Atomic-Level Binding Affinity Estimation via Physics-Informed Neural Networks: AI-Enhanced Computational Methods for Protein-Ligand Interaction Prediction

Dr. Matej Rojc, Professor of Computer Science, University of Ljubljana, Slovenia

1. Introduction

Predicting the binding affinity, preferred orientation, and stereochemistry of the complex between a protein receptor and a small-molecule ligand is a critical step in structure-driven drug discovery and development. Traditionally, this has been achieved using computational methods that explicitly treat the electronic structure of the atoms in the model. A variety of challenges have been encountered in the process. This introductory section reviews how combining artificial intelligence and computational means could alleviate some of these problems. We are primarily concerned with the revision of the role that quantum and molecular mechanical scoring functions play, and the consequences when AI strategies are used to leverage them. The value of new methodology in terms of feasibility in silico and cost avoidance is analyzed, especially in the context of rare events like drug-induced liver injury or abnormal heart rhythms, which often do not occur until drugs enter a clinical population numbering in the tens of thousands.

There is a growing interest in leveraging machine learning techniques to solve biophysical problems. Leaderboards tracking shared tasks demonstrate overall improvement over many benchmarking databases. The success of AI body lixiviation thus seems to surpass that of QM, at least on the short concept recognition. More importantly, advances in machine learning in recent years mean that these sorts of systems are now within the reach of most computational research labs that have the computing resources to be active in a drug design context. The essay does not give a comprehensive overview of systems for LM scores – not even beyond AI-enhanced and hybrid methods.

1.1. Background and Significance

Historical context and abstractions about protein-ligand interactions The interactions between proteins, the primary functional molecules in living organisms, and other molecules are called ligand interactions. Proteins mediate specific biological processes through successive alternative ligand binding and the subsequent disaggregation of ligand-receptor subunit complexes. There is a current consensus about the importance of the predictability of protein-ligand binding in drug-receptor and antigen-antibody-based immunogenicity studies. Specifically, protein-ligand binding is related to adverse effects that may occur after the alteration of a protein via a biological drug or the off-target effects with small molecule drugs. The combination of the atomic-level structure of proteins and the structures of their ligands is called the individual dataset. A variety of atomic-level structures hidden in these datasets may provide a more comprehensive and detailed description of the interactions between proteins and ligands. The bioactivity of the ligand is usually verified through an assay; the merits of positive and negative bioactivity are abstracted as non-bioactive and bioactive, respectively. The combination of protein targets and their bioactive ligands constitutes a helpful small-molecule pharmacology dataset. The experimental determination of protein-ligand binding events is labor-intensive and time-consuming. Computational techniques have been adopted to predict protein-ligand interactions rather than traditional screening. Structure-based drug design, a computer-aided drug design strategy, has been used for more than a decade to modernize drug development. Artificial intelligence, which is also another term for machine learning, has inspired a transformation in various fields. The use of AI in computational biophysics has allowed for a shift from traditional computational biophysics into the fourth era in silico medicine. AI and other computational techniques have been linked in a variety of ways to extract knowledge from big data. Proteinographic knowledge of the ligand, protein, the cellular environment, the concentration of reactants, solvent, pH, salt, temperature, pressure, and so on has been accumulated. An understanding of proteins requires a multidisciplinary approach that describes proteins in terms of biology, mathematics, computing, and statistics. In addition, drug discovery is usually meaningful to show basic results, such as expected binding free energies and the final stable structures of the protein-ligand complexes; for a large-scale first screening, such accurate results are not always necessary. Therefore, a suitable multidisciplinary solution should be able to

obtain the required overall knowledge in a trade-off between investment time, effort payback, system complexity, and number of sample applications for parameter restraints.

