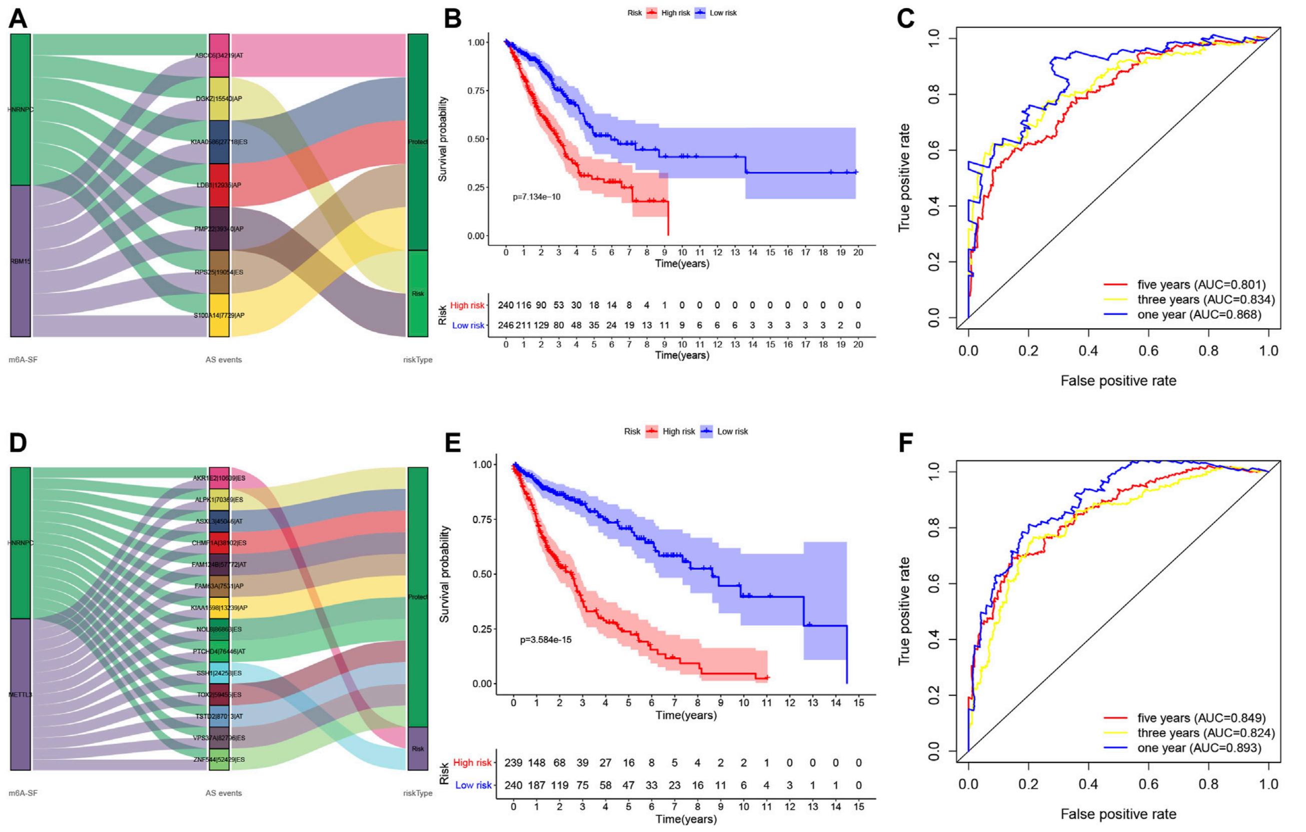


FIGURE 6 | Construction of the prognosis-related AS events signature. **(A)** The Sankey diagram. The data show the potential prognostic value of the m6A-related prognostic AS events in LUAD. **(B)** Kaplan–Meier curves stratified by high- and low-risk groups of LUAD. **(C)** The ROC curves in LUAD. **(D)** The Sankey diagram. The data show the potential prognostic value of the m6A-related prognostic AS events in LUSC. **(E)** Kaplan–Meier curves stratified by high- and low-risk groups of LUSC. **(F)** The ROC curves in LUSC. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; ROC, receiver operating characteristic curve.

mRNA splicing events in LUAD and 46,020 in LUSC and m6A regulators could regulate mRNA splicing. We then constructed the NSCLC prognosis-related AS events signature and divided the patients into high- *vs.* low-risk groups using seven and 14 AS genes in LUAD and LUSC, respectively. The data showed that DGKZ|15540|AP and PMP22|39340|AP were the risky m6A-related AS events in LUAD, whereas ABCC6|34219|AT, KIAA0586|27718|ES, LDB1|12935|AP, RPS25|19054|ES, and S100A14|7729|AP were the protective m6A-related AS events in LUAD. Similarly, AKR1E2|10639|ES and SSH1|24258|ES were the risky m6A-related AS events in LUSC, whereas ALPK1|70369|ES, FAM63A|7531|AP, CHMP1A|38102|ES, TSTD2|87013|AT, KIAA1598|13239|AP, ASXL3|45046|AT, VPS37A|82796|ES, TOX2|59455|ES, ZNF544|52429|ES, NOL8|86863|ES, FAM124B|57772|AT, and PTCHD4|76446|AT were the m6A-related protective AS events in LUSC. Further analyses showed that the LUAD risk signature was associated with the gender of patients and tumor T, N, and TNM stages. In addition, the risk signature and TNM stage were independent prognostic predictors in LUAD and the risk signature and T stage were independent prognostic predictors in LUSC. In conclusion, our current study demonstrated the usefulness of this AS prognostic signature in the prediction of LUAD and LUSC prognosis. Further study will verify this AS signature in a prospective dataset from NSCLC patients.

M6A modification and GAS occur most commonly in mRNA transcripts and their alterations play an important role in the development and progression of human cancers (Cherry and Lynch, 2020; Sun et al., 2019). Accumulated evidence suggests that m6A regulators-mediated gene methylation played a critical role in NSCLC development (ref); however, the underlying molecular mechanisms of m6A regulator actions in cancer development remain to be fully elucidated. Recently, the m6A regulators have been shown to act as an important splicing factor during GAS events (Kasowitz et al., 2018; Yoshimi et al., 2019; Geng et al., 2020), although research of the m6A regulator regulating AS events is still in the early stage in the field of cancer research, including lung cancer. Therefore, our current study conducted the GO terms and KEGG pathway analyses of these m6A and related GAS events in NSCLC and found that m6A regulators were significantly enriched in the regulation mRNA splicing spliceosome biology process. We also found that the expression of some of them, including METTL3, HNRNPC, and RBM15, could predict NSCLC prognosis,

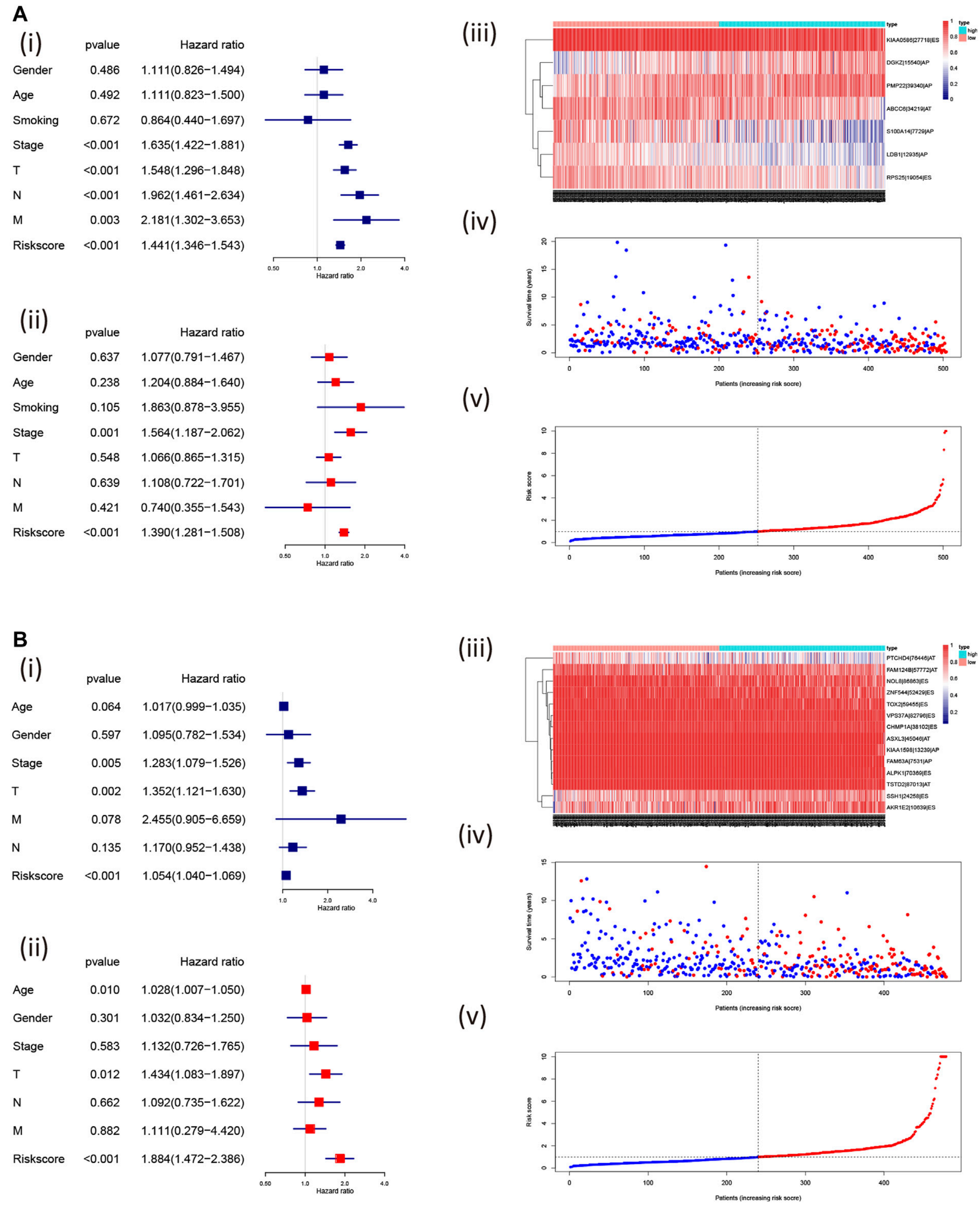


FIGURE 7 | Assessment of the independent prognostic value using the risk scores. The univariate and multivariate Cox regression analyses of the risk signature in LUAD **(A)**; i, univariate and ii, multivariate Cox regression analyses and in LUSC **(B)**; i, univariate and ii, multivariate Cox regression analyses risk signature. The visualized gene percent spliced index value in LUAD (**A**; iii) and LUSC (**B**; iii) and risk scores were associated with NSCLC survival [LUAD: **(A)**: iv, v; LUSC: **(A)**: iv, v; the red dot or line represents the deceased, while the blue dot or line represents alive]. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PSI, percent spliced in index.

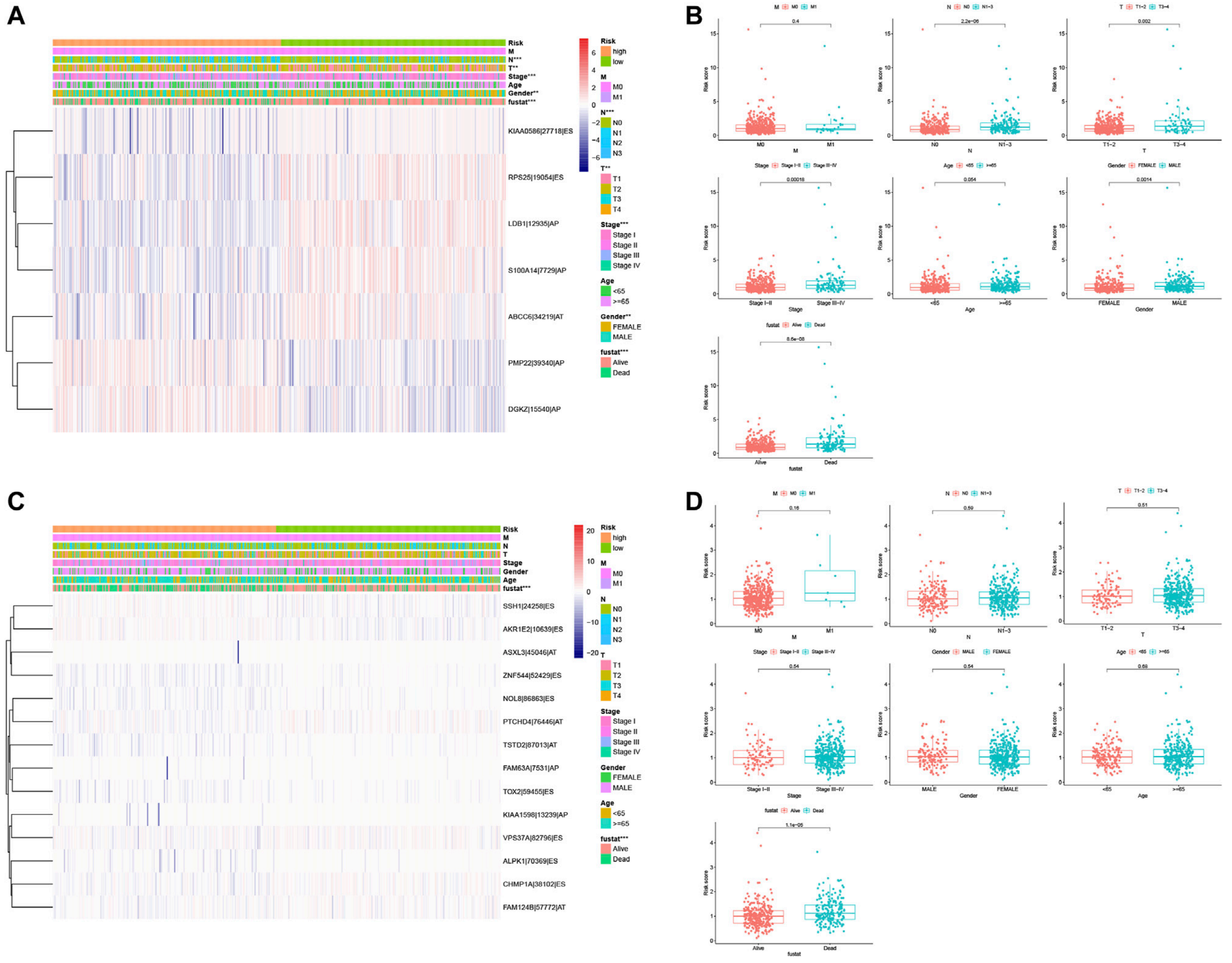


FIGURE 8 | Association of the risk signature with clinicopathological features from both LUAD and LUSC. **(A)** Heat map of different clinicopathological characteristics from high- *vs.* low-risk LUADs stratified by percent spliced index (PSI) of the AS events. **(B)** Association of the risk scores with clinicopathological features from LUAD. **(C)** Heat map of different clinicopathological features of from high- *vs.* low-risk stratified by the PSI of the AS events. **(D)** Association of the risk scores with clinicopathological features from LUSC. ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$. AS, alternative splicing; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PSI, percent spliced in index.

although the hazard ratios suggest that their prediction of NSCLC survival might be marginal. A previous study from Sun et al. (2020) showed the usefulness of the m6A regulators as the risk signature in LUAD (AUC, 0.65–0.82). Our results also suggested that m6A regulators play an important role in NSCLC development. Other previous studies (Lee et al., 2010; Gruber et al., 2016; Li et al., 2020) reported that HNRNPC was an RNA-binding protein (the “reader”), which could regulate RNA splicing, 3-terminal processing, and translation (Gruber et al., 2016; Lee et al., 2010; Li et al., 2020). HNRNPC overexpression was observed in a variety of human cancers, including lung cancer (Park et al., 2012) HNRNPC, as a protein-coding gene, could also interact with KHSRP to activate the IFN-α-JAK-p-STAT1 signaling pathway and promoted NSCLC cell proliferation, migration, and invasion (Yan et al., 2019). It can also regulate the GAS as an “m6A switcher” (Alarcón et al., 2015; Dai et al., 2018; Li et al., 2017; Wang et al., 2020). Furthermore, RBM15, as a “writer,” can bind to METTL3 and WTAP and direct them to specific RNA sites for m6A modification (Wang et al., 2020), although it does not possess any catalytic functions (Chen et al., 2019). RBM15 was also shown to interact with the METTL3 complex and depletion of these adapters could also reduce the m6A level (Pendleton et al., 2017). Further investigation of RBM15 and GAS events revealed that RBM15 was able to bind to specific intron regions to recruit the splicing factor SF3B1AS (Zhang et al., 2015). In addition, *METTL3*, containing highly conserved sequences, is the most important component of the m6A methyltransferase complex and was shown to be an S-adenosyl methionine (SAM)–binding protein and catalyze m6A modification (Wang et al., 2016).

TABLE 3 | Clinicopathological features from LUAD and LUSC subgroups stratified by the AS events signature.

	LUAD	*Adj-p*	LUSC	*Adj-p*
Gender		*		8.96E-01
Male	224		353	
Female	262		126	
Age		1.38E-01		9.50E-01
<65	215		166	
≥65	271		313	
TNM stage		***		5.04E-01
I	263		235	
II	118		156	
III	80		81	
IV	25		7	
Tumor (T)		**		7.25E-01
T1	165		110	
T2	256		280	
T3	44		68	
T4	21		21	
Lymph node (N)		***		6.30E-01
N0	319		310	
N1	91		125	
N2	70		39	
N3	6		5	
Metastasis (M)		5.17E-01		9.96E-01
M0	460		472	
M1	26		7	
Status		***		**
Dead	120		293	
Alive	366		186	
Total case	486		479	

*LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; TNM: tumor-node metastasis; *** represents* $p < 0.001$*; ** represents* $p < 0.01$*; * represents* $p < 0.05$.

METL3 expression was high in LUAD and promoted the translation of the epidermal growth factor receptor (EGFR) mRNA and hippo pathway effector TAZ mRNA in lung cancer cells, for induction of cell growth, survival, and invasion (Lin et al., 2016). METTL3 was also shown to interact with GAS events of the skipped exons and alternative first exon (Alarcón et al., 2015), and METTL3 dysregulation was reported to indeed affect GAS events (Katz et al., 2015; Liu et al., 2014). METTL3 silence significantly affected gene expression and alternative splicing patterns, leading to modulation of the p53 pathway and cell apoptosis (Dominissini et al., 2012). Taken altogether, these three m6A RNA methylation regulatory genes were important in the regulation of GAS events in NSCLC.

Indeed, AS events is an important mRNA modification process and produce a large number of mRNA and protein isoforms with different regulatory functions (Bonnal et al., 2020; Liu et al., 2020). The prognostic value of the AS events in NSCLC has well been documented, for example, Zhao et al. (2020) built a predictive model of aberrant AS events and predicated NSCLC prognosis. Indeed, alternation in splicing factor expression could alter many AS events in NSCLC (Coomer et al., 2019). For instance, QKI was shown to one of the most downregulated splicing factors in NSCLC, while QKI-5 was able to competitively bind to NUMB with SF1 protein to induce the NUMB exon 11 skip and, therefore, inhibited the Notch signaling (Zong et al., 2014; de Miguel et al., 2016). In lung cancer, QKI expression was significantly reduced, increasing in the abnormal splicing of num exon 11 to, in turn, activate the Notch signaling pathway and tumor cell proliferation (Zong et al., 2014; de Miguel et al., 2016). The AS events also influenced p53 expression in NSCLC and MDM2-B, an AS product of MDM2, was able to promote p53-independent cell growth and inhibition of apoptosis (Coomer et al., 2019). In this regard, the AS events are important in NSCLC development and progression (Bonnal et al., 2020; Sciarrillo et al., 2020).

Furthermore, the weighted gene co-expression network analysis (WGCNA) is a widely used data mining method, especially used for studying the biological networks based on pairwise correlations between variables (Langfelder and Horvath, 2008). In the current study, we used WGCNA to select the AS events that are highly correlated with the NSCLC survival-related m6A RNA regulators. After that, we performed the GO and KEGG pathways enrichment analysis to identify genes of m6A-related AS events to significantly participate in gene pathways that play an important role in NSCLC tumorigenesis, progression, drug sensitivity, and metastasis. Indeed, some of the abnormal AS events were associated with drug sensitivity and resistance of NSCLC (Motegi et al., 2019; Pilié et al., 2019) as well as cell adhesion molecule binding process (Song et al., 2013; Hintermann and Christen, 2019). Our KEGG analysis showed that the genes in the m6A-related AS events significantly participated in the GPCR signaling in LUAD, and the latter is mediated by three major G protein subclasses and each subclass also has multiple proteins that are products due to the AS events (Kang et al., 2015; Kallifatidis et al., 2020). Similarly, we found the m6A-related AS events in the TGF-β signaling in LUSC. The TGF-β signaling pathway was frequently downregulated in human cancers (Syed, 2016), whereas this pathway activation could also promote tumorigenesis, metastasis, and chemoresistance (Colak and Ten Dijke, 2017; Zi, 2019). In this regard, genes of the m6A-related AS event-led activation of the TGF-β signaling could promote LUSC tumorigenesis. However, further study is needed to confirm this speculation.

In addition, in our current study, we constructed the risk signature using these altered genes in m6A and AS events to associate with NSCLC prognosis, and the AUC of the ROC curves showed the sensitivity and specificity of LUAD and LUSC, respectively, which are better than other recent studies (Zhao et al., 2020) (Liu et al., 2020). In these risk signatures, a previous study showed that S100A14 overexpression was able to promote LUAD cell migration and invasion (Ding et al., 2018). In all recent studies of the AS events in NSCLC, Li et al. (2017) were the first to construct an AS risk signature for the prediction of NSCLC prognosis, while Zhao et al. (2020) constructed the AS risk signature stratified by gender of patients. Liu et al. (2020) formed an AS signature for LUSC. Our current study also explored abnormal expression of the splicing factors in NSCLC as well as the C-index (**Supplementary Table S7**). However, our current study does have some limitations, for example, the AS events database is relatively simple and lacks all other relevant datasets for us to verify our data. In addition, the relationship of m6A regulators with the AS events and the mechanism by which they play a role in NSCLC development

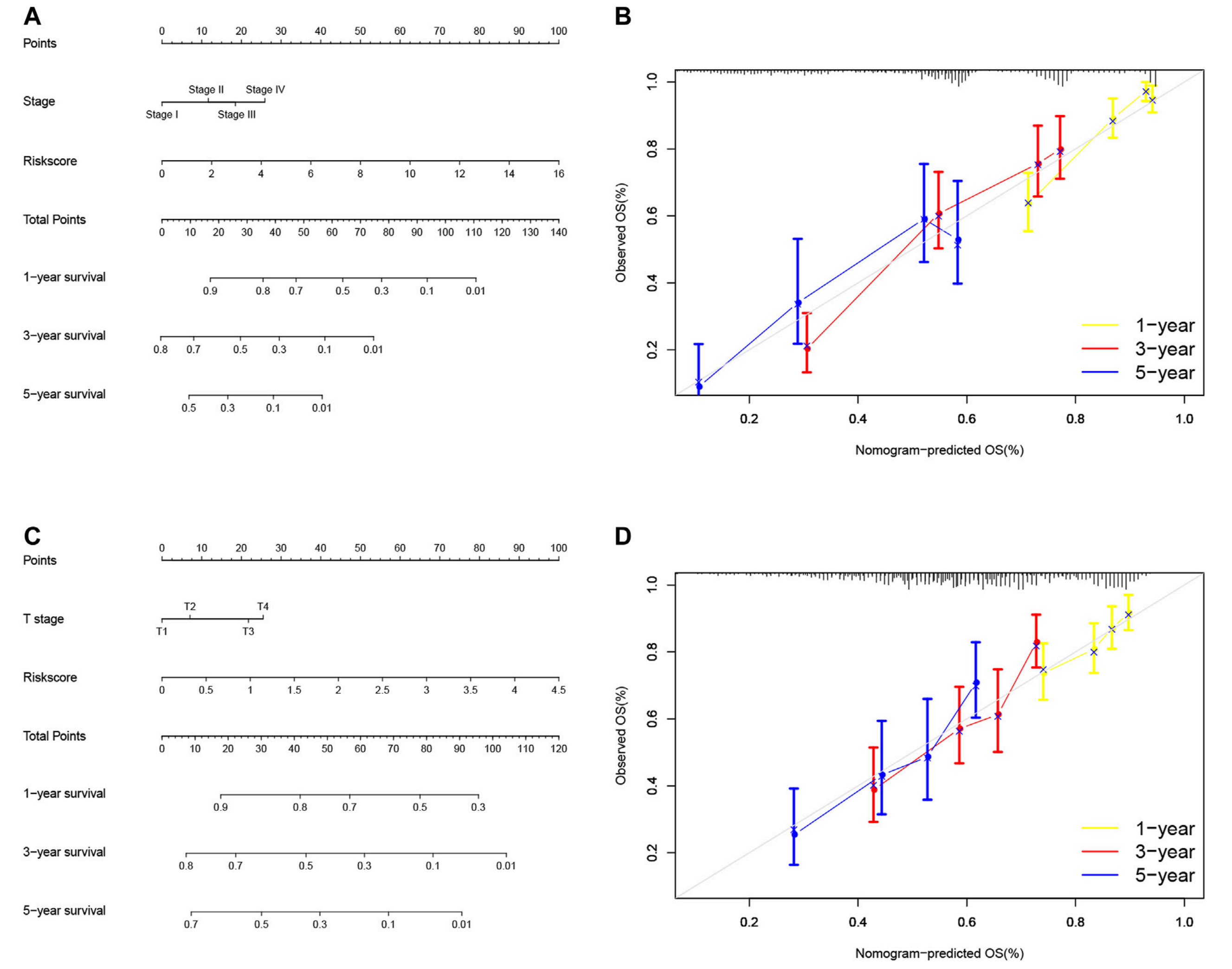


FIGURE 9 | Nomogram prediction of overall survival (OS) of LUAD and LUSC patients. **(A)** Construction of AS clinicopathological nomograms using AS risk signature and pathological stage. The nomogram was then used to predict the 1-, 3-, and 5-year OS of LUAD patients. **(B)** The nomogram AS clinicopathological nomogram calibration plot in LUAD. It predicts 1-, 3-, and 5-year prognoses. **(C)** Construction of AS clinicopathological nomograms using AS risk signature and T stage. The nomogram was then used to predict 1-, 3-, and 5- OS of LUSC patients. **(D)** The AS clinicopathological nomogram calibration plot. It predicts the 1-, 3-, and 5-year prognosis of LUSC. AS, alternative splicing; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OS, overall survival.

remain; thus, more studies are needed to clarify the true biological role of the AS events in NSCLC tumorigenesis.

CONCLUSIONS

Our current study assessed the role of m6A-related AS events in NSCLC as a signature in the prediction of NSCLC prognosis. The current study revealed the regulation of AS events by some key m6A regulators may play an important role in NSCLC development and progression. This study might provide a novel insight into the mechanism of NSCLC tumorigenesis, which may lead to novel strategies in future control of NSCLC.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the TCGA-LUSC and TCGA-LUAD database (https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

XW conceived and designed the work. ZZ and QC carried out software coding and data analysis. ZZ and PZ formatted the tables and figures. ZZ, BH, WP, and GT wrote the manuscript. FY and LW critically reviewed the codes and the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by grants from the National Natural Science Foundation of China (#81672308, #81101767, and #81972195), the Hunan Provincial Key Area R&D Program (#2019SK2253), and the Scientific Research Program of Hunan Provincial Health Commission (#20201047). Independent Exploration and Innovation Project of Graduate Student of Central South University (#2021zzts0385).

ACKNOWLEDGMENTS

The authors thank Yongguang Tao for their comments and suggestions in this study. Thank you for the revision by *Medjaden* Bioscience Limited (Hong Kong, China).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.657087/full#supplementary-material

SUPPLEMENTARY FIGURE S1 | (A) The upstate diagrams of LUAD and LUSC. **(A)** The upstate diagram of alternative splicing events in LUAD. **(B)** The upstate diagram of alternative splicing events in LUSC. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

SUPPLEMENTARY FIGURE S2 | (A) The cluster dendrogram of the LUAD patients. **(B)** The cluster dendrogram of the LUSC patients. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

SUPPLEMENTARY FIGURE S3 | Cluster analysis of samples in LUAD **(A)** and LUSC **(B)** in the detection of the outliers. The white-to-red linear gradient color indicates the association with the corresponding clinical variables, while the gray indicates missing data. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

SUPPLEMENTARY FIGURE S4 | Bioinformatical analysis of AS modules Associated with m6A RNA methylation regulatory genes. **(A)** The eigengene dendrogram and heat map. They identify groups of the correlated eigengenes termed meta-modules in LUAD. **(B)** Module–trait association in LUAD. Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the legend. **(C)** The eigengene dendrogram and heat map. They identify groups of the correlated eigengenes termed meta-modules in LUSC. **(D)** Module–trait associations in LUSC. Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the legend. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

SUPPLEMENTARY FIGURE S5 | The parameter of the adjacency function in the weighted gene correlation network analysis algorithm. **(A)** Analysis of the soft threshold power and the average connectivity of various soft threshold powers in LUAD. **(B)** Analysis of the soft threshold power and the average connectivity of the various soft threshold powers in LUSC. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

SUPPLEMENTARY FIGURE S6 | Scatterplot of the gene or clinical traits significance (the y-axis) vs. module membership (the x-axis) in the most significant module of LUAD **(A)** and LUSC **(B)**. In modules related to a trait of interest, genes with high module membership often also have had high gene significance. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma. ADDIN

REFERENCES

Alarcón, C. R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S., and Tavazoie, S. F. (2015). HNRNPA2B1 Is a Mediator of m6A-Dependent Nuclear RNA Processing Events. *Cell* 162 (6), 1299–1308. doi:10.1016/j.cell.2015.08.011

Amado, F. M., Barros, A., Azevedo, A. L., Vitorino, R., and Ferreira, R. (2014). An Integrated Perspective and Functional Impact of the Mitochondrial Acetylome. *Expert Rev. Proteomics* 11 (3), 383–394. doi:10.1586/14789450.2014.899470

Balata, H., Fong, K. M., Hendriks, L. E., Lam, S., Ostroff, J. S., Peled, N., et al. (2019). Prevention and Early Detection for NSCLC: Advances in Thoracic Oncology 2018. *J. Thorac. Oncol.* 14 (9), 1513–1527. doi:10.1016/j.jtho.2019.06.011

Bokar, J. A., Shambaugh, M. E., Polayes, D., Matera, A. G., and Rottman, F. M. (1997). Purification and cDNA Cloning of the AdoMet-Binding Subunit of the Human mRNA (N6-Adenosine)-Methyltransferase. *Rna* 3 (11), 1233–1247.

