Proteome-Wide Binding Site Identification Through Graph Convolutional Networks: AI-Enhanced Computational Methods for Drug Target Prediction

Dr. José Barreto, Professor of Informatics, University of Lisbon, Portugal

1. Introduction

Accurate drug target prediction is the uppermost priority for enlightening novel therapeutic opportunities in the pharmaceutical sector. The simplicity and cost-effectiveness of the computational methods initiated the transformation from conventional to in-silico drug identification approaches in the last few decades. Recent advancements in high-throughput biological properties of chemical attributes have led to a massive amount of data. Diverse structural, biophysical, and chemical computational approaches have widespread applications in drug target prediction. Nevertheless, each of these methods harbors some limitations in delimiting drug side effects. Computational predictions suffering from iterative overlapping may result in false positive results, posing dangerous threats to human health.

Recent progress in artificial intelligence has predominantly influenced the tendency of drug discovery by modifying the computational methods involved in estimating drug targets. In comparison to other conventional methods, artificial intelligence has more precise, reliable, and significant predictive power, free from iterative overlapping. These AI-based computational methods also reduce the computational cost by integrating various algorithms and techniques. Therefore, it is mandatory to comprehend and discuss the updated version of AI-enhanced computational methods used in drug target prediction. In the current review, we will discuss the state-of-the-art AI-enhanced computational methods in drug target predictions, emphasizing the available software tools, existing libraries, and updated algorithms.

2. Fundamentals of Drug Target Prediction

Drug target prediction aims to find proteins that can be bound by specific drugs to mediate related phenotypes or motivations. Since a drug can correspond to a set of genes or proteins, another definition of drug target prediction can be described as finding the molecular targets linked to drug-induced diseases. With the progress in systems biology and systems pharmacology, more drug targets are utilized in the design of new drugs and the development of novel therapeutic strategies. In fact, drug targets can modify the biological system in some way to produce a therapeutic effect or potentially interfere with normal biological functions, and can indirectly serve as predictors of relevant biological traits. Understanding drug targets can indirectly help uncover the chemical and genetic basis of complex biological traits. With the development of genome biology, large-scale phenotypic information and multi-dimensional biological data make it possible to predict candidate drug targets for further drug development.

Drugs perform their functions via their targets in biological systems. Drugs can physically interact with a variety of molecules, including proteins, nucleic acids, and lipids. Because proteins have important biological functions, about 75% of actionable drug targets are proteins. Proteins govern the vast majority of biological processes, and the functions of proteins are involved in various biological functions, while dysfunction of these proteins is the primary cause of diseases. Therefore, the association between diseases and proteins can be described as diseases being caused by the dysfunctions of proteins. Molecular biology and bioinformatics have contributed to a deeper understanding of drugs and proteins. Advances in molecular biology have been instrumental in constructing the single level of biological networks. High-throughput technologies have successfully provided large-scale genomic data and proteomic data. Several methods have been developed to predict candidate drug targets based on gene and protein data. Typically, the method can be divided into two parts. The first part is to find the genes associated with specific diseases or phenotypes. The second part is to predict the proteins that correspond to these genes as potential drug targets. Because protein-protein interactions provide valuable insights into the underlying cellular machinery, more screening methods have been proposed to explore drug targets based on PPI data with optimization. Furthermore, database-matching and literature-based methods have been developed through the analysis of known drugs and their targets.

Although some drugs have potential side effects or unexpected toxicities, we cannot deny that these methods have greatly facilitated the discovery and development of new drugs. With more and more protein and small molecule information accumulated in databases, various computational approaches have been developed to predict drug-target interactions, such as supervised learning approaches and network-based methods. Pathway analysis is based on the idea that examining the structure and dynamics of molecular networks is important and can help systematically understand the entire biological process. Regulatory information is necessary for studying biological pathways. The regulatory information comes from the expression levels of genes/proteins and the co-expression patterns. Gene set enrichment analysis assumes that a pre-defined statistical function can measure the association of a gene/protein set with a given disease or phenotype. In addition, functional similarity analysis has shown that drug targets of similar drugs are more likely to be located in the same pathways. The biological network-based approaches also support this phenomenon. Evidence strongly shows that drugs used in combination with other drugs are more effective than when used alone.