1.2. Scope and Objectives

While the key aim of this review is to outline the recent trend in the development of AI-driven computational methods and differentiate them from conventional computational methods based on a theoretical framework, it also includes a brief section on the theoretical framework. Currently, the bottleneck for the application of AI technologies to drug design is user-unfriendly, and it is difficult for a researcher who is not familiar with AI technologies to evaluate synergistic effects between drugs and obtain information that can be translated into understanding at the molecular level. By integrating all these considerations, the scope of this review is focused on the theoretical concept, principal algorithms, descriptive performance, and applications of AI methodologies. In essence, this review provides a comprehensive and systematic assessment of AI-based techniques in the study of protein-ligand interactions, an expanded explanation of theoretical knowledge, and the comparative descriptive quantitative statistical performance of existing models. This review is intended to provide a deep analysis of the principal methodologies in AI-driven techniques and their specific applications, standardization, performance evaluation, and novelty. Furthermore, I am honored to say that the idea has the potential to provide new insight into this subject matter and that the review may contain particular rules, ideas, and problems of AI-enhanced computational methods for predicting protein-ligand interactions. Moreover, the discovered knowledge could present the boundaries of AI-for-PLI studies.

2. Foundations of Protein-Ligand Interactions

Forge does require a moment's study of some basic biological principles, and so we provide a brief overview. The protein-ligand concept is based on the structural and biochemical foundations, which determine protein-ligand recognition and association, and there are many excellent comprehensive reviews of supramolecular biochemistry that describe the process of molecular recognition. In general terms, proteins bind specifically and nonspecifically to hydrophobic and electrostatic surfaces, and all such binding events depend upon the identity (electron distribution) and surroundings

(solvent versus hydrophobic interior) of amino acid residues whose limitation of motion, or rigidity with internal conformational flexibility, determine the characteristic way that each protein structure binds ligands. Seven different classes of protein structures are known, each with unique functions in biological systems.

Non-specific Interactions of Protein with Ligand: Seven classes of domain structures of proteins have been classified. The criteria invoke methods to describe the structure with solvent or hydrophobic interiors that control interaction properties, creating unique functions and structures. The nature of a ligand is determined by the strength of the intermolecular forces that hold the structure (shape) of the molecule together, which create, in turn, the unique ability to carry out a particular function. Structure, then, is related to function manifest initially by ligand binding, and although many properties of a ligand determine the strength of binding to a protein, we will consider a few of these structure-property correlations on the next page.

2.1. Protein Structure and Function

The machinery of a living cell is a vast collection of specialized macromolecules. These molecules, the basic building blocks of life, perform a wide variety of functions, from powering cellular processes to generating cellular structure. Proteins are a class of macromolecules that encode functional information. As well as the potential for great complexity, at the most basic level of organization, a protein is made up of strings of amino acids that were once the genetic code. The three-dimensional arrangement of those amino acids, which classify its biological function and biochemical behavior, is a protein's primary determinant. Proteins are also the macromolecules that have the power to determine the level of molecular specificity with which an enzyme or other functional protein will react with other molecules. The interaction between proteins and other macromolecules – nucleic acids, other proteins, and small molecules – is often transient, but highly controlled. The catalytic activity of enzymes is the best characterized example of this, but it is by no means the only one; protein machines, gated pores, and molecular-level conformational switches change. The assembly of a protein complex from component pieces is a critical function that some proteins provide. Mutations or chemical modifications can disrupt these abilities and lead to disease. Of particular therapeutic relevance, therefore, are the interactions between proteins and small, drug-like molecules at chemical sites within the protein known as binding sites. In

this manner, drugs can be developed to block the metabolism, biological membrane, molecular transportation, and biochemical signal transduction mediated by proteins. The binding pockets are generally parts of protein structure, but the types of them vary with the proteins. Additional important features of the structure include the physical and electrostatic properties of intermolecular contacts and the manner in which the surface area and volume of the pocket establish its physicochemical properties and accessibility to the solvent. Since the protein structures and the shape of the ligand must match, the protein active states or conformational changes should be regarded. Since the side chains of residues near the active site are often disordered, the side chains of flexible residues that form the binding pocket should also be included. In order to clarify proteins' roles in cellular processes, information on structural features must be analyzed. Because both point mutations and deletions/insertions of amino acids can alter protein 3D structures at the atomic and binding region levels, they should be included in the analysis. Considerations about structural variation and standardization are important for understanding cellular processes, pathogenicity, and reviewing drug trials. In addition, the structural variation of ligand-binding sites should be considered in combinatorial diversity.