Bonnal, S. C., López-Oreja, I., and Valcárcel, J. (2020). Roles and Mechanisms of Alternative Splicing in Cancer - Implications for Care. *Nat. Rev. Clin. Oncol.* 17 (8), 457–474. doi:10.1038/s41571-020-0350-x

Chen, X.-Y., Zhang, J., and Zhu, J.-S. (2019). The Role of m6A RNA Methylation in Human Cancer. *Mol. Cancer* 18 (1), 103. doi:10.1186/s12943-019-1033-z

Cherry, S., and Lynch, K. W. (2020). Alternative Splicing and Cancer: Insights, Opportunities, and Challenges from an Expanding View of the Transcriptome. *Genes Dev.* 34 (15-16), 1005–1016. doi:10.1101/gad.338962.120

Colak, S., and Ten Dijke, P. (2017). Targeting TGF-β Signaling in Cancer. *Trends Cancer* 3 (1), 56–71. doi:10.1016/j.trecan.2016.11.008

Coomer, A. O., Black, F., Greystoke, A., Munkley, J., and Elliott, D. J. (2019). Alternative Splicing in Lung Cancer. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1862 (11-12), 194388. doi:10.1016/j.bbagrm.2019.05.006

Dai, D., Wang, H., Zhu, L., Jin, H., and Wang, X. (2018). N6-methyladenosine Links RNA Metabolism to Cancer Progression. *Cell Death Dis.* 9 (2), 124. doi:10.1038/s41419-017-0129-x

David, C. J., and Manley, J. L. (2008). The Search for Alternative Splicing Regulators: New Approaches Offer a Path to a Splicing Code. *Genes Dev.* 22 (3), 279–285. doi:10.1101/gad.1643108

de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G. M. (2020). Global burden of Cancer Attributable to Infections in 2018: a Worldwide Incidence Analysis. *Lancet Glob. Health* 8 (2), e180–e190. doi:10.1016/s2214-109x(19)30488-7

de Miguel, F. J., Pajares, M. J., Martínez-Terroba, E., Ajona, D., Morales, X., Sharma, R. D., et al. (2016). A Large-Scale Analysis of Alternative Splicing Reveals a Key Role of QKI in Lung Cancer. *Mol. Oncol.* 10 (9), 1437–1449. doi:10.1016/j.molonc.2016.08.001

Deng, X., Su, R., Feng, X., Wei, M., and Chen, J. (2018). Role of N6-Methyladenosine Modification in Cancer. *Curr. Opin. Genet. Dev.* 48, 1–7. doi:10.1016/j.gde.2017.10.005

Ding, F., Wang, D., Li, X.-K., Yang, L., Liu, H.-Y., Cui, W., et al. (2018). Overexpression of S100A14 Contributes to Malignant Progression and Predicts Poor Prognosis of Lung Adenocarcinoma. *Thorac. Cancer* 9 (7), 827–835. doi:10.1111/1759-7714.12654

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the Human and Mouse m6A RNA Methylomes Revealed by m6A-Seq. *Nature* 485 (7397), 201–206. doi:10.1038/nature11112

Frankiw, L., Baltimore, D., and Li, G. (2019). Alternative mRNA Splicing in Cancer Immunotherapy. *Nat. Rev. Immunol.* 19 (11), 675–687. doi:10.1038/s41577-019-0195-7

Geng, Y., Guan, R., Hong, W., Huang, B., Liu, P., Guo, X., et al. (2020). Identification of m6A-Related Genes and m6A RNA Methylation Regulators in Pancreatic Cancer and Their Association with Survival. *Ann. Transl Med.* 8 (6), 387. doi:10.21037/atm.2020.03.98

Graedel, T. E. (2019). Material Flow Analysis from Origin to Evolution. *Environ. Sci. Technol.* 53 (21), 12188–12196. doi:10.1021/acs.est.9b03413

Gruber, A. J., Schmidt, R., Gruber, A. R., Martin, G., Ghosh, S., Belmadani, M., et al. (2016). A Comprehensive Analysis of 3′ End Sequencing Data Sets Reveals Novel Polyadenylation Signals and the Repressive Role of Heterogeneous Ribonucleoprotein C on Cleavage and Polyadenylation. *Genome Res.* 26 (8), 1145–1159. doi:10.1101/gr.202432.115

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hintermann, E., and Christen, U. (2019). The Many Roles of Cell Adhesion Molecules in Hepatic Fibrosis. *Cells* 8 (12), 1503. doi:10.3390/cells8121503

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Jr., Wu, Y.-L., et al. (2017). Lung Cancer: Current Therapies and New Targeted Treatments. *The Lancet* 389 (10066), 299–311. doi:10.1016/s0140-6736(16)30958-8

Ji, P., Wang, X., Xie, N., and Li, Y. (2018). N6-Methyladenosine in RNA and DNA: An Epitranscriptomic and Epigenetic Player Implicated in Determination of Stem Cell Fate. *Stem Cell Int.* 2018, 1–18. doi:10.1155/2018/3256524

Kallifatidis, G., Mamouni, K., and Lokeshwar, B. L. (2020). The Role of β-Arrestins in Regulating Stem Cell Phenotypes in Normal and Tumorigenic Cells. *Ijms* 21 (23), 9310. doi:10.3390/ijms21239310

Kang, Y., Zhou, X. E., Gao, X., He, Y., Liu, W., Ishchenko, A., et al. (2015). Crystal Structure of Rhodopsin Bound to Arrestin by Femtosecond X-ray Laser. *Nature* 523 (7562), 561–567. doi:10.1038/nature14656

Kasowitz, S. D., Ma, J., Anderson, S. J., Leu, N. A., Xu, Y., Gregory, B. D., et al. (2018). Nuclear m6A Reader YTHDC1 Regulates Alternative Polyadenylation and Splicing during Mouse Oocyte Development. *Plos Genet.* 14 (5), e1007412. doi:10.1371/journal.pgen.1007412

Katz, Y., Wang, E. T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., et al. (2015). Quantitative Visualization of Alternative Exon Expression from RNA-Seq Data. *Bioinformatics* 31 (14), 2400–2402. doi:10.1093/bioinformatics/btv034

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Lee, E. K., Kim, H. H., Kuwano, Y., Abdelmohsen, K., Srikantan, S., Subaran, S. S., et al. (2010). hnRNP C Promotes APP Translation by Competing with FMRP for APP mRNA Recruitment to P Bodies. *Nat. Struct. Mol. Biol.* 17 (6), 732–739. doi:10.1038/nsmb.1815

Li, F., Wang, H., Huang, H., Zhang, L., Wang, D., and Wan, Y. (2020). m6A RNA Methylation Regulators Participate in the Malignant Progression and Have Clinical Prognostic Value in Lung Adenocarcinoma. *Front. Genet.* 11, 994. doi:10.3389/fgene.2020.00994

Li, S., Hu, Z., Zhao, Y., Huang, S., and He, X. (2019). Transcriptome-Wide Analysis Reveals the Landscape of Aberrant Alternative Splicing Events in Liver Cancer. *Hepatology* 69 (1), 359–375. doi:10.1002/hep.30158

Li, Y., Sun, N., Lu, Z., Sun, S., Huang, J., Chen, Z., et al. (2017). Prognostic Alternative mRNA Splicing Signature in Non-small Cell Lung Cancer. *Cancer Lett.* 393, 40–51. doi:10.1016/j.canlet.2017.02.016

Lin, K.-T., and Krainer, A. R. (2019). PSI-sigma: a Comprehensive Splicing-Detection Method for Short-Read and Long-Read RNA-Seq Analysis. *Bioinformatics* 35 (23), 5048–5054. doi:10.1093/bioinformatics/btz438

Lin, S., Choe, J., Du, P., Triboulet, R., and Gregory, R. I. (2016). The M 6 A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Mol. Cel.* 62 (3), 335–345. doi:10.1016/j.molcel.2016.03.021

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3-METTL14 Complex Mediates Mammalian Nuclear RNA N6-Adenosine Methylation. *Nat. Chem. Biol.* 10 (2), 93–95. doi:10.1038/nchembio.1432

Liu, Y., Jia, W., Li, J., Zhu, H., and Yu, J. (2020). Identification of Survival-Associated Alternative Splicing Signatures in Lung Squamous Cell Carcinoma. *Front. Oncol.* 10, 587343. doi:10.3389/fonc.2020.587343

Maconachie, R., Mercer, T., Navani, N., and McVeigh, G. (2019). Lung Cancer: Diagnosis and Management: Summary of Updated NICE Guidance. *Bmj* 364, l1049. doi:10.1136/bmj.l1049

Meng, L.-B., Shan, M.-J., Qiu, Y., Qi, R., Yu, Z.-M., Guo, P., et al. (2019). TPM2 as a Potential Predictive Biomarker for Atherosclerosis. *Aging* 11 (17), 6960–6982. doi:10.18632/aging.102231

Motegi, A., Masutani, M., Yoshioka, K.-i., and Bessho, T. (2019). Aberrations in DNA Repair Pathways in Cancer and Therapeutic Significances. *Semin. Cancer Biol.* 58, 29–46. doi:10.1016/j.semcancer.2019.02.005

Niemira, M., Collin, F., Szalkowska, A., Bielska, A., Chwialkowska, K., Reszec, J., et al. (2019). Molecular Signature of Subtypes of Non-small-cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-expression Network Analysis (WGCNA). *Cancers* 12 (1), 37. doi:10.3390/cancers12010037

Papasaikas, P., and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends Biochem. Sci.* 41 (1), 33–45. doi:10.1016/j.tibs.2015.11.003

Park, S. H., Goo, J. M., and Jo, C.-H. (2004). Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J. Radiol.* 5 (1), 11–18. doi:10.3348/kjr.2004.5.1.11

Park, Y. M., Hwang, S. J., Masuda, K., Choi, K.-M., Jeong, M.-R., Nam, D.-H., et al. (2012). Heterogeneous Nuclear Ribonucleoprotein C1/C2 Controls the Metastatic Potential of Glioblastoma by Regulating PDCD4. *Mol. Cell Biol.* 32 (20), 4237–4244. doi:10.1128/mcb.00443-12

Paschalis, A., Sharp, A., Welti, J. C., Neeb, A., Raj, G. V., Luo, J., et al. (2018). Alternative Splicing in Prostate Cancer. *Nat. Rev. Clin. Oncol.* 15 (11), 663–675. doi:10.1038/s41571-018-0085-0

Pendleton, K. E., Chen, B., Liu, K., Hunter, O. V., Xie, Y., Tu, B. P., et al. (2017). The U6 snRNA M 6 A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* 169 (5), 824–835. e814. doi:10.1016/j.cell.2017.05.003

Pilié, P. G., Tang, C., Mills, G. B., and Yap, T. A. (2019). State-of-the-art Strategies for Targeting the DNA Damage Response in Cancer. *Nat. Rev. Clin. Oncol.* 16 (2), 81–104. doi:10.1038/s41571-018-0114-z

Pio, R., and Montuenga, L. M. (2009). Alternative Splicing in Lung Cancer. *J. Thorac. Oncol.* 4 (6), 674–678. doi:10.1097/JTO.0b013e3181a520dc

Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E. F., Faivre-Finn, C., et al. (2018). Metastatic Non-small Cell Lung Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 29 (Suppl. 4), iv192–iv237. doi:10.1093/annonc/mdy275

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Rizvi, A. A., Karaesmen, E., Morgan, M., Preus, L., Wang, J., Sovic, M., et al. (2019). Gwasurvivr: an R Package for Genome-wide Survival Analysis. *Bioinformatics* 35 (11), 1968–1970. doi:10.1093/bioinformatics/bty920

Sciarrillo, R., Wojtuszkiewicz, A., Assaraf, Y. G., Jansen, G., Kaspers, G. J. L., Giovannetti, E., et al. (2020). The Role of Alternative Splicing in Cancer: From Oncogenesis to Drug Resistance. *Drug Resist. Updates* 53, 100728. doi:10.1016/j.drup.2020.100728

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Sholl, L. (2017). Molecular Diagnostics of Lung Cancer in the Clinic. *Transl. Lung Cancer Res.* 6 (5), 560–569. doi:10.21037/tlcr.2017.08.03

Soh, J. Y., Jung, S.-H., Cha, W. C., Kang, M., Chang, D. K., Jung, J., et al. (2019). Variability in Doctors' Usage Paths of Mobile Electronic Health Records Across Specialties: Comprehensive Analysis of Log Data. *JMIR Mhealth Uhealth* 7 (1), e12041. doi:10.2196/12041

Song, Y., Zhu, Z., An, Y., Zhang, W., Zhang, H., Liu, D., et al. (2013). Selection of DNA Aptamers against Epithelial Cell Adhesion Molecule for Cancer Cell Imaging and Circulating Tumor Cell Capture. *Anal. Chem.* 85 (8), 4141–4149. doi:10.1021/ac400366b

Sun, L., Liu, W.-K., Du, X.-W., Liu, X.-L., Li, G., Yao, Y., et al. (2020). Large-scale Transcriptome Analysis Identified RNA Methylation Regulators as Novel Prognostic Signatures for Lung Adenocarcinoma. *Ann. Transl Med.* 8 (12), 751. doi:10.21037/atm-20-3744

Sun, T., Wu, R., and Ming, L. (2019). The Role of m6A RNA Methylation in Cancer. *Biomed. Pharmacother.* 112, 108613. doi:10.1016/j.biopha.2019.108613

Sun, Y., and Ma, L. (2019). New Insights into Long Non-coding RNA MALAT1 in Cancer and Metastasis. *Cancers* 11 (2), 216. doi:10.3390/cancers11020216

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Syed, V. (2016). TGF-β Signaling in Cancer. *J. Cel. Biochem.* 117 (6), 1279–1287. doi:10.1002/jcb.25496

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017). The Spike-And-Slab Lasso Cox Model for Survival Prediction and Associated Genes Detection. *Bioinformatics* 33 (18), 2799–2807. doi:10.1093/bioinformatics/btx300

Tanoue, L. T., Tanner, N. T., Gould, M. K., and Silvestri, G. A. (2015). Lung Cancer Screening. *Am. J. Respir. Crit. Care Med.* 191 (1), 19–33. doi:10.1164/rccm.201410-1777CI

Urbanski, L. M., Leclair, N., and Anczuków, O. (2018). Alternative-splicing Defects in Cancer: Splicing Regulators and Their Downstream Targets, Guiding the

Way to Novel Cancer Therapeutics. *WIREs RNA* 9 (4), e1476. doi:10.1002/wrna.1476

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., et al. (2020). ROC Curves for Clinical Prediction Models Part 1. ROC Plots Showed No Added Value Above the AUC When Evaluating the Performance of Clinical Prediction Models. *J. Clin. Epidemiol.* 126, 207–216. doi:10.1016/j.jclinepi.2020.01.028

Wan, Q., Tang, J., Han, Y., and Wang, D. (2018). Co-expression Modules Construction by WGCNA and Identify Potential Prognostic Markers of Uveal Melanoma. *Exp. Eye Res.* 166, 13–20. doi:10.1016/j.exer.2017.10.007

Wang, P., Doxtader, K. A., and Nam, Y. (2016). Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases. *Mol. Cel.* 63 (2), 306–317. doi:10.1016/j.molcel.2016.05.041

Wang, T., Kong, S., Tao, M., and Ju, S. (2020). The Potential Role of RNA N6-Methyladenosine in Cancer Progression. *Mol. Cancer* 19 (1), 88. doi:10.1186/s12943-020-01204-7

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., and Zhao, J. C. (2014). N6-methyladenosine Modification Destabilizes Developmental Regulators in Embryonic Stem Cells. *Nat. Cel Biol.* 16 (2), 191–198. doi:10.1038/ncb2902

Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., et al. (2016). Nuclear M 6 A Reader YTHDC1 Regulates mRNA Splicing. *Mol. Cel.* 61 (4), 507–519. doi:10.1016/j.molcel.2016.01.012

Xie, H., and Xie, C. (2019). A Six-Gene Signature Predicts Survival of Adenocarcinoma Type of Non-small-cell Lung Cancer Patients: A Comprehensive Study Based on Integrated Analysis and Weighted Gene Coexpression Network. *Biomed. Res. Int.* 2019, 1–16. doi:10.1155/2019/4250613

Yan, M., Sun, L., Li, J., Yu, H., Lin, H., Yu, T., et al. (2019). RNA-binding Protein KHSRP Promotes Tumor Growth and Metastasis in Non-small Cell Lung Cancer. *J. Exp. Clin. Cancer Res.* 38 (1), 478. doi:10.1186/s13046-019-1479-2

Yang, Q., Zhao, J., Zhang, W., Chen, D., and Wang, Y. (2019). Aberrant Alternative Splicing in Breast Cancer. *J. Mol. Cel Biol.* 11 (10), 920–929. doi:10.1093/jmcb/mjz033

Yoshimi, A., Lin, K. T., Wiseman, D. H., Rahman, M. A., Pastore, A., Wang, B., et al. (2019). Coordinated Alterations in RNA Splicing and Epigenetic Regulation Drive Leukaemogenesis. *Nature* 574 (7777), 273–277. doi:10.1038/s41586-019-1618-0

Zhang, L., Tran, N.-T., Su, H., Wang, R., Lu, Y., Tang, H., et al. (2015). Cross-talk between PRMT1-Mediated Methylation and Ubiquitylation on RBM15 Controls RNA Splicing. *Elife* 4, e07938. doi:10.7554/eLife.07938

Zhang, S., Hu, Z., Lan, Y., Long, J., Wang, Y., Chen, X., et al. (2020a). Prognostic Significance of Survival-Associated Alternative Splicing Events in Gastric Cancer. *Aging* 12 (21), 21923–21941. doi:10.18632/aging.104013

Zhang, S., Tong, Y. X., Zhang, X. H., Zhang, Y. J., Xu, X. S., Xiao, A. T., et al. (2019). A Novel and Validated Nomogram to Predict Overall Survival for Gastric Neuroendocrine Neoplasms. *J. Cancer* 10 (24), 5944–5954. doi:10.7150/jca.35785

Zhang, Y., Geng, X., Li, Q., Xu, J., Tan, Y., Xiao, M., et al. (2020c). m6A Modification in RNA: Biogenesis, Functions and Roles in Gliomas. *J. Exp. Clin. Cancer Res.* 39 (1), 192. doi:10.1186/s13046-020-01706-8

Zhang, Y., Liu, X., Liu, L., Li, J., Hu, Q., and Sun, R. (2020b). Expression and Prognostic Significance of m6A-Related Genes in Lung Adenocarcinoma. *Med. Sci. Monit.* 26, e919644. doi:10.12659/msm.919644

Zhang, Y., Wang, W., Xu, X., Li, Y., Zhang, H., Li, J., et al. (2021). Impact of Radiotherapy Pattern on the Prognosis of Stage IV Lung Adenocarcinomas Harboring EGFR Mutations. *Cmar* 13, 3293–3301. doi:10.2147/cmar.s299563

Zhao, D., Zhang, C., Jiang, M., Wang, Y., Liang, Y., Wang, L., et al. (2020). Survival-associated Alternative Splicing Signatures in Non-small Cell Lung Cancer. *Aging* 12 (7), 5878–5893. doi:10.18632/aging.102983

Zi, Z. (2019). Molecular Engineering of the TGF-β Signaling Pathway. *J. Mol. Biol.* 431 (15), 2644–2654. doi:10.1016/j.jmb.2019.05.022

Zong, F.-Y., Fu, X., Wei, W.-J., Luo, Y.-G., Heiner, M., Cao, L.-J., et al. (2014). The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *Plos Genet.* 10 (4), e1004289. doi:10.1371/journal.pgen.1004289

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fmolb.2021.658906

Predicting the Specificity-Determining Positions of Receptor Tyrosine Kinase Axl

*Tülay Karakulak[1,2,3,4,5], Ahmet Sureyya Rifaioglu[6], João P. G. L. M. Rodrigues[7] and Ezgi Karaca[1,2]**

[1] Izmir Biomedicine and Genome Center, Izmir, Turkey, [2] Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Izmir, Turkey, [3] Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland, [4] Department of Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland, [5] Swiss Institute of Bioinformatics, Lausanne, Switzerland, [6] Department of Electrical – Electronics Engineering, İskenderun Technical University, Hatay, Turkey, [7] Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, United States

Owing to its clinical significance, modulation of functionally relevant amino acids in protein-protein complexes has attracted a great deal of attention. To this end, many approaches have been proposed to predict the partner-selecting amino acid positions in evolutionarily close complexes. These approaches can be grouped into sequence-based machine learning and structure-based energy-driven methods. In this work, we assessed these methods' ability to map the specificity-determining positions of Axl, a receptor tyrosine kinase involved in cancer progression and immune system diseases. For sequence-based predictions, we used SDPpred, Multi-RELIEF, and Sequence Harmony. For structure-based predictions, we utilized HADDOCK refinement and molecular dynamics simulations. As a result, we observed that (i) sequence-based methods overpredict partner-selecting residues of Axl and that (ii) combining Multi-RELIEF with HADDOCK-based predictions provides the key Axl residues, covered by the extensive molecular dynamics simulations. Expanding on these results, we propose that a sequence-structure-based approach is necessary to determine specificity-determining positions of Axl, which can guide the development of therapeutic molecules to combat Axl misregulation.

Keywords: protein selectivity, sequence analysis, molecular dynamics, Axl, HADDOCK

OPEN ACCESS

Edited by:
Tunca Dogan,
Hacettepe University, Turkey

Reviewed by:
Elodie Laine,
Université Pierre et Marie Curie,
France
Marc F. Lensink,
Centre National de la Recherche
Scientifique (CNRS), France

****Correspondence:***
Ezgi Karaca
ezgi.karaca@ibg.edu.tr

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *26 January 2021*
Accepted: *20 April 2021*
Published: *14 June 2021*

Citation:
Karakulak T, Rifaioglu AS,
Rodrigues JPGLM and Karaca E
(2021) Predicting the Specificity-
Determining Positions of Receptor
Tyrosine Kinase Axl.
Front. Mol. Biosci. 8:658906.
doi: 10.3389/fmolb.2021.658906

INTRODUCTION

The functional identification of proteins is essential to understand the grounds of innate cellular processes. Several computational tools have been deployed to annotate protein function from ever-accumulating protein sequences (Friedberg, 2006). These approaches aim to define functionally important residues through comparative sequence analysis (Whisstock and Lesk, 2004). Resolving the functionally key amino acids is particularly interesting, as modulation of these residues holds a great potential to design protein-based therapeutics (Moll et al., 2016). Such key amino acids can be identified upon searching for conserved positions across different species. Alternatively, within a species, one could look for the differentially mutated amino acid positions of closely-related protein families, i.e., paralogs (Gogarten and Olendzenski, 1999; Mirny and Gelfand, 2002; Chagoyen et al., 2016). In paralogs, some mutations are evolved to act as specificity-determining positions

(SDPs) for regulating selective protein interactions (Rausell et al., 2010; Sloutsky and Naegle, 2016). Thus, SDPs are often ascribed to the specialized functions of proteins (Capra and Singh, 2008; Chakraborty and Chakrabarti, 2015; Wong et al., 2015). SDPs can either select a binding partner (partner-selecting) or tune the affinity of a protein toward different ligands (affinity-tuning) (Chagoyen et al., 2016; Sloutsky and Naegle, 2016; Pitarch et al., 2020).

During the last three decades, several sequence-based SDP predictors have been proposed (Pirovano et al., 2006; Chakrabarti and Panchenko, 2008; Chakraborty and Chakrabarti, 2015; Chagoyen et al., 2016). These methods rely on the application of different machine learning techniques, which can be grouped into entropy-, evolution-, and feature-based (Teppa et al., 2012). The majority of these methods expand on the use of a precalculated multiple sequence alignment (MSA) file. The entropy-based methods compute the variability of specific amino acid positions in an alignment of related protein sequences, allowing the identification of highly varying positions (Kalinina et al., 2004; Ye et al., 2006; Feenstra et al., 2007). As an example, SDPpred uses mutual information entropy scores to predict SDPs (Kalinina et al., 2004). The evolutionary-based methods, on the other hand, use substitution matrices or phylogenetic trees to calculate residue-based variability scores (del Sol Mesa et al., 2003; Pazos et al., 2006; Capra and Singh, 2008). The evolutionary-based method Xdet, for example, combines the substitution matrix with GO or EC annotations, together with the available interactome data (Pazos et al., 2006). Different than the other sequence-based methods, Xdet can provide partner-specific SDPs, though, it only works on large protein families (Pitarch et al., 2020). Finally, the feature-based methods perform feature extraction of each amino acid position. The extracted feature vectors are fed into a classifier, such as random forest, support vector machine or neural network (Ahmad and Sarai, 2005; Wong et al., 2015). For instance, Ahmad and Sarai proposed a position-specific scoring matrix-based SDP prediction of DNA binding proteins (Ahmad and Sarai, 2005). Here, each residue is represented as a feature vector by using its and its neighbors' conservation scores. Then, the feature vectors are processed by a neural network classifier to categorize the input residues as SDP or non-SDP for DNA binding. As the sequence-based SDP prediction methods do not use heavy input data, they are computationally efficient. However, the application of these methods is rather limited as they are mostly trained with small sequence datasets with classical machine learning algorithms.

The available structure-based SDP prediction methods make use of the core-support-rim model, as proposed by Levy. According to this model, the protein-protein interaction surface can be dissected into three, as: (i) the core; the amino acids, which get buried upon complexation, (ii) the support; the residues, which are buried in the uncomplexed state and become more buried upon complexation, (iii) the rim; the amino acids, which stay solvent accessible both in free and complexed states (Levy, 2010). In a recent work of Ivanov et al., this definition was used to discriminate SDPs of four paralog protein families (Ivanov et al., 2017). Here, the authors structurally modeled and analyzed all paralog interactions, for which the experimental affinities were at hand. Their analysis showed that SDPs are located at the rim, where they form strong electrostatic (charge-charge) interactions (Chakrabarti and Janin, 2002; Ivanov et al., 2017). Other groups utilized atomistic molecular dynamics simulations to trace partner-selecting paralog interactions. For example, van Wijk et al. demonstrated that a single salt bridge is the key determinant for selective ubiquitin-conjugating enzyme (E2) and ubiquitin ligase (E3) interactions (van Wijk et al., 2012). Being at the rim of E2-E3 surface, the partner-selecting role of this salt bridge was validated by mutagenesis and yeast two-hybrid screening. Another recent example explored how protocadherins specifically find their partners to polymerize, which is an essential mechanism for neuronal development. For this, Nicoludis et al. combined molecular dynamics simulations with evolutionary coupling information (Nicoludis et al., 2019). Compared to the sequence-based SDP prediction methods, the structure-based approaches provide a refined and thus an experimentally testable SDP set. However, these approaches generally require expertise in computational structural biology tools and depending on the size of the system, they could be computationally intensive.

As the sequence- and structure-based methods have different advantages, we chose a model system to map the prediction landscape of these approaches. For this, we concentrated on a paralogous protein receptor tyrosine kinase family (TAM), made by Tyro3, Axl, and Mer proteins. TAM receptors, like the other receptor tyrosine kinases, are activated through their interactions with extracellular proteins, triggering receptor dimerization and autophosphorylation of their kinase domains (Rothlin and Lemke, 2010). Earlier studies identified two related proteins, the growth arrest-specific protein 6 (Gas6) and vitamin K-dependent protein S (Pros1) as TAM ligands (Hafizi and Dahlbäck, 2006). The binding of these ligands to TAM leads to downstream activation of diverse signaling pathways (Wium and Paccez, 2018). Besides Gas6/Pros1, three other ligands (tubby, tubby-like protein and galactin-3) were shown to bind to TAM proteins (Myers et al., 2019). These structures are neither sequence- nor structure-wise related to Gas6 and Pros1. This suggests that they bind to TAM family by using a different mechanism compared to Gas6 and Pros1. As there is little information on the binding profiles of these new ligands, in this work, we focused only on TAM:Gas6/Pros1 interactions.

TAM receptors share 52–57%, while Gas6/Pros1 share 40% pairwise sequence similarity. Across TAM members, Pros1 binds to Tyro3 and Mer, while it cannot bind to Axl. Gas6 binds to all three receptors with the highest affinity toward Axl (Hafizi and Dahlbäck, 2006; Yanagihashi et al., 2017). Among the different combinations, the Axl:Gas6 interaction is particularly interesting given its involvement in numerous types of signaling pathways (e.g., tumor-cell growth, metastasis, epithelial to mesenchymal transition, drug resistance, etc.) (Zhu et al., 2019). Relatedly, Axl aberrant regulation was shown to lead to different types of cancer and infectious diseases (Van Der Meer et al., 2014), as well as to promote SARS-CoV-2 entry into cell (Wu et al., 2017; Wium and Paccez, 2018; Wang et al., 2021). Although the structure Axl:Gas6 complex is resolved, Axl's ligand-selecting residues is still unknown. To help to close this knowledge gap, we used three

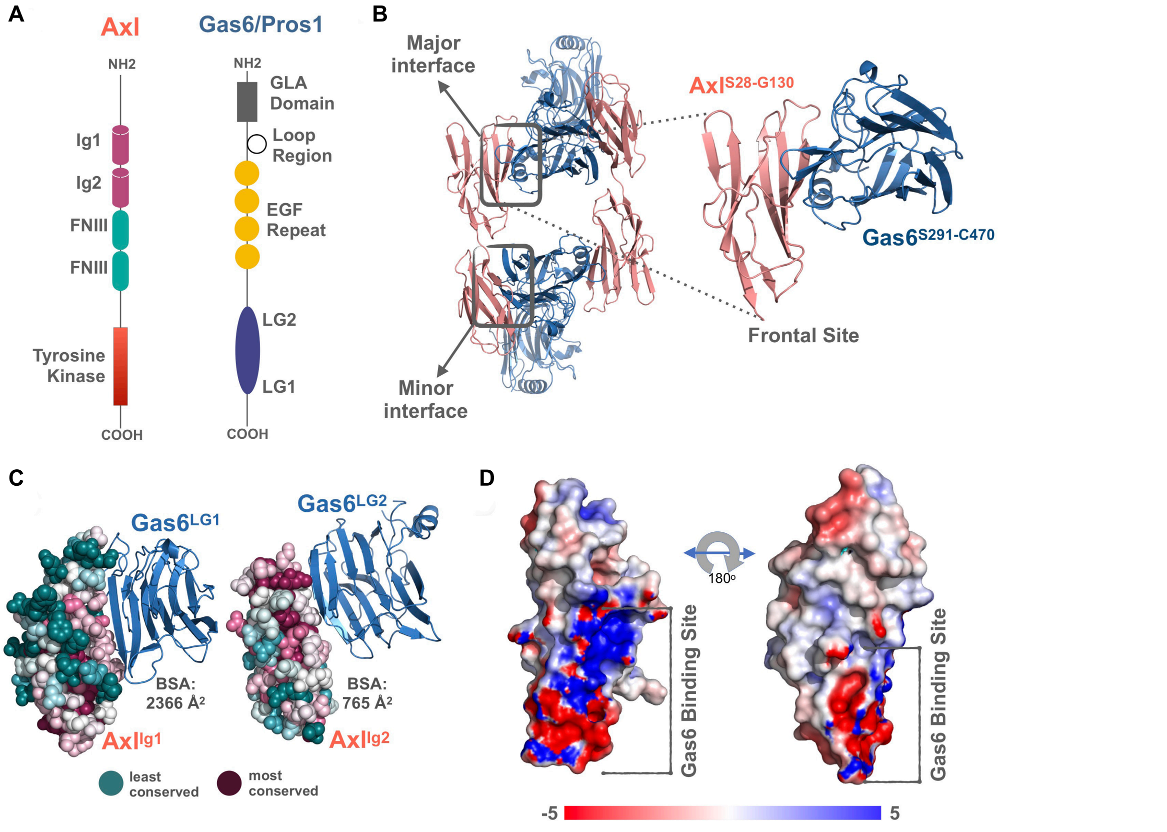


FIGURE 1 | (A) The domain organization of TAM family and its ligands, Gas6 and Pros1. TAM family consists of Ig1, Ig2, two FNIII, and tyrosine kinase domains (Linger et al., 2008; Lemke, 2013). Gas6 and Pros1 are composed of GLA domain, loop region, EGF Repeat, and LG2, LG1 domains (Linger et al., 2008; Lemke, 2013). **(B)** Axl(Ig1-Ig2):Gas6(LG1-LG2) interaction involves two interfaces: The major interface is formed between Axl-Ig1:Gas6-LG1 and the minor one is established among Axl-Ig2:Gas6-LG1 [PDB ID: 2C5D, Sasaki et al. (2006)]. The inset represents the charged frontal side of the major interface (Axl is depicted in pink cartoon, whereas Gas6 is represented in purple cartoon). **(C)** Conservation scores of Axl residues predicted via ConSurf webserver (Glaser et al., 2003; Landau et al., 2005; Ashkenazy et al., 2016). The most conserved sites are colored with deep purple and the least conserved ones with deep teal. **(D)** Electrostatic potential of Axl:Gas6 interacting site. The color scale ranges from -5 (red) to 5 (blue). One side of Axl's Gas6 binding surface is heavily charged, while the other side is composed of neutral amino acids.

sequence-based SDP predictors, SDPpred, Multi-RELIEF (both feature-based), and Sequence Harmony (entropy-based) to map Axl partner-selecting SDPs. In addition, we analyzed the selective Axl:ligand interactions, by using simple refinement and extensive molecular dynamics simulations.

RESULTS

Axl:Gas6 Interface

TAM receptors share two immunoglobulin (Ig)-like, two fibronectin type III domains (FNIII), followed by a single-pass transmembrane helix, and an intracellular kinase domain (**Figure 1A**). TAM ligands, Gas6 and Pros1 contain an N-terminal gamma-carboxyglutamic acid (GLA) domain, four epidermal growth factor-like (EGF) repeats, and two laminin G (LG)-like domains (**Figure 1A**). The crystal structure of Axl:Gas6 interaction is the only available TAM:ligand structure [PDB ID: 2C5D (Sasaki et al., 2006)]. In the Axl:Gas6 structure, two Ig-like domains of Axl interact with two LG-like domains of Gas6, without involving any receptor-receptor or ligand-ligand interactions (**Figure 1B**). Axl and Gas6 interact through two symmetric copies of major and minor interfaces, burying 2366 Å^2 and 765 Å^2 surface areas, respectively (**Figures 1B,C**). While the minor interface is highly conserved across TAM, the major interface is not. The major interface is spatially segregated into a frontal site, involving a series of charged residues, and a hydrophobic distal site (**Figure 1D**; Sasaki et al., 2006). The segregated characteristics of the major interface contribute to its ligand selection, as well as to Axl's high affinity toward Gas6 (Sasaki et al., 2006). Thus, for studying the ligand selectivity of Axl, we focused on the major Axl:Gas6 interface (**Figure 1B-inset**).

Sequence-Based Axl SDP Predictions Agree in One Residue

Among the available sequence-based SDP predictors, we selected three methods to probe Axl ligand selectivity (**Supplementary Table 1**). These algorithms, i.e., SDPpred, Sequence Harmony,

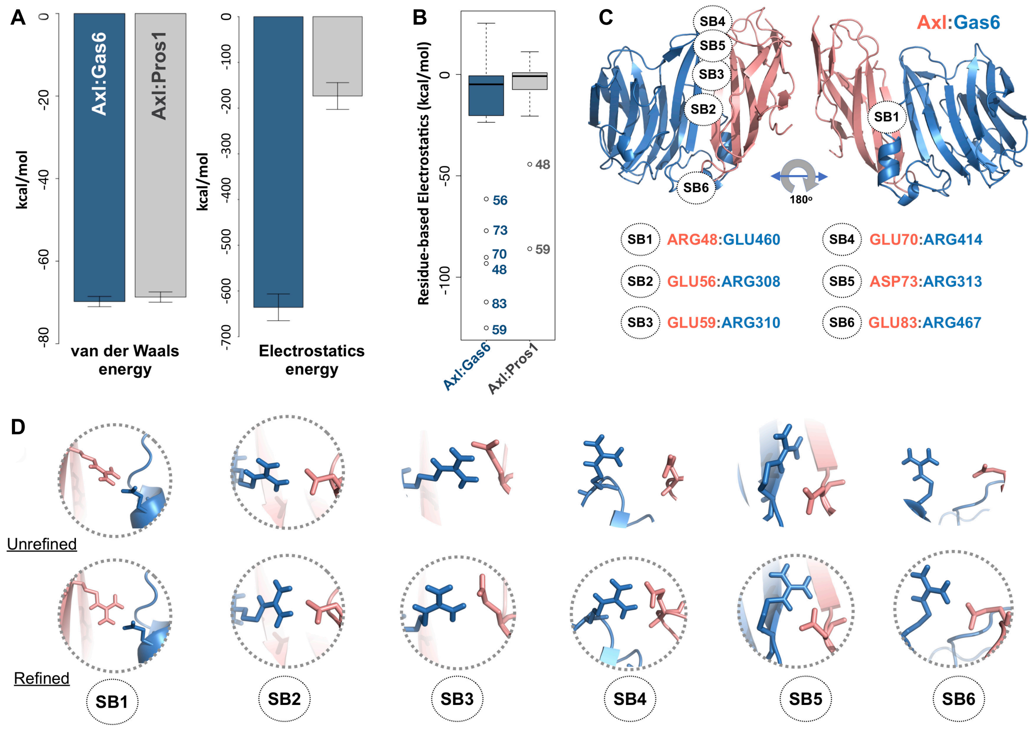


FIGURE 2 | (A) The mean van der Waals and electrostatics energetics of the top scoring Axl:Gas6 and Axl:Pros1 major interfaces calculated by HADDOCK. The mean van der Waals energies of Axl:Gas6 (blue) and Axl:Pros1 (gray) are –69.8 ± 1.2 kcal/mol and –68.9 ± 2.2 kcal/mol, respectively. The mean electrostatics of Axl:Gas6 and Axl:Pros1 are –635.8 ± 29.3 kcal/mol and –173.5 ± 21.4 kcal/mol, respectively. **(B)** Residue-based electrostatics contribution of Gas6-facing Axl residues. Axl residues behaving differently than the rest of the population are R48, E59, E70, D73, E83 in the case of Axl:Gas6 (blue), and E48, E59 in the case of Axl:Pros1 (gray). The whiskers were computed with a whisker length of 2xIQR. **(C)** The distribution of the salt bridges (SBs) across Axl:Gas6 interface. Axl is represented in light pink and Gas6 in marine blue. Only SB1 is located on the back and rather neutral side of the complex. This and all the structural images were generated with PyMOL molecular visualization software (Schrödinger, 2015). **(D)** The arrangement of the potential SB-making residues in the case of crystal (first row, pdb id: 2C5D) and HADDOCK-refined (second row) Axl:Gas6 complex. The pairs forming SBs are encircled.

and Multi-RELIEF, were selected based on their widespread use and their availability as a web service (Kalinina et al., 2004; Feenstra et al., 2007; Ye et al., 2008). Initially, to analyze the TAM sequences, the mammalian (human, mouse, rat, pig, chimpanzee) TAM Ig1 sequences were retrieved from UniProtKB (The UniProt Consortium, 2018). These sequences were grouped into Axl and Tyro3 & Mer sequence groups. The MSA of each group was constructed with Clustal Omega (Sievers and Higgins, 2017). For each approach, MSAs were formatted according to the requirements of the webservers. As earlier studies showed that partner-selecting SDPs are located at the rim of protein-protein interfaces, we filtered out the sequence-based SDP predictions by keeping the positions corresponding to the rim of the Axl:Gas6 complex (Ivanov et al., 2017). Within this framework, SDPpred predicted 19 SDPs, five of which (T46, R48, Q50, D84, K96) were at the rim of Axl:Gas6. The majority of the SDPpred predictions corresponded to the non-interacting regions of the Axl:Gas6 complex (as calculated by the EPPIC web server, Duarte et al., 2012). The same trend was observed for Multi-RELIEF, which contained two rim Axl amino acids out of 15 SDP predictions (R48, E70). In the case of Sequence Harmony, the minority of the predictions (4/15) were located at the rim of Axl:Gas6 (T46, R48, Q50, K96). As such, the combined Axl SDP list, predicted by these three webservers became T46, R48, Q50, E70, D84, K96, where they only agreed on R48. The complete list of the SDP predictions is provided under **Supplementary Table 2**.

TABLE 1 | Positional sequence comparison of R48, E56, E70, D73, E83. E56 and D73 are conserved in both Axl and Tyro3 (shown in bold).

Axl	Tyro3	Mer
R48	N63	N114
E56	**E70**	Q124
E70	Q85	L138
D73	**D87**	H141
E83	-	D151

Axl Selectivity Is Regulated by Salt Bridges

To study partner-selecting Axl SDPs, we modeled the three-dimensional structure of Axl:Pros1 (Ig1:LG1) complex, to use it as the negative (non-binder) control. We refined the two Axl:ligand complexes with HADDOCK 2.2 webserver (van Zundert et al., 2016). We chose HADDOCK, since it provides a user friendly web service to carry out the analysis proposed here.