3. Traditional Methods vs. AI-Enhanced Methods

Standard statistical and computational methods have been extensively used in the drug discovery process for target identification and drug repurposing. Correlation analysis, sparse canonical support recovery, and biclustering techniques have been applied to integrated omics data and clinical data to identify disease-specific biological modules corresponding to drug targets, disease biomarkers, and additional information for biological and cellular information. Cut-off value determined by absolute clustering association studies and conventional principal component analysis have been used to identify post-translationally modified proteins corresponding to the problem of single drug resistance. Probabilistic matrix factorization and alternating least squares, Bayesian matrix factorization, and collective matrix factorization have been used in drug-target interaction prediction and in identifying subnetwork molecular signatures resolving tumor heterogeneity. Nevertheless, the application of these traditional techniques is limited by reliance on prior knowledge to specify fixed coefficients and other parameters.

Traditional statistical and computational methods are often subject to model inaccuracy due to nonlinearity, high dimensionality, and sparsity of the high-throughput datasets. The lack of nonresponse prevention skills and the missing at random data assumption produces biased preliminary output and shows limited recall rate of the positive data. The longer processing time required to handle large datasets represents another issue. To address these limitations, artificial intelligence enhanced methods, particularly in the form of cutting-edge machine learning and transfer learning strategies, make target prediction feasible and available. To compare the merits of the traditional established statistical method and newly applied AI-enhanced methods, we compared the results of three modern target prediction methods to the traditionally implemented method for treatment in the clinical section.

4. Machine Learning Techniques for Drug Target Prediction

When complete profiles of gene products are available, computational methods could be used to predict novel drug targets with higher efficiency than traditional experimental methods. By using machine learning techniques, we could identify potential drug targets by exploiting large-scale datasets and obtain more significant discoveries than rule-based methods. Supervised learning, unsupervised learning, and deep learning are three kinds of machine learning techniques widely used in drug target prediction. Supervised learning can automatically figure out hidden associations between drugs and targets using labeled training datasets. Unsupervised learning could be used to explore the hidden structure of the network and to find out subtypes related to drug-target interactions. Deep learning is flexible in coping with various input features in drug target prediction. Deep learning models could predict drug-target interactions from drug chemical structures, drug-protein networks, drug side effects, transcriptional responses in cell lines, etc. Powerful predictive capability and flexibility in feature representation make deep learning a promising framework in drug target prediction. Employing effective machine learning methods to construct accurate prediction models for identifying novel drug targets has attracted much attention due to the rapid emergence of large-scale and complex biological data. An ideal model for drug target prediction is expected to achieve three criteria: high predictive and interpretive performance. Prediction accuracy improvements are usually pursued by applying advanced statistical or machine learning methods, resulting in more accurate feature selection, sample partitioning, and model validation, making the analysis of high-

dimensional data with noise and redundancy computationally feasible. Integration of various features can improve predictive performance. By scoring and integrating the features of drug chemical structures, similarity between drugs and targets in a bipartite network, the validation marks of human proteins, unsupervised probabilities of interactions between drugs and targets, and tissue expression profiles of human genes, the team has reduced the expected gaps between new experimental deterministic results and the machine learning-based predictions. Also, such a system could be used to predict adverse effects of drugs and to guide individual applications.

4.1. Supervised Learning

The most prevalent type of DTI approaches that are used until today are supervised learning approaches. In supervised learning, a model is trained using a dataset that labels inputs with the corresponding expected outputs. This plain explanation means that we need to have labeled data, which in our case means that we should have a dataset that contains compounds with their interacting proteins.

Supervised learning has several algorithms for creating a predictive model, such as decision trees, support vector machines, and discriminant analysis. Several machine learning techniques have been used to provide qualitative and/or quantitative classification of the chemical compounds as DTIs or non-DTIs. Another important example is the binary kernel classifier using a few kernels to predict the compound–target interaction. There have been many successful case studies of using these techniques to develop computational models and predict new DTIs, and many have reported substantial advantages compared to other non-AI computational techniques.

Machine learning models can have several limitations such as overfitting since these models need a lot of data for training. Today, machine learning is one of the key technologies used in pharmaceutical research and public health that accrue large collections of data, which makes ML models practical. That is why there are now so many deep learning applications in bioinformatics as well.