2.2. Ligand Binding

Proteins are versatile in that they can bind with a wide variety of ligands, including both polar and nonpolar small and large molecules. Some of the key points that promote specific interaction are the polarity of the binding site with adjacent functional groups, the complementary aspect, and the symmetry of the reactive functional groups of both protein and ligand. There are inner forces that control the combination between protein and ligand, such as van der Waals interactions, hydrogen bonding, and hydrophobic packing. The binding process could be altered by factors like ionic strength, temperature, ratio of ligand or protein, and pH. When a ligand combines with a protein, this is an intermolecular process that affects the laws of thermodynamics. Ligand binding is rapid, but this combination process is reversible to some degree. Some ligands and receptors bind at a speed approaching the diffusion-controlled limit. The interactions that occur in ligand binding with proteins are very specific and have high affinity. Affinity between proteins and ligands is defined as the ability of a protein binding site to bind to a ligand. In pharmacology, the affinity is also called the strength

of a bond or the bond-dissociation number. Affinity values may be evaluated at equilibrium. This is carried out by means of inhibition practices.

Ligand binding with the protein can also cause the ligand to massively change form. The ligand can have – in essence – no conformational change, but proteins might have changed significantly. Transformations are reversible. After the binding of the ligand to the protein, the shape of the ligand could also change. The characteristics of the ligand change during reversible reactions. In theory, a method for ligand binding with a particular protein might be described by induced-fit or lock-and-key theories. According to the lock-and-key hypothesis, the ligand matches precisely into the form of the pocket of a protein without modifications in the conformation of the protein taking place. According to the induced-fit hypothesis, ligands perfectly match into a three-dimensional site of a protein by changes in the protein conformation. Proteins that have the highest fit are the ligands that generally show the most powerful binding affinity. An association between ligands and proteins is due to the position and the type of ions surrounding the binding site. Both reduced and increased affinity through the ions can appear. Some ligands fit very closely with the protein binding site, while others only fit close to it. In the case of high affinity between a ligand and a particular protein, the conformation of the ligand binding domain can be reshaped dramatically. The changes in conformation also occur with intermediates in the reversible reactions. When modeling interactions, it is important to understand the changes in conformation. Despite the profound theoretical nature, both models provide unique insights that help researchers better comprehend the importance of protein-ligand interactions. Overall, an understanding of ligand binding in biological systems is vital and can be put into use for pharmacologists and chemists to design medications, as well as to develop and increase memory and transportation in next-generation electro-mechanical systems. Such knowledge can revolutionize industries in the future. Protein-ligand binding is essential, and therefore progress in the computational prediction of protein and ligand interactions is crucial.

3. Traditional Computational Approaches in Predicting Protein-Ligand Interactions

For decades prior to the formulation of AI-empowered strategies, researchers explored traditional computational approaches for predicting protein-ligand interactions. Molecular docking simulations, in particular, aim to study the relative positions and

conformations of small molecules, referred to as ligands, within binding sites. More computationally extensive molecular dynamics simulations are pursued following the identification of top docking conformers to further investigate key insights into ligand binding. Based on these binding conformations, binding free energies can be calculated and binding affinities for different ligands can be subsequently predicted, mirroring recent deep learning-based frameworks. More applied purposes focusing on directly investigating and predicting binding conformations, such as SBVS and IBIS methodology, are also quite relevant to the early drug discovery phase. Pharmacophore modeling, on the other hand, posits that the fundamental elements dictating a receptor-ligand interaction are the steric and electronic features of the participating compounds. Computationally, a pharmacophore can be estimated using three key approaches: database superposition, ligand-based or structure-based alignment, and field point approaches. Generally, the first two strategies align compounds based on common pharmacophore features, like their bioactivities. The latter approach calculates field points for a given ligand, such as hydrophobicity and hydrogen bonding. The use of traditional computational methods has emboldened cheminformatics and bioinformatics research, facilitating the adoption of established molecular inhibitors and the identification of new targets for clinical protein-ligand studies. Nonetheless, it has been consistently argued that such techniques are inaccurate in outcomes and are, especially in the absence of computing resources and data deficiencies, computationally inefficient. Consequently, researchers across multiple disciplines have looked towards alternative techniques for predicting and studying protein-ligand interactions, leading to the current consensus of integrating machine learning techniques.