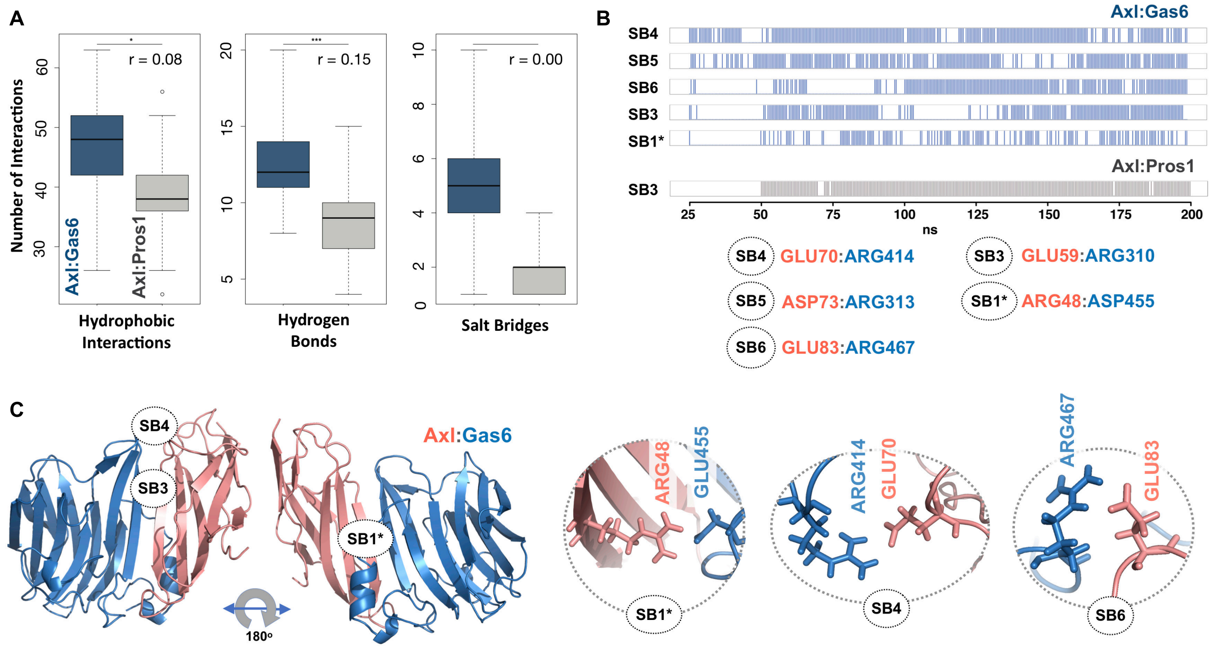


FIGURE 3 | (A) The distributions of hydrophobic contact (left), hydrogen bond (middle) and SB numbers (right) for Axl:Gas6 (marine blue) and Axl:Pros1 (light gray) simulations. The distributions are interpreted with box-and-whisker statistics. The associations between two Axl:Gas6 and Axl:Pros1 distributions were reflected in r values, according to which the SB distributions do not have anything in common. **(B)** Consistently and stably observed SBs formed in Axl:Gas6 and Axl:Pros1 simulations. Each row indicates the observation frequency of the indicated SB. The frequency data came from replica 1 simulations (**Table 2**). The SB numbering follows the ones used in **Figure 2C**. SB1 is denoted with * as in this case R48 couples with a different Gas6 residue than observed in **Figure 2C**. **(C)** The positioning of potential SDPs on Axl:Gas6 structure predicted by molecular dynamics analysis.

When used for refinement, HADDOCK skips docking stages and performs several independent short molecular dynamics simulations in explicit solvent. The top-scoring Axl complexes, ranked by the HADDOCK score, differed mostly in interface electrostatics: Axl:Gas6 has ~3.6 times better electrostatics energy than Axl:Pros1 (−635.8 ± 29 kcal/mol vs. −173.5 ± 21 kcal/mol) (**Figure 2A**). Other interface features and energy terms, such as buried surface area and van der Waals energies, were comparable between the complexes. These results underscore that Axl selectivity is mainly driven by the electrostatics interactions. We analyzed per-residue electrostatics of interfacial Axl residues (31 for the Axl:Gas6 complex and 27 for Axl:Pros1) (**Figure 2B**). In the case of Axl:Gas6, Axl R48, E56, E59, E70, E73, E83 contributed to the interface electrostatics the most (**Figure 2B**). Being at the rim of the Axl:Gas6 complex, these residues formed six different salt bridges (**Figure 2C** and **Supplementary Table 2**). As introduced earlier, previous studies have shown that the ligand-selecting SDPs are rim amino acids, capable of forming opposing charge interactions. This made these salt bridge forming residues the perfect SDP candidates. Among these six salt bridges (SBs), SB2-6 were located on the charged frontal side of the complex (**Figure 2C**, left). Interestingly, SB1 and SB3 (mediated by R48, E59) were also present at the Axl:Pros1 interface. We, therefore, eliminated R48, E59 from the initial Axl SDP list. This left E56, E70, D73, E83 Axl residues as the strongest partner-selecting SDPs. Here, we should note that SB3-SB6 were not present in the Axl:Gas6 crystal structure (**Figure 2D**). The proper establishment of these salt bridges was secured only after the HADDOCK refinement. Finally, if E56, E70, D73, E83 were Gas6-selective, their positions should be substituted with different amino acids in Tyro3 and Mer. This turned out to be the case for E70 and E83, leaving those as the final HADDOCK-based Axl SDP predictions (**Table 1**).

To explore the time-dependent interaction profiles of Axl:Gas6 and Axl:Pros1, we carried out molecular dynamics (MD) simulations of the HADDOCK-refined Axl complexes. Even though running MD simulations requires expertise, we used it to gain the highest resolution information on our system. For each complex, we ran four independent (replica) MD simulations, totaling 1.6 microseconds. The analysis of these trajectories showed that the Axl:Gas6 complex is more stable than Axl:Pros1, as reflected in the lower root mean square deviation (RMSD) (0.15 ± 0.01 nm vs. 0.23 ± 0.03 nm) (**Supplementary Figure 1**), and radius of gyration profiles (**Supplementary Figure 2**). To perform a more in-depth analysis of the interactions between Axl and its ligands, we calculated the inter-molecular hydrophobic, hydrogen bonds and salt bridges formed during the simulations by using the *interfacea* python package (**Figure 3A**). When we pooled the interaction data of each Axl complex, we observed that Axl:Pros1 contained a fewer number of contacts in all interaction types. The most significant difference between Axl:Gas6 and Axl:Pros1 interaction distributions was observed in the case of salt bridges. Axl:Gas6 trajectories reflected, on average, four to five stable salt bridges, where this number dropped to two in the case of Axl:Pros1 (**Figure 3A**, right panel). We then looked for the salt bridges, which were seen in four different trajectories consistently for more than 25% of the simulation time (**Table 2**). Here, our assumption was that the SDP positions should form stable salt bridges within a trajectory and should be observed consistently across four trajectories. These criteria left us with five salt bridges, four of which were the same as the ones selected by the HADDOCK refinement: E70:R414^{Gas6} (SB4), D73:R313^{Gas6} (SB5), E83:R467^{Gas6} (SB6), E59:R310^{Gas6} (SB3) (listed in the decreasing observation frequency in **Figure 2B**). E56-mediated SB2, coming from our HADDOCK refinement analysis was observed only in one replica, indicating that it could be coincidental (**Table 2**). As another surprising outcome, R48 of SB1 formed a stable and consistent salt bridge with D455^{Gas6}, instead of E460^{Gas6}, which was suggested by the HADDOCK refinement. Interestingly, E460^{Gas6} has also a glutamic acid correspondence on Pros1, while in D455^{Gas6} matches with an alanine in Pros1. In the case of Axl:Pros1, only E59:K314^{Pros1} was observed in a statistically significant manner, which corresponds to SB3 of Axl:Gas6 (**Figure 3B** and **Table 2**). These observations left us with four possible selective salt bridges, three of which were formed by the positions unique to Axl: R48, E70, E83 (**Table 1**). Our across-ortholog comparison revealed that R48, E70, E83 are all conserved, supporting the SDP candidacy of these positions (**Supplementary Figure 3**). The spatial distribution of the final list of salt bridges formed by these residues is illustrated in **Figure 3C**.

TABLE 2 | SBs formed in parallel **(A)** Axl:Gas6 and **(B)** Axl:Pros1 simulations.

(A)	Axl Resi	Gas6 Resi	Axl:Gas6-replica #1 (%)	Axl:Gas6-replica #2 (%)	Axl:Gas6-replica #3 (%)	Axl:Gas6-replica #4 (%)
SB4	**70**	**414**	**81.14**	**34.57**	**55.41**	**70.57**
SB5	**73**	**313**	**72.86**	**81.71**	**69.43**	**76.86**
SB6	**83**	**467**	**60.29**	**48.00**	**68.57**	**71.14**
SB3	**59**	**310**	**59.43**	**32.86**	**83.14**	**56.86**
SB1*	**48**	**455**	**43.71**	**44.00**	**53.43**	**30.57**
SB2	56	308	–	–	–	25.14
(B)	**Axl**	**Pros1**	**Axl-Pros1-replica #1 (%)**	**Axl-Pros1-replica #2 (%)**	**Axl-Pros1-replica #3 (%)**	**Axl-Pros1-replica #4 (%)**
SB3	**59**	**314**	**97.67**	**97.67**	**94.33**	**94.33**
SB2	59	316	–	–	–	66.66
SB1	48	465	64.67	51.67	–	33.00

Each row indicates the observation frequency of the denoted salt bridge. The SB numbering follows the ones used in ***Figure 2C****. The consistent and stable SBs are marked in bold. * corresponds to the SB1* presented in* ***Figure 3B****.*

DISCUSSION

In this work, we used three sequence-based SDP predictors, namely, SDPpred, Multi-RELIEF, and Sequence Harmony to map Axl's ligand-selecting SDPs. Next to these approaches, we also carried simple refinement and extensive MD simulations of Axl:ligand interactions. As the primary outcome of this exercise, we found that the sequence-based SDP predictors largely overpredict the potential SDP positions. Hence, we used available literature data to filter out the structurally non-viable ligand-selecting SDPs. As a result, the three methodologies in combination proposed six SDPs, where they agreed only on R48. Our HADDOCK-refinement-based approach suggested R48 as a strong electrostatic contributor to the Axl:Gas6 interface. Though, by only following HADDOCK refined structures, we had to eliminate R48 from the potential SDP list, as it significantly contributed to the Axl:Pros1 interaction energetics too. Elaborate MD simulations were necessary to rescue R48's SDP candidacy. During our MD simulations, R48 formed a new salt bridge, which was neither observed in the crystal nor in the HADDOCK refined complexes. In the end, HADDOCK refinement proposed four selective SBs, three of which were supported by the MD simulations. Checking the evolutionary variance of MD-deduced SDP positions suggested R48, E70, and E83 as the strongest Axl SDP candidates. Strikingly, Multi-RELIEF (plus the rim information) could predict two of these (R48, E70) without running extensive simulations.

To validate R48, E70, and E83 as the ligand-selecting Axl SDPs, we artificially mutated the SB1, SB4, and SB6 forming Gas6 residues their Pros1 counterparts, and vice versa, by using EvoEF1 (Pearce et al., 2019; **Figures 3**, **4**). EvoEF1 is a machine learning approach, poised to calculate the impact of point mutations across protein-protein interfaces. According to EvoEF1, Axl:Gas6 interaction stability was significantly reduced when individual and combined Gas6-to-Pros1 and Gas6-to-alanine mutations were imposed. On the other hand, individual and combined Pros1-to-Gas6 mutations led to a significant increase in the stability of Axl:Pros1 complex (**Figure 4**). These findings underscore the vitality of SB1, SB4, and SB6 to the formation of Axl:Gas6 complex (**Figure 3C**).

Future of the SDP Prediction Field

Given the importance of the knowledge of SDPs for protein design, it is essential to use an economically feasible and accurate predictor. To this end, using machine learning (ML) methodologies in SDP prediction is very suitable, as ML tools would allow calculating dozens of SDP predictions in seconds. Though, the current ML-based approaches face many challenges. As an example, the majority of the sequence-based SDP prediction methods require a precalculated MSA file, together with subfamilies or subgroups definition. Here, caution should be taken as different MSA algorithms produce different alignment results based on varying parameters, which in the end will affect the final SDP list. Besides, dividing protein families into subfamilies requires expert knowledge. As another important limitation, the experimentally determined SDP datasets are rather small, which, in turn, prevents creating large-scale training of the feature-based methods. Construction of such large-scale SDP training datasets will make it possible to use deep learning algorithms, which have outperformed state-of-the-art methods in similar problems (LeCun et al., 2015;

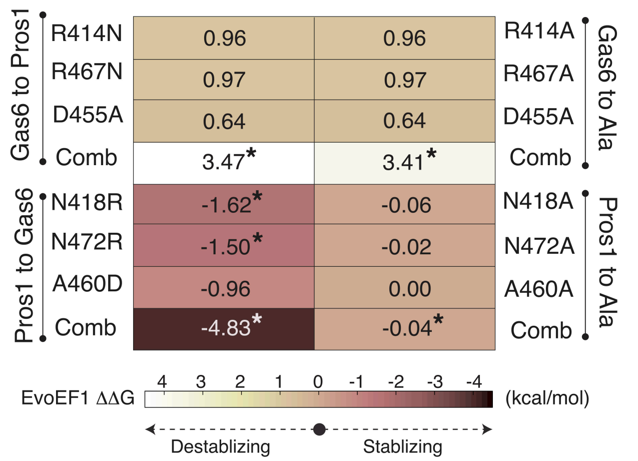


FIGURE 4 | EvoEF1 △△G predictions for the selective Gas6-to-Pros1 or Pros1-to-Gas6 mutations. R414, R467, and D455 correspond to SB4, SB6, and SB1* forming residues, as presented in **Figure 3B**. The significant changes are marked with *. The stabilizing and destabilizing mutation color scheme ranges from –4 (brown) to 4 (yellow). Gas6-to-Ala and Pros1-to-Ala mutations were run as a control. Comb refers to combined mutations.

Zamora-Resendiz and Crivelli, 2019; Gao et al., 2020; Dai and Bailey-Kellogg, 2021). The energy-based methods, as presented in this work under the umbrella of HADDOCK refinement and MD simulations, could offer a refined SDP list, which can be tested experimentally. These approaches, however, take much longer time as they explicitly use structures and calculate forces acting on these structures. As an example, HADDOCK refinement of complexes can take up to half an hour, depending on the available computing resources. MD simulations, on the other hand, can take up to days or weeks, based on the dedicated number of computing cores used. Considering the pros and cons of both approaches, it is evident that new SDP prediction methods, which combine the advantages of both sequence- and structure-based methodologies, should be developed. However, until then, to predict SDPs, conservation-filtered HADDOCK refinement can be used in combination with structurally-filtered Multi-RELIEF predictions. Both of these approaches are easily accessible through web services. Their combination covers all of the conservation-filtered MD-based SDP predictions, without the requirement of heavy calculations.

METHOD

Sequence-Based Methods

SDPpred is an entropy-based SDP prediction method which utilizes mutual information to determine well-conserved residues within the same groups but differ between them (Kalinina et al., 2004). The equation to mutual information score for a column p in the alignment is given below:

$$I_p \ = \sum_{i \ =1}^{N} \sum_{a \ =1}^{20} f_p(\alpha,\, i)\log(\frac{f_p(a,\, i)}{f_p(a)f_p(i)})$$

In this equation, N is the number of specificity groups, a is the amino acid type, $f_p(i)$ ratio of protein sequences belonging to group i. $f_p(a)$ is the number of occurrences of residue a in the whole alignment at position p. $f_p(a, i)$ is the number of occurrences of residue a in group i at position p. SDPpred calculates column-wise scores for each position in the MSA and outputs SDPs over the protein sequences. The server can be reached at http://monkey.belozersky.msu.ru/~psn/query.htm.

Multi-RELIEF a machine-learning based SDP prediction method which employs RELIEF algorithm to identify specificity determining residues (Kononenko, 2005; Ye et al., 2008). Multi-RELIEF algorithm requires predefined groups and their MSA as input. The aim of this method is to calculate a weight vector for each position in MSA. The weight vector is initialized with zeros at the beginning. At each iteration, a random sequence *seq* is selected and its nearest neighbors from the same class (i.e., hit(*seq*)) and opposite class (i.e., miss(*seq*)) are determined based on the Hamming distance. Subsequently, the weight of each residue is calculated with the following equation:

$$\mathrm{w}\,[i] = w\,[i] - \frac{diff\,\left(seq\,[i]\,,\, miss\,(seq)\,[i]\right)}{m}$$

$$+\frac{diff\,\left(seq\,[i]\,,\, hit\,(seq)\,[i]\right)}{m}$$

where

$$diff\ \ (a,\ b)\ \ = \begin{cases} 0, & a \ = b \\ 1, & a \ \neq b \end{cases}$$

In the above equation, i represents the ith position in the weight vector or sequences and m represents the number of sequences. The algorithm outputs a weight vector whose length is the same as the number of positions in the alignment. Higher weights indicates the higher probability of being SDP for the corresponding position.

Sequence Harmony is another entropy-based SDP prediction method (Feenstra et al., 2007). It takes MSA and two user-specified groups as input and calculates relative entropy scores for each residue that shows degree of conservations. Sequence Harmony provides ranking of the entropy scores as outputs. Sequence Harmony and Multi-RELIEF methods are merged under the Multi-Harmony web server at https://www.ibi.vu.nl/programs/shmrwww/ (Brandt et al., 2010).

Template-Based Modeling of Axl:ligand Complexes

LG1 domain of Pros1 was modeled with i-TASSER (Roy et al., 2010). Pros1-to-Gas6 structural alignment was carried out with FATCAT web-tool (Ye and Godzik, 2003) (by using the Gas6 coordinates of 2C5D). The final Axl:Pros1 coordinates were visualized and saved in PyMOL (Schrödinger, 2015). All Axl:ligand complexes were water refined with HADDOCK2.2 web server (van Zundert et al., 2016). The standard HADDOCK refinement protocol samples 20 models. These models slightly differ from each other as each one is refined with molecular dynamics simulation starting with a different initial velocity.

In the end, the generated models are ranked with the HADDOCK score, which is a sum of electrostatics (E_Elec), van der Waals (E_vdW) and desolvation terms (E_desolv): 1.0. E_vdW+ 0.2. E_elec + 1.0. E_desolv. The top ranking four models, i.e., the best four models with the lowest HADDOCK scores, are offered as the final complex states. We generated 200 refined structures for each Axl:ligand complex. The top four ranking models were isolated as the final solutions.

HADDOCK refinement outputs residue-based energy scores of each complex (expressed in E_Elec, E_vdW and E_elec+E_vdW), deposited in *ene-residue.disp* file (can be found under HADDOCK output folder: structures/it1/water/analysis). This file describes the contributions of each interface amino acid to the intermolecular interaction. These residue-based HADDOCK energies were analyzed by using R (R Core Team, 2013) and Rstudio (RStudio Team, 2020).

Molecular Dynamics Simulations

GROMACS 5.1.4 software and its tools were used to run molecular dynamics simulations (MD) and quality controls (e.g., temperature, pressure, RMSD, Rg analyses) (Van Der Spoel et al., 2005). The AMBER99SB-ILDN force field (Lindorff-Larsen et al., 2010) was used to parameterize the protein molecules, while the TIP3P water model was used to represent the solvent (Jorgensen et al., 1983). The simulation was run in a rhombic dodecahedron unit cell. The minimum periodic distance to the simulation box was set to be 1.4 nm. The mdp simulation files were adapted from https://github.com/haddocking/molmod-data (Rodrigues et al., 2016).

Before the production run, each complex was minimized in vacuum by using the steepest descent algorithm (Mandic, 2004). They were then solvated with the TIP3P water, together with neutralizing ions (51 NA+ and 48 CL- ions were added to neutralize Axl:Gas6, while 58 NA+ and 49 CL- ions were added to neutralize Axl:Pros1). The relevant topology files were edited according to the newly included NA+ and CL- ions. The second cycle of energy minimization was performed on the solvated systems. The solvent and hydrogen atoms were relaxed with a 20 ps long molecular dynamics simulation under constant volume where the temperature was equilibrated to 300 K (NVT). This was followed by 20 ps long molecular dynamics simulation under constant pressure where the pressure is equilibrated to 1 bar (NPT). As a last step before the production run, position restraints were released upon reduction of its force constant from 1,000 to 100, 100 to 10, and 10 to 0. To generate a parallel run of a given complex, random seed was changed before running the NVT step. The coordinates were written in every 10 ps. The integration time step was set to 2 fs.

For each Axl complex, we ran four independent (replica) MD simulations, totaling 1.6 microseconds. In each simulation, upon reaching 200 ns, the periodic boundary conditions were corrected. The system was stripped off solvent and ion atoms. The Root Mean Square Deviations (RMSDs) were calculated by using the average coordinates as a reference. After leaving the equilibration periods out (25 ns for Axl:Gas6 and 50 ns for Axl:Pros1), 350 snapshots for Axl:Gas6 and 300 snapshots from Axl:Pros1 were extracted.

Interface Analysis

The interfacial hydrophobic contacts, hydrogen and salt bridges were calculated with *interfacea* python library (https://github.com/JoaoRodrigues/interfacea) (Rodrigues et al., 2019). *interfacea* classifies an inter-monomer interaction as hydrophobic, if there are at least two non-polar atoms within 4.4 Å. It considers a pairwise contact as a hydrogen bond, if a hydrogen donor (D) and acceptor (A) gets within 2.5 Å. It then filters D-H-A triplets with a minimum angle threshold (default 120 degrees). Finally, it classifies an interaction as a salt bridge if there are oppositely charged groups within 4.0 Å. For the interface classification as core and rim, EPPIC webserver was used (Duarte et al., 2012). Core residues are the ones that are buried at least in the protein structure (>95%). The rest of interface residues were counted as rim residues.

For comparing different simulations, the box-and-whisker statistics were generated with the standard boxplot function of R. The salt bridges were classified as stable if they were observed for >25% of a simulation time. They were classified as consistent if they were observed to be stable in all simulations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in https://github.com/CSB-KaracaLab/Paralog_SDP.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptualization of this work, as well as to the writing of the manuscript. AR and TK performed the sequence-based analysis. TK performed simulations and all the structure-based analysis. JR devised and developed *interfacea* package. EK participated at all levels and supervised the work.

FUNDING

This work was supported by the EMBO Installation Grant granted to EK (project # 4421).

ACKNOWLEDGMENTS

We wholeheartedly thank Mehmet Öztürk and Tuğçe Batur for their assistance on Axl biology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.658906/full#supplementary-material

REFERENCES

Ahmad, S., and Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 6:33. doi: 10.1186/1471-2105-6-33

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* 44, W344–W350.

Brandt, B. W., Feenstra, K. A., and Heringa, J. (2010). Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.* 38, W35–W40.

Capra, J. A., and Singh, M. (2008). Characterization and prediction of residues determining protein functional specificity. *Bioinformatics* 24, 1473–1480. doi: 10.1093/bioinformatics/btn214

Chagoyen, M., García-Martín, J. A., and Pazos, F. (2016). Practical analysis of specificity-determining residues in protein families. *Brief. Bioinform.* 17, 255–261. doi: 10.1093/bib/bbv045

Chakrabarti, P., and Janin, J. L. (2002). Dissecting protein-protein recognition sites. *Proteins* 47, 334–343. doi: 10.1002/prot.10085

Chakrabarti, S., and Panchenko, A. R. (2008). Coevolution in defining the functional specificity. *Proteins Struct. Funct. Bioinform.* 75, 231–240. doi: 10.1002/prot.22239

Chakraborty, A., and Chakrabarti, S. (2015). A survey on prediction of specificity-determining sites in proteins. *Brief. Bioinform.* 16, 71–88. doi: 10.1093/bib/bbt092

Dai, B., and Bailey-Kellogg, C. (2021). Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*

del Sol Mesa, A., Pazos, F., and Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* 326, 1289–1302. doi: 10.1016/s0022-2836(02)01451-1

Duarte, J. M., Srebniak, A., Schärer, M. A., and Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinformatics* 13:334–316. doi: 10.1186/1471-2105-13-334

Feenstra, K. A., Pirovano, W., Krab, K., and Heringa, J. (2007). Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.* 35, W495–W498.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief. Bioinform.* 7, 225–242. doi: 10.1093/bib/bbl004

Gao, W., Mahajan, S. P., Sulam, J., and Gray, J. J. (2020). Deep learning in protein structural modeling and design. *Patterns* 1:100142. doi: 10.1016/j.patter.2020.100142

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., et al. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19, 163–164. doi: 10.1093/bioinformatics/19.1.163

Gogarten, J. P., and Olendzenski, L. (1999). Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* 9, 630–636. doi: 10.1016/s0959-437x(99)00029-5

Hafizi, S., and Dahlbäck, B. (2006). Gas6 and protein S: vitamin K-dependent ligands for the Axl receptor tyrosine kinase subfamily. *FEBS J.* 273, 5231–5244. doi: 10.1111/j.1742-4658.2006.05529.x

Ivanov, S. M., Cawley, A., Huber, R. G., Bond, P. J., and Warwicker, J. (2017). Protein-protein interactions in paralogues: electrostatics modulates specificity on a conserved steric scaffold.Srinivasan N, editor. *PLoS One* 12:e185928. doi: 10.1371/journal.pone.0185928

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869

Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.* 32, W424–W428.

Kononenko, I. (2005). *Estimating Attributes: Analysis and Extensions of RELIEF. In: Machine Learning: ECML-94*, Vol. 784. Berlin: Springer, 171–182.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., et al. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* 33, W299–W302.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lemke, G. (2013). Biology of the TAM receptors. *Cold Spring Harb Perspect Biol.* 5:a009076. doi: 10.1101/cshperspect.a009076

Levy, E. D. (2010). A Simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403, 660–670. doi: 10.1016/j.jmb.2010.09.028

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinform.* 78, 1950–1958. doi: 10.1002/prot.22711

Linger, R. M., Keating, A. K., Earp, H. S., and Graham, D. K. (2008). TAM receptor tyrosine kinases: biologic functions, signaling, and potential therapeutic targeting in human cancer. *Adv. Cancer Res.* 100, 35–83. doi: 10.1016/S0065-230X(08)00002-X

Mandic, D. P. (2004). Descent algorithm. *Signal. Process.* 11, 115–118.

Mirny, L. A., and Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* 321, 7–20. doi: 10.1016/s0022-2836(02)00587-9

Moll, M., Finn, P. W., and Kavraki, L. E. (2016). Structure-guided selection of specificity determining positions in the human Kinome. *BMC Genomics* 17(Suppl. 4), 431–339. doi: 10.1186/s12864-016-2790-3

Myers, K. V., Amend, S. R., and Pienta, K. J. (2019). Targeting Tyro3, Axl and MerTK (TAM receptors): implications for macrophages in the tumor microenvironment. *Mol. Cancer* 18:94.

Nicoludis, J. M., Green, A. G., Walujkar, S., May, E. J., Sotomayor, M., Marks, D. S., et al. (2019). Interaction specificity of clustered protocadherins inferred from sequence covariation and structural analysis. *Proc. Natl. Acad. Sci. USA* 116, 17825–17830. doi: 10.1073/pnas.1821063116

Pazos, F., Rausell, A., and Valencia, A. (2006). Phylogeny-independent detection of functional residues. *Bioinformatics* 22, 1440–1448. doi: 10.1093/bioinformatics/btl104

Pearce, R., Huang, X., Setiawan, D., and Zhang, Y. (2019). EvoDesign: designing protein-protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function. *J. Mol. Biol.* 431, 2467–2476. doi: 10.1016/j.jmb.2019.02.028

Pirovano, W., Feenstra, K. A., and Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.* 34, 6540–6548. doi: 10.1093/nar/gkl901

Pitarch, B., Ranea, J. A. G., and Pazos, F. (2020). Protein residues determining interaction specificity in paralogous families. *Bioinformatics* doi: 10.1093/bioinformatics/btaa934

R Core Team (2013). *R: A Language and Environment for Statistical Computing [Internet]*. Vienna: R Core Team.

Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1995–2000. doi: 10.1073/pnas.0908044107

Rodrigues, J. P. G. L. M., Melquiond, A. S. J., and Bonvin, A. M. J. J. (2016). Molecular dynamics characterization of the conformational landscape of small peptides: a series of hands-on collaborative practical sessions for undergraduate students. *Biochem. Mol. Biol. Educ.* 44, 160–167. doi: 10.1002/bmb.20941

Rodrigues, J., Valentine, C., and Jimenez, B. (2019). JoaoRodrigues/interfacea: first beta version of the API.

Rothlin, C. V., and Lemke, G. (2010). TAM receptor signaling and autoimmune disease. *Curr. Opin. Immunol.* 22, 740–746. doi: 10.1016/j.coi.2010.10.001

RStudio Team (2020). *RStudio: Integrated Development for R*. Boston, MA: RStudio, PBC. Available online at: http://www.rstudio.com/

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Sasaki, T., Knyazev, P. G., Clout, N. J., Cheburkin, Y., Göhring, W., Ullrich, A., et al. (2006). Structural basis for Gas6 – Axl signalling. *EMBO J.* 25, 80–87. doi: 10.1038/sj.emboj.7600912

Schrödinger, L. L. C. (2015). *The PyMOL Molecular Graphics System, Version˜1.8.*

Sievers, F., and Higgins, D. G. (2017). Clustal Omega for making accurate alignments of many protein sequences. *Prot. Sci.* 27, 135–145. doi: 10.1002/pro.3290

Sloutsky, R., and Naegle, K. M. (2016). High-resolution identification of specificity determining positions in the laci protein family using ensembles of sub-sampled alignments. *PLoS One* 11:e0162579. doi: 10.1371/journal.pone.0162579

Teppa, E., Wilkins, A. D., Nielsen, M., and Buslje, C. M. (2012). Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics* 13:235–238. doi: 10.1186/1471-2105-13-235

The UniProt Consortium (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.

Van Der Meer, J. H. M., Van Der Poll, T., and Van't Veer, C. (2014). TAM receptors, Gas6, and protein S: roles in inflammation and hemostasis. *Blood* 123, 2460-2469. doi: 10.1182/blood-2013-09-528752

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291

van Wijk, S. J. L., Melquiond, A. S. J., de Vries, S. J., Timmers, H. T. M., and Bonvin, A. M. J. J. (2012). Dynamic control of selectivity in the ubiquitination pathway revealed by an d to e substitution in an intra-molecular salt-bridge network. *PLoS Comput. Biol.* 8:e1002754 doi: 10.1371/journal.pcbi.100275

van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., et al. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* 428, 720–725. doi: 10.1016/j.jmb.2015.09.014

Wang, S., Qiu, Z., Hou, Y., Deng, X., Xu, W., Zheng, T., et al. (2021). AXL is a candidate receptor for SARS-CoV-2 that promotes infection of pulmonary and bronchial epithelial cells. *Cell Res.* 31, 126–140. doi: 10.1038/s41422-020-00460-y

Whisstock, J. C., and Lesk, A. M. (2004). Prediction of protein function from protein sequence and structure. *Quart. Rev. Biophys.* 36, 307–340.

Wium, M., and Paccez, J. D. (2018). The dual role of tam receptors in autoimmune diseases and cancer : an overview. *Cells* 7:166. doi: 10.3390/cells7100166

Wong, K.-C., Li, Y., Peng, C., Moses, A. M., and Zhang, Z. (2015). Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res.* 43, 10180–10189.

Wu, G., Ma, Z., Hu, W., Wang, D., Gong, B., Fan, C., et al. (2017). Molecular insights of Gas6 / TAM in cancer development and therapy. *Cell Death Dis.* 8:e2700. doi: 10.1038/cddis.2017.113

Yanagihashi, Y., Segawa, K., Maeda, R., Nabeshima, Y.-I., and Nagata, S. (2017). Mouse macrophages show different requirements for phosphatidylserine receptor Tim4 in efferocytosis. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8800–8805. doi: 10.1073/pnas.1705365114

Ye, K., Feenstra, K. A., Heringa, J., Ijzerman, A. P., and Marchiori, E. (2008). Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics* 24, 18–25. doi: 10.1093/bioinformatics/btm537

Ye, Y., and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 (Suppl 2):ii246-55. doi: 10.1093/bioinformatics/btg1086

Ye, K., Lameijer, E.-W. M., Beukers, M. W., and Ijzerman, A. P. (2006). A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins Struct . Funct. Bioinform.* 63, 1018–1030. doi: 10.1002/prot.20899

Zamora-Resendiz, R., and Crivelli, S. (2019). Structural learning of proteins using graph convolutional neural networks. *bioRxiv* [Preprint]. doi: 10.1101/610444

Zhu, C., Wei, Y., and Wei, X. (2019). AXL receptor tyrosine kinase as a promising anti-cancer approach: functions, molecular mechanisms and clinical applications. *Mol. Cancer* 18:153.

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/fmolb.2021.637355

Integrative Predictive Modeling of Metastasis in Melanoma Cancer Based on MicroRNA, mRNA, and DNA Methylation Data

*Ayşegül Kutlay and Yeşim Aydin Son**

Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey

Introduction: Despite the significant progress in understanding cancer biology, the deduction of metastasis is still a challenge in the clinic. Transcriptional regulation is one of the critical mechanisms underlying cancer development. Even though mRNA, microRNA, and DNA methylation mechanisms have a crucial impact on the metastatic outcome, there are no comprehensive data mining models that combine all transcriptional regulation aspects for metastasis prediction. This study focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation.

Method: We developed multiple machine learning models to distinguish the metastasis by integrating miRNA, mRNA, and DNA methylation markers. We used the TCGA melanoma dataset to differentiate between metastatic melanoma samples by assessing a set of predictive models. For this purpose, machine learning models using a support vector machine with different kernels, artificial neural networks, random forests, AdaBoost, and Naïve Bayes are compared. An iterative combination of differentially expressed miRNA, mRNA, and methylation signatures is used as a candidate marker to reveal each new biomarker category's impact. In each iteration, the performances of the combined models are calculated. During all comparisons, the choice of the feature selection method and under and oversampling approaches are analyzed. Selected biomarkers of the highest performing models are further analyzed for the biological interpretation of functional enrichment.

Results: In the initial model, miRNA biomarkers can identify metastatic melanoma with an 81% F-score. The addition of mRNA markers upon miRNA increased the F-score to 92%. In the final integrated model, the addition of the methylation data resulted in a similar F-score of 92% but produced a stable model with low variance across multiple trials.

Conclusion: Our results support the role of miRNA regulation in metastatic melanoma as miRNA markers model metastasis outcomes with high accuracy. Moreover, the integrated evaluation of miRNA with mRNA and methylation biomarkers increases the model's power. It populates selected biomarkers on the metastasis-associated pathways of

Edited by:
Arzucan Ozgur,
Boğaziçi University, Turkey

Reviewed by:
Slavica Jonic,
UMR7590 Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), France
Rene Snajder,
German Cancer Research Center (DKFZ), Germany

***Correspondence:**
Yeşim Aydin Son
yesim@metu.edu.tr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

***Received:** 03 December 2020*
***Accepted:** 30 July 2021*
***Published:** 23 September 2021*

Citation:
Kutlay A and Aydin Son Y (2021) Integrative Predictive Modeling of Metastasis in Melanoma Cancer Based on MicroRNA, mRNA, and DNA Methylation Data.
Front. Mol. Biosci. 8:637355.
doi: 10.3389/fmolb.2021.637355

melanoma, such as the "osteoclast", "Rap1 signaling", and "chemokine signaling" pathways.

Source Code: https://github.com/aysegul-kt/MelonomaMetastasisPrediction/

Keywords: machine learning, melanoma, metastasis, metastatic molecular signatures, miRNA, mRNA, DNA methylation

INTRODUCTION

Melanoma, a cancer with a rapid increase in incidence and high mortality, is a malignant tumor of skin pigmentation cells with a high mortality rate. Melanoma can develop anywhere on the body but is most commonly observed in areas exposed to the sun, such as the back, legs, arms, and face. With nearly 300,000 cases, melanoma is one of the most common cancer types worldwide (World Cancer Research Fun, 2021).

