Deep Learning Architectures for Phenotypic Bioactivity Profiling in High-Throughput Chemical Screening Assays

Dr. Xiaojing Wang, Professor of Electrical and Computer Engineering, University of Illinois Urbana-Champaign (UIUC)

1. Introduction

As high-throughput screening (HTS) has always been one of the key technologies in drug discovery, it can identify a potential lead compound from millions of compounds rapidly and has attracted much attention in the pharmaceutical industry. In recent decades, high-throughput screening has grown profoundly, accompanied by the development of automation technology. As experimental technology has made rapid progress, almost all fields of science, particularly the life sciences, are producing large-scale data sets. Automated workflows have come a long way, offering a plethora of assays collectively to detect a range of biological activities that may be affected by particular treatments utilizing a wide range of techniques, including microscopy, mass spectrometry, and DNA arrays. Additionally, due to their specificity and sensitivity, biological assays such as in vitro cell-based assays, cell-free assays, and data obtained from genomics, proteomics, and structural biology have led to the discovery of many drug compounds. In recent years, the use of computational tools to extract information from increasingly large and complex data has been exacerbated by advances in experimental technology and the willingness of experimentalists and drug discoverers to analyze and interpret them.

The success of experimental and computational biology efforts to analyze, interpret, and predict a meaningful result largely depends on the ability of the HTS data analysis platform to answer the scientific question of interest in an efficient and timely way. However, the analysis of data derived from HTS encounters challenges and provides a platform for analysis, especially in distinguishing and discriminating both desirable and undesirable objects, such as compounds, functional classes, properties, and phenotypes, and in mining interactive data patterns for holes. These HTS data are usually massive,

containing both systematic and stochastic errors, with a large number of low-dimensional observations. Moreover, subjects such as rare events and unbalanced classes of data must be referred to as well. Therefore, to be of benefit to individual researchers in drug discovery efforts, it requires an efficient and precise analysis of data. Aiming to narrow this gap, there is a continuous demand to capture optimal processing capabilities. This is evident from the convergence of machine learning, computer-aided diagnosis, and bioinformatics to resolve such intricate issues. In essence, machine learning has changed the automation of HTS data analysis and quality scoring in drug discovery by providing novel approaches to analytical data and HTS assessment in drug discovery.

1.1. Overview of High-Throughput Screening (HTS) in Drug Discovery

High-throughput screening (HTS) is one of the key components in the multi-stage drug discovery process. HTS is a powerful technology that facilitates researchers in running thousands of compounds through biological tests to identify their activity, extract meaningful data, and classify them for further analyses. Compound libraries are screened against a known target. Assay development is the first step of HTS designed to prepare for the screening of libraries. HTS may cover single-point assays, multi-point assays, and cytotoxicity assays. After primary HTS, selected hits go through the cycle of testing and refinement.

However, conventional HTS still involves some limitations due to low sensitivity and a high false hit rate. Interestingly, these issues are not stand-alone; rather, they are two sides of a coin. While sensitivity refers to the percentage of true positives, a high false hit rate may result in false positives or false negatives, both of which could adversely affect sensitivity. With the advent of technologies and methodologies in drug discovery, there arises the belief that high-volume automation technologies in HTS could become absolutely indispensable. At present, a lot of biological data are uploaded to publicly available databases, including molecular structure, biological activity profiles, and target annotations. It is observed that off-the-shelf, ready-for-use, field-hardened machine learning algorithms and models are increasingly available, hence the time is right for the application of artificial intelligence in HTS to advance the natural history of the field.

Ultimately, the main goal of HTS is to identify as quickly as possible those compounds that stimulate or obstruct a particular cell-based assay or biochemical assay. The

principal advantage of conducting an HTS is the rapidity of the process. This rapidity makes HTS the first step in today's search for lead compound identification, and in some circumstances, HTS is effectively the only step. It is common practice to investigate hundreds of compounds initially, and leaps to HTS can reduce timelines for identifying lead compounds from years to just months.