3.1. Docking Simulations

Docking simulations, also referred to as virtual screening, intend to predict a suitable binding mode and pose of a ligand or drug in the active site of a macromolecule, using computational models underpinned by molecular mechanics. However, the true value of any docked pose remains unknown, since generating reliable binding poses involves approximate methods. The reduced cost and simplified systems in docking simulations offer a trade-off in accuracy for the sake of managing large datasets. There are several possible approaches in molecular docking, which differ from one another by the exact models of the physical processes involved in ligand–macromolecule binding. Methods can also include various techniques to search both the ligand's conformational and

orientational degrees of freedom in or around the active site. For the most part, docking simulations revolve around two linked stages, each with its conformational search technique, be it genetic algorithms, exhaustive search, and so on. Docking protocols are completed by scoring functions that usually combine electrostatic, van der Waals, hydrogen bond, and desolvation contributions in order to identify which poses generated at the previous stage are most likely to be those that the molecule would assume in an experiment. In the parallel docking pose ensembles generated for each ligand, scores are also used to sort these poses, and the lowest energy ones are returned as the best predictions. Overall, docking simulations are the computational methods that, despite a wide range of valid criticisms, offer the best balance between cost and accuracy available for exploring large datasets of compound models to make potentially testable predictions about interactions with drug targets. They are extensively used as a lead optimization tool in drug design in the industry and have been at the core of several recently successful case studies. Their primary limitation stems from a poor performance when it comes to the correct reproduction of a ligand pose that is physically reasonable for biological activity. A great deal of hands-on involvement is also called for with this technique. The methods used in molecular dynamics and Monte Carlo simulation, which can be used to enhance docking simulations, require more knowledge to operate and are much more computationally expensive to perform.

3.2. Pharmacophore Modeling

Computational methods to predict protein-ligand interactions rely on the identification of molecular properties and their relationship with biological activity. A pharmacophore is a concept represented as polyhedral simplexes indicating the spatial arrangement of chemical functionalities that are responsible for the interaction with a target macromolecule. It represents the ensemble of molecular properties that are necessary to achieve an interaction with a target protein and helps in the determination of which atoms should bind to the target protein. The three-dimensional geometrical expression is derived from three key pharmacophore elements or features that indicate how an interacting ligand binds to a macromolecular receptor.

The pharmacophore modeling method can be classified in terms of the data set from which features can be derived as correspondent and non-correspondent techniques. The correspondent technique is particularly useful in drug discovery, where it uses a set of

known active compounds from which a pharmacophore model can be built. Once the active compounds have been identified, they can be used for the construction of a pharmacophore model, and subsequently, the pharmacophore model can be used to screen large chemical libraries. An attraction of this technique is that by having identified a large set of active compounds, it can be possible to screen a small subgroup of these for in-depth analysis through more computationally and time-consuming methods such as docking. One of the limits of this method is the ability to appropriately select the active compounds required to build the model. Moreover, since the pharmacophore model is basically a ligand-based approach, it does not consider information associated with the pharmacological activity of the protein active site; thus, it cannot make assumptions about ligand binding modes. Obtained hypotheses require further validation with techniques such as docking. Nevertheless, since the pharmacophore tends to be used for a complex target or protein-ligand system, by its conjunction with virtual screening, specificity and sensitivity can be improved.

4. Machine Learning in Predicting Protein-Ligand Interactions

Predicting protein-ligand interactions is a challenging task in chemical and structural biology. In recent years, machine learning has become an indispensable tool in this area. By considering subtle interactions between various types of macromolecules and small molecules, machine learning techniques provide several advantages over traditional methods, which often require structural resolution of the complex. Advanced machine learning strategies can not only predict protein-ligand interactions more precisely, but also process complex omics data effectively. These methods have boosted the computational study of protein-ligand interactions. In particular, feature extraction from a large amount of structural and omics data and hidden patterns or relations among these features are central for inferring molecular interactions.