According to CDC statistics, 85,000 new cases are reported in the United States on a yearly basis, where 8,000 people die annually (United States Cancer Stat, 2021). In the European Union, on the other hand, melanoma cancer incidence reaches 14,000 annual cases. It is considered one of the fastest rising forms of cancer, albeit with hot spots in Europe, those being Scandinavian countries, Switzerland, and Austria (American Cancer Society, 2016). In addition, 16,000 new melanoma cases have been reported in the United Kingdom, which corresponds to 4% of all cancer types, and it has had a rising incidence rate of 135% over 30 years (Cancer Research UK, 2017).

Both distant and regional metastases are possible in melanomas. The most common metastasis sites in melanoma cases are bone, the brain, the liver, the lung, and skin. The presence of skin metastasis may be the first outward sign of lymphatic or hematogenous spreading. So in melanoma, rather than diagnosis, the prognosis is a critical concern. It is possible to detect at least suspicious cases *via* visual examination or short screening. Early diagnosis leads to high cure rates, but there is still no effective treatment in later stages, where metastasis is observed frequently (Damsky et al., 2010).

Signatures for Metastasis

When the balance between cell growth and death is disrupted as a result of either "uncontrolled cell growth" or "loss of apoptosis (programmed cell death)", tumorigenesis starts (Ma and Weinberg, 2008; Oppenheimer, 2006). At the initial stage, a malignant tumor presents at the site of the initial conversion of a normal cell to a tumor cell, called a primary tumor. This primary tumor may stay stable in this originated tissue (benign) or spread to the other parts of the body (malignant) by invasion or metastasis (Carter, 1974; Oppenheimer, 1982; Tonini et al., 2003; Shen et al., 2013). Understanding the molecular basis of carcinogenesis is essential in preventing, diagnosing, and treating cancer and its metastasis (Harris, 1991).

Many different markers have been proposed to describe the molecular foundation of metastasis. DNA methylation, gene expression profiles, and microRNAs are frequent biomarkers for predicting metastasis for most cancer types.

MicroRNAs are noncoding RNAs and regulate proliferation, cell cycle control, apoptosis, differentiation, migration, and metabolism (Kasinski and Slack, 2011; Jansson and Lund, 2020; Stahlhut and Slack, 2013). So, it is not surprising that microRNAs play a crucial role as suppressors or promoters of carcinogenesis or metastasis by controlling their target mRNA (Shalaby et al., 2014). Based on this understanding, microRNAs became the main focus in cancer biology and were proven to be crucial components of the normal and pathologic states of cells (Stahlhut and Slack, 2013; Hayes et al., 2014).

DNA methylation is a chemical process in which DNA binds with a methyl group. This process modifies the functionality of the DNA itself. It is an important regulator that plays a crucial role in genomic imprinting, X-chromosome inactivation, repression of repetitive elements, and aging. DNA methylation associates with many types of cancer (Zhang et al., 2011). Global hypomethylation also implicates cancer development and progression through different mechanisms (Craig and Wong, 2011). Typically, there is hypermethylation of tumor suppressor genes and hypomethylation of oncogenes (Gonzalo, 2020; Melchers et al., 2015).

Predictive Models for Metastasis for Other Cancer Types

Although predictive machine learning models for melanoma metastasis are limited, many studies propose predictive biomarkers for different metastatic cancers. While most of the studies target specific markers such as microRNA or protein expression, recent studies (Souza et al., 2017) investigate the integrated usage of miRNA and mRNA signatures.

Binary logistic regression, which uses miRNA-331 and miRNA-195 as markers, is able to distinguish between metastasis and local breast cancer (sensitivity = 0.95 and specificity = 0.76) (McAnena et al., 2019). A study conducted by Souza et al. (2017) developed an integrated model using the expression levels of 27 miRNAs and 81 target mRNAs to classify prostate cancer patients from controls with 67% sensitivity and 75% specificity. Another study reports a statistical model with 71.4% accuracy for forecasting lymph node metastasis with independent test cases (Moriya et al., 2009). Besides, the SVM (support vector machine) classifier, which uses gene expression profiling with microarrays, predicts metastasis with 78% accuracy for breast cancer (Burton et al., 2012). To predict the lymph node metastasis of primary lung cancer tumors, computerized tomography (CT) and mRNA expression profiling are combined *via* statistical analysis (Chang et al., 2008). This method increased the accuracy from 55% (CT) to 86% (CT and mRNA). A statistical model built with ANOVA and hierarchical clustering predicts future metastasis in head and neck squamous cell carcinoma (HNSCC) with an accuracy of 77% (Rickman et al., 2008). The research conducted by Kan et al. (2004) proposed a predictive model for "lymph node metastasis" by using artificial neural networks based on gene expression profiles of primary tumors with an accuracy of 77%.

Chen and colleagues (Chen et al., 2009) studied "cancer metastasis networks". In that study, a large set of patient data and the prediction of progression patterns generated a system network for the primary tumor and the sides of metastasis. By using these networks (which are constructed by hierarchical clustering), they have tried to predict the primary site of the tumor after a sequence of metastasis multinomial logistic regression with an overall accuracy of 51% (prostate, 84%; colon, 80%; lung and bronchus, 69%; ovary, 64%; larynx, 61%; and female breast, 56%).

Roessler et al. (2010) have generated a risk classifier tool to predict hepatocellular tumors by using gene expression levels with a combination of serum AFP levels or BCLC staging. Among six different prediction algorithms—support vector machines (SVMs), nearest centroid (NC), 3-nearest neighbor (3-NN), 1-nearest neighbor (1-NN), linear discriminant analysis (LDA), or compound covariate predictor (CCP)—CCP achieved the best sensitivity and specificity (76 and 60%, respectively) on cases from the "Liver Cancer Institute". They also tested the model on another case set from the "Laboratory of Experimental Carcinogenesis". The model predicts the risk with a sensitivity of 84% and a specificity of 65%.

Another study (Watanabe et al., 2010) proposed a model to predict liver metastasis with a primary colorectal tumor by using gene expression profiles of DNA microarray samples with the k-nearest neighbor (KNN) method and 10-fold cross-validation. The model predicts metastasis with 86.2% accuracy. Zemmour et al. (2015) developed three models (elastic net, LASSO, and CoxBoost) to predict early breast cancer metastasis using DNA microarray data. The study used a publicly available dataset as a training set. Then they validated the results on two different datasets (van de Vijver's and Desmedt's). The model predicts metastasis with 66% accuracy on the previous and 59% accuracy on the other dataset.

Predictive Models for Melanoma Metastasis

Unlike other cancers, there are limited studies on modeling melanoma metastasis. Recently, serum levels of the cytokines IL-4, GM-CSF, and DCD and the Breslow thickness were proposed as markers to predict melanoma metastasis, where a linear regression achieved the best balance accuracy (80%) in the test set (Mancuso et al., 2020). A deep convolutional neural network (DCNN) study to predict BAP1 mutation also identified decisive prognostic factors for predicting metastatic risk *via* whole slide images with an area under the curve of 0.90 (Zhang et al., 2020). Additionally, mir-205-5p was found to be a significant biomarker for metastatic melanoma by Valentini et al. (2019). Also, Wei et al. (2019) indicated that TRIM44-tripartite motif-containing protein-44, regulated by miR-26b-5p, was identified as amplified on melanoma tissues. The same study reported miR-26-5p as downregulated on melanoma. The study conducted by Kinslechner et al. (2019) showed that scavenger receptor class B type 1 (SR-BI) protein expression contributes to metastatic melanoma. Wang et al. (2019) proposed long noncoding RNA TUG1 as a prognostic biomarker of metastatic melanoma. Besides, they have also indicated that miR-29c-3p, which is the target for G-protein signaling 1 (RGS1), suppresses the expression of TUG1.

Overall, transcriptional regulation is one of the critical mechanisms underlying cancer development. Even though mRNA, microRNA, and DNA methylation mechanisms have a critical impact on metastatic outcomes, there are no comprehensive data mining models that combine all aspects of transcriptional regulation for metastasis prediction. In this study, we focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation. We used differentially expressed miRNA, mRNA, and methylation signatures on the TCGA melanoma dataset to distinguish metastatic melanoma samples by assessing a set of predictive models. The highest performing model is selected, and its biomarkers are further analyzed for the biological interpretation of functional enrichment and to determine regulatory networks.

We used the TCGA Skin Cutaneous Melanoma (SKCM) dataset, which has been analyzed in various studies on the overall survival and identification of prognostic markers based on genomics data. Xiong et al. (2019) used clinical data and miRNA sequencing data to associate the observed survival rate. Similarly, Chen et al. (2017), Yang et al. (2018), Ma et al. (2017), and Xue et al. (2020) studied RNA sequencing data and proposed noncoding RNAs for SKCM prognosis. Guo et al. (2015) combined miRNA and mRNA sequencing data and proposed 15 miRNAs and 5 mRNAs for prognosis. Additionally, Jiang et al. presented the integration of mutation, copy number variation, methylation, and mRNA expression data for identifying prognostic markers. Selitsky et al. (2019) used mRNA sequencing data and applied machine learning models to measure the relative similarity of gene expression profiles of bulk tumor samples and different B cell phenotypes.

MATERIALS AND METHODS

In the study, opened data for Skin Melanoma (SKCM) (The Cancer Genome Atlas N, 2015) of the TCGA (The Cancer Genome Atlas) database are used, which is a part of the TCGA dataset served on the Cancer Genomics Cloud (CGC). The Cancer Genomics Cloud (CGC) (Institute, 2020) hosts a large genomic dataset and provides tools for searching and analyzing genomic data, serving as a computational environment on the cloud. The data browser tool provided by the CGC is used to search for TCGA cases and CCLE cell lines. On TCGA, a melanoma dataset with 470 cases composed of 352 metastatic and 97 primary tumor samples is used during this study, with three experimental strategies in the dataset, namely, miRNA expression, mRNA expression, and methylation.

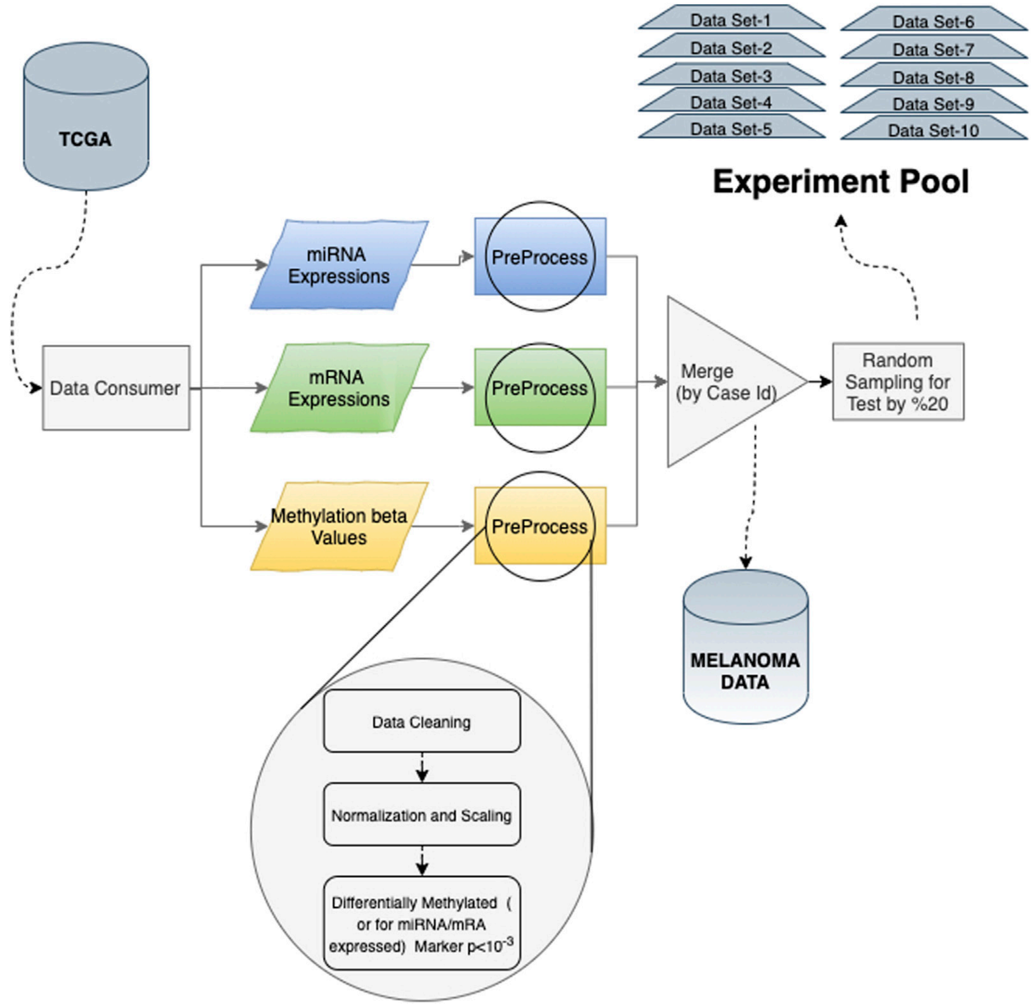


FIGURE 1 | Experimental pool generation process: each method is evaluated using a sample experimental pool under the same circumstances. miRNA, mRNA, and methylation data consumed through TCGA were processed separately and merged to generate the whole melanoma marker dataset. Then, through random splinting, 10 individual sample datasets are constructed. Each random split is saved by applying both undersampling and oversampling (SMOTE) techniques.

We have collected the melanoma data for miRNA sequencing, RNA sequencing, and methylation array for this study's systematical analysis. For 470 different cases with primary and metastatic melanoma, tissue samples are compared to distinguish the metastatic melanoma from the primary tumor. We finalized the predictive model input preprocessing by applying data cleaning, normalization, and scaling preprocessing steps for the remaining 449 cases (**Figure 1**). Overall, 470 distinct cases and 11,265 opened files have been found by using four filters:

1. Primary Site (Skin)
2. Project (TCGA-SKCM)
3. Experimental Strategy (miRNA-Seq; Methylation array; RNA-Seq)
4. File Access (Open)

We generated a subset of cases, which contains all data for "miRNA sequences", "methylation array", and "RNA sequences". In the current interface of the GDC Data Portal, the following search query provides the data files in the repository:

> Cases.primary_site in ("skin") and cases.project.program.name in ("TCGA") and cases.project.project_id in ("TCGA-SKCM") and files.access in ("open") and files.experimental_strategy in ("Methylation Array", "RNA-Seq", "miRNA-Seq").

TCGA provides various attributes for "miRNA sequences", "methylation array", and "RNA sequences". For miRNA, we used "miRNA Expression Quantification", which are miRNA expressions provided as a table that associate miRNA IDs with a read count and a normalized count in reads per million miRNA mapped. Raw read counts, the number of reads aligned to each gene, calculated using the HT-Seq algorithm, are used for mRNA. Ensemble gene ID represents these data and the number of read-aligned mRNAs. For methylation analysis, TCGA provides beta-

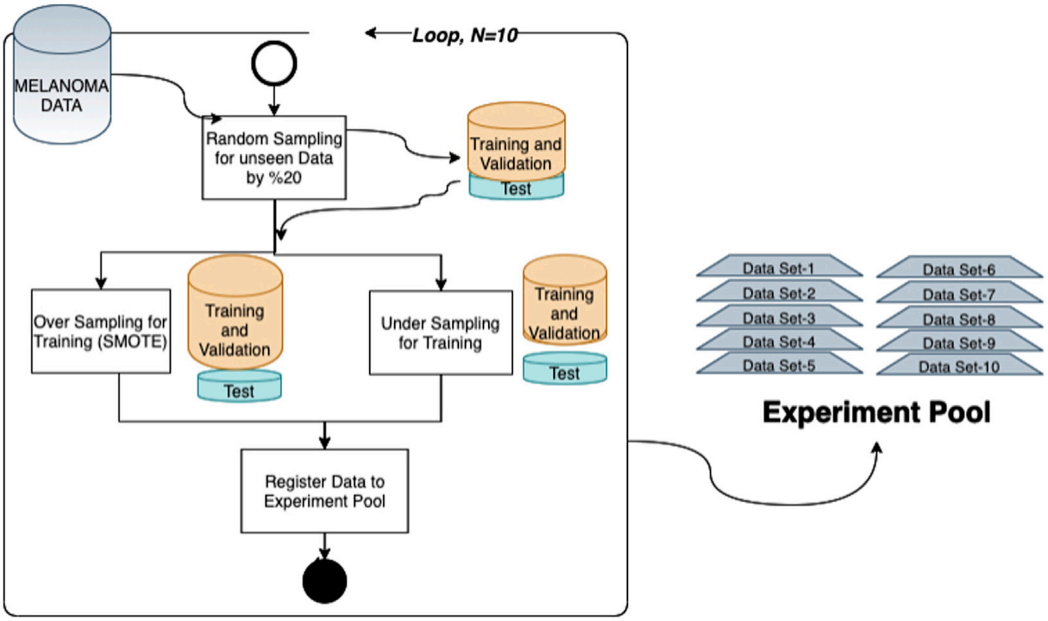


FIGURE 2 | Training validation and unseen test data generation in each trial; for this purpose, the following steps are followed for both undersampling and oversampling. 1) The significant variables listed in the given category are selected from the dataset. 2) A set selected randomly from whole data with an 80% ratio of each class is kept for unseen data. 3) A technique is applied to solve the curse of the dimensionality problem (for dimensional reduction, principal component analysis is applied, and for oversampling runs, the SMOTE algorithm is used with K = 3). 4) Steps 1–3 are repeated for each data split in the experiment pool.

values, which approximate the percentage of methylation of the gene (**Figure 1**).

The data analysis is started with data preprocessing and variable selection. miRNA expression is used for the initial cycle of the spiral analysis method. Then, 11,265 separate files that contain miRNA and mRNA expressions for each case are downloaded from TCGA with a manifest file that contains metadata for the specific case. The manifest file is used to read and combine case files to generate a data pool. The final data pool contains 472 observations with 60,492 properties for mRNA, 450 observations with 1,904 properties for miRNA, and 483 observations with 34,014 variables for methylation. We only chose the cases which have all three experiments, namely, miRNA, mRNA, and methylation.

The sample type property is used for the class variable, which is a categorical variable with four levels, namely, “Primary Tumor”, “Solid Tissue Normal”, “Metastatic”, and “Additional Metastatic”. Only the samples with “Primary Tumor” and “Metastatic” are selected for further analysis.

There were variables for miRNA and mRNA expressions with a constant (1 or 0) value for all samples. These attributes have been removed from the dataset. The remaining samples are subject to a significance test concerning class variables; log normalization and Z-score normalization used for relevant markers. Markers are scaled in the 0–1 range. The t-test has been used as a significance test (p-value is defined as 0.001). As a result of the test, 425 miRNAs, 2061 mRNAs, and 8,698 methylation variables were significantly expressed between the two groups (“Primary Tumor” and “Metastatic”).

For a detailed analysis of the results, all possible miRNA patterns and their target mRNA and gene methylation are calculated. Then, depending on the significance level, different patterns are defined *via* evaluation with each other.

Random selection is applied for each class with a 20% ratio to separate unseen data for testing during the analysis. We repeated this randomization process to create 10 different splits, which are used as a separate trial. By generating more than one split, we aim to decrease the bias due to random splitting and test the repeatability. So, as an experiment environment, we created an experiment pool constructed by 10 random partitions for the test set and the training set generated by applying both undersampling and oversampling (SMOTE) (Fernández et al., 2018) techniques for addressing class imbalance issues. So, 80% of the data are used for training and validation (**Figure 2**). In each trial, both dimensional reduction and feature selection techniques were applied separately to solve the curse of the dimensionality problem for both undersampling and oversampling methodologies, and different machine learning techniques were evaluated with 10-fold cross-validation. Final models are tested against the unseen data separated at the beginning. All these processes were repeated 10 times for each data set in the experimental pool. Finally, the mean values of prediction parameters are calculated for the results reported in this study.

Each test/training subset listed in the experiment pool was trained and tested for different models by adding miRNA expressions, mRNA expressions, and methylation beta-values iteratively. Besides, to address the curse of dimensionality, we tried both dimensional reduction and feature selection techniques. Seven methods, namely, SVMs with linear, radial,

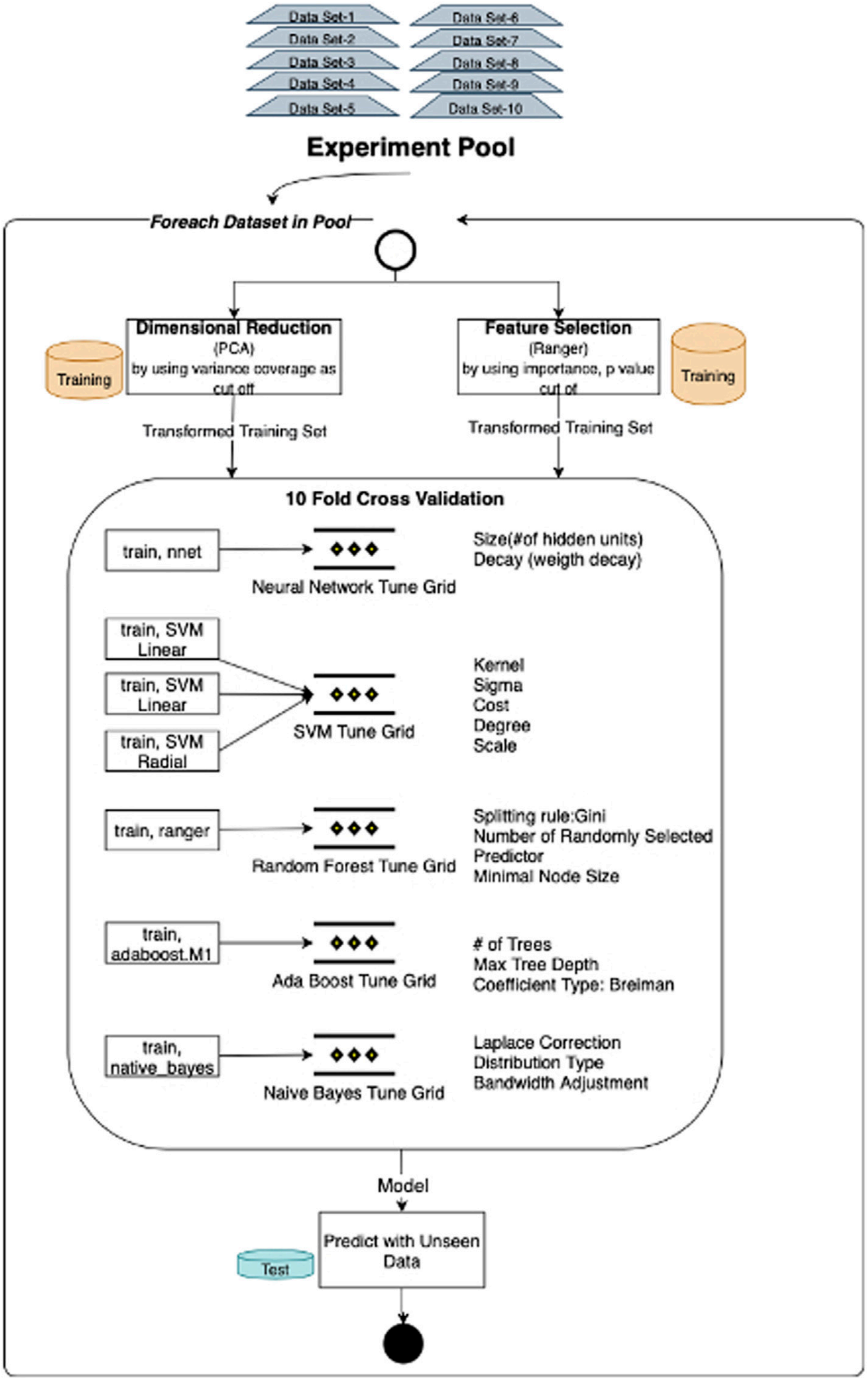


FIGURE 3 | Model training and testing process: experiment flow initiated by applying alternative dimensionality solutions, namely, PCA and feature selection. Through each experiment flow, models are trained with seven (SVMs with linear, radial, and polynomial kernels, neural networks, random forests, AdaBoost, and Naive Bayes) machine learning algorithms and tested with the same unseen data. Overall flow is repeated for each data subset in the experiment pool.

and polynomial kernels, neural networks, random forests, AdaBoost, and Naive Bayes, have been applied to generate and test the predictive model (**Figure 3**). Neural networks and support vector machines are frequent models that have been applied to similar classification models. But as we search the literature, we did not see any research which applied bagging, boosting, and probabilistic methods. So, we chose at least one representative of various classification algorithm categories, namely, artificial neural networks, bagging methods, boosting methods, and probabilistic models, one or more. Apart from support vector machines and neural networks, we included adaptive boosting, an ensemble method that composes a robust classifier from various weak classifiers, and random forest, which relies on bagging techniques to increase classification performance more than one decision tree. Apart from all these, Naïve Bayes also chooses an alternative since it is a fundamental model based on probabilistic techniques. The mean F-score and the mean *p*-value are evaluated as performance indicators for validation and test dataset classifications. Box plot distribution of classification scores is investigated for each dataset in the experimental pool. The best model for each category is made by comparing mean F-scores and mean *p*-values. If these results are the same for two or more best model candidates, we have reviewed the box plot of significance and sensitivity distributes.

This study follows the following coding mechanism to map the alternative scenarios of class imbalance and dimensionality solution techniques for each category. This annotation is used as the naming convention of the given result set in the following sections:

- a1: miRNA biomarkers modeled with feature selection and undersampling
- b1: miRNA biomarkers modeled with feature selection and SMOTE
- c1: miRNA biomarkers modeled with PCA and undersampling
- d1: miRNA biomarkers modeled with PCA and SMOTE
- a2: miRNA and mRNA biomarkers modeled with feature selection and undersampling
- b2: miRNA and mRNA biomarkers modeled with feature selection and SMOTE
- c2: miRNA and mRNA biomarkers modeled with PCA and undersampling
- d2: miRNA and mRNA biomarkers modeled with PCA and SMOTE
- a3: miRNA, mRNA, and methylation biomarkers modeled with feature selection and undersampling
- b3: miRNA, mRNA, and methylation biomarkers modeled with feature selection and SMOTE
- c3: miRNA, mRNA, and methylation biomarkers modeled with PCA and undersampling
- d3: miRNA, mRNA, and methylation biomarkers modeled with PCA and SMOTE

All preprocessing, training, validation, and testing are done using R studio using various R packages.

- Neural Network (package: nnet) (Ripley, 2002; Ripley and Venables, 2021)
- AdaBoost (package: adabag) (Alfaro et al., 2013; Alfaro et al., 2018)
- Random Forest (package: ranger) (Wright and Ziegler, 2017; (Wright et al., 2021)
- Naïve Bayes (package: naivebayes) (Majka and Majka, 2020)
- Support Vector Machine (package: kernlab) (Karatzoglou et al., 2004; Karatzoglou et al., 2016)
- Smote (smotefamily) (Siriseriwan, 2019a; Siriseriwan, 2019b)

During the collection and evaluation of the results, we followed a systematic cross-comparison technique. First, we collected the prediction scores for different classification models to find the best algorithm. Evaluation of the successors within each feature category identified the winner. Finally, model progress and the contributions of adding new feature categories are assessed based on these collected results. The illustration of this process is summarized in **Figure 4**.

The experiment is repeated for each subset in the data pool to find the prediction scores, and mean values were calculated. We have assessed the predictive algorithm using seven different machine learning models, including representatives of various classification algorithm categories, namely, artificial neural networks, bagging methods, boosting methods, and probabilistic models. We applied 10-fold cross-validation for each subset and calculated the mean F-score; the mean *p*-value was used to evaluate each category's best model. If the results are the same for two or more model candidates, we have reviewed the box plot of significance and sensitivity distributes to choose the one with low variance.

$$F\ Score = 2* \frac{(Precision*Recall)}{Precision + Recall} \tag{1}$$

As a final step, we performed functional and pathway enrichment analysis by using DAVID (Huang et al., 2007; Dennis et al., 2003). The KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers are compared for sets of "miRNA", "miRNA and mRNA", and "miRNA, mRNA, and methylation" to better understand the contributing factors behind the higher precision and consistency after including methylation data in the models.

RESULTS

In this study, we have evaluated the potential genetic biomarkers of melanoma metastasis. In addition, we developed multiple predictive models to predict the metastatic outcome by integrating miRNA, mRNA, and DNA methylation markers by using the TCGA melanoma dataset. This study's experimental strategy is composed of a 3-cycle evaluation, each of which targets different feature categories. In each cycle, the evaluation of different techniques to solve dimensionality and the class imbalance problem is applied. **Figure 5** summarizes the results of all evaluation techniques for each cycle.

At the first step of the initial cycle, we have implemented a predictive model (a1) with a microRNA biomarker model using

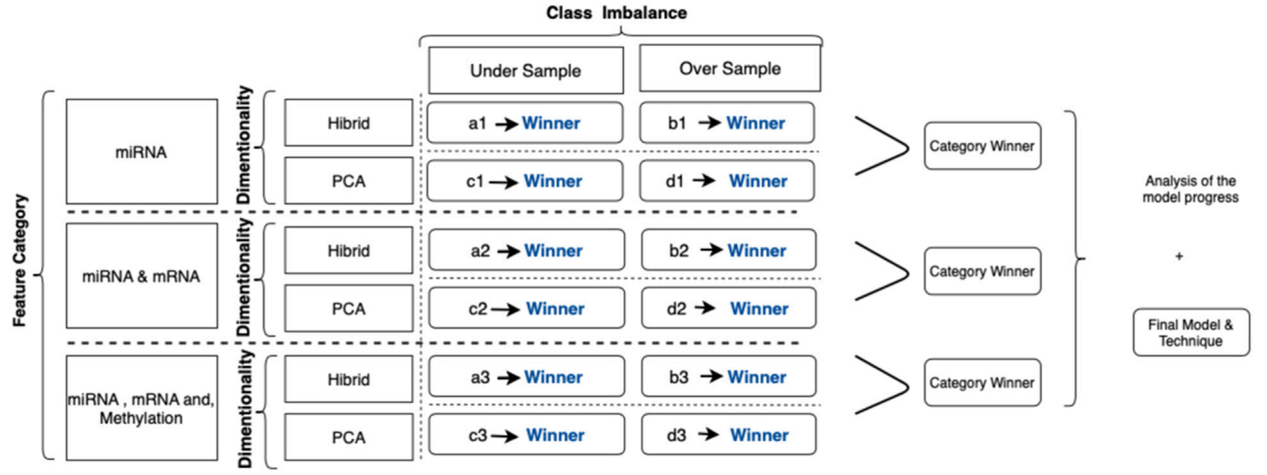


FIGURE 4 | Illustration for category-based analysis with techniques applied: each evaluation criterion is represented with a code. For example, (a1) represents the predictive models by using miRNA signatures with the hybrid method, that is, the random forest, to calculate feature importance undersampling for the class imbalance solution. Similarly, (d3) represents the outcomes of models applied to predict metastasis using significant miRNA, mRNA, and methylation biomarkers using PCA as a dimensional solution and SMOTE as a class imbalance solution.

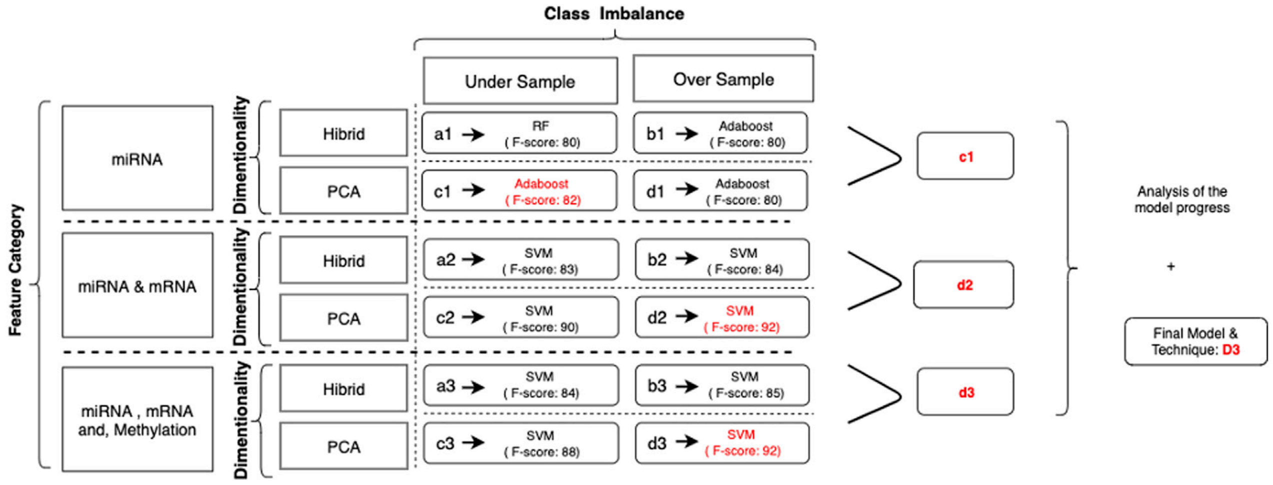


FIGURE 5 | Illustration for results of category-based analysis with techniques applied to solve significant issues: as a result of the evaluation process, (c1) is selected as the successor model for miRNA markers. When two markers, miRNA and mRNA, are combined, the winner is identified as (d2). In the final cycle, the merge of all biomarkers resulted in (d3) as the successor. Among all, (d3) was the winner to predict the metastatic outcome.

feature selection through importance (the hybrid model) and the class imbalance solution through undersampling. The predictive model with adaptive boosting (AdaBoost) demonstrates the best results among all the trials with the highest F-score and accuracy. Besides, the variance of the results for the different datasets in the experiment was also low compared to other models. Similarly, the random forest has the second-best results among all trials (an F-score of 80%). In the second scenario (b1), when we replace the class imbalance solution with SMOTE, the random forest demonstrates similar results, with an F-score of 79%. In parallel, adaptive boosting (AdaBoost) presents a comparable performance (an F-score of 80%) to that of the random forest model with a slightly higher score. In the third trial (c1), we have used undersampling and dimensional reduction with PCA. According to our results, adaptive boosting (AdaBoost) showed better scores (an F-score of 80%), but for this time, the SVM with the linear kernel (an F-score of 78%) was better than the random forest (an F-score of 72%), demonstrating the second-best results. Finally, we applied SMOTE to address the class imbalance issues (d1). The results were similar to those of the first trial; adaptive boosting showed the best results (F-score: 80%), and the random forest also had better results (F-score: 79%) than other models (**Figure 6**).

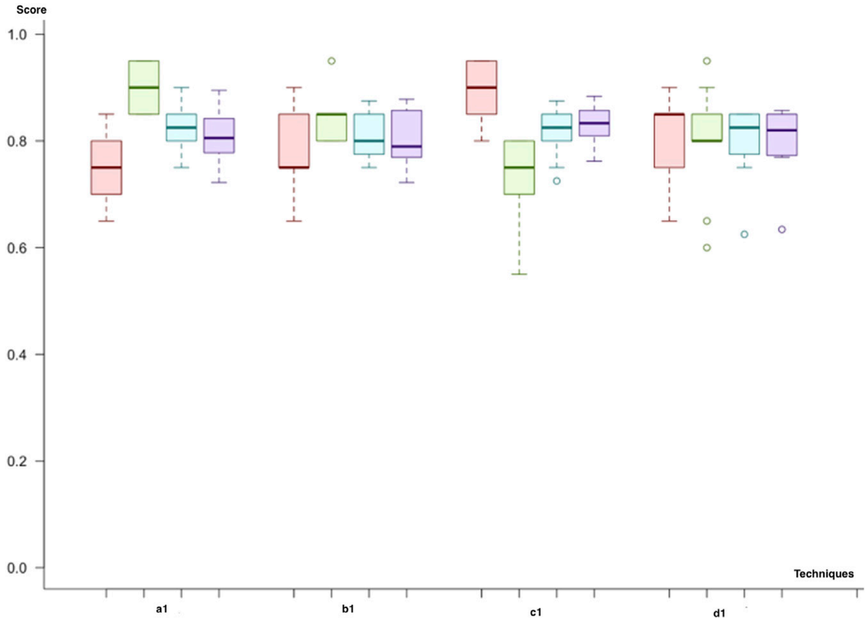


FIGURE 6 | Model comparison of techniques used for miRNA biomarkers (red, sensitivity; green, predictivity; blue, accuracy; purple, F-score): Category 1, which uses a hybrid model of feature selection and an AdaBoost classifier, has the best results among all scenarios.