4.2. Unsupervised Learning

Unsupervised learning methods can also be used for drug target prediction. In these methods, one typically analyzes the data obtained from cell-based assays, chemoinformatics, or bioinformatics without considering any labeled responses. In the

field of drug target prediction, unsupervised learning methods are usually used to explore the data. Such exploration includes clusters in the data, or the identification of salient features and a dimensionality of the targeted features. Two commonly used unsupervised learning techniques in the field of drug target prediction are cluster analysis, also known as clustering, which reduces the dimension of the features, or simply called dimensionality reduction.

Cluster analysis is typically used to detect the similarity structures of biological entities or states in many fields, including biostatistics and genomics. Clustering might be used to group individual subjects based on the gene expression data to detect novel subgroups. In the field of drug discovery, for example, one could use clustering to group proteins or genes based on their cellular response to drugs to discover novel drug targets. However, although clustering has a variety of uses in the field of biostatistics and bioinformatics, it provides little guidance for the target discovery process. Another advantage of unsupervised learning techniques is that they do not require training data. Unsupervised learning is useful for finding hidden patterns in complex data. For example, clustering can be used to discover and differentiate patient subpopulations. A novel drug that is useful for some but not all patients can be focused on populations for whom it is expected to work. However, a challenge of unsupervised learning is that it can be difficult to interpret the results of the analysis since the data are not labeled. The results of clustering analysis can be affected by different parameter settings. For example, different metrics for distance measurement or methods for aggregation of individual features could affect the grouping of cells. While various methods are available, to date, few studies have applied unsupervised learning techniques to the problem of drug target prediction. Researchers used non-negative matrix factorization, an unsupervised learning technique borrowed from the field of image processing, to classify DNA microarray data based on gene expression levels. In this model, the authors classified yeast genes as direct or indirect targets of the factor. The results showed that unsupervised learning techniques can be useful for the detection of novel drug targets. It is anticipated that their usage will broaden in computational methods for drug target prediction.

4.3. Deep Learning

Deep learning is a part of machine learning methods based on artificial neural networks that contain many computational layers to learn data representations and high-level abstractions. An artificial neural network can potentially achieve the desired result by transforming the inputs through a series of computational layers. Deep learning methods can handle a large amount of biological data that has high dimensionality, including genomic data, as well as the inherent complexity and uncertainty usually encountered in the modeling of biological systems. With some data transformation and dimensionality reduction techniques, deep learning models can also be applied to tackle structured drug information that is utilized in predicting drug-target interactions.

Several studies have demonstrated successful applications of deep learning in predicting drug-target interactions when utilizing a significant amount of well-organized biological data, including textual and chemical structure-related data. Despite its advantages, the application of deep learning to computational drug target prediction also has the challenges of high computational requirements and the lack of good understanding and interpretability of the model. Recently, a number of case studies and review papers shed light on the potential uses of deep learning technologies and data generation methods in pharmaceutical research. Furthermore, AI-based methodologies and drug discovery platforms by leading international pharmaceutical companies in collaboration with technology companies have been presented as being a game changer in drug discovery. At present, deep learning methods are not commonly used by the research community for drug target prediction purposes due to the unavailability of large-scale annotated data. As deep learning requires a massive amount of data to train, it is utilized predominantly in the prediction of protein targets using high-throughput molecular descriptors and large-scale feature data. In contrast, traditional computational methods are more widely applicable for a comprehensive variety of drug target predictions, including species-specific and multi-target drugs.

5. Challenges and Limitations of AI-Enhanced Drug Target Prediction

The use of conventional computational methods has introduced some issues in the field of drug target prediction, revolving around the high-throughput production of data and the excessive optimization or standardization of modular parts of the workflow. These issues, however, can be turned into opportunities by the sophisticated implementation

of AI techniques. Still, some concerns regarding regulatory and cybersecurity have arisen. Specifically, it emerges that there is a set of hurdles being faced right now: data quality from several sources, explainability still far from the expectations for trustworthiness, and the need for improvement of prediction algorithms, interpretation tools, and visualization methods.