In the context of protein-ligand interactions, several machine learning strategies and algorithms have been developed specifically. Some machine learning methods focus on constructing algorithms for processing protein-ligand data efficiently, whereas others are adept at utilizing a variety of omics data for better predictions. Current studies have developed a wide array of machine learning models that are proficient at various facets of protein-ligand interactions. To deepen a user's understanding of the subject, this review begins with an introduction to the biological background and data preprocessing

steps. This section is crucial in making strides in model optimization and performance enhancement. Given the current trends, different machine learning algorithms for application in this field have been designed. These models aid in the prediction of protein-ligand interactions from a variety of vantage points. The advancements in machine learning techniques can be tailored to suit various purposes. After data sampling and feature generation, indispensable machine learning strategies can be employed. These comprise supervised and unsupervised learning, among others. As such, a broader perspective on protein-ligand interactions can be tackled with these strategies. Given the contributions of machine learning technologies in modern drug discovery, we bring forth a few case studies to highlight the recent trends. Practical challenges when using machine learning approaches, such as data scarcity and model overfitting, have also been addressed.

4.1. Feature Engineering

Fulfilling the promise of machine learning-based predictive ensemble models to provide quick, accurate predictions of protein-ligand activities depends on the selection and transformation of relevant input features. Protein-ligand interactions are complex and may be shaped and/or enriched by multidimensional chemistry as well as by allostery and distant ligand-binding sites. It may also require computing. Molecular descriptors, fingerprints, kernel methods, and approximately 1D representations are commonly used in predictive benchmark studies. Many state-of-the-art predictive models involve feature sets assembled through extensive clustering and manipulation of input feature sets, including molecular descriptors, proteins, and multicomponent protein environments, and complex hierarchical clusters from intermediate representations of a protein 3D structure.

Just as one critical decision involves model selection and ensemble modeling, feature set creation and selection require deep domain expertise, and these are the critical decisions that can help optimize a prediction model's performance and representativeness of its broad biological context. Continually, they generate a substantially low-dimensional representation for their system of interest encoded into a feature set. Domain-specific intuition about the various interactions that result in good chemical series compounds and sufficient experimental data to validate if used as feature input are indispensable but not nearly adequate. In summary, we chose features that capture the influence of

individual interactions contained in a larger feature set and tried to limit the size of the feature set without losing much information. Dynamic residues are predicted to change ligand activities and are thus incorporated into pharmacophoric representations.

4.2. Supervised Learning Algorithms

An effective approach to predict protein-ligand interactions is incorporating computational methods into the machine learning paradigm. The perspectives of machine learning in predicting protein-ligand interactions mainly include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In this paper, the supervised learning methodologies, including the appropriate algorithms for prediction and their applications in drug discovery, are specifically and concisely introduced. In supervised learning paradigms, the support vector machine, the decision tree, random forests, gradient boosting tree, multilayer perceptron, the convolutional neural network, and the recurrent neural network are commonly used supervised learning paradigms. Typically, reliable models are constructed based on two necessary datasets: the training dataset, which is employed to train the constructed model; and the testing dataset, which is utilized to evaluate the fitting accuracy of the model. Cross-validation or splitting validation, which are performed using the training dataset, are fundamental for validating the outer fitting of the model. Besides the selection of the training and testing datasets, the hyperparameters of the model are also vital for the predictive accuracy of the model. Furthermore, several classical case studies have testified that a suitable machine learning algorithm with a discriminative dataset could make the model reliable for practical drug discovery. However, there are still some challenges for the practical application of the machine learning algorithm, such as overfitting and data imbalance. It is worth noting that the integration of professional domain knowledge will accelerate the model's optimal selection.

5. Applications and Case Studies

5.1 Drug Repurposing In drug repurposing, it is essential to model the binding of small molecules to a selected target. Once a protein or a particular area of the protein that modulates a disease of interest is selected, identifying binding small molecules can be conceived as a function of predicting which known drugs or compound libraries are likely to bind the pocket with high affinity.