As a result of the initial cycle, microRNA biomarkers predict the primary tumor's metastatic outcome with an F-score of almost 80%. In predictive models, by using miRNA markers, all workflows showed similar classification power. We selected (c1) the adaptive boosting with the PCA and undersampling as it results in the highest F-score. Both the random forest and adaptive boosting (AdaBoost) demonstrated better results in each workflow (**Figure 6**).

In the second cycle, we utilized both miRNA and mRNA as biomarkers. Like in the previous cycle, we first used (a2) feature selection through importance (the hybrid model) and the class imbalance solution through undersampling. When we compare the predictive models, the results were quite similar, with varying F-scores between 81% and 83%. However, the random forest produces the best results with regard to the mean F-score (83%) and the mean p-value (8.26×10^{-05}). The SVM with a polynomial kernel was the second-best model to predict the metastatic outcome, with the same F-score but a lower p-value (9.34×10^{-05}). As a second trial (b2), we have replaced the class imbalance solution with SMOTE. The results for each model, which vary in the range of 80–84% for the F-score, were quite similar. The neural network showed the best F-score (84%) and p-value (2.41×10^{-05}). The SVM with linear and polynomial kernels also had the same F-score (84%), and the neural network showed a higher significance. Adaptive boosting and the random forest demonstrate better results for the miRNA–mRNA cycle and predict the metastatic outcome with equal mean F-scores of 81%. In the third trial (c2), undersampling for class imbalance and dimensional reduction with PCA are applied. The SVM with the linear kernel was the best model with the highest F-score (90%). The neural network was the second-best model to predict metastasis with an F-score of 89%. Nevertheless, this time, adaptive boosting (F-score: 82%) and the random forest (F-score: 75%) are left behind. As the final trial (d2), we have applied SMOTE and dimensional reduction with PCA (d2). The neural network and the SVM with the linear kernel produced the best results compared to the rest with F-scores of 91 and 92%, respectively. On the other hand, adaptive boosting and the random forest showed high variance across different trials (**Figure 7**).

At the end of the second cycle, we saw that models using miRNA and mRNA marker winner models had F-scores ranging between 83 and 92%. The prediction scores for both boosting and bagging techniques were not as good as they were in the first cycle. Since the F-score for (d2), the SVM using PCA and SMOTE, has the highest scores, it is selected.

In the third cycle, we combined all miRNA, mRNA, and methylation biomarkers. Similar to previous cycles, we applied a combination of each class imbalance and dimensionality solution techniques. We decided on neural networks since model significance demonstrated improvement in our results. Firstly, all Neural network, SVM with linear and polynomial kernel predicts metastasis with an F-score of 83% by using under-sampling and feature selection through importance techniques (a3). Both SVM with radial kernel and random forest predict with similar F-scores (83%). So, the results of the prediction model were close to each other for this trial. However, the lowest variance across different trials was observed with SVM (linear kernel). In the second trial (b3), we have replaced the class imbalance solution technique with SMOTE. Both SVM with linear kernel and the polynomial kernel were the two best performing models with F-scores of 84% and 85%. In the

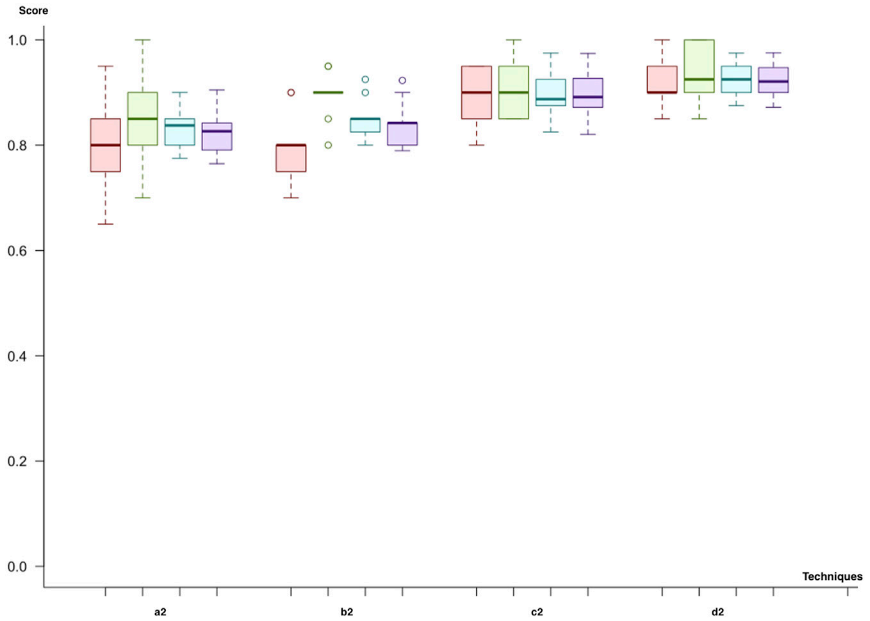


FIGURE 7 | Model comparison of techniques used for miRNA and mRNA biomarkers (red, sensitivity; green, predictivity; blue, accuracy; purple, F-score): the model listed in 4, which applies (d2), is selected as the successor model for the second cycle.

TABLE 1 | Summary for iterative progress on model precision scores.

	Best method	Tuning grid	Best tune hyperparameters	Validation	Test
miRNA	PCA and undersample AdaBoost	Max depth: [2:8] # of trees: [1:16]	Max depth: 6 Number of trees: 12 Co-efficiency of learning: Breiman	Accuracy: 86% F-score: 86%	Accuracy: 81% F-score: 82%
miRNA and mRNA	PCA SMOTE SVM (linear kernel)	Cost: $10^{(-4)}$ × (20:150))	Cost: 0.0025	Accuracy: 81% F-score: 0.82%	Accuracy: 93% F-score: 92%
miRNA, mRNA, and methylation	PCA SMOTE SVM (linear kernel)	Cost: $10^{(-4)}$ × (20:150))	Cost: 0.0027	Accuracy: 82% F-score: 83%	Accuracy: 93% F-score: 92%

The miRNA model applied by feature selection through importance (the hybrid model) and the class imbalance solution through undersampling is the method to be applied for prediction. For both "miRNA–mRNA" and the "miRNA–mRNA–methylation" triple model, principal component analysis for dimensionality and SMOTE for the class imbalance solution was the best method to increase predictive power and stability of the model.

third trial (b4), sampling and dimensional reduction with PCA are applied. SVM was the best model regardless of the selected kernel (F-score; 88%). Finally, when we applied SMOTE instead of under-sampling (d3), SVM with linear kernel demonstrated slightly higher scores (F-score 92%). In contrast, SVM with polynomial kernel and Neural network had an F-score of 91% and 90%. The best predictive model was SVM, trained by using dimensional reduction with PCA and SMOTE (d3). Like the second cycle, both SVM and Neural Network models resulted in better results in all trials. In addition, both under-sampling and oversampling techniques produced similar results (**Figure 8**).

As a result of all evaluations (see **Supplementary Material**), we came up with successors for each biomarker category (**Table 1**; **Figure 9**). First of all, the random forest with (a1) feature selection and undersampling achieved best results for miRNA markers (F-score = 81%, sensitivity = 75%, specificity = 90%, accuracy = 82%, and $p = 1.7 \times 10^{-4}$). In addition, the SVM (d2) with PCA and SMOTE was the most successful technique for a combination of miRNA and mRNA markers (F-score = 92%, sensitivity = 92%, specificity = 93.5%, accuracy = 93%, and $p = 1.0 \times 10^{-7}$). Finally, by using all miRNA, mRNA, and methylation markers (d3), the SVM reached the same results as the previous one, with higher consistency across different trials (F-score = 92%, sensitivity = 92%, specificity = 93%, accuracy = 92%, and $p = 1.05 \times 10^{-7}$) (**Figure 9**).

In the third model, 10 miRNA biomarkers, namely, hsa-mir-142, hsa-mir-29c, hsa-mir-3124, hsa-mir-3130, hsa-mir-326, hsa-mir-331, hsa-mir-4419b, hsa-mir-4444, hsa-mir-4474, hsa-mir-4491, hsa-mir-4523, hsa-of mir-625, and hsa-mir-766, are found to be upregulated and 1 miRNA, hsa-mir-203a, was found to be downregulated. Hence, 11 miRNA markers have been used as a biomarkers in our successor model to predict metastasis. In addition, 163 methylation and 1770 mRNA markers are selected in the final triple-biomarker model. All miRNA biomarkers and their target miRNA and methylation information in their target genes are presented in **Supplementary Tables S2, S3**.

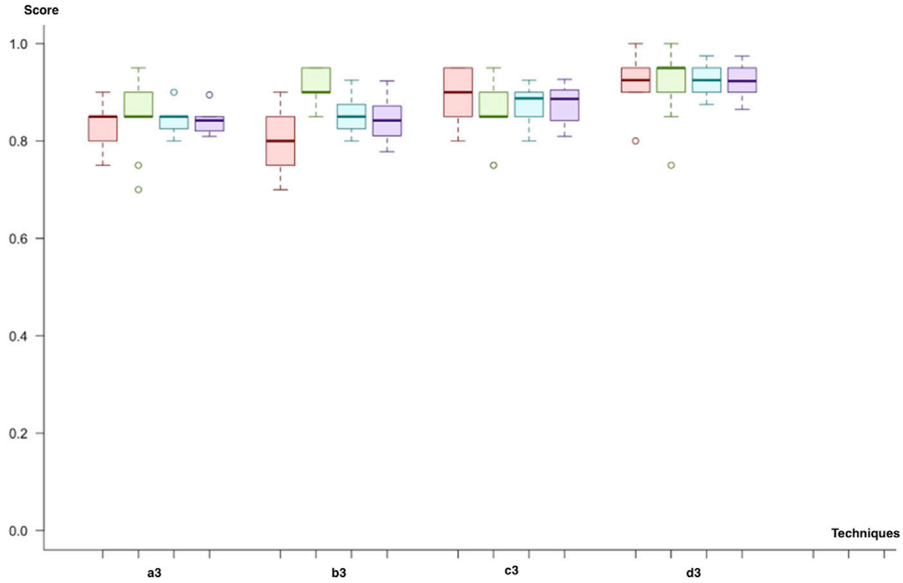


FIGURE 8 | Model comparison of techniques used for miRNA, mRNA, and methylation biomarkers (red, sensitivity; green, predictivity; blue, accuracy; purple, F-score). The model listed in 4, which applies (d3), is selected as the successor model for the final cycle.

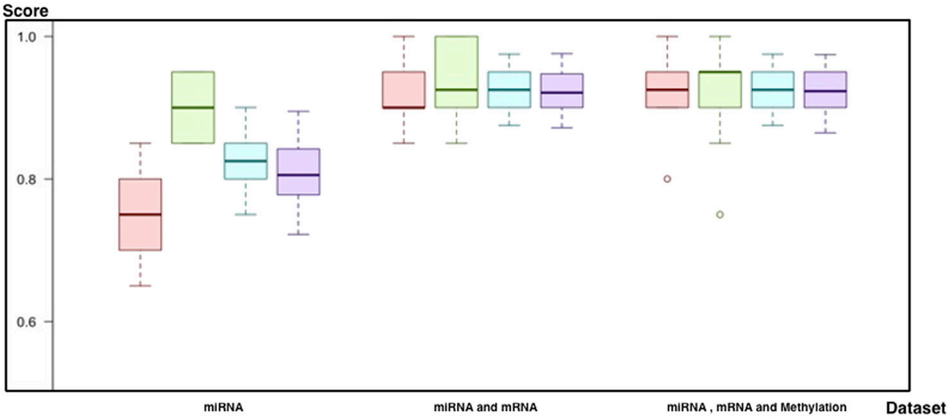


FIGURE 9 | Comparison of best models for each biomarker set (red, sensitivity; green, predictivity; blue, accuracy; purple, F-score): 1) the performance of the predictive model by using miRNA, 2) the performance of the predictive model by using miRNA and mRNA markers, and 3) the performance of the predictive model by using miRNA, mRNA, and methylation markers.

Evaluation of the overall results at the functional level is completed with an enrichment analysis. We used DAVID [(Huang et al., 2007; Dennis et al., 2003)] tools for biological interpretation of selected features used in the selected "miRNA and mRNA" classification and "miRNA, mRNA, and methylation" classification.

Using the functional enrichment analysis, the KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers are compared for "miRNA and mRNA" with "miRNA, mRNA, and methylation" to examine the reason for higher precision and consistency of addition of methylation. In the model with methylation markers, the significance of the osteoclast, Rap1 signaling pathway, and chemokine signaling pathways increased (**Figure 10**). Osteoclast differentiation also appealed within the top 15 pathways when all 3 biomarker categories are combined. In addition, the Rap1 signaling pathway and chemokine signaling were listed in the top 3 among the most significant pathways (**Table 2**).

DISCUSSIONS

Melanoma can be distinguished with visual assessment or through a short screening. Although there is an opportunity for a cure when detected in the early stages, treatment is challenging in later stages. Likewise, metastasis is an undesired

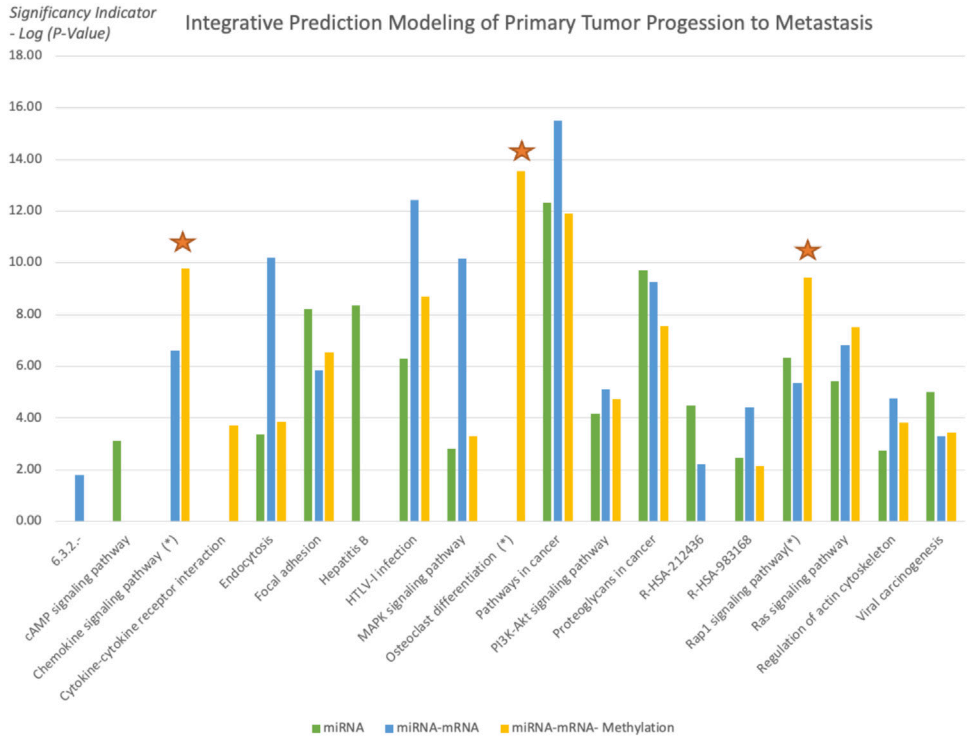


FIGURE 10 | Significant pathways functionally enriched in all three feature sets. As the new biomarker set is added, the significance of the pathways is evaluated. Osteoclast, Rap1 signaling pathway, and chemokine signaling pathways showed a significant increase in the third model.

TABLE 2 | Comparison of the top 15 pathways of different biomarker sets.

	***p*-value**		
	miRNA	**miRNA–mRNA**	**miRNA–mRNA–methylation**
6.3.2.-	—	1.60×10^{-02}	—
cAMP signaling pathway	7.40×10^{-04}	—	—
Chemokine signaling pathway (*)	—	2.40×10^{-07}	$1.60\ 10^{-10}$
Cytokine–cytokine receptor interaction	—	—	$1.90\ 10^{-04}$
Endocytosis	4.30×10^{-04}	6.30×10^{-11}	$1.40\ 10^{-04}$
Focal adhesion	6.10×10^{-09}	$1.40\ 10^{-06}$	$2.80\ 10^{-07}$
Hepatitis B	4.40×10^{-09}	—	—
HTLV-I infection	5.10×10^{-07}	3.60×10^{-13}	2.00×10^{-09}
MAPK signaling pathway	1.60×10^{-03}	6.70×10^{-11}	4.90×10^{-04}
Osteoclast differentiation (*)	—	—	2.90×10^{-14}
Pathways in cancer	4.60×10^{-13}	3.10×10^{-16}	1.20×10^{-12}
PI3K-Akt signaling pathway	7.00×10^{-05}	8.00×10^{-06}	1.90×10^{-05}
Proteoglycans in cancer	2.00×10^{-10}	5.70×10^{-10}	2.70×10^{-08}
R-HSA-212436	3.40×10^{-05}	6.00×10^{-03}	—
R-HSA-983168	3.60×10^{-03}	3.80×10^{-05}	7.30×10^{-03}
Rap1 signaling pathway (*)	4.80×10^{-07}	4.30×10^{-06}	3.70×10^{-10}
Ras signaling pathway	3.80×10^{-06}	1.50×10^{-07}	3.00×10^{-08}
Regulation of actin cytoskeleton	1.80×10^{-03}	1.70×10^{-05}	1.50×10^{-04}
Viral carcinogenesis	9.70×10^{-06}	5.10×10^{-04}	3.80×10^{-04}

**p-values of the osteoclast, Rap one signaling pathway, and chemokine signaling pathways gradually increased after adding a new biomarker set. In addition, the Rap1 signaling pathway and chemokine signaling were listed among the top three pathways with increasing significance with osteoclast differentiation. Other pathways with increasing significance, such as cytokine–cytokine receptor interaction and the Ras signaling pathway, are also observed.*

outcome in such cases, and differential diagnosis is crucial for the treatment decision. So, the opportunity for the diagnosis of metastatic melanomas in earlier stages may support therapeutic decisions and advice for more frequent and in-depth screening, providing a higher chance for cure or prevention of further metastatic progress.

This study shows that miRNA plays an essential role in the metastatic progression of primary melanoma and predicts metastasis outcomes with high accuracy. miRNA biomarkers anticipated metastatic results with an F-score of 82%. Expansion of mRNA markers upon miRNA reached an F-score of 92%. The ultimate model, which includes DNA methylation, results in a comparative F-score of 92% but delivered a steady model with low variation over different trials. Moreover, the integrated evaluation of miRNA with mRNA and methylation biomarkers increases the model's predictive power. Another remarkable finding in this study is that the boosting and bagging model's performance was better for miRNA signatures. However, when we added new mRNA and DNA methylation, we got higher prediction scores for neural networks and support vector machine classifiers.

One limitation of the study was the data imbalance and small sample size. We validated and tested our models in a restricted data size since we could not access additional datasets on the GEO or CGC, combining all three markers at the time of the study. We utilized oversampling techniques and ran the overall process multiple times to reduce the bias to address this limitation. Additionally, we were able to compare various machine learning models as they were appropriate for the data size in the study. However, we realize that deep learning methods would be competitive with these techniques. Therefore, repetition of the study with a balanced or more extensive dataset in the future can further validate the biomarkers reported here.

In machine learning studies, undersampling techniques are also used to deal with class imbalance issues. So we performed oversampling and undersampling methods and evaluated their outcomes. The SMOTE, a synthetic minority oversampling method based on the k-nearest neighbors, has been tested with different k values between 3 and 6, and the final k was chosen as 3. We used the 1:2 ratio for oversampling of the minority class. Under the given circumstances, we generated similar results for both undersampling and oversampling. Overall, our results present satisfactory evidence that the synthetic minority oversampling technique can also be applicable for prediction studies for genomics data.

As our model is based on the differences between primary and metastatic melanomas, the markers identified here can be used for differential diagnosis. We believe that it will become possible to predict melanomas with metastatic potential (prediction of prognosis). In those cases, several actions can be taken in the clinic, such as intensive scanning for metastasis or frequent follow-ups with patients. In the future, patients with higher risk can be offered prevention from metastasis with gene therapies based on emerging technologies like miRNA therapies or gene editing.

In this study, we focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation. We used the TCGA melanoma dataset to predict metastatic melanoma samples by assessing a set of predictive models. Differentially expressed miRNA, mRNA, and methylation signatures are used as biomarkers throughout the study. The highest performing models' selected biomarkers are further analyzed for the biological interpretation of functional enrichment and determining regulatory networks. So we focused on gradually including new feature sets. To reveal our evaluation pattern for including new biomarker sets, we performed functional enrichment analysis. The functional enrichment of the KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers are compared for sets of "miRNA", "miRNA and mRNA", and "miRNA, mRNA, and methylation", and we tried to search for the reason behind the higher precision and consistency achieved after addition of methylation. The osteoclast, Rap1 signaling pathway, and chemokine signaling pathways significantly increased and listed the top 15 pathways when all 3 biomarker sets are used for modeling. So the combined model populates selected biomarkers on the metastasis-associated pathways of melanoma.

Osteoclasts are multinucleated cells responsible for bone resorption. Molecular pathways involved in osteoclast proliferation, differentiation, and survival are essential players in bone metastasis. Osteoclast differentiation is a systemic pathway that controls bone renovation. Since the main metastasis sites for melanoma cancer include bone, the liver, the lung, and skin/muscle (Wright et al., 2021), functional enrichment of osteoclast-related pathways within top-level pathways is a supporting finding for our study design.

Ras-associated protein-1 (Rap1) is an important regulator for basic cell functions such as cellular migration and polarization. This pathway is an important factor for tumor metastasis, so such an increase in the significance level is also critical for the metastatic outcome (Zhang et al., 2017).

Chemokines are involved in controlling the migration of cells during normal processes of tissue maintenance or development. The chemokine-receptor system plays critical roles in various physiological processes, including immune homeostasis, inflammatory responses, and cancer progression. Chemokines have essential roles in tumor progression and are involved in the growth of many cancers and metastasis (Sarvaiya et al., 2013).

Since the initial discovery of the relationship between cancer and miRNA signatures, many studies have shown that miRNA has a critical role in the regulation of genes, and thus, has a critical role in tumorigenesis. Today, many techniques for the early detection of and diagnosis of tumors are available. Still, when invasive procedures are required for diagnosis or treatment, it is vital to know the tumor's metastatic potential to estimate the risks vs. the benefits of the procedure. Also, in the later stages of tumor development, any information about the metastatic status of late-stage tumors is required for deciding between therapy choices. Hence, the miRNAs reported in this study can be candidates for therapeutic targets of melanoma metastasis.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study can be found in the TCGA (The Cancer Genome Atlas Network, 2015). The names of the datasets and search methods can be found in the article. Additional information can be found in the **Supplementary Material**.

AUTHORS' CONTRIBUTIONS

AK collected and processed the TCGA data. AK constructed the models and finalized the comparisons. YA coordinated the research. AK and YA commented on the results and wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The results shown in this study are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.637355/full#supplementary-material

REFERENCES

Alfaro, E., Gáamez, M., and García, N. (2013). Adabag: An R Package for Classification with Boosting and Bagging. *J. Stat. Softw.* 54 (2), 1–35. doi:10.18637/jss.v054.i02

Alfaro, E., Gamez, M., and Garcia, N. (2018). CRAN - Package Adabag," CRAN R Project. Online. Available: https://cran.r-project.org/web/packages/adabag/index.html (Accessed Jun 23, 2021).

American Cancer Society (2016). *European Commission Melanoma Skin Cancer*. Atlanta.

Burton, M., Thomassen, M., Tan, Q., and Kruse, T. A. (2012). Prediction of Breast Cancer Metastasis by Gene Expression Profiles: A Comparison of Metagenes and Single Genes. *Cancer Inform.* 11, 193–217. doi:10.4137/cin.s10375

Cancer Research UK (2017). Melanoma Skin Cancer Incidence Statistics. Online. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer#:~:text=Melanoma skin cancer incidence,new cancer cases (Accessed 03 Jun, 2021).

Carter, R. L. (1974). The Spread of Tumours in the Human Body. *J. Clin. Pathol.* 27 (5), 432–433. doi:10.1136/jcp.27.5.432-c

Chang, J. W., Yi, C. A., Son, D.-S., Choi, N., Lee, J., Kim, H. K., et al. (2008). Prediction of Lymph Node Metastasis Using the Combined Criteria of Helical CT and mRNA Expression Profiling for Non-small Cell Lung Cancer. *Lung Cancer* 60 (2), 264–270. doi:10.1016/j.lungcan.2007.09.026

Chen, L. L., Blumm, N., Christakis, N. A., Barabási, A.-L., and Deisboeck, T. S. (2009). Cancer Metastasis Networks and the Prediction of Progression Patterns. *Br. J. Cancer* 101 (5), 749–758. doi:10.1038/sj.bjc.6605214

Chen, X., Guo, W., Xu, X.-J., Su, F., Wang, Y., Zhang, Y., et al. (2017). Melanoma Long Non-coding RNA Signature Predicts Prognostic Survival and Directs Clinical Risk-specific Treatments. *J. Dermatol. Sci.* 85 (3), 226–234. doi:10.1016/J.JDERMSCI.2016.12.006

Craig, J. M., and Wong, N. C. (2011). *Epigenetics: A Reference Manual*. Victoria, Australia: Caister Academic Press.

Damsky, W. E., Jr., Rosenbaum, L. E., and Bosenberg, M. (2010). Decoding Melanoma Metastasis. *Cancers (Basel)* 3 (1), 126–163. Mar. 2011. doi:10.3390/cancers3010126

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4 (5), P3. doi:10.1186/gb-2003-4-5-p3

Fernández, A., García, S., Herrera, F., and Chawla, N. V. SMOTE: Synthetic Minority Over-sampling Technique, 01-Apr-2018. [Online]. *J. Artif. Intelligence Res.* 16 (No 1). Available: https://dl.acm.org/doi/10.5555/1622407.1622416 (Accessed Dec 26, 2019).

Gonzalo, S. (2020). Epigenetic Alterations in Aging. *J. Appl. Physiol.* 109 (2), 586–597. doi:10.1152/japplphysiol.00238.2010

Guo, J., Yang, M., Zhang, W., Lu, H., and Li, J. (2015). A Panel of miRNAs as Prognostic Indicators for Clinical Outcome of Skin Cutaneous Melanoma. *Int. J. Clin. Exp. Med.* 9 (1), 28–39.

Harris, C. C. (1991). Molecular Basis of Multistage Carcinogenesis. *Princess Takamatsu Symp.* 22, 3–19.

Hayes, J., Peruzzi, P. P., and Lawler, S. (2014). MicroRNAs in Cancer: Biomarkers, Functions and Therapy. *Trends Mol. Med.* 20 (8), 460–469. doi:10.1016/j.molmed.2014.06.005

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., et al. (2007). DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists. *Nucleic Acids Res.* 35 (Suppl. 2), W169–W175. doi:10.1093/nar/gkm415

Institute, N. C. (2020). *The Cancer Genome Atlas Program - National Cancer Institute*. Online. Rockville, MD: National Institute of Health. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (Accessed 22 Jun, 2021).

Jansson, M. D., and Lund, A. H. (2012). MicroRNA and Cancer*Mol. Oncol.*, 6 (6), 590–610. doi:10.1016/j.molonc.2012.09.006

Kan, T., Shimada, Y., Sato, F., Ito, T., Kondo, K., Watanabe, G., et al. (2004). Prediction of Lymph Node Metastasis with Use of Artificial Neural Networks Based on Gene Expression Profiles in Esophageal Squamous Cell Carcinoma. *Ann. Surg. Oncol.* 11 (12), 1070–1078. doi:10.1245/ASO.2004.03.007

Karatzoglou, A., Hornik, K., Smola, A., and Zeileis, A. (2004)Kernlab - an S4 Package for Kernel Methods in R. *J. Stat. Softw.* 11 (1), 1–20. doi:10.18637/jss.v011.i09

Karatzoglou, A., Smola, A., and Hornik, K. 2016 *Kernlab: Kernel-Based Machine Learning Lab. R Package Kernlab. Version 0.9-29*. Comprehensive R Archive Network (CRAN), 12. (Accessed 22 June 2021).

Kasinski, A. L., and Slack, F. J. (2011). MicroRNAs en route to the clinic: Progress in validating and targeting microRNAs for cancer therapy. *Nat. Rev. Cancer* 11 (12), 849–864. doi:10.1038/nrc3166

Kinslechner, K., Schütz, B., Pistek, M., Rapolter, P., Weitzenböck, H., Hundsberger, H., et al. (2019). Loss of SR-BI Down-Regulates MITF and Suppresses Extracellular Vesicle Release in Human Melanoma. *Ijms* 20 (5), 1063. doi:10.3390/ijms20051063

Ma, L., and Weinberg, R. A. (2008). Micromanagers of Malignancy: Role of microRNAs in Regulating Metastasis. *Trends Genet.* 24 (9), 448–456. doi:10.1016/j.tig.2008.06.004

Ma, X., He, Z., Li, L., Yang, D., and Liu, G. (2017). Expression Profiles Analysis of Long Non-coding RNAs Identified Novel lncRNA Biomarkers with Predictive Value in Outcome of Cutaneous Melanoma. *Oncotarget* 8 (44), 77761–77770. doi:10.18632/oncotarget.20780

Majka, M., and Majka, M. (2020). *CRAN - Package Na*. Online. Available: https://cran.r-project.org/web/packages/naivebayes/index.html (Accessed Jun 23, 2021). Ivebayes," CRAN R Project.

Mancuso, F., Lage, S., Rasero, J., Díaz-Ramón, J. L., Apraiz, A., Pérez-Yarza, G., et al. (2020). Serum Markers Improve Current Prediction of Metastasis Development in Early-stage Melanoma Patients: a Machine Learning-based Study. *Mol. Oncol.* 14 (8), 1705–1718. doi:10.1002/1878-0261.12732

McAnena, P., Tanriverdi, K., Curran, C., Gilligan, K., Freedman, J. E., Brown, J. A. L., et al. (2019). Circulating microRNAs miR-331 and miR-195 Differentiate Local Luminal a from Metastatic Breast Cancer. *BMC Cancer* 19 (1), 436. doi:10.1186/s12885-019-5636-y

Melchers, L., Clausen, M., Mastik, M., Slagter-Menkema, L., Van Der Wal, J., Wisman, G., et al. (2015). Identification of Methylation Markers for the Prediction of Nodal Metastasis in Oral and Oropharyngeal Squamous Cell Carcinoma. *Epigenetics* 10 (9), 850–860. doi:10.1080/15592294.2015.1075689

Moriya, Y., Iyoda, A., Kasai, Y., Sugimoto, T., Hashida, J., Nimura, Y., et al. (2009). Prediction of Lymph Node Metastasis by Gene Expression Profiling in Patients with Primary Resected Lung Cancer. *Lung Cancer* 64, 86–91. doi:10.1016/j.lungcan.2008.06.022

Oppenheimer, S. B. (1982). *Cancer: A Biological and Clinical Introduction*. Boston: Allyn & Bacon.

Oppenheimer, S. B. (2006). Cellular Basis of Cancer Metastasis: A Review of Fundamentals and New Advances. *Acta Histochem.* 108 (5), 327–334. doi:10.1016/j.acthis.2006.03.008

Rickman, D. S., Millon, R., De Reynies, A., Thomas, E., Wasylyk, C., Muller, D., et al. (2008). Prediction of Future Metastasis and Molecular Characterization of Head and Neck Squamous-Cell Carcinoma Based on Transcriptome and Genome Analysis by Microarrays. *Oncogene* 27, 6607–6622. doi:10.1038/onc.2008.251

Venables, W. N., and Ripley, B. (2002). *Feed-Forward Neural Networks and Multinomial Log-Linear Models, Modern Applied Statistics with S, Fourth edition. Springer, New York*.

Ripley, B., and Venables, W. (2021). CRAN - Package Nnet," CRAN R Project. Online. Available: https://cran.r-project.org/web/packages/nnet/index.html (Accessed Jun 23, 2021).

Roessler, S., Jia, H.-L., Budhu, A., Forgues, M., Ye, Q.-H., Lee, J.-S., et al. (2010). A Unique Metastasis Gene Signature Enables Prediction of Tumor Relapse in Early-Stage Hepatocellular Carcinoma Patients. *Cancer Res.* 70 (24), 10202–10212. doi:10.1158/0008-5472.CAN-10-2607

Sarvaiya, P. J., Guo, D., Ulasov, I., Gabikian, P., and Lesniak, M. S. (2013). Chemokines in Tumor Progression and Metastasis. *Oncotarget* 4 (12), 2171–2185. doi:10.18632/oncotarget.1426

Selitsky, S. R., Mose, L. E., Smith, C. C., Chai, S., Hoadley, K. A., Dittmer, D. P., et al. (2019). Prognostic Value of B Cells in Cutaneous Melanoma. *Genome Med.* 11 (1). doi:10.1186/S13073-019-0647-5

Shalaby, T., Fiaschetti, G., Baumgartner, M., and Grotzer, M. (2014). MicroRNA Signatures as Biomarkers and Therapeutic Target for CNS Embryonal Tumors: the Pros and the Cons. *Ijms* 15 (11), 21554–21586. doi:10.3390/ijms151121554

Shen, J., Stass, S. A., and Jiang, F. (2013). MicroRNAs as Potential Biomarkers in Human Solid Tumors. *Cancer Lett.* 329 (2), 125–136. Feb-2013. doi:10.1016/j.canlet.2012.11.001

Siriseriwan, W. (2019a). *Package "smotefamily" Title A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE*. Available: https://cran.r-project.org/package=smotefamily (Accessed June 23, 2021).

Siriseriwan, W. (2019b). CRAN - Package Smotefamily," CRAN R Project. Online. Available: https://cran.r-project.org/web/packages/smotefamily/index.html (Accessed Jun 23, 2021).

Souza, M. F. d., Kuasne, H., Barros-Filho, M. D. C., Cilião, H. L., Marchi, F. A., Fuganti, P. E., et al. (2017). Circulating mRNAs and miRNAs as Candidate Markers for the Diagnosis and Prognosis of Prostate Cancer. *PLoS ONE* 12 (9), e0184094. doi:10.1371/journal.pone.0184094

Stahlhut, C., and Slack, F. J. (2013). MicroRNAs and the Cancer Phenotype: Profiling, Signatures and Clinical Implications. *Genome Med.* 5 (12), 111. doi:10.1186/gm516

The Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. *Cell* 161 (7), 1681–1696. doi:10.1016/j.cell.2015.05.044

Tonini, T., Rossi, F., and Claudio, P. P. (2003). Molecular Basis of Angiogenesis and Cancer. *Oncogene* 22 (43), 6549–6556. doi:10.1038/sj.onc.1206816

United States Cancer Statistics CDC United States Cancer Statistics Data Visualizations. Online. Available: https://www.cdc.gov/cancer/uscs/dataviz/index.htm (Accessed 03 Jun, 2021).