The first obstacle regards the increasing volume of data produced nowadays, which has to be certified and correlated in a coherent and proper workflow. In addition, there is the issue of the lack of comprehensive and adaptable data, with a perfect balance between the chemical and biological realms, for a complete functional pharmaceutical profile, making the training of the learning model not exhaustive. This situation could have consequences in terms of imbalanced data distributions, conscious or unconscious biases, and the partial reproduction of modular components, such as ligand-target binding predictions, which should, in reality, be part of the first steps of an integrated pipeline for drug discovery. The data selection and very likely overfitting could be an additional limitation. Nonetheless, whereas in traditional methods these data need to be perfect or have as few errors as possible, in AI, a cascade of errors could help to re-adapt contexts, correct the wrongly used algorithms, or address them in the right areas, if not previously considered.

6. Case Studies and Applications

This section aims at providing clear views of what and when AI-laden drug target predictions have practical implications based on real-world case studies. The preliminary research is dedicated to the rise of AI innovations that are now proceeding with numerous advances across academia and industry. It is anticipated that successes such as those presented will affect other research routes for predicting drug targets.

A collection of diverse research studies successfully applying AI-enhanced or AI-induced computational methods for identifying high-quality candidate drug targets in a practical drug discovery scenario is tabulated. Among the case studies, four works resulted in experimental validations of novel leads for future in vivo analysis. Many of the successful AI applications target human protein drug discovery-related processes illustrating first use cases in the European Science Hub and further emphasizing potential improvements in terms of flagging off-target complications and first-in-class leads in small molecule drug development. A recent drug discovery context pertains to

big pharma challenges in tuberculosis treatments complementing yet another research contribution of this article. The most active collaboration underlines an artificial intelligence relationship-driven fourth revolution as a next stage in productivity growth for computer-enabled discoveries. Key dynamic collaborations include but are not limited to computational biologists working closely with auxiliary pharmacologists. The multidisciplinary author list is considered to open up chances in ideating more original applications for drug discovery across diverse therapeutic areas since problem statement insight and practical hands-on experiences in conceptualizing AI-borne solutions are interrelated.

Lessons Learned: Two recent success stories underline how large pharma has already used AI-borne deep learning methods for achieving experimental success in computational drug discovery as a prelude to entered clinical phase therapeutics. Insights from the case studies initially position our review at enhancing researchers' thinking on implementing timely and cutting-edge drug target prediction research targeting clinically relevant therapeutic areas. Suitably, the case studies cater to broad therapeutic applications spanning cancers, oncology biomarker discoveries, and rare diseases without a marketed treatment, in tandem with improving antimicrobial stewardship through the early detection of resistant pathogens based on multi-center cases in computational drug development.

7. Future Directions and Emerging Trends

The machine learning algorithms are anticipated to have profound enhancements, and these updated technologies are expected to be widely and intensively implemented. Along with machine learning and AI, genomics and proteomics are expected to be an enriched source of information, and integration of the two can provide a novel insight into drug targets. However, there exist several challenges that need to be addressed. Outsourcing research will require a multidisciplinary approach and collaborative ventures among artificial intelligence experts, molecular biologists, chemists, and pharmaceutical scientists for driving innovative ideas. The accumulation of authentic big data is an essential requirement without which AI or deep learning approaches can return unexpected results unless significant overfitting and noisy data can be upgraded. Drugs arising from AI-enhanced approaches could pose novel ethical and legal implications. We are in an era of highly personalized drugs, intelligent systems, and

precision therapeutics. A variety of exciting opportunities and possibilities exist in this field, which unfortunately remains relatively unexplored. In terms of future applications and exploring novel domains of drug target space, we anticipate AI to be combined with other emergent fields, such as intelligent designs of new proteins or drug candidates, development of more precise and safe therapeutic drugs, administrative AI networks, and isolating more capable and shorter paths and strategies to modulate multiple levels of biological systems for conflict resolution. With these many precursors in view, this emphasis on a wide-ranging review serves to fill the gap in information about dispatched approaches that apply AI to scrutinize and prioritize putative drug targets. Here, we explain the know-how and prospects for extrapolating from human genetics as a means of steering attentive AI studies to conspicuously produce some upcoming drug targets.