5.1.2 Case Study: Modifying the Approved Drug Clofazimine to Combat DTNBP1 Variants This case study provides an example of conducting a preliminary DTNBP1 mutation-based protein-ligand prediction followed by a closer examination to extract candidate molecules (clofazimine was used as input for the case). Results reveal that most of the identified compound candidates are clofazimine derivatives, emphasizing the power of computational predictions in decision-making processes regarding drug repositioning. In this section, we will provide the results of molecular dynamics simulations of clofazimine in the structures of wild-type and mutant duck DTNBP1, as well as the sequence enrichment profiling. Predictions related to the potential target site in DTNBP1 will also be discussed.

5.2 De Novo Drug Design Generally speaking, de novo approaches can be split into different types of protocols, each generated based on different strategies and main molecular interactions: docking-based, fragment-based, structure-based, and stand-alone molecular design tools. While the list above is not exhaustive, it highlights the fact that new compounds can be designed flexibly and using a wide range of tools, in order to focus the exploration of chemical space in directions that are relevant for determining the protein–drug binding energy and thus can be used for the creation of pharmacological drugs.

5.2.2 Case Study: Applying a Docking-Based De Novo Drug Design In this case study, we will show an example of a de novo drug design approach. More details for each of these potential targets are described in the following sections, highlighting the current state of their biochemistry and their importance as ideal drug targets. Lastly, possible challenges, pitfalls, and other computational datasets that could/should/need to be collected prior to these docking calculations are discussed. Although different strategies can be designed to choose the most relevant starting point for this particular de novo drug design approach, we selected two cost parameters to design the library of candidate molecules because they are widely used in real-life drug discovery efforts.

5.1. Drug Repurposing

Drug repurposing, the process of repurposing drugs approved for one indication for the treatment of another or a wider range of diseases, has seen increasing interest in recent years. Drug repurposing, as an innovative approach to find new uses for existing drugs for novel therapeutic, diagnostic, or prognostic indications, offers multiple advantages,

including reduced development times, costs, and risks compared to traditional drug discovery and development. Surprisingly, despite its potential, the prior art of all concepts of drug repurposing has a history that spans at least 50 years, and to date, the process of finding suitable candidates to repurpose has relied on serendipitous findings or fortuitous errors rather than a systematic strategy. Progress in informatics, particularly in the applications of advanced computational techniques to the problem, is accelerating the development process.

Computational methods such as similarity searching, graph-based methods, and ligand-based descriptor methods applied to rank, select, or prioritize the compounds are being reported in increasing numbers. In particular, in silico docking and scoring methods are used for novel target identification and for the ranking and prioritization of virtual hits generated by library similarity search. A number of case studies are used to demonstrate successful applications of these techniques to generate or validate hypotheses on new uses for drugs in different therapeutic areas, including neurology, cardiology, gastroenterology, infectious disease, and inborn errors of metabolism. In this paper, we demonstrate that computational methods can create in silico hypotheses for the repurposing of compounds and predict new indications for known drugs. However, we note that in silico models require in vitro validation before in vivo confirmation. Drug repurposing may offer numerous advantages over new treatment development to the patient, the healthcare system, and society as a whole. A number of withdrawal drugs have been promulgated as being suitable for drug repurposing, including sildenafil, fluoxetine, candesartan, and riluzole. Perhaps the best-documented example of drug repurposing to date is sildenafil. Interdisciplinary collaboration is paramount for successful drug repurposing identification programs characterized by cross-fertilization of ideas, drug screening, and knowledge. There are challenges to be met, including regulatory approval, safety, toxicology, mechanisms of action, and market acceptance, as well as finding funding for off-patent drugs that offer no grant of market exclusivity at the end of the research process.