Valentini, V., Zelli, V., Gaggiano, E., Silvestri, V., Rizzolo, P., Bucalo, A., et al. (2019). MiRNAs as Potential Prognostic Biomarkers for Metastasis in Thin and Thick Primary Cutaneous Melanomas. *Anticancer Res.* 39 (8), 4085–4093. doi:10.21873/anticanres.13566

Wang, Y., Liu, G., Ren, L., Wang, K., and Liu, A. (2019). Long Non-coding RNA TUG1 Recruits miR-29c-3p from its T-arget G-ene RGS1 to P-romote P-roliferation and M-etastasis of M-elanoma C-ells. *Int. J. Oncol.* 54 (4), 1317–1326. doi:10.3892/ijo.2019.4699

Watanabe, T., Kobunai, T., Yamamoto, Y., Kanazawa, T., Konishi, T., Tanaka, T., et al. (2010). Prediction of Liver Metastasis after Colorectal Cancer Using Reverse Transcription-Polymerase Chain Reaction Analysis of 10 Genes. *Eur. J. Cancer* 46 (11), 2119–2126. doi:10.1016/j.ejca.2010.04.019

Wei, C.-Y., Wang, L., Zhu, M.-X., Deng, X.-Y., Wang, D.-H., Zhang, S.-M., et al. (2019). TRIM44 Activates the AKT/mTOR Signal Pathway to Induce Melanoma Progression by Stabilizing TLR4. *J. Exp. Clin. Cancer Res.* 38 (1), 137. doi:10.1186/s13046-019-1138-7

World Cancer Research Fund Skin Cancer | World Cancer Research Fund International. Online. Available: https://www.wcrf.org/dietandcancer/skin-cancer/ (Accessed 03 Jun, 2021).

Wright, M. N., Wager, S., and Probst, P. (2021). CRAN - Package ranger," CRAN R Project. Online. Available: https://cran.r-project.org/web/packages/ranger/index.html (Accessed Jun 23, 2021).

Wright, M. N., and Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* 77 (1). doi:10.18637/jss.v077.i01

Xiong, J., Bing, Z., and Guo, S. (20192019). Observed Survival Interval: A Supplement to TCGA Pan-Cancer Clinical Data Resource. *Cancers* 11 (3), 280. doi:10.3390/cancers11030280

Xue, D., Cheng, P., Jiang, J., Ren, Y., Wu, D., and Chen, W. (2020). Systemic Analysis of the Prognosis-Related RNA Alternative Splicing Signals in Melanoma. *Med. Sci. Monit.* 26, e921133. Mar. 2020. doi:10.12659/MSM.921133

Yang, S., Xu, J., and Zeng, X. (2018). A Six-Long Non-coding RNA Signature Predicts Prognosis in Melanoma Patients. *Int. J. Oncol.* 52 (4), 1178–1188. doi:10.3892/ijo.2018.4268

Zemmour, C., Bertucci, F., Finetti, P., Chetrit, B., Birnbaum, D., Filleron, T., et al. (2015). Prediction of Early Breast Cancer Metastasis from Dna Microarray Data Using High-Dimensional Cox Regression Models. *Cancer Inform.* 14s2, CIN.S17284–138. doi:10.4137/CIN.S17284

Zhang, F. F., Cardarelli, R., Carroll, J., Zhang, S., Fulda, K. G., Gonzalez, K., et al. (2011). Physical Activity and Global Genomic DNA Methylation in a Cancer-free Population. *Epigenetics* 6 (3), 293–299. doi:10.4161/epi.6.3.14378

Zhang, H., Kalirai, H., Acha-Sagredo, A., Yang, X., Zheng, Y., and Coupland, S. E. (2020). Piloting a Deep Learning Model for Predicting Nuclear BAP1 Immunohistochemical Expression of Uveal Melanoma from Hematoxylin-And-Eosin Sections. *Transl. Vis. Sci. Technol.* 9 (2), 50–13. doi:10.1167/tvst.9.2.50

Zhang, Y. L., Wang, R. C., Cheng, K., Ring, B. Z., and Su, L. (2017). Roles of Rap1 Signaling in Tumor Cell Migration and Invasion. *Cancer Biol. Med.* 14 (1), 90–99. doi:10.20892/j.issn.2095-3941.2016.0086

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 October 2021
doi: 10.3389/fmolb.2021.666705

Performance Assessment of the Network Reconstruction Approaches on Various Interactomes

M. Kaan Arici[1,2] *and Nurcan Tuncbag*[3,4]*

[1]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey, [2]Foot and Mouth Diseases Institute, Ministry of Agriculture and Forestry, Ankara, Turkey, [3]Chemical and Biological Engineering, College of Engineering, Koc University, Istanbul, Turkey, [4]School of Medicine, Koc University, Istanbul, Turkey

Beyond the list of molecules, there is a necessity to collectively consider multiple sets of omic data and to reconstruct the connections between the molecules. Especially, pathway reconstruction is crucial to understanding disease biology because abnormal cellular signaling may be pathological. The main challenge is how to integrate the data together in an accurate way. In this study, we aim to comparatively analyze the performance of a set of network reconstruction algorithms on multiple reference interactomes. We first explored several human protein interactomes, including PathwayCommons, OmniPath, HIPPIE, iRefWeb, STRING, and ConsensusPathDB. The comparison is based on the coverage of each interactome in terms of cancer driver proteins, structural information of protein interactions, and the bias toward well-studied proteins. We next used these interactomes to evaluate the performance of network reconstruction algorithms including all-pair shortest path, heat diffusion with flux, personalized PageRank with flux, and prize-collecting Steiner forest (PCSF) approaches. Each approach has its own merits and weaknesses. Among them, PCSF had the most balanced performance in terms of precision and recall scores when 28 pathways from NetPath were reconstructed using the listed algorithms. Additionally, the reference interactome affects the performance of the network reconstruction approaches. The coverage and disease- or tissue-specificity of each interactome may vary, which may result in differences in the reconstructed networks.

Keywords: protein-protein interactions, interactome, network reconstruction, heat diffusion, personalized PageRank, prize-collecting Steiner forest, pathway reconstruction

OPEN ACCESS

Edited by:
Arzucan Ozgur,
Boğaziçi University, Turkey

Reviewed by:
Hilal Kazan,
Antalya Bilim University, Turkey
Cesim Erten,
Antalya Bilim University, Turkey
Eileen Marie Hanna,
Lebanese American University,
Lebanon

****Correspondence:***
Nurcan Tuncbag
ntuncbag@ku.edu.tr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *10 February 2021*
Accepted: *14 July 2021*
Published: *05 October 2021*

Citation:
Arici MK and Tuncbag N (2021) Performance Assessment of the Network Reconstruction Approaches on Various Interactomes. Front. Mol. Biosci. 8:666705. doi: 10.3389/fmolb.2021.666705

INTRODUCTION

Computational approaches improve our understanding about the mechanisms of perturbations, effects of drugs, and functions of genes in the biological system by interpreting multiple "omic" data and reducing their complexity (Liu et al., 2020; Paananen and Fortino, 2020). Integrative network analysis approaches are used to interpret the complex interactions between "omic" entities as a whole beyond the list of molecules. The impact of an alteration in any omic entity, for example, upregulated

Abbreviations: APSP, all-pairs shortest paths; CDGs, cancer driver genes; FPR, false positive rate; HD, heat diffusion; HDF, heat kernel diffusion with flux; MCC, Matthew's correlation coefficient; MI, MINT inspired; PCA, principal component analysis; PCSF, prize-collecting Steiner forest; PPR, personalized PageRank; PRF, personalized PageRank with flux or downregulated genes or mutated or phosphorylated proteins, may not be local; rather, it diffuses to the distant sites of the interactome.

Many pathway databases cataloged the molecular interactions. Each database explains interactions *via* different approaches. KEGG (Kanehisa et al., 2017) provides annotated pathways, while Reactome (Jassal et al., 2020) gives detailed information on components and the reactions. Additionally, integrated interactomes such as HIPPIE, ConsensusPathDB, and STRING combine multiple resources to come up with a weighted interactome. There are several scoring schemas to measure the reliability of interactions such as MI-score and IntScore. These methods combine different weights including the number of publications, detection method, or network topology (Turinsky et al., 2011; Kamburov et al., 2012; Kamburov et al., 2013; Alanis-Lobato et al., 2017; Szklarczyk et al., 2019). Although the combination of multiple resources improves the quality of the interactomes, it still does not completely solve the bias toward well-studied proteins or the artifacts from high-throughput experiments (Žitnik et al., 2013; Caraus et al., 2015; Skinniderid et al., 2018; Vitali et al., 2018). Besides the false positives, interactomes are not complete and have false negatives which are the undetected interactions. To complete the missing parts in the interactome and to detect spurious interactions, several prediction approaches have been employed using network topology (Alkan and Erten, 2017), link prediction, protein structures (Singh et al., 2006; Tuncbag et al., 2012; Mosca et al., 2014; Segura et al., 2015; Yerneni et al., 2018; Ietswaart et al., 2021), or additional data such as gene expression (Cannistraci et al., 2013; Lei and Ruan, 2013; Hulovatyy et al., 2014; Szklarczyk et al., 2021). For example, Interactome3D uses the structural knowledge in PDB and homology-based prediction to construct a highly accurate interactome (Mosca et al., 2013). The main limitation of proteome-wide structural interactome construction is the number of structurally resolved protein complexes.

Network reconstruction approaches aim to transform the list of seed genes/proteins into their interactome-wide impact based on the topological proximity. Steiner trees/forests, statistical models, and network propagation with random walk or heat diffusion systems have been frequently used in omics data integration with the molecular interactions (Leiserson et al., 2015; Cowen et al., 2017; SeahSen et al., 2017) or identifying disease-associated pathways, subnetworks, or modules (Paull et al., 2013; Kim et al., 2015; Silverbush et al., 2019). These approaches construct context-specific subnetworks under a certain condition such as disease association or for revealing the impact of an external stimulus such as drug treatment or pathogen infection (Braunstein et al., 2019; Tabei et al., 2019). Recently, DriveWays (Baali et al., 2020), MEXCOwalk (Ahmed et al., 2020), iCell (Malod-Dognin et al., 2019), ModulOmics (Silverbush et al., 2019), and Omics Integrator (Tuncbag et al., 2016b) predicted the cancer driver modules. MEXCOwalk implements a random walk on the reference interactome by using mutation frequencies and their mutual exclusivity for the identification of the cancer driver modules. ModulOmics uses protein–protein, regulatory, and gene co-expression networks together with mutual exclusivity of mutations to identify highly functional driver modules. Omics Integrator solves the prize-collecting Steiner forest problem to construct optimal subnetworks from the single- or multi-omic datasets. Omics Integrator was applied to several conditions from cancer driver network construction (Dincer et al., 2019) and to viral infection modules in the host organisms (Sychev et al., 2017). iCell uses the matrix factorization to integrate multi-omics datasets with tissue-specific interactomes. In this study, we compared the performance of four network reconstruction approaches, all-pairs shortest path (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF), on six different interactomes. A conceptual representation of these methods is illustrated in **Figure 1**. We did not consider the methods in this comparison that modify the underlying interactome or reconstruct regulatory networks using gene expression, such as ARACNe (Lachmann et al., 2016), GENIE (Fontaine et al., 2011), and INFERELATOR (Madar et al., 2009). APSP merges the shortest paths between pairs of nodes in the seed list. HDF implements the heat diffusion process by transferring the initial heat of the seed list to their neighbors. PRF applies a random walk to find the nodes most relevant to the seed list. We calculate the edge flux in both HDF and PRF based on the resulting node weights. PCSF finds an optimal forest that connects the seeds either directly or by adding intermediate nodes. We evaluated the performance of these algorithms on a gold standard dataset containing 32 curated pathways in the NetPath database using different metrics such as precision, recall, and MCC values. The performance of each network reconstruction approach is highly dependent on the reference interactomes. Additionally, each method has its own strengths and limitations. We found that the interactomes have some critical differences that can significantly affect the performance of network reconstruction approaches, such as their edge weight distributions, the bias toward some well-studied proteins, their coverage of disease-associated proteins, and the structurally resolved interactions. APSP has the highest recall and the lowest precision, while PRF, HDF, and PCSF have more balanced and comparable performance in precision and recall. Among them, PCSF performed the best in terms of the F1 score, which represents the balance between the precision and the recall. Overall, our study presents an extensive comparison of the selected network reconstruction approaches and shows the impact of the input interactome in their performance. This comparison presenting the strong and weak aspects of the interactomes and reconstruction approaches has the potential to be beneficial to the field.

METHODS

Reference Interactomes

We used PathwayCommons v12 (Rodchenkov et al., 2019), iRefWeb v13 (Turinsky et al., 2011), HIPPIE v2.2 (Alanis-Lobato et al., 2017), ConsensusPathDB (Kamburov et al., 2013), STRING (Szklarczyk et al., 2019, 2021), and OmniPath

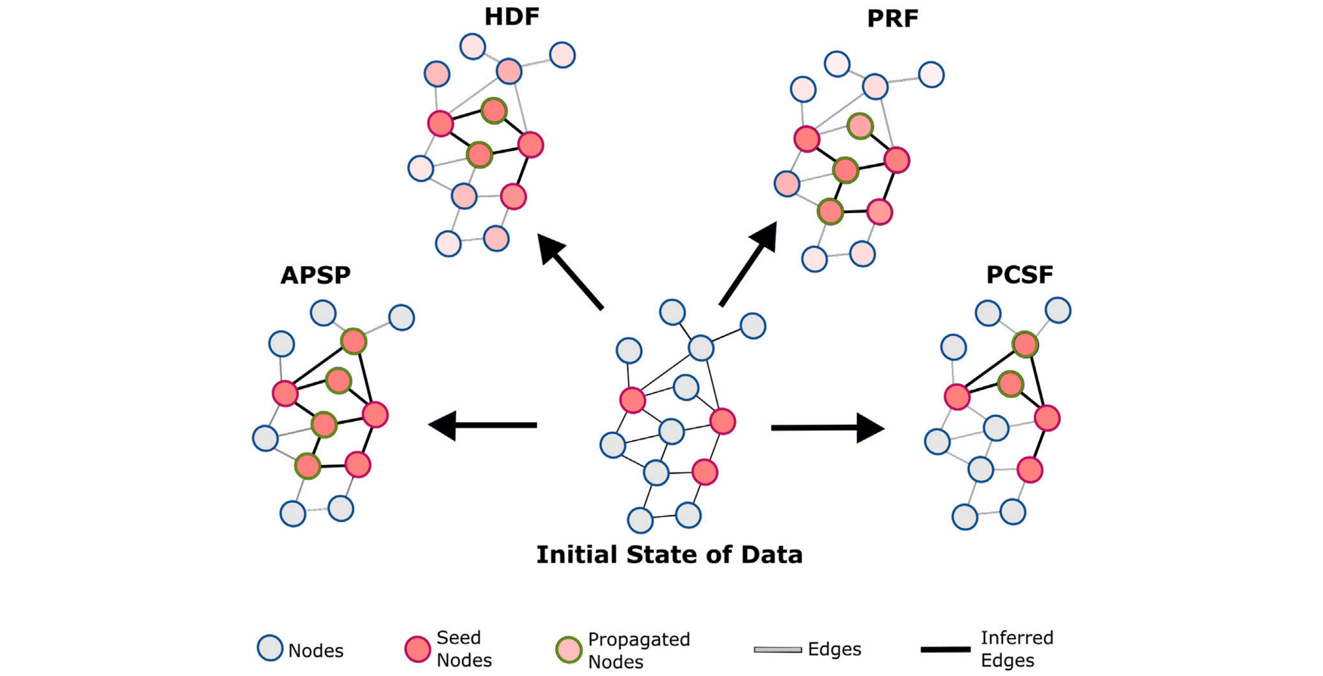


FIGURE 1 | Conceptual representation of reconstruction algorithms: all-pair shortest paths (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF). In the APSP, the subnetwork is reconstructed with the union of all shortest paths between seed nodes. HDF diffuses the heat that initially belongs to seed nodes. After limited steps of transfer, the heat of the nodes is used for flux score calculations for edges. PRF uses a personalized PageRank algorithm to find the probability of nodes after randomly walking in the reference interactome and calculates flux scores. PCSF finds the optimum forest to link seed nodes either directly or through intermediate nodes. The union of optimum forests reconstructs subnetworks.

(Ceccarelli et al., 2020) for the interactome comparison and the assessment of subnetwork inference approaches. We mapped the names of proteins in interactomes (nodes) to their reviewed Uniprot identifiers (The UniProt Consortium, 2019). The statistics of the interactomes are listed in **Table 1**. Some interactomes have confidence scores, which represent how real an interaction is. PathwayCommons and OmniPath do not have confidence scores. iRefWeb uses the MI-scoring scheme, which considers multiple parameters including experimental detection methods. HIPPIE v2.2 and ConsensusPathDB (Release 34) have confidence scores on edges calculated based on their own scheme (Kamburov et al., 2012; Alanis-Lobato et al., 2017). We filtered the STRING interactome by recalculating confidence scores considering only the experiment and database scores (von Mering et al., 2005).

Interactome Comparison Metrics

We compared the reference interactomes at both the node and edge levels using different metrics, namely, the overlap coefficient, correlation of edge confidence scores, inclusion of disease-associated proteins, and overlap with the pathway edges. The overlap coefficient is a similarity measure for two given datasets, S_1 and S_2, which can be node sets or edge sets of graphs or information coming from a database. The overlap coefficient was calculated using **Eq. 1** for pairwise comparison of interactomes and coverage of varied knowledge (Simpson, 1966; Kuzmin et al., 2016) as follows:

$$\textbf{overlap}\,(S_1, S_2) = \frac{|S_1 \cap S_2|}{\min\,(|S_1|, |S_2|)} \qquad (1)$$

We compared each pair of interactomes, $G(V_G, E_G, c(e_G))$ and $H(V_H, E_H, c(e_H))$, where V is the node set and E is the edge set, and $0 \le c(e) \le 1$, where $c(e)$ is the confidence score of an edge. The node-level similarity of the given interactomes were calculated using the overlap coefficient by applying **Eq.1** where V_G is used as S_1 and V_H is S_2. Likewise, the edge level overlap coefficient in each pair of interactomes is determined using **Eq. 1** where E_G and E_H are assigned as S_1 and S_2, respectively.

Next, we explored the structurally known protein-protein interactions in each reference interactome using the overlap coefficient. Interactome INSIDER has 4,150 interactions from PDB and 2,901 interactions from Interactome3D (Meyer et al., 2018). The edge level overlap coefficient between each reference interactome (G) and each structural interactome (H) is calculated using **Eq. 1**.

The interactomes and network reconstruction methods are frequently used for revealing cancer driver modules. We downloaded the 568 cancer driver genes (CDGs) from intOGen (Martínez-Jiménez et al., 2020). The overlap coefficient between CDGs (S_1) and proteins in each reference interactome (S_2) is calculated using **Eq. 1**. Additionally, the number of publications about each CDG and the degree centrality of the CDGs are analyzed to find out the bias of the interactomes toward well-studied or cancer-associated proteins.

TABLE 1 | Reference interactomes and their statistics.

Interactome	Number of proteins	Number of interactions	Confidence score
iRefWeb v13.0	11,295	80,351	Yes
PathwayCommons v12	18,536	1,126,072	No
HIPPIE v2.2	15,984	369,584	Yes
ConsensusPathDB	17,269	359,201	Yes
STRING v11	8,992	229,306	Yes
OmniPath	6,549	35,684	No

The overlap of each reference interactome (**G**) with the known interactions in 171 pathways in KEGG (**H**) is calculated using **Eq. 1** (Kanehisa et al., 2017). Modeling a small-sized network is a challenging task because a small number of molecular interactions limit the overall dynamic range of the signals (Tkačik et al., 2009; Azpeitia et al., 2020). Therefore, we discarded KEGG pathways having less than 30 edges from the interactome evaluations.

Among the selected interactomes, iRefWeb, HIPPIE, ConsensusPathDB, and STRING have edge confidence scores that are calculated with different scoring approaches. We applied an all-pair comparison of the given interactomes ($\boldsymbol{G, H}$) with the Pearson correlation analysis on the confidence scores in the intersection of edge sets in interactome pairs ($\boldsymbol{E_G \cap E_H}$)

Biological networks follow the scale-free power law distribution, $\boldsymbol{P(k) = k^{-\gamma}}$, where k is the degree of a node, and γ is the power coefficient (Barabási and Albert, 1995; Alm and Mack, 2016). To linearize the representation of both degree distribution and publication distribution, the logarithm of distribution was used as $\boldsymbol{\log(P(k)) = -\gamma \log(k)}$. We collected the number of publications about each protein from UniProt. The correlation between the degree and the number of publications of the nodes was evaluated using the Pearson correlation test on a log scale.

Network Reconstruction Methods

We used four reconstruction approaches, the shortest path, heat diffusion, PageRank, and PCSF. Selected interactomes are separately employed as the reference network, $\boldsymbol{G(V, E, c(e))}$, where $\boldsymbol{V}$ is the node set, $\boldsymbol{E}$ is the undirected edge set, and $\boldsymbol{c(e)}$ is the weight of an edge. These networks are weighted by confidence scores in the interactions, $\boldsymbol{0 \le c(e) \le 1}$. Network reconstruction algorithms infer the subnetwork, $\boldsymbol{R(V_R, E_R)}$, where $\boldsymbol{V_R \subseteq V}$ and $\boldsymbol{E_R \subseteq E}$, by connecting the seed node set, $\boldsymbol{V_I \subseteq V}$. The given node set is weighted with uniform $\boldsymbol{1/|V_I|}$ where $\boldsymbol{|V_I|}$ is the number of seed nodes, while the remaining node set is weighted as $\boldsymbol{0}$, so that $\boldsymbol{w(v)}$ can be defined for reconstruction algorithms.

All-Pairs Shortest Paths

We found out all shortest paths between each pair of nodes, $\boldsymbol{u}$ and $\boldsymbol{v \in V_I, u \neq v}$. When there are multiple shortest paths between u and v, we included all of them. Finally, we merged all shortest paths to obtain the final subnetwork. We did not put any edge weight–based filtering or path length threshold.

Personalized PageRank

The PageRank algorithm was normally designed for propagation in directed graphs. Personalized PageRank (PPR) is adapted to undirected graphs by converting each edge into both directed edges. The PageRank score of each node, $\boldsymbol{p(v)}$, in the reference interactome, $\boldsymbol{G}$, represents the probability of being at the node at a certain time step ($\boldsymbol{t}$) that is calculated using the following iterative formula:

$$p_{t+1}(y) = \frac{1-\lambda}{N} + \lambda \sum_{x_i \to y} \frac{p_t(x_i)}{\deg(x_i)} \tag{2}$$

where **Eq. 2** includes the probability of node $\boldsymbol{y \in V}$ that is calculated using the damping factor ($\boldsymbol{\lambda}$) defining the probability of walking from neighbor nodes ($\boldsymbol{x_i}$) to $\boldsymbol{y}$, and $\boldsymbol{N}$ is the number of nodes (Page et al., 1998; Langville and Meyer, 2005). Initial probabilities of nodes were taken from $\boldsymbol{w(v)}$. We iterated **Eq.2** 100 times by default to obtain $\boldsymbol{p(v)}$.

Heat Diffusion

In the heat diffusion (HD), seed nodes having uniform heats prioritize their related nodes *via* heat transfer, which is formulated as follows:

$$p(v) = p_0 \left(I + \frac{-\boldsymbol{\alpha}}{N} \mathrm{L} \right)^N \tag{3}$$

In **Eq. 3**, $\boldsymbol{L = I - W}$, where $\boldsymbol{I}$ represents an identity matrix and $\boldsymbol{W = D^{-1}A}$ in which $\boldsymbol{D}$ and $\boldsymbol{A}$ are defined as the diagonal degree matrix and the adjacency matrix, respectively. $\boldsymbol{p_0}$ is the initial heat vector in which nodes were weighted from $\boldsymbol{w(v)}$. $\boldsymbol{N}$ and $\boldsymbol{\alpha}$ are, respectively, the number of iterations and the heat diffusion rate. $\boldsymbol{N = 3}$ is set as the default (Nitsch et al., 2010). At the end of heat diffusion, nodes have the diffused heat $\boldsymbol{p(v)}$ as the weight.

Edge Selection Over Flux Scores

Personalized PageRank with flux (PRF) and heat kernel diffusion with flux (HDF) are calculated over $\boldsymbol{deg(v)}$, which is defined as the number of interactions in $\boldsymbol{G}$, and node scores $\boldsymbol{0 \le p(v) \le 1}$, which come from PPR or HD. In our study, unlike TieDie and HotNet with heat diffusion algorithms and flux on a random walk with restart, the threshold value is employed to eliminate uncritical nodes (Vandin et al., 2011; Creighton et al., 2013; Rubel and Ritz, 2020). The related nodes with $\boldsymbol{p(v_i) \ge 1/n}$ where $\boldsymbol{n}$ is the number of nodes in the interactome are considered for subnetwork reconstruction. We calculated the directional flux scores $\boldsymbol{f_{u\rightarrow t}}$ using **Eq. 4** where $\boldsymbol{u, t} \in \boldsymbol{V}$, $\boldsymbol{p(u)}$ is the score that comes from PPR or HD, and $\boldsymbol{deg(u)}$ is the number of neighbors of node $\boldsymbol{u}$. Likewise, we calculated $\boldsymbol{f_{t\rightarrow u}}$ using **Eq. 5**. We determined the final flux of the edge as the minimum of $\boldsymbol{f_{u\rightarrow t}}$ and $\boldsymbol{f_{t\rightarrow u}}$ (**Eq. 6**).

$$f_{u\rightarrow t}(u, t) = \frac{p(u)\times c(e)}{deg(u)} \tag{4}$$

$$f_{t\rightarrow u}(t, u) = \frac{p(t)\times c(e)}{deg(t)} \tag{5}$$

$$f(e) = min\left(f_{u\rightarrow t}(u, t), f_{t\rightarrow u}(t, u)\right) \tag{6}$$

Edges are ranked from the highest flux score to the lowest by taking the negative logarithm of the flux. A total flux ($\boldsymbol{F}$) is calculated among the related nodes as follows:

$$F = \sum f(e) \tag{7}$$

$\boldsymbol{0 \leq \tau \leq 1}$**,** where $\boldsymbol{\tau}$ is a flux threshold value that is the selection percentage of $\boldsymbol{F}$. Edges are selected by summing flux scores from the highest to the lowest until the targeted flux amount, $\boldsymbol{\tau x F}$. The edges having low flux scores are excluded from reconstructed subnetworks (Rubel and Ritz, 2020).

Prize-Collecting Steiner Forest

We used the PCSF algorithm implemented in Omics Integrator2. The seed nodes, $\boldsymbol{v_i} \in \boldsymbol{V_I}$, are weighted uniformly, and the edge costs are calculated using the cost function implemented in Omics Integrator 2 which combines the edge confidence score, $c(e)$, and a penalty calculated from node degrees scaled with the γ parameter. If the reference interactome does not have confidence scores, $\boldsymbol{c(e) = 1}$ is uniformly defined. PCSF also penalizes the nodes based on their degrees (Tuncbag et al., 2016a). The new version, Omics Integrator 2, penalizes the edges based on the degrees of the node pair. The following function finds an optimum forest, $\boldsymbol{F(V, E)}$, by minimizing the objective function (Tuncbag et al., 2013):

$$f'(F) = \sum \beta.p(v) + \sum cost(e) + \omega.\kappa \tag{8}$$

In **Eq. 8**, $\boldsymbol{\kappa}$ is the number of connected components, $\boldsymbol{\beta}$ controls the relative weight of the node prizes, and $\boldsymbol{\omega}$ controls the cost of adding an additional tree to the solution network.

PCSF provides an optimum forest for each parameter set and an augmented forest which includes all the edges in the interactome that are present between the nodes in the optimal forest. We obtained the final reconstructed networks with the intersection of the optimal augmented forests that were generated using multiple parameter sets.

Performance Analysis

NetPath is the curated human signaling pathway database that is composed of immune signaling pathways and cancer signaling pathways. In this study, 32 pathways in NetPath were used as a plausible dataset (Kandasamy et al., 2010). Since the computational cost of reconstruction was expensive for all pathways in NetPath with all parameter sets, first, optimum parameter sets were determined before performance analysis.

Parameter Tuning

Parameters of reconstruction algorithms were separately optimized for each reference interactome. Thus, Wnt, TCR, TNFα, and TGFβ pathways on NetPath were used for parameter selection. Nodes in each pathway were independently shuffled and split into five-fold. Each fold was, respectively, removed from the complete pathway node list, and network reconstruction was executed with the remaining folds. Parameters of reconstruction algorithms were separately tuned for each reference interactome to maximize the F1 score (**Eq. 12**). In the APSP, all identified shortest paths among seed node sets were inserted into a reconstructed pathway without any parameter tuning, so we do not adjust any parameter. We tuned the parameters in the given interval in **Table 2** for PRF, HDF, and PCSF and for each reference interactome. Parameter sets of PRF and HDF were tuned in a two-dimensional grid *via* the mean of parameters that pooled the 10 highest F1 scores (**Supplementary Figures 1, 2**). In the PCSF, the union of all parameters that achieve the best coverage of the seed nodes, $\boldsymbol{V_I}$, for each pathway was used as optimum parameter sets.

The Calculation of Performance Scores

After tuning the parameters on four pathways, the remaining 28 pathways in NetPath, listed in **Supplementary Table 1**, were used for performance evaluation with five-fold cross-validation. We evaluated each reconstruction algorithm separately on each reference interactome by calculating the F1 score, Matthew's correlation coefficient (MCC), recall and precision values, and false positive rate (FPR) in **Eqs 9–13** as follows:

$$recall(TP, TN) = \frac{|TP|}{(|TP| + |FN|)} \tag{9}$$

$$precision(TP, FN) = \frac{|TP|}{(|TP| + |FP|)} \tag{10}$$

$$FPR(TP, FN) = \frac{|FP|}{(|FP| + |TN|)} \tag{11}$$

$$F1_{score} = \frac{2\times precision\ x\ recall}{precision + recall} \tag{12}$$

$$MCC(TP, TN, FP, FN) = \frac{(|TP|\times|TN|) - (|FP|\times|FN|)}{\sqrt{(|TP| + |FP|)(|TP| + |FN|)(|TN| + |FP|)(|TN| + |FN|)}} \tag{13}$$

Seed nodes were not counted in the performance calculation. However, all edges in the reconstructed network were used in the performance evaluation since interactions were not used in the initial input. For a given reference in interactome **G(**$\boldsymbol{V, E}$**)** and an seed node set $\boldsymbol{(V_I)}$ from a pathway $\boldsymbol{T(V_T, E_T)}$, a network is reconstructed, $\boldsymbol{R(V_R, E_R)}$, using the listed methods, where $\boldsymbol{V_T}$;, $\boldsymbol{V_R}$ and $\boldsymbol{V_I} \subseteq \boldsymbol{V}$, and $\boldsymbol{E_T}$ and $\boldsymbol{E_R} \subseteq \boldsymbol{E}$. Node-level true positives $\boldsymbol{(TP_V)}$ and edge-level true positives ($\boldsymbol{TP_E}$) are obtained from $|\boldsymbol{V_R} \cap \boldsymbol{V_T}|$ and $|\boldsymbol{E_R} \cap \boldsymbol{E_T}|$, respectively. Node-level true negatives $\boldsymbol{(TN_v)}$ and edge-level true negatives $\boldsymbol{(TN_E)}$ are obtained from $|\boldsymbol{V} \setminus \boldsymbol{(V_R} \cup \boldsymbol{V_T)}|$ *and* $|\boldsymbol{E} \setminus \boldsymbol{(E_R} \cup \boldsymbol{E_T)}|$, respectively. False positives $\boldsymbol{FP_V}$ and $\boldsymbol{FP_E}$ are

TABLE 2 | Tuning ranges of parameter sets in PageRank flux (PRF), heat diffusion flux (HDF), and prize-collecting Steiner forest.

Reconstruction algorithm	Parameter	Range	Increment
PRF	Damping factor (λ)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
HDF	Heat diffusion rate(α)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
PCSF	Dummy edge weight (ω)	0–5	0.5
	Edge reliability (β)	0–5	0.5
	Degree penalty (γ)	0–10	0.5

equal to $|V_R \setminus V_T|$ and $|E_R \setminus E_T|$, respectively. False negatives FN_V and FN_E, are equal to $|V_T \setminus V_R|$ and $|E_T \setminus E_R|$, respectively.

We performed principal component analysis (PCA) to figure out critical scores that explain the highest variance across all pathways. We statistically assessed overall performance data including both edge- and node-based scores by individually grouping reference interactomes and reconstruction methods.

Data Availability Statement

Codes and datasets used for this study are publicly available at the online repository https://github.com/metunetlab/Interactome_Network_Reconstruction_Assessment_2021. We downloaded NetPath, http://netpath.org/browse, and PathwayCommons, https://www.pathwaycommons.org/archives/PC2/v12/PathwayCommons12.All.hgnc.txt.gz, iRefWeb, http://wodaklab.org/iRefWeb/search/index, HIPPIE, http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php, STRING, https://string-db.org/cgi/download, ConsensusPathDB, http://cpdb.molgen.mpg.de/, Reference Human Proteome from UniProtDB, https://www.uniprot.org/, using the query https://www.uniprot.org/uniprot/?query=proteome:UP000005640%20reviewed:yes, INSIDER, http://interactomeinsider.yulab.org/downloads.html, and KEGG, https://www.kegg.jp/kegg/download/ and http://rest.kegg.jp/get/+'pathwayid'+'/kgml'. OmniPath and the signaling pathways in Glioblastoma (WP2261) were retrieved from WikiPathway using Cytoscape 3.8.0.

RESULTS

Systematic Evaluation of Reference Human Interactomes

Network reconstruction algorithms are highly dependent on the quality and coverage of the reference interactome. Therefore, we systematically explored the properties of iRefWeb, PathwayCommons, HIPPIE, ConsensusPathDB, OmniPath, and STRING databases. Among them, iRefWeb, HIPPIE, ConsensusPathDB, and STRING provide the measure of confidence in interactions as scores. First, we compared the pairs of interactomes to determine how similar they are in terms of their node and edge sets. PathwayCommons is the largest network in size, so it has the highest fraction of node and edge overlap compared to all other interactomes. iRefWeb, PathwayCommons, HIPPIE, and ConsensusPathDB are the most similar interactomes to each other based on the node and edge overlaps (**Figure 2A**). On the other hand, STRING and OmniPath have fewer common nodes and edges with other interactomes. We need to note that the raw data in STRING contain more than one million interactions in human interactomes, and we used only the experimental and database interactions which resulted in a relatively small-sized interactome with medium or high confidence edges. Before using the network reconstruction algorithms, obtaining the reference interactome with measurements of interaction confidence is fundamental to decreasing the impact of the false positives. Because network reconstruction algorithms leverage the edge confidence scores and the topology of the reference interactomes during the propagation or optimization, confidence scores may substantially affect the accuracy of the resulting network. Even two topologically equivalent interactomes may produce different subnetworks as a result of network reconstruction if their confidence score distributions are different from each other. In **Figure 2B**, the number of edges in each reference interactome is shown, which are categorized as low, medium, and high confidence edges based on the interaction scores. ConsensusPathDB contains predominantly high confidence interactions, while HIPPIE and iRefWeb interactions are accumulated in medium and low confidence intervals. HIPPIE and iRefWeb use MINT-inspired (MI) confidence score calculation, while ConsensusPathDB uses the IntScore tool (Braun et al., 2009; Turner et al., 2010; Kamburov et al., 2011; Kamburov et al., 2012; Turinsky et al., 2011; Schaefer et al., 2012; Alanis-Lobato et al., 2017). We recalculated the confidence scores in STRING by considering only the experiment and database scores. PathwayCommons and OmniPath do not provide confidence scores. Edge confidence scores can be computed in various ways. Different scoring schemes lead to variation in the confidence score distributions across the interactomes. As expected, the correlation of confidence scores between HIPPIE and iRefWeb is the highest (r = 0.67, $p <$ 0.01) because both use MI-Score. The correlation between confidence scores in iRefWeb and ConsensusPathDB is very low (r = 0.25, p < 0.01) (**Figure 2C**) because ConsensusPathDB uses a different scoring scheme, IntScore. While MI-Score considers homologous interactions, the detection method, and the number of publications about the interactions, IntScore includes topological properties, literature evidence, and similarities in annotation of proteins.