2. Machine Learning in Drug Discovery

Machine learning has revolutionized many fields due to the ability of AI-based algorithms to explain complex data and make predictions. Basically, machine learning algorithms learn patterns by training on a large amount of data until prediction accuracy is high. In general, machine learning can be supervised or unsupervised, semi-supervised or active learning; here, we mostly focus on supervised and unsupervised learning, which are the most widely exploited in drug discovery processes. In supervised learning, the algorithm is trained on a set of input-output pairs. In the context of discovering new small molecules, this method can be applied to predict properties of chemical compounds (representing the input) by exploiting data on known molecules characterized by those properties of interest (output). Conversely, unsupervised learning does not utilize output data and may help in identifying trends, structures, or clusters in the data.

Machine learning has immediately met great expectations for drug discovery, as it can help to face the complexity of biological phenomena and the data generated. For example, it emerged as a valuable tool for the most labor-consuming phases of the drug discovery process, i.e., compound screening and optimization. Machine learning has been exploited to enhance hit expansion and chemical lead optimization. Currently available models can predict compound activity towards different biological targets, assuming the three-dimensional structure of the target is known. The prediction of binding affinity of protein-ligand has emerged as a key challenge addressed by state-of-the-art models. Moreover, models can be trained to predict on- and off-target effects, e.g., to solve the selectivity/specificity challenge. In fact, a major goal in drug discovery is to develop drugs that interact only with the target to maximize efficacy, avoid off-target effects, prevent side effects, and thus increase safety. In this context, several in silico multitask machine learning models can predict how molecules act towards different biological targets. For a drug candidate, in addition to its specific mechanism of

action, it is extremely important to understand the potential drugs it can interact with, to avoid drug-drug interactions. In this context, predictive models can be applied to the prediction of drug-drug, drug-food, or drug-disease-nutrient interactions. Compounds' clearance represents another critical aspect in pharmaceutical research. Several models can be computed to predict compound metabolic profiles, such as the sites susceptible to metabolism, the metabolic reactions, active enzymes, and CYPs.

2.1. Fundamentals of Machine Learning

Machine learning (ML) is a group of methods that use algorithms to identify patterns, associations, or features in input data and develop models according to learned understandings of the data to make predictions, decisions, or assignments on new, unseen cases. The inputs used by these algorithms are typically called training data, which consist of a set of input records, each with a set of corresponding outcome labels. A wide range of model evaluation metrics for assessing the effectiveness of models in different learning tasks exists. Basically, ML consists of three main learning paradigms: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is used for predicting future outcomes with labeled training examples, the outcome of interest being continuous or categorical. In the case of unsupervised learning, no labeled training examples are given, and the objective is to learn the data structure hidden inside the input space and group similar records. Reinforcement learning considers the state of an environment and has agents taking actions to maximize some notion of cumulative reward.

Depending on the measurable tasks we are facing, there are many off-the-shelf ML algorithms that we can employ. A few examples of the algorithms and methods that have been widely used in drug discovery are support vector machines, decision trees, random forests, neural networks, and deep learning. In addition to selecting appropriate algorithms and methods, the quality of the data and the process of selecting predictor features in the input data can significantly impact the effectiveness of a machine learning model. This is referred to as the feature engineering process. However, machine learning is not without challenges. In practice, a lot of effort in developing and applying machine learning is spent on training models and evaluating model performance. The primary difficulty associated with training and evaluating machine learning models for data scientists and researchers involves data, model, and implementation issues.

2.2. Applications of Machine Learning in Drug Discovery

Some decades ago, the idea of using computers to process and analyze chemical and biological entities was far from logical. At present, machine learning techniques are applied in almost every area of drug discovery, from early stages such as HTS to late stages such as sales prediction. Machine learning methods are used to enhance the technique of hit identification by building complex, non-obvious SARs in big HTS datasets. As such data is represented in a multidimensional space, techniques such as neural networks are usually used in competitive modeling. In addition, machine learning is used in lead optimization to predict ADMET properties, and those that are not the result of the SAR are used in secondary screens to predict, for example, off-target effects. Other domains that exploit the benefits of machine learning techniques include the study of scaffold hopping. Not only can these methods be used in synergy, but there are several examples of successful applications in the development of new compounds. Another well-suited process for machine learning technologies is toxicity prediction. Therefore, there are plenty of technologies that can be linked together to create synergy in drug discovery