5.2. De Novo Drug Design

5.2. De Novo Drug Design. De novo refers to creating something from scratch. De novo design begins with the discussion of drugs that have never existed. To some extent, the design and optimization of a new specific therapeutic compound are equivalent to

navigating an incredibly vast chemical space. Thus, traditional design strategies are often slow and inefficient. Computer-aided drug design, where computational methods influence the medicinal chemistry process, has greatly accelerated and improved drug discovery in recent decades. Iterations of Structure-Based and/or Ligand-Based Design lead to learning cycles. Iteration of the Learning Cycle leads to feedback-directed iterations of the de novo design strategy. In this way, the technology envisaged enables a transition from classic design paradigms to an AI-driven interdisciplinary endeavor. Several de novo design approaches for candidate ligands for novel targets exist. Structure- or pharmacophore-based approaches design fragments that are elongated and linked to a final ligand. Ligand-based approaches use 2D or 3D representations of the protein target and design compounds that maximize the likelihood of being active. Some of the latest methods explore the use of self-consistent field estimations to examine the geometry of the ligand-bound protein and the electronic topology of the potential ligand-interaction region directly. In each case, predicted bioactivity is validated but with a clear preference for prediction of bioactivity of molecules with minimal chemical similarities to any of the input ligands. In essence, the predicted bioactivity of newly designed compounds can only be verified by time-consuming experiments. Thus, experimental estimation of de novo compound-binding affinity will remain as the unknown in the learning scheme for this and similar de novo design methods.

6. Challenges and Future Directions

In the past decade, the field of predicting protein-ligand interactions has witnessed remarkable progress, in part due to the use of data- and knowledge-intensive machine learning methods, including various forms of artificial intelligence and machine learning. However, several major challenges and limitations remain that need to be addressed in order for this field to continue to progress. These include: (1) a limited volume of high-quality training datasets, (2) the slow speed of training and constraints of training sets, and limits in generalization, model interpretability, and explainability, and (3) a growing risk of potential overfitting of models. In addition, while there is hope that there will be deep collaborations with the experimental molecular biology community, who hold the real potential to improve the capability of the resulting methodologies if provided with clear guidance that includes insights and explanations, previously developed tools are not structured in ways to directly support such efforts.

Future directions and opportunities include the potential and need for actively addressing the critical challenges that largely represent current limitations in the related fields. In recent years, several innovative sources of experimental data have emerged. These are of diverse types, including integrative multi-scale capabilities as well as capturing cell- and tissue-specific biological variability. These include library-centric approaches that relate increasing information to the creation and generation of datasets; drug repurposing and pharmacological landscapes at increasing scale for integration between genetic and physical variability. Advances along these lines have clear implications and relevance for the future of predictive methodologies of this sort. Furthermore, data integration and the use of machine learning should be able to augment draft models with currently existing mathematics. Specifically, the data collected through more advanced tools will undoubtedly allow for updates and corrections of heterogenic annotative subcomponents in the coding used in related fields. Such updating and advancements of training datasets will allow for the development of more comprehensive algorithms using such updated and/or corrected annotations; the addition of a few seminal examples to the training corpus can fundamentally change the trajectories of wide-ranging mathematical and computational predictive algorithms. This demonstrates advances in data collection technologies to be a broad engine of progress in these fields. Another critical direction for the future is the need to focus on the development of more interpretable AI representations with highly informative AI representations that could also be useful in guiding future experiments. New explainable machine learning models that allow for bringing value to hypothesis testing mechanisms are sorely needed in these fields. These new data collection and modeling directions, if accomplished, could help address identified weaknesses in existing training methods. In a field that generally requires interpretability and trust to be sensitive to tools, heuristics, and algorithms, an investment in developing models that provide these capabilities is crucial.

6.1. Data Availability and Quality

In these historically data-starved approaches to studying protein-ligand interactions, there are many sources of errors, including: (1) the unavailability of biological data, (2) limited quality of the available data, (3) unexplained variation in the datasets, to name a few. Exclusive reliance upon proprietary datasets is problematic, as we are unable to validate our findings from independent and diverse datasets. Data is conventionally

available from open-access databases and historical work with proprietary datasets. All datasets of relevance have issues. Consequently, it is imperative to caveat that findings from our work may absolve or exacerbate errors underlying the datasets. Curation and enhanced data artistry are required to both address the issue of data quality and limited availability. One approach to make data more available is to encourage collaboration via datasets that are available.

The quality of the data is an important consideration as it can determine the performance of the algorithm. In machine learning, data is often considered an important part of any pipeline; some would even argue it is foundational to any predictive system. Data includes biological properties of known protein-ligand pairs, or proteins alone, as well as drug-drug similarities, which can also contribute to improving the accuracy of a machine learning model. Collecting data can be difficult, so the lack thereof, or ownership of a 'data bank', can become a choking point for predictions based on machine learning or related models. Some sources of disaccord between databases may be deliberate side effects of methodologies used. Furthermore, caution should be employed when including such data to train machine learning algorithms, particularly when attributing a fundamental property of protein-ligand interactions based on the changing and subjective nature of analytics. These will only in some cases be relevant to individual targets or research problems.

6.2. Interpretability and Explainability

Interpretability refers to the ability to understand the decision-making process of a model, explaining what features it uses to issue a given prediction. One recent concept regarding the most complex black-box models is that there may not be any "underlying logic" per se that can be interpreted without a more sophisticated definition of interpretability. Explainability, and the closely aligned concept of explainable AI, pertains to a "communicating process" between model and human. Most importantly, interpretability and explainability, or the lack thereof, have implications. For instance, if a clinician cannot comprehend or interpret a model's features, their trust in the technology would drop, and the model would not see adoption. Besides the penalizing evidence of currently skirting the issue of interpretability, such as the temporary discontinuation of certain technologies and the concerns surrounding others, not providing tools for understanding underlying neural networks would preclude research

and development for the field at large. One major hurdle precluding the interpretation of molecular predictions is the enrichment of recent, state-of-the-art predictive models, due to the layering of different types of neural networks that encode disparate representations of one another.

Several groups have employed a variety of steps to improve the interpretability of their predictors. Most common among these include investigating feature importances, as well as entertaining simpler predicates of complex models. The techniques involved showing an overview of which features were the most important in driving predictions and comparing predicates to elicit discrepancies between explanations and the predictions of the larger, more complex model. Most directly, the larger field of explainable AI, which leverages machine learning and domain expertise to develop an understanding of AI systems, could be appropriated to give experts better intuition and insights into the direction of predictions. So, building confidence in AI algorithms used for drug discovery will require ethics-based learning through interdisciplinary collaboration between domain experts and data scientists. Furthermore, forthcoming research into convolutional neural networks, autoencoders, and electrolyte solvation free energies may provide insight into how to build models that are accurate and ascertainable. More fundamentally, building a molecular recognition model that is state-of-the-art and interpretable will likely be the result of an increase in the fundamental understanding of the physical interactions at play, e.g., co-evolution, hydrogen bonding, or lone pairs in the case of d-orbitals.

7. Conclusion

In developing AI-enhanced computational methods to predict protein-ligand interactions, we have seen significant advancements in integrating ML and DL with traditional computational methodologies. These advancements show remarkable improvements over traditional methods in scoring functions, de novo design, virtual screening, and binding free energy calculations. These improvements possess noteworthy potential for transformative contributions to the drug discovery process. Despite the substantial advances in the field, much remains before AI models become the first choice application for investigating PPIs. A primary challenge is the quality of the available data, including data with imprecise experimental uncertainty, data varying in the quality of the reported structural complexes, and data only being available for

comparably limited protein-ligand families. The interactivity of an AI model beyond its predictive potential remains a rare design feature to this day. In summary, numerous applications in AI-enhanced computational methods towards PPI prediction tasks deserve continuous exploration. Many exceedingly promising strategies remain to be evaluated, with substantial potential in improving different application areas of relevance. ML and DL will continue to alter the drug discovery process; new sustainable therapeutic methods are made possible through AI. Moreover, targeted applications support the interdisciplinary nexus between physics, computer science, and biology, with the latest results increasingly being referred to in structural and cell biology. Overall, AI-empowered approaches to predict protein-ligand interactions are being increasingly reported in academic research, providing valuable resources and insights to the research community, where these models will be applied in the near future for rapid and effective discovery of therapeutics.