Confidence scores do not completely solve the bias in the interactomes despite being a powerful measurement to filter out false positives. Therefore, we additionally analyzed the interactomes based on the bias toward well-studied proteins using different features, namely, the number of publications about the proteins, coverage of the cancer driver genes, and the number of interactions having structural details. Well-studied proteins, such as TP53 and EGFR, have hundreds of high confidence interactions in the interactomes (Schaefer et al., 2015; Chen et al., 2018; Porras et al., 2020). Indeed, there is a trade-off between the interaction confidence scores of certain proteins and systematic study bias. We used the number of publications and the degree centrality of proteins in each

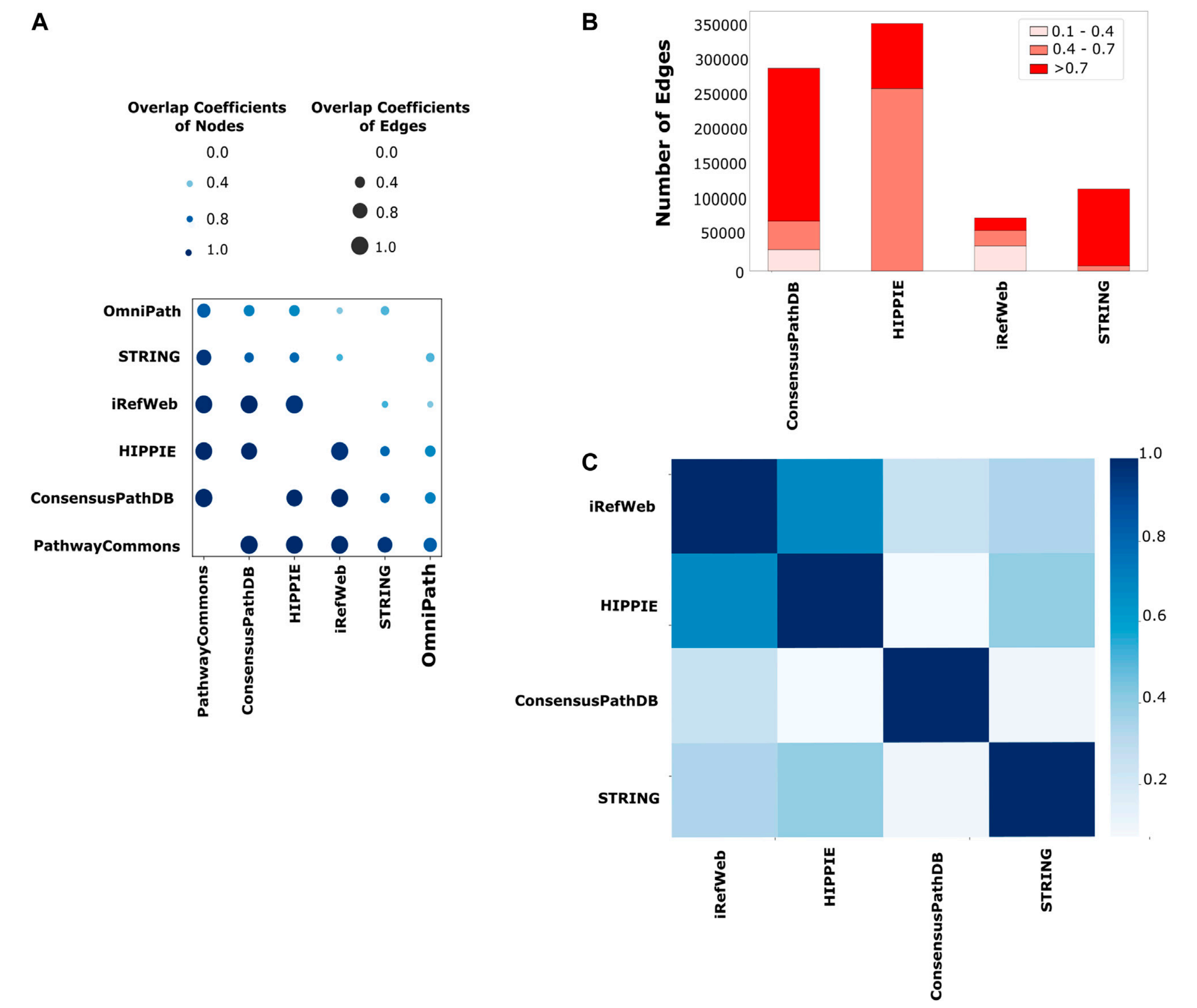


FIGURE 2 | Comparison of the reference interactomes. **(A)** Node- and edge-level commonalities between different interactomes; the node overlap score is displayed on a light-to-dark blue-color scale, while the edge similarity score is shown on a size scale where the higher similarity is represented with the larger circle. **(B)** Scores are categorized as low confidence between 0.1 and 0.4, medium confidence between 0.4 and 0.7, and high confidence between 0.7 and 1.0 for each interactome. PathwayCommons and OmniPath are not demonstrated here due to the lack of a confidence score for their edges. Edges with low confidence scores are mostly seen in iRefWeb and small portions of edges in ConsensusPathDB have low confidence scores, while low confidence edges are not seen in the filtered STRING and HIPPIE. **(C)** Different strategies on confidence score calculation are used in interactomes. Using their common edges, the correlation coefficients among their confidence scores over interactomes are demonstrated in the heatmap. The highest correlation between HIPPIE and iRefWeb is seen with the darkest blue. The same confidence score calculation strategy, MI scoring, is used in both interactomes.

reference interactome to explore if highly connected proteins are also well-studied ones.

Each analyzed interactome is a scale-free network so that their degree distributions follow the power law (**Supplementary Figure 3**) (Barabási and Albert, 1995; Vidal et al., 2011). The number of publications about proteins follows the power law distribution as their degree distribution (**Supplementary Figure 4**). Thus, the number of publications and degrees were analyzed using log-based values to find out their correlation. The number of publications and the degrees of proteins are positively correlated in all interactomes (**Figure 3A**). We observed the highest correlation in PathwayCommons (r = 0.62, p < 0.01) and HIPPIE (r = 0.61, p < 0.01), which implies the bias toward well-studied proteins in these interactomes. iRefWeb, STRING, and OmniPath have moderate correlation between the degree and the number of publications, which implies relatively less biased interactomes (**Supplementary Table 2**). We note that this comparison is performed on the whole interactome without

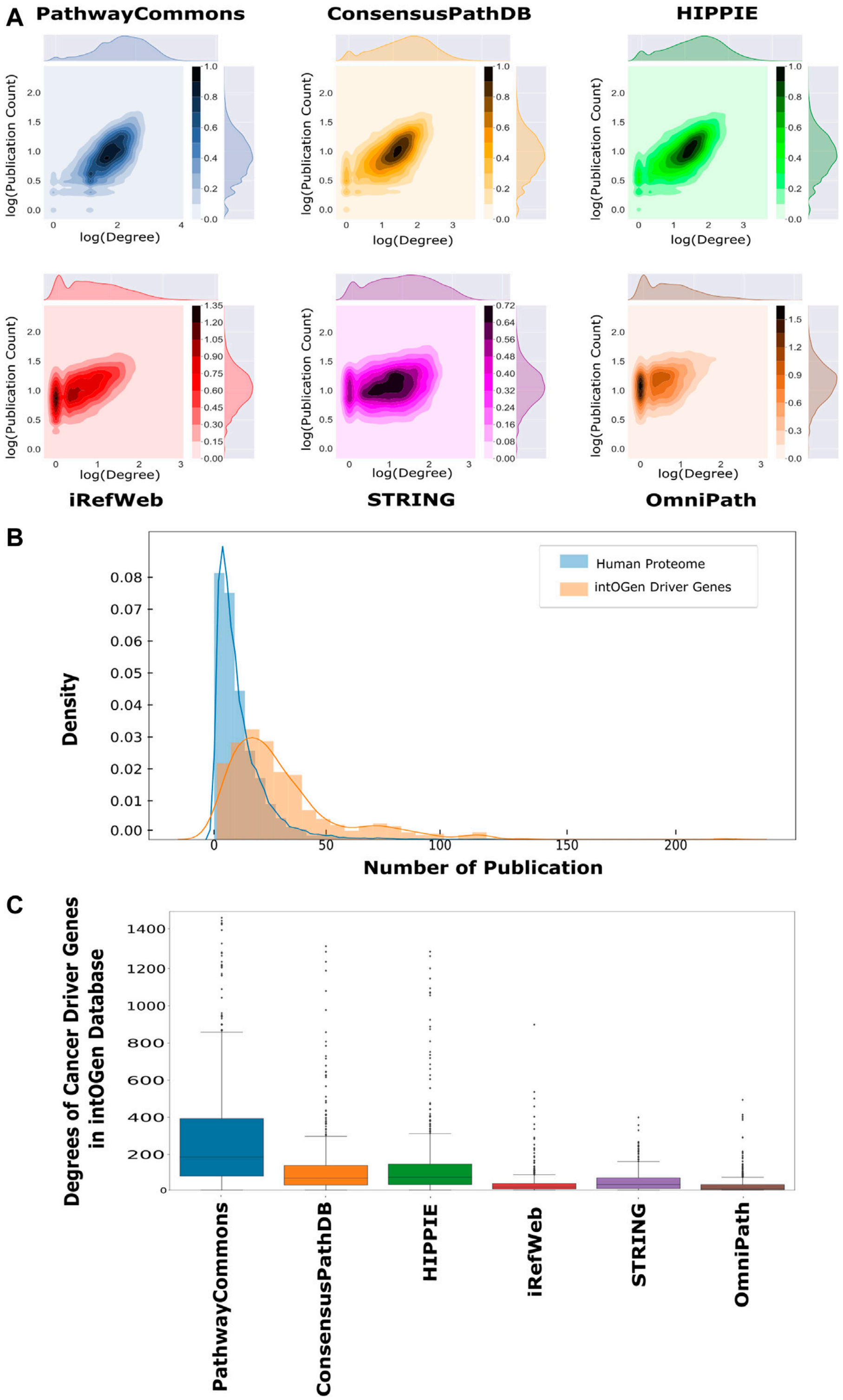


FIGURE 3 | Correlation between publication counts and degrees over interactomes. **(A)** Log–log scale joint graphs of publication distribution and degree distribution for each interactome were drawn since both follow a power-law distribution. While all interactomes have a positive correlation between the degree and publication number, PathwayCommons, HIPPIE, ConsensusPath, and iREF have well-studied hubs. On the other hand, hubs in iRefWeb and OmniPath are not composed of relatively well-studied proteins (p-values <0.001 and $r_{PathwayCommons} = 0.622$, $r_{ConsensusPathDB} = 0.556$, $r_{HIPPIE} = 0.614$, $r_{iRefWeb} = 0.508$, $r_{STRING} = 0.250$ and $r_{OmniPath} = 0.400$). **(B)** Distributions of the number of publications and cancer driver genes in the intOGen database are shown, respectively, in blue and orange. The probability of well-studied cancer driver genes (CDGs) is higher than the probability of well-studied proteins. **(C)** Driver gene degrees in the interactomes are demonstrated in the boxplot in which driver genes in PathwayCommons have more connection than other interactomes. OmniPath and iRefWeb do not have highly connected driver genes as many as ConsensusPath, HIPPIE, STRING, and PathwayCommons.

any confidence score–based filtering, except STRING. We expect that if only the high or medium confidence interactions in other interactomes would be considered, the correlations may be dramatically reduced and the bias toward well-studied proteins may be dumped. Reconstruction algorithms are also adapted to overcome this inherent bias toward the nodes and edges in interactomes. For example, heat diffusion and random walk, together with the edge flux calculation, use node degrees for normalization, while PCSF penalizes highly connected proteins (Creixell et al., 2015; Tuncbag et al., 2016a; Rubel and Ritz, 2020). In this way, false-positive edges belonging to hub nodes are excluded from the final subnetwork.

One application area of network reconstruction algorithms is the discovery of disease-associated pathways, especially in cancer, by inferring the seed proteins/genes. The resulting networks are used for patient stratification, biomarker discovery, or the analysis of drug mechanisms of action (Mo et al., 2018; Huang et al., 2019; Koh et al., 2019; Wang et al., 2021). Therefore, we searched for the coverage of the cancer driver genes (CDGs) in each interactome. CDGs provide growth advantage to the tumor cells and alter signaling pathways. Additionally, CDGs are important markers in tumor stratification, characterization, and drug development (Waks et al., 2016; Bailey et al., 2018; Zsákai et al., 2019). We obtained the list of CDGs from the intOGen database (Martínez-Jiménez et al., 2020). We found that significantly more publications are present for CDGs than for the rest of the proteomes, as shown in **Figure 3B** ($p < 0.01$). The presence of driver genes and their edges help in accurately reconstructing the driver pathways in cancer. All analyzed interactomes are highly inclusive of driver genes, especially PathwayCommons, ConsensusPathDB, and HIPPIE (**Supplementary Figure 5**). However, the degrees of CDGs in the PathwayCommons interactome are significantly higher than others (**Figure 3C**).

In terms of protein interactions, the most accurate and confident interactions can be caught by their structural identification. Structures of protein–protein complexes uncover the binding sites, domain contacts, and many more (Schmidt et al., 2014; Nero et al., 2018; Hicks et al., 2019). The only drawback is the availability of limited structural data. Despite the exponential increase in PDB with the help of the X-ray, CryoEM, and NMR techniques, the number of protein complexes can still only cover around 16% of the whole interactome (Berman et al., 2000; Mosca et al., 2013; Venko et al., 2017). Many structure-based predictive approaches are also employed to accurately identify protein–protein interactions. Therefore, we further analyzed each interactome based on the representation of structurally annotated interactions. For this purpose, we used the complexes in PDB and Interactome3D. We found that HIPPIE has the highest coverage of structurally known protein–protein interactions (**Figure 4A**). HIPPIE is followed by PathwayCommons and ConsensusPathDB. iRefWeb, OmniPath, and the filtered STRING interactome have the lowest coverages.

Another source of confident interactions is the curated pathways, despite being incomplete. Generated subnetworks are required to be biologically meaningful so that their downstream analysis can sign proper biological functions (Vidal et al., 2011; Sevimoglu and Arga, 2014). Therefore, we explored the coverage of interactomes based on the curated pathways retrieved from KEGG, which is one of the most frequently used databases for pathway annotations. We found that KEGG pathways are relatively less represented in iRefWeb, while PathwayCommons and filtered STRING highly covered them (**Figure 4B**). We need to note that some individual pathways are better covered in some interactomes although their overall coverage is relatively low (**Figure 5**). For example, the MAPK and RAS signaling pathways are better represented in OmniPath, although OmniPath has a moderate coverage of all pathways. Individual pathway coverage of each interactome is listed in **Supplementary Table**.

Performance of Network Reconstruction Algorithms

As evidenced in detail, each interactome has its own strengths and weaknesses. These properties have a direct effect on the performance of network reconstruction algorithms. Therefore, we used each interactome as the reference for each network reconstruction algorithm to monitor the variance in the performance. We used four well-established network reconstruction algorithms, the all-pair shortest paths (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF) algorithms, to evaluate their performance on the gold standard dataset of 32 curated pathways retrieved from NetPath. Four pathways are used for parameter tuning, and the rest (28 pathways) is used for performance evaluation.

We collected both node- and edge-level performance metrics for each pair of interactomes and reconstruction methods on each pathway. We found that node-level performance is relatively more robust to different interactomes or different pathways in each approach than the edge-level performance. The largest variation is in the edge-level F1 scores, in that the balance between the recall and precision values is highly variable across pathways and interactomes (**Supplementary Figure 6**). The F1 scores ($p < 0.001$) and precision ($p < 0.001$) scores of the reconstructed pathways that are inferred from PathwayCommons are mostly lower than the scores of HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb (**Figure 6A**). The second highest variation is in the edge-level MCC, used for binary classification over imbalanced data (Boughorbel et al., 2017; Magnano and Gitter, 2021). This result implies that the algorithms do not perform well with a relatively very large reference interactome because of the potential dominance of false positives over the true-positive interactions. Based on the F1 score and the precision value, we did not find a significant difference in performance when HIPPIE, ConsensusPathDB, OmniPath, or iRefWeb interactomes are used. Therefore, we continued with HIPPIE as a reference interactome for further assessments since it has the most balanced features based on the comparison in the previous part, including coverage of structurally known interactions. The comparison of edge-based performance scores showed that APSP significantly has the lowest precision values ($p <$

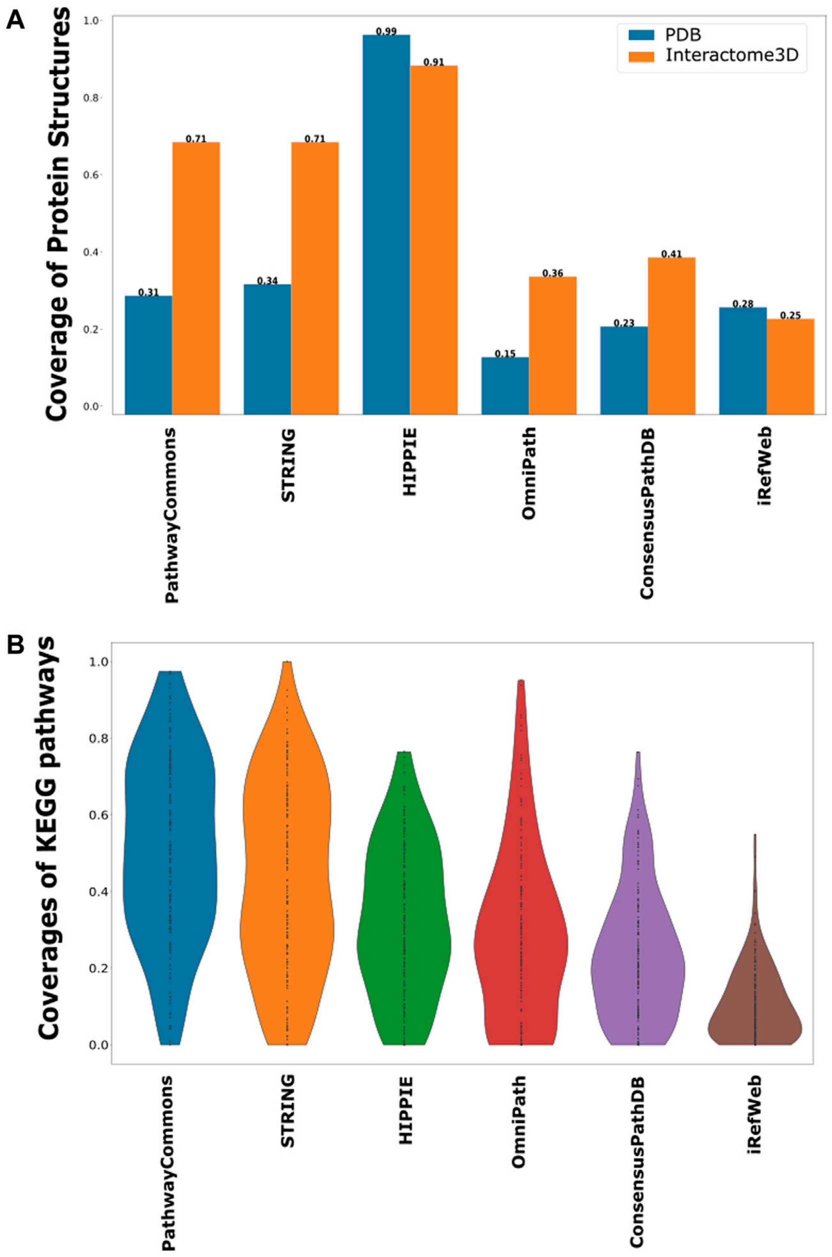


FIGURE 4 | Coverage of structurally known interactions and pathway interactions in each interactome. **(A)** Structural information is demonstrated in two groups, as known interactions in PDB in blue and predicted interactions in Interactome3D in orange. **(B)** Overlaps between the interactions in KEGG pathways and each interactome are shown as a violin-plot.

0.001) and the highest recall values among all reconstruction approaches when the performance across all pathways is evaluated. There is no significant difference in precision values between HDF, PRF, and PCSF (**Figure 6B**). The recall values of the reconstructed pathways do not significantly differ between HDF and PRF, while PCSF ($p < 0.001$) has significantly higher recall scores ($p < 0.001$) than HDF and PRF (**Figure 6C**). The trade-off between the precision and recall scores can be noticed in the results of reconstruction methods. Insertion of all shortest paths between the seed nodes in the APSP algorithm causes both the reduction in precision values and the increase in recall values. The significantly high FPR in APSP ($p < 0.001$) indicates that false-positive edges dominate the true-positive edges (**Supplementary Figure 7**). Therefore, F1 scores of the APSP-reconstructed pathways are significantly lower than those of other methods ($p < 0.001$) (**Figure 6D**). On the other hand, PCSF-reconstructed pathways have moderately high recall and precision scores and the highest F1 score by a considerable margin, optimizing the trade-off between the precision and recall values. Interestingly, the interval of recall scores in the reconstructed pathways in PCSF is not variable in a

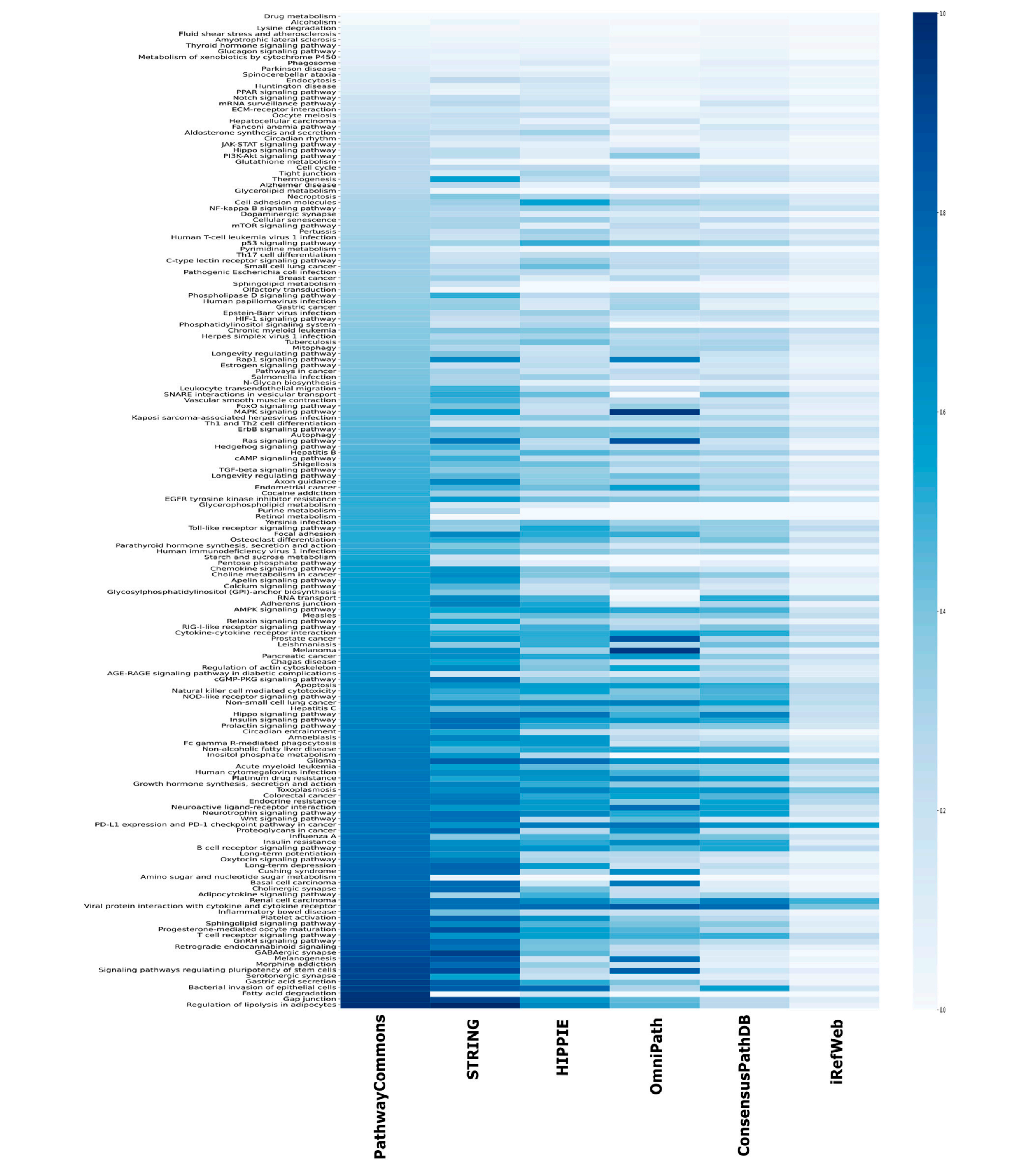


FIGURE 5 | Overlap between each pathway in KEGG and each interactome. The sorted coverage of interactomes for 171 KEGG pathways is shown as a heatmap.

wide interval as in other methods; rather, it fluctuates around 0.65. The PCSF approach gives an optimum forest as an output together with an augmented forest which includes all the edges in the interactome that are present between the nodes in the optimal forest. We obtained the final network of PCSF by taking the intersection of augmented forests from multiple parameters. In

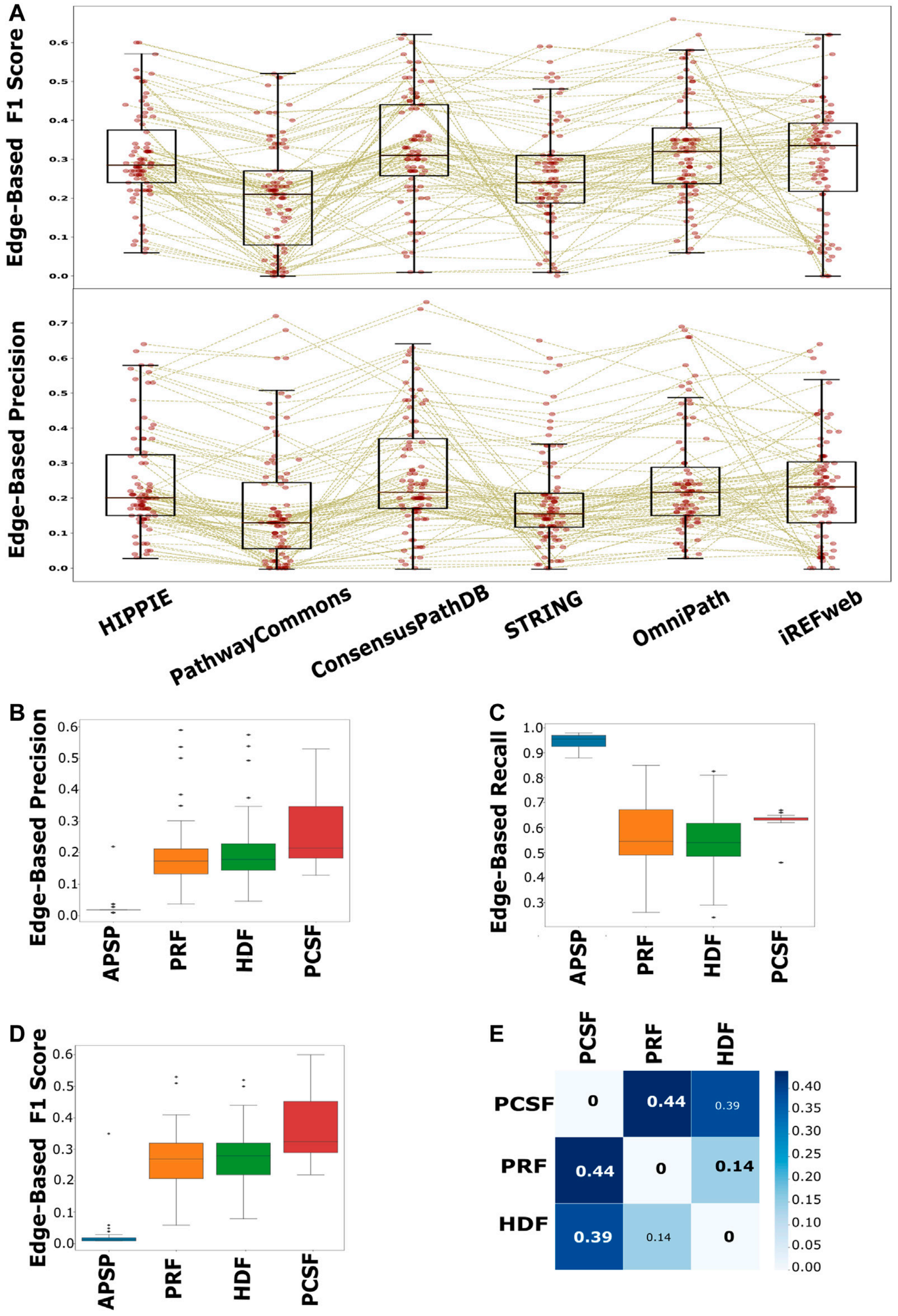


FIGURE 6 | Performance evaluation of each interactome and method in pathway reconstruction. **(A)** Boxplot of edge-based precision and F1 scores, for each interactome, shows that PathwayCommons and STRING are significantly lower than HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb, while there is not any distinct difference among HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb. The performance values for each reconstructed network is represented with red points in the boxplots. Brown lines connect the performance scores of the same pathway across the interactomes. **(B)** Edge-based precision, **(C)** edge-based recall, and **(D)** edge-based F1 scores are separately demonstrated for reconstruction algorithms. **(E)** HDF, PRF, and PCSF were compared in terms of the reconstructed pathways. The heatmap shows that the reconstructed pathways by PCSF are different, having %44 and %39 different edges, respectively, than the ones reconstructed by PRF and HDF.

this way, adding an edge to the final network was made very stringent. We computed the Jaccard similarity matrix among HDF, PRF, and PCSF to demonstrate the variation on the edge-level performance in the reconstructed pathways (**Figure 6E**; Ricotta et al., 2016). PCSF penalizes highly connected nodes, which reduces the dominance of well-studied or highly connected nodes in the reconstructed networks. In this way, important but low-degree nodes are also successfully included in the reconstructed pathways. As a result, PCSF has balanced precision and recall values, and its reconstructed pathways have the highest dissimilarity compared to the reconstructed pathways from other methods. Overall, the performance of the algorithms is highly affected by the parameter selection along with the used background interactome. To illustrate the reconstructed networks intuitively and to distinguish their commonalities and differences for each algorithm, we selected two case studies; one is selected from the NetPath database and the other is selected from WikiPathways.

Case Studies: Reconstruction of the Notch Pathway and Glioblastoma Disease Pathway

Our first case study is the Notch signaling pathway to intuitively illustrate the performance of each approach. The Notch signaling pathway plays a critical role in cell fate determination by regulating differentiation, apoptosis, proliferation, and morphogenesis. Its signaling cascades are associated with many human cancers (Sjölund et al., 2005; Bazzoni and Bentivegna, 2019; Guo et al., 2019). The APSP method recovers many true-positive edges, but it also introduces many false positives in the Notch pathway (**Supplementary Figure 7**). Therefore, only PRF, HDF, and PCSF results inferred from a set of seeds selected from the Notch pathway are illustrated in **Figure 7**. Notch receptors are single-pass transmembrane proteins, receiving signals from transmembrane ligands such as JAG1, JAG2, DLL1, and DLL4. The given protein list includes Notch receptors and CNTN1, JAG2, and DLL4. All reconstruction algorithms successfully identified JAG1 and the interaction between Notch receptors and their ligands except for DLL. True-positive nodes having a low degree in the reference interactome were caught better by PCSF than by PRF and HDF. Additionally, PCSF accurately included nodes such as CNTN1, WDR12, LEF1, RBX1, SIN3A, and many other true positives in the final reconstructed network. Although PCSF performs well in recovering low-degree nodes, it could not include some other nodes such as AKT1, SKP1, SPEN, and TCF3 in the pathway. PCSF successfully found the interactions between Furin–Notch receptors that regulate the Notch pathway in cancer progression where Furin, a low-degree ligand, generates biologically active heterodimer receptors (Qiu et al., 2015). On the other hand, PCSF fails to construct the interactions including low-degree nodes such as JAK2 and WDR12. HDF and PRF mostly reveal the interaction between high-degree nodes such as MAML1 and Notch receptors since the heat diffusion and the PageRank algorithm tend to give high scores to these nodes.

The Notch pathway has cross talk with other critical pathways in cancer such as the PI3K-AKT-mTOR and JAK-STAT signaling pathways (Chan et al., 2007; Hillmann and Fabbro, 2019). The cross talk is mediated by the nodes with low-degree and high betweenness centrality in the reference interactome such as PIK3R1, LCK, and JAK2. Although we could reveal these intermediate nodes that are important in cross talk between multiple pathways with PCSF, we could not achieve the same performance in the added edges. Despite correctly identifying PIK3R1 interaction with Notch1 and LCK, interactions with PIK3R2 and AKT were not found. In the JAK-STAT and Notch pathway cross talk (Rawlings et al., 2004; Liu et al., 2010), we accurately found intermediate nodes such as JAK2, HES1, and HES5, but we failed in recovering their interactions with STAT3 in the PCSF-reconstructed pathway.

Our second case study is the glioblastoma (GBM) disease pathway. Disease-related pathways are mostly composed of multiple signaling pathways. GBM is the most aggressive type of brain cancer. Multiple signaling pathways such as the PI3K/AKT/mTOR, EGFR/RAS/MAPK, P53, and RB pathways have abnormal activity in GBM tumors (Ohgaki and Kleihues, 2007). Disease-related pathways are mostly composed of multiple signaling pathways. The presence of cross talk *via* intermediate molecules is the reason why multiple pathways are related to a disease. In this regard, signaling pathways in GBM, retrieved from WikiPathways, were reconstructed by multiple algorithms using HIPPIE as the reference interactome. Multiple signaling pathways such as the PI3K/AKT/mTOR, EGFR/RAS/MAPK, P53, and RB pathways are associated with GBM. Alterations on these pathways may lead to more aggressive and invasive phenotype by disturbing DNA repair, apoptosis, and G1/S progression and enhancing cell cycle progression and cell migration (Ohgaki and Kleihues, 2007). Some nodes such as PIK3CG and CDK1NA and their interactions, mediating the cross talk between multiple pathways, were not efficiently revealed by reconstruction algorithms. CDKN1A is responsible for the inhibition of the RB signaling pathway by transducing signals coming from the PI3K/AKT/mTOR pathway. Even though the reconstructed subnetwork recovers the RB signaling pathway, all four algorithms failed in reconstructing the edges connecting two signaling pathways (**Figure 8**). Thus, these algorithms are good at revealing the mediator nodes in cross talk between pathways, but they fail in revealing the connection between them. The HDF and PRF methods ranked some nodes as important, such as APOH, FBLN5, AFP, and MMP12. Although these proteins are not present in the studied pathway, their association with GBM was previously discovered in transcriptomic or proteomic studies (Varma Polisetty et al., 2012; Kros et al., 2015; Trojan et al., 2020).

DISCUSSION

In this study, we comprehensively explored the properties of interactomes from seven sources and the performance of four network reconstruction algorithms on known pathways. Our comparison reveals that PathwayCommons, having the highest number of nodes and edges, has the highest coverage of nodes and edges across all interactomes, including CDGs, and known

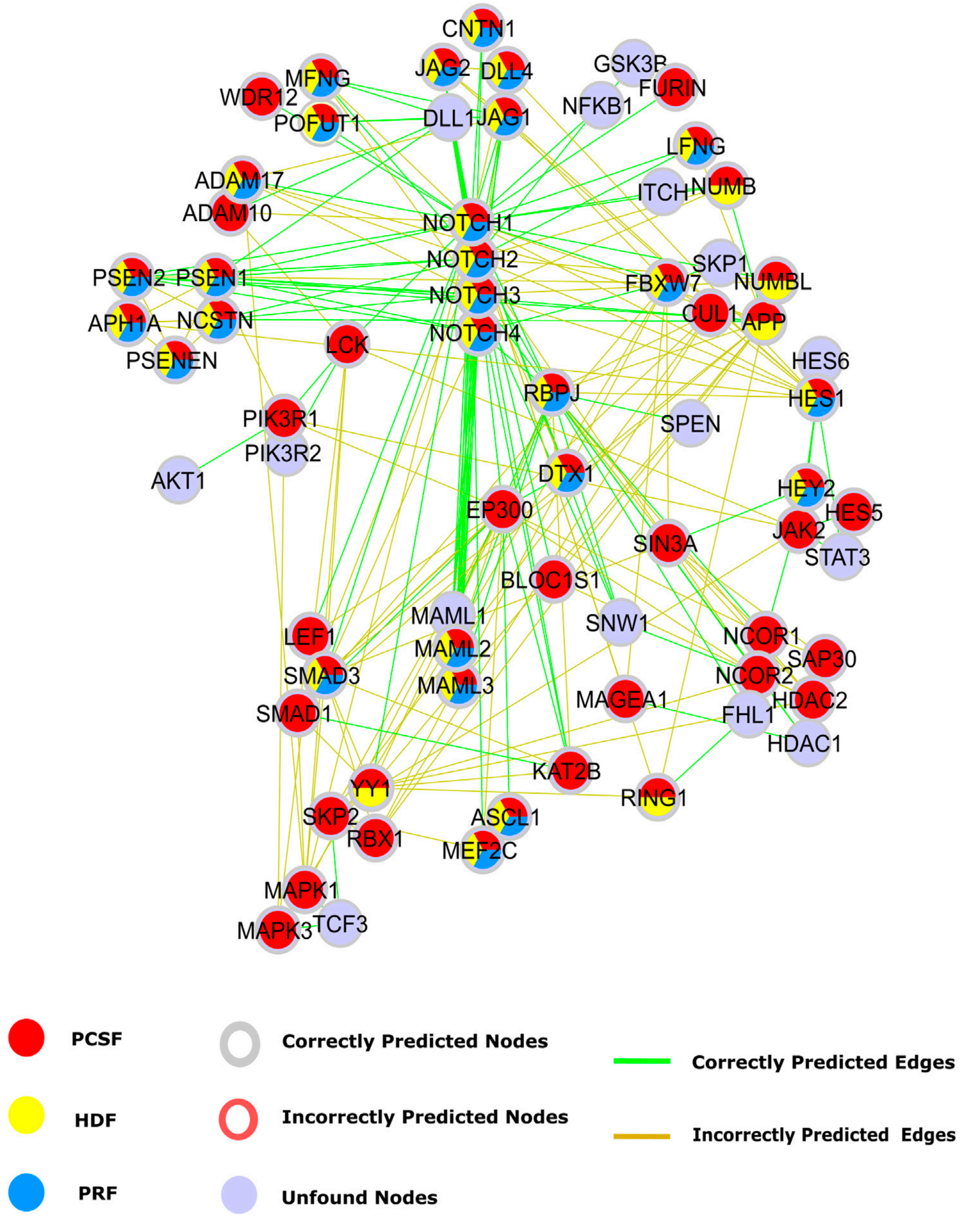


FIGURE 7 | Reconstructed Notch pathway. Nodes that are present in the pathway, but are not found by any algorithms are colored light blue. Nodes that are found by PCSF, PRF, and HDF are colored red, yellow, and cyan, respectively. Green edges are present in Notch pathway in NetPath, while incorrectly included edges by any algorithm are shown in brown.

pathways. However, precision values of the reconstruction methods are significantly lower than the others when PathwayCommons is used as the reference interactome. We did not observe a significant difference in recall values among all interactomes. The significant correlation between the degree and the number of publications of the nodes in PathwayCommons shows a bias toward well-studied proteins. Interestingly, although HIPPIE and ConsensusPathDB have a similar bias, the precision of the algorithms on these interactomes is better than that of PathwayCommons. These results imply that HIPPIE and ConsensusPathDB have a good balance in down-weighting the false positives and preserving high confidence edges.

The results of the different network reconstruction algorithms may include disjoint edges. The highest recall scores in APSP come along with the highest FPR score because the APSP algorithm adds many false-positive edges besides the true positives. Some studies, such as PathLinker (Ritz et al., 2016), use a distance threshold during shortest path calculation, a limited number of shortest paths between the source and the target, or additional data including orientation of the signal from the receptors to the transcription factors so that the false positive

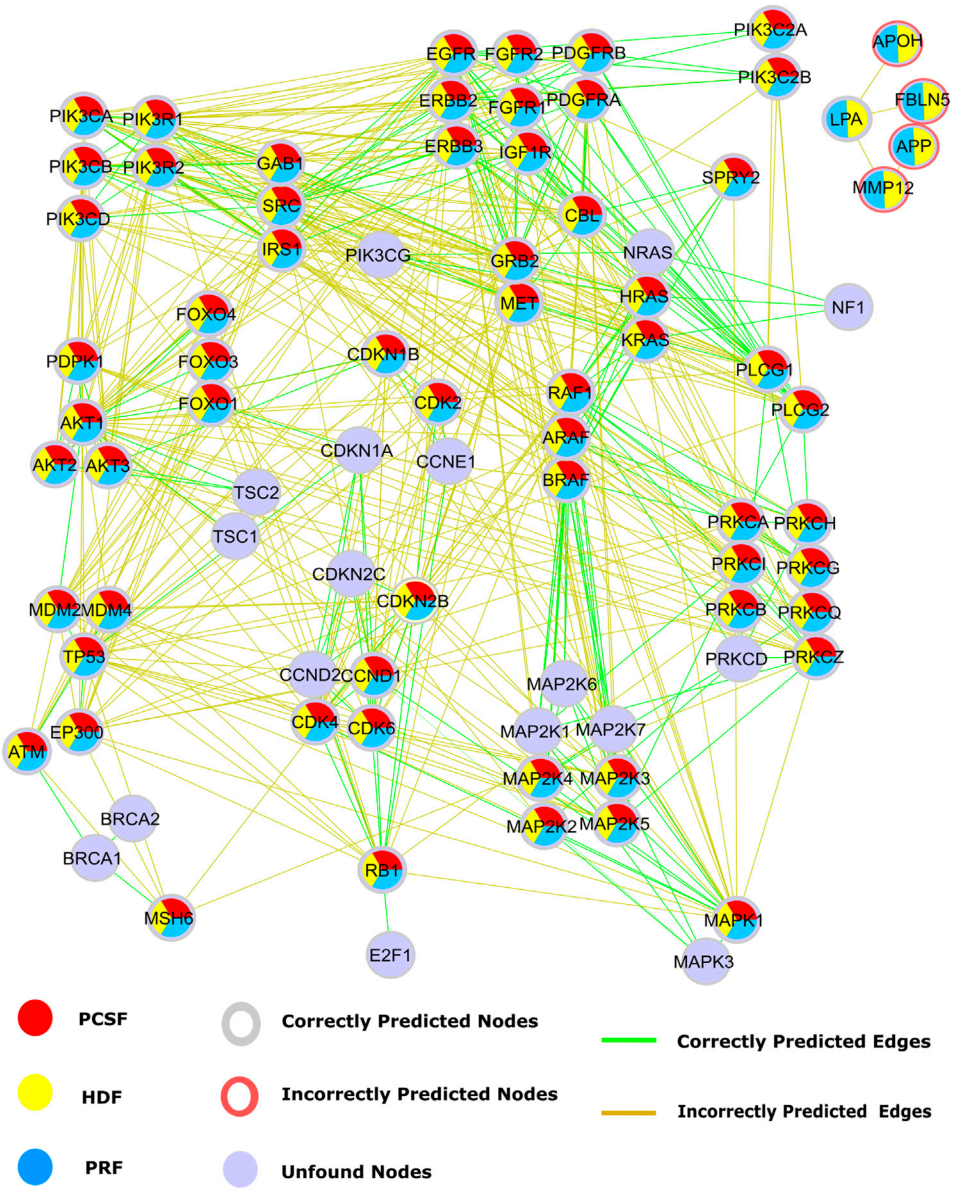


FIGURE 8 | Reconstructed signaling pathways in glioblastoma (WP2261) are predicted by PCSF are red, PRF are yellow, HDF are cyan, and nodes that are not found by any algorithm are colored light blue. Additionally, red rings represent the nodes that are incorrectly propagated. Green edges are the correct ones that are present in GBM pathway, while incorrect edges included by any algorithm are colored brown.

rate is controlled. We need to note that we did not apply any distance-based threshold, additional data, or refinement in the APSP algorithm. Thus, F1 scores and precision scores are extremely low in APSP. On the other hand, PRF, HDF, and PCSF have similar performances of false positive and true positive edges. PCSF has the highest F1 score compared to PRF and HDF. Interactomes are imbalanced datasets where true-negative edges are significantly more than true-positive edges. Naturally, the precision scores seem relatively low in the pathways formed by our algorithms since the FPR gets higher in such imbalanced datasets. The reconstructed Notch pathway shows that PCSF is better at finding weakly connected nodes. However, PCSF does not perform well in revealing the intermediate nodes and their edges achieving the cross talk between the Notch pathway and the PI3K-AKT-mTOR and JAK-STAT signaling pathways. Moreover, the intermediate nodes that links signaling pathways in GBM cannot construct completely true edges. In our study, the nodes are proteins; however, pathways may include small molecules and non-peptide nodes. Therefore, the reconstruction algorithms probably add false edges to include true terminals. The lack of some nodes in reference interactomes may be one of the reasons for the low precision scores.

Network reconstruction algorithms are highly dependent on topological properties and edge weights of the reference interactomes (Janjić and Pržulj, 2017; Liu et al., 2017). Among the evaluated approaches, the highest recall values are achieved by using the APSP algorithm together with the lowest precision values. The APSP algorithm adds many false-positive edges, besides the true positives. On the other hand, PRF, HDF, and PCSF have similar performances, while PCSF has a higher F1 score than PRF and HDF. High recall scores together with low precision scores are the result of the unbalanced data where the number of edges in the target pathway is dramatically lower than that in the rest of the interactome (Saito and Rehmsmeier, 2015). The low precision score with the moderate recall score is common among reconstruction algorithms of human signaling networks (Atias and Sharan, 2011; Ritz et al., 2016; Grimes et al., 2019). Additionally, edge-based performances of reconstruction algorithms are not as good as their node-based performance. We also observe a similar pattern of performances in our evaluation.

In a recent study, the performance of flux algorithms was shown to exceed the performance of PCSF with default parameters (Rubel and Ritz, 2020). However, the selected set of parameters significantly affects the performance of reconstruction algorithms, especially in PCSF. Automating parameter tuning that considers topological properties of reconstructed subnetworks can improve the performance (Magnano and Gitter, 2021). Therefore, in this study, we reconstructed pathways by extensively tuning the parameter set, followed by merging multiple optimal forests to reach the best performance. Parameter sets of other reconstruction algorithms were also tuned to find the optimum parameters. We can explain the overperformance of PCSF compared to other methods with detailed parameter tuning and considering multiple optimal solutions.

Several methods use topological properties of reference interactomes to predict new links and to filter out false-positive interactions (Cannistraci et al., 2013; Lei and Ruan, 2013; Hulovatyy et al., 2014; Alkan and Erten, 2017). Additionally, functional annotations, protein structures, and domain–domain interactions were also used to identify missing protein associations (Singh et al., 2006; Segura et al., 2015; Yerneni et al., 2018; Ietswaart et al., 2021). We need to note that we did not use the methods that modify the underlying interactome (Alanis-Lobato et al., 2018) and the methods that construct regulatory networks (Madar et al., 2009; Fontaine et al., 2011; Lachmann et al., 2016) in our evaluation. The performance of the APSP, HDF, PRF, and PCSF algorithms may change upon any modification or refinement of the reference interactomes. These reference interactomes are undirected graphs, but signaling pathways are intrinsically directed graphs. Indeed, the directionality of the edges can be incorporated either with the known or with the predicted ones. Orientation of the reconstructed networks can improve the mechanistic understanding of biological pathways. Therefore, using a directed reference interactome can boost the performance of each algorithm. Finally, biomolecular interactions are temporally and spatially diverse. Interactomes are incomplete sets of interactions, and the time dimension is not considered in our evaluation. Subnetwork reconstruction algorithms may be improved in the future to include biological annotations and temporal and spatial interactions of proteins.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Conceptualization: MA and NT. Data curation: MA. Formal analysis: MA and NT. Methodology: MA and NT. Project administration: NT. Supervision: NT. Visualization: MA.

FUNDING

NT has received support from the Career Development Program of TUBITAK under the project number 117E192. MA has been financially supported with the TUBITAK-2211 fellowship.

ACKNOWLEDGMENTS

NT acknowledges the support from the UNESCO-L'Oreal National for Women in Science Fellowship and the UNESCO-L'Oréal International Rising Talent Fellowship and TUBA-GEBIP.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.666705/full#supplementary-material

REFERENCES

Ahmed, R., Baali, I., Erten, C., Hoxha, E., and Kazan, H. (2020). MEXCOwalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules. *Bioinformatics* 36 (3), 872–879. doi:10.1093/bioinformatics/btz655

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing Meaningfulness and Reliability of Protein-Protein Interaction Networks. *Nucleic Acids Res.* 45, D408–D414. doi:10.1093/nar/gkw985

Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. (2018). The Latent Geometry of the Human Protein Interaction Network. *Bioinformatics* 34 (16), 2826–2834. doi:10.1093/bioinformatics/bty206

Alkan, F., and Erten, C. (2017). RedNemo: Topology-Based PPI Network Reconstruction via Repeated Diffusion with Neighborhood Modifications. *Bioinformatics* 33 (4), btw655–544. doi:10.1093/bioinformatics/btw655

Alm, J. F., and Mack, K. M. L. (2016). Degree-correlation, Robustness, and Vulnerability in Finite Scale-free Networks. *Asian Res. J. Maths.* 2 (5), 1–6. http://arxiv.org/abs/1606.08768.

Atias, N., and Sharan, R. (2011). An Algorithmic Framework for Predicting Side Effects of Drugs. *J. Comput. Biol.* 18 (3), 207–218. doi:10.1089/cmb.2010.0255

Azpeitia, E., Balanzario, E. P., and Wagner, A. (2020). Signaling Pathways Have an Inherent Need for Noise to Acquire Information. *BMC Bioinformatics* 21 (1). doi:10.1186/s12859-020-03778-x

Baali, I., Erten, C., and Kazan, H. (2020). DriveWays: a Method for Identifying Possibly Overlapping Driver Pathways in Cancer. *Sci. Rep.* 10 (1), 1–14. doi:10.1038/s41598-020-78852-8

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173 (2), 371–e18. doi:10.1016/j.cell.2018.02.060

Barabási, A.-L., and Albert, R. (1995). Emergence of Scaling in Random Networks. *Mat. Res. Soc. Symp. Proc.* 286, 509. doi:10.1126/science.286.5439.509

Bazzoni, R., and Bentivegna, A. (2019). Role of Notch Signaling Pathway in Glioblastoma Multiforme Pathogenesis. *Cancers* 11 (3), 292. doi:10.3390/cancers11030292

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLOS ONE* 12 (6), e0177678. doi:10.1371/journal.pone.0177678

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., et al. (2009). An Experimentally Derived Confidence Score for Binary Protein-Protein Interactions. *Nat. Methods* 6 (1), 91–97. doi:10.1038/nmeth.1281

Braunstein, A., Ingrosso, A., and Muntoni, A. P. (2019). Network Reconstruction from Infection Cascades. *J. R. Soc. Interf.* 16 (151), 20180844. doi:10.1098/rsif.2018.0844

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013). Minimum Curvilinearity to Enhance Topological Prediction of Protein Interactions by Network Embedding. *Bioinformatics* 29, i199–209. doi:10.1093/bioinformatics/btt208

Caraus, I., Alsuwailem, A. A., Nadon, R., and Makarenkov, V. (2015). Detecting and Overcoming Systematic Bias in High-Throughput Screening Technologies: a Comprehensive Review of Practical Issues and Methodological Solutions. *Brief. Bioinform.* 16 (6), 974–986. doi:10.1093/bib/bbv004

Ceccarelli, F., Turei, D., Gabor, A., and Saez-Rodriguez, J. (2020). Bringing Data from Curated Pathway Resources to Cytoscape with OmniPath. *Bioinformatics* 36 (8), 2632–2633. doi:10.1093/bioinformatics/btz968

Chan, S. M., Weng, A. P., Tibshirani, R., Aster, J. C., and Utz, P. J. (2007). Notch Signals Positively Regulate Activity of the mTOR Pathway in T-Cell Acute Lymphoblastic Leukemia. *Blood* 110 (1), 278–286. doi:10.1182/blood-2006-08-039883

Chen, Z., Oh, D., Dubey, A. K., Yao, M., Yang, B., Groves, J. T., et al. (2018). EGFR Family and Src Family Kinase Interactions: Mechanics Matters? *Curr. Opin. Cel Biol.* 51, 97–102. doi:10.1016/j.ceb.2017.12.003

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network Propagation: A Universal Amplifier of Genetic Associations. *Nat. Rev. Genet.* 18 (9), 551–562. doi:10.1038/nrg.2017.38

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Robertson, G., et al. (2013). Comprehensive Molecular Characterization of clear Cell Renal Cell Carcinoma. *Nature* 499 (7456), 43–49. doi:10.1038/nature12222

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., et al. (2015). Pathway and Network Analysis of Cancer Genomes. *Nat. Methods* 12 (7), 615–621. doi:10.1038/nmeth.3440

Dincer, C., Kaya, T., Keskin, O., Gursoy, A., and Tuncbag, N. (2019). 3D Spatial Organization and Network-Guided Comparison of Mutation Profiles in Glioblastoma Reveals Similarities across Patients. *Plos Comput. Biol.* 15 (9), e1006789. doi:10.1371/journal.pcbi.1006789

Fontaine, J.-F., Priller, F., Barbosa-Silva, A., and Andrade-Navarro, M. A. (2011). Génie: Literature-Based Gene Prioritization at Multi Genomic Scale. *Nucleic Acids Res.* 39 (Suppl. 2), W455–W461. doi:10.1093/nar/gkr246

Grimes, T., Potter, S. S., and Datta, S. (2019). Integrating Gene Regulatory Pathways into Differential Network Analysis of Gene Expression Data. *Sci. Rep.* 9 (1). doi:10.1038/s41598-019-41918-3

Guo, J., Li, P., Liu, X., and Li, Y. (2019). NOTCH Signaling Pathway and Non-coding RNAs in Cancer. *Pathol. Res. Pract.* 215 (11), 152620. doi:10.1016/j.prp.2019.152620

Hicks, M., Bartha, I., Di Iulio, J., Venter, J. C., and Telenti, A. (2019). Functional Characterization of 3D Protein Structures Informed by Human Genetic Diversity. *Proc. Natl. Acad. Sci. USA* 116 (18), 8960–8965. doi:10.1073/pnas.1820813116

Hillmann, P., and Fabbro, D. (2019). PI3K/mTOR Pathway Inhibition: Opportunities in Oncology and Rare Genetic Diseases. *Int. J. Mol. Sci.* 20 (22), 5792. doi:10.3390/ijms20225792

Huang, L., Brunell, D., Stephan, C., Mancuso, J., Yu, X., He, B., et al. (2019). Driver Network as a Biomarker: Systematic Integration and Network Modeling of Multi-Omics Data to Derive Driver Signaling Pathways for Drug Combination Prediction. *Bioinformatics* 35 (19), 3709–3717. doi:10.1093/bioinformatics/btz109

Hulovatyy, Y., Solava, R. W., and Milenković, T. (2014). Revealing Missing Parts of the Interactome via Link Prediction. *PLoS ONE* 9 (3), e90073. doi:10.1371/journal.pone.0090073

Ietswaart, R., Gyori, B. M., Bachman, J. A., Sorger, P. K., and Churchman, L. S. (2021). GeneWalk Identifies Relevant Gene Functions for a Biological Context Using Network Representation Learning. *Genome Biol.* 22 (1), 55. doi:10.1186/s13059-021-02264-8

Janjić, V., and Pržulj, N. (2017). The Topology of the Growing Human Interactome Data. *J. Integr. Bioinformatics* 11 (2), 27–42. doi:10.1515/jib-2014-238

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48 (D1), D498–D503. doi:10.1093/nar/gkz1031

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: Toward a More Complete Picture of Cell Biology. *Nucleic Acids Res.* 39 (Suppl. 1), D712–D717. doi:10.1093/nar/gkq1156

Kamburov, A., Stelzl, U., and Herwig, R. (2012). IntScore: A Web Tool for Confidence Scoring of Biological Interactions. *Nucleic Acids Res.* 40 (W1), W140–W146. doi:10.1093/nar/gks492

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB Interaction Database: 2013 Update. *Nucleic Acids Res.* 41 (D1), D793–D800. doi:10.1093/nar/gks1055

Kandasamy, K., Mohan, S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., et al. (2010). NetPath: A Public Resource of Curated Signal Transduction Pathways. *Genome Biol.* 11 (1), R3. doi:10.1186/gb-2010-11-1-r3

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092

Kim, Y.-A., Cho, D.-Y., Dao, P., and Przytycka, T. M. (2015). MEMCover: Integrated Analysis of Mutual Exclusivity and Functional Network Reveals Dysregulated Pathways across Multiple Cancer Types. *Bioinformatics* 31 (12), i284–i292. doi:10.1093/bioinformatics/btv247

Koh, H. W. L., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iOmicsPASS: Network-Based Integration of Multiomics Data for Predictive Subnetwork Discovery. *Npj Syst. Biol. Appl.* 5 (1), 22. doi:10.1038/s41540-019-0099-y

Kros, J. M., Huizer, K., Hernández-Laín, A., Marucci, G., Michotte, A., Pollo, B., et al. (2015). Evidence-based Diagnostic Algorithm for Glioma: Analysis of the Results of Pathology Panel Review and Molecular Parameters of EORTC 26951 and 26882 Trials. *J. Clin. Oncol.* 33 (17), 1943–1950. doi:10.1200/JCO.2014.59.0166

Kuzmin, K., Gaiteri, C., and Szymanski, B. K. (2016). Synergy Landscapes: A Multilayer Network for Collaboration in Biological Research. *Adv. Netw. Sci.* 9564, 205–212. doi:10.1007/978-3-319-28361-6_18

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016). ARACNe-AP: Gene Network Reverse Engineering through Adaptive Partitioning Inference of Mutual Information. *Bioinformatics* 32 (14), 2233–2235. doi:10.1093/bioinformatics/btw216

Langville, A. N., and Meyer, C. D. (2005). A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Rev.* 47 (1), 135–161. doi:10.1137/S0036144503424786

Lei, C., and Ruan, J. (2013). A Novel Link Prediction Algorithm for Reconstructing Protein-Protein Interaction Networks by Topological Similarity. *Bioinformatics* 29 (3), 355–364. doi:10.1093/bioinformatics/bts688

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *Nat. Genet.* 47 (2), 106–114. doi:10.1038/ng.3168

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y.-C., Cheng, F., et al. (2020). Computational Network Biology: Data, Models, and Applications. *Phys. Rep.* 846, 1–66. doi:10.1016/j.physrep.2019.12.004

Liu, G., Wang, H., Chu, H., Yu, J., and Zhou, X. (2017). Functional Diversity of Topological Modules in Human Protein-Protein Interaction Networks. *Sci. Rep.* 7 (1), 16199. doi:10.1038/s41598-017-16270-z

Liu, W., Singh, S. R., and Hou, S. X. (2010). JAK-STAT Is Restrained by Notch to Control Cell Proliferation of theDrosophilaintestinal Stem Cells. *J. Cel. Biochem.* 109 (5), a–n. doi:10.1002/jcb.22482

Madar, A., Greenfield, A., Ostrer, H., Vanden-Eijnden, E., and Bonneau, R. (2009). The Inferelator 2.0: A Scalable Framework for Reconstruction of Dynamic Regulatory Network Models. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2009, 5448–5451. doi:10.1109/IEMBS.2009.5334018

Magnano, C. S., and Gitter, A. (2021). Automating Parameter Selection to Avoid Implausible Biological Pathway Models. *Npj Syst. Biol. Appl.* 7 (1), 1–12. doi:10.1038/s41540-020-00167-1

Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemingway, H., Ketteler, R., et al. (2019). Towards a Data-Integrated Cell. *Nat. Commun.* 10 (1), 805. doi:10.1038/s41467-019-08797-8

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A Compendium of Mutational Cancer Driver Genes. *Nat. Rev. Cancer* 20, 555–572. doi:10.1038/s41568-020-0290-x

Meyer, M. J., Beltrán, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., et al. (2018). Interactome INSIDER: a Structural Interactome Browser for Genomic Studies. *Nat. Methods* 15 (2), 107–114. doi:10.1038/nmeth.4540

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A Fully Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-type Omics Data. *Biostatistics* 19 (1), 71–86. doi:10.1093/biostatistics/kxx017

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* 10 (1), 47–53. doi:10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: A Catalog of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucl. Acids Res.* 42 (D1), D374–D379. doi:10.1093/nar/gkt887

Nero, T. L., Parker, M. W., and Morton, C. J. (2018). Protein Structure and Computational Drug Discovery. *Biochem. Soc. Trans.* 46 (5), 1367–1379. doi:10.1042/BST20180202

Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B., and Moreau, Y. (2010). Candidate Gene Prioritization by Network Analysis of Differential Expression Using Machine Learning Approaches. *BMC Bioinformatics* 11, 460. doi:10.1186/1471-2105-11-460

Ohgaki, H., and Kleihues, P. (2007). Genetic Pathways to Primary and Secondary Glioblastoma. *Am. J. Pathol.* 170 (5), 1445–1453. doi:10.2353/ajpath.2007.070011

Paananen, J., and Fortino, V. (2020). An Omics Perspective on Drug Target Discovery Platforms. *Brief. Bioinform.* 21 (6), 1937–1953. doi:10.1093/bib/bbz122

Page, L. B., Brin, S., Motwani, R., and Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web*.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering Causal Pathways Linking Genomic Events to Transcriptional States Using Tied Diffusion through Interacting Events (TieDIE). *Bioinformatics* 29 (21), 2757–2764. doi:10.1093/bioinformatics/btt471

Porras, P., Barrera, E., Bridge, A., del-Toro, N., Cesareni, G., Duesbury, M., et al. (2020). Towards a Unified Open Access Dataset of Molecular Interactions. *Nat. Commun.* 11 (1), 1–12. doi:10.1038/s41467-020-19942-z

Qiu, H., Tang, X., Ma, J., Shaverdashvili, K., Zhang, K., and Bedogni, B. (2015). Notch1 Autoactivation via Transcriptional Regulation of Furin, Which Sustains Notch1 Signaling by Processing Notch1-Activating Proteases ADAM10 and Membrane Type 1 Matrix Metalloproteinase. *Mol. Cel Biol.* 35 (21), 3622–3632. doi:10.1128/mcb.00116-15

Rawlings, J. S., Rosler, K. M., and Harrison, D. A. (2004). The JAK/STAT Signaling Pathway. *J. Cel Sci.* 117 (8), 1281–1283. doi:10.1242/jcs.00963

Ricotta, C., Podani, J., and Pavoine, S. (2016). A Family of Functional Dissimilarity Measures for Presence and Absence Data. *Ecol. Evol.* 6 (15), 5383–5389. doi:10.1002/ece3.2214

Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N., Simmons, K., Powell, A., et al. (2016). Pathways on Demand: Automated Reconstruction of Human Signaling Networks. *Npj Syst. Biol. Appl.* 2 (1), 1–9. doi:10.1038/npjsba.2016.2

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., et al. (2019). Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data. *Nucleic Acids Res.* 48, 489–497. doi:10.1093/nar/gkz946

Rubel, T., and Ritz, A. (2020). Augmenting Signaling Pathway Reconstructions. *Proc. 11th ACM Int. Conf. Bioinformatics Comput. Biol. Health Inform.* 10, 1–10. doi:10.1145/3388440.3412411

Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative Than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* 10 (3), e0118432. doi:10.1371/journal.pone.0118432

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). Hippie: Integrating Protein Interaction Networks with experiment Based Quality Scores. *PLoS ONE* 7 (2), e31826. doi:10.1371/journal.pone.0031826

Schaefer, M. H., Serrano, L., and Andrade-Navarro, M. A. (2015). Correcting for the Study Bias Associated with Protein-Protein Interaction Measurements Reveals Differences between Protein Degree Distributions from Different Cancer Types. *Front. Genet.* 6, 260. doi:10.3389/fgene.2015.00260

Schmidt, T., Bergner, A., and Schwede, T. (2014). Modelling Three-Dimensional Protein Structures for Applications in Drug Design. *Drug Discov. Today* 19 (7), 890–897. doi:10.1016/j.drudis.2013.10.027

SeahSen, C. S., Kasim, S., Fudzee, M. F. M., Law Tze Ping, J. M., Mohamad, M. S., Saedudin, R. R., et al. (2017). An Enhanced Topologically Significant Directed Random Walk in Cancer Classification Using Gene Expression Datasets. *Saudi J. Biol. Sci.* 24 (8), 1828–1841. doi:10.1016/j.sjbs.2017.11.024

Segura, J., Sorzano, C. O. S., Cuenca-Alba, J., Aloy, P., and Carazo, J. M. (2015). Using Neighborhood Cohesiveness to Infer Interactions between Protein Domains. *Bioinformatics* 31 (15), 2545–2552. doi:10.1093/bioinformatics/btv188

Sevimoglu, T., and Arga, K. Y. (2014). The Role of Protein Interaction Networks in Systems Biomedicine. *Comput. Struct. Biotechnol. J.* 11 (18), 22–27. doi:10.1016/j.csbj.2014.08.008

Silverbush, D., Cristea, S., Yanovich-Arad, G., Geiger, T., Beerenwinkel, N., and Sharan, R. (2019). Simultaneous Integration of Multi-Omics Data Improves the Identification of Cancer Driver Modules. *Cel Syst.* 8 (5), 456–466.e5. doi:10.1016/j.cels.2019.04.005

Simpson, G. (1966). Notes on the Measurement of Faunal Resemblance. *Am. J. Sci.* 258-A, 300–311. http://earth.geology.yale.edu/~ajs/1960/ajs_258A_11.pdf/300.pdf.

Singh, R., Xu, J., and Berger, B. (2005). Struct2Net: Integrating Structure into Protein-Protein Interaction Prediction. *Pac. Symp. Biocomput* 2006, 403–414. doi:10.1142/9789812701626_0037

Sjölund, J., Manetopoulos, C., Stockhausen, M.-T., and Axelson, H. (2005). The Notch Pathway in Cancer: Differentiation Gone Awry. *Eur. J. Cancer* 41 (17), 2620–2629. doi:10.1016/j.ejca.2005.06.025

Skinnider, M. A., Stacey, R. G., Foster, L. J., and Iakoucheva, L. M. (2018). Genomic Data Integration Systematically Biases Interactome Mapping. *Plos Comput. Biol.* 14, e1006474. doi:10.1371/journal.pcbi.1006474

Sychev, Z. E., Hu, A., DiMaio, T. A., Gitter, A., Camp, N. D., Noble, W. S., et al. (2017). Integrated Systems Biology Analysis of KSHV Latent Infection Reveals Viral Induction and reliance on Peroxisome Mediated Lipid Metabolism. *Plos Pathog.* 13 (3), e1006256. doi:10.1371/journal.ppat.1006256

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tabei, Y., Kotera, M., Sawada, R., and Yamanishi, Y. (2019). Network-based Characterization of Drug-Protein Interaction Signatures with a Space-Efficient Approach. *BMC Syst. Biol.* 13 (S2), 39. doi:10.1186/s12918-019-0691-1

The UniProt Consortium (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506. doi:10.1093/nar/gky1049

Tkačik, G., Walczak, A. M., and Bialek, W. (2009). Optimizing Information Flow in Small Genetic Networks. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 80 (3), 1–18. doi:10.1103/PhysRevE.80.031920

Trojan, A., Kasprzak, H., Gutierrez, O., Penagos, P., Briceno, I., O. Siachoque, H., et al. (2020). "Neoplastic Brain, Glioblastoma, and Immunotherapy," in *Brain and Spinal Tumors - Primary and Secondary* (The Shard: IntechOpen). doi:10.5772/intechopen.84726

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., et al. (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner forest Problem. *J. Comput. Biol.* 20 (2), 124–136. doi:10.1089/cmb.2012.0092

Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016a). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *Plos Comput. Biol.* 12 (4), e1004879. doi:10.1371/journal.pcbi.1004879

Tuncbag, N., McCallum, S., Huang, S.-s. C., and Fraenkel, E. (2012). SteinerNet: a Web Server for Integrating 'omic' Data to Discover Hidden Components of Response Pathways. *Nucleic Acids Res.* 40 (W1), W505–W509. doi:10.1093/nar/gks445

Tuncbag, N., Milani, P., Pokorny, J. L., Johnson, H., Sio, T. T., Dalin, S., et al. (2016b). Network Modeling Identifies Patient-specific Pathways in Glioblastoma. *Sci. Rep.* 6 (1), 1–12. doi:10.1038/srep28668

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2011). Interaction Databases on the Same page. *Nat. Biotechnol.* 29, 391. doi:10.1038/nbt.1867

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: Interactive Analysis of Consolidated Protein Interaction Data and Their Supporting Evidence. *Database (Oxford)* 2010, baq023. doi:10.1093/database/baq023

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for Detecting Significantly Mutated Pathways in Cancer. *J. Comput. Biol.* 18 (3), 507–522. doi:10.1089/cmb.2010.0265

Varma Polisetty, R., Gautam, P., Sharma, R., Harsha, H. C., Nair, S. C., Kumar Gupta, M., et al. (2012). LC-MS/MS Analysis of Differentially Expressed Glioblastoma Membrane Proteome Reveals Altered Calcium Signalling and Other Protein Groups of Regulatory Functions Running Title-Glioblastoma Membrane Proteins. Available at: https://www.mcponline.org.

Venko, K., Roy Choudhury, A., and Novič, M. (2017). Computational Approaches for Revealing the Structure of Membrane Transporters: Case Study on Bilitranslocase. *Comput. Struct. Biotechnol. J.* 15, 232–242. doi:10.1016/j.csbj.2017.01.008

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell* 144 (6), 986–998. doi:10.1016/j.cell.2011.02.016

Vitali, F., Marini, S., Pala, D., Demartini, A., Montoli, S., Zambelli, A., et al. (2018). Patient Similarity by Joint Matrix Trifactorization to Identify Subgroups in Acute Myeloid Leukemia. *JAMIA Open* 1 (1), 75–86. doi:10.1093/jamiaopen/ooy008

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2004). STRING: Known and Predicted Protein-Protein Associations, Integrated and Transferred across Organisms. *Nucleic Acids Res.* 33 (DATABASE ISS.), D433–D437. doi:10.1093/nar/gki005

Waks, Z., Weissbrod, O., Carmeli, B., Norel, R., Utro, F., and Goldschmidt, Y. (2016). Driver Gene Classification Reveals a Substantial Overrepresentation of Tumor Suppressors Among Very Large Chromatin-Regulating Proteins. *Sci. Rep.* 6 (1), 1–12. doi:10.1038/srep38988

Wang, Y., Yang, Y., Chen, S., and Wang, J. (2021). DeepDRK: a Deep Learning Framework for Drug Repurposing through Kernel-Based Multi-Omics Integration. *Brief. Bioinform.* 00 (August 2020), 1–10. doi:10.1093/bib/bbab048

Yerneni, S., Khan, I. K., Wei, Q., and Kihara, D. (2018). IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks. *Ieee/acm Trans. Comput. Biol. Bioinf.* 15 (4), 1247–1258. doi:10.1109/tcbb.2015.2476809

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering Disease-Disease Associations by Fusing Systems-Level Molecular Data. *Scientific Rep.* 3 (1), 1–9. doi:10.1038/srep03202

Zsákai, L., Sipos, A., Dobos, J., Erős, D., Szántai-Kis, C., Bánhegyi, P., et al. (2019). Targeted Drug Combination Therapy Design Based on Driver Genes. *Oncotarget* 10 (51), 5255–5266. doi:10.18632/oncotarget.26985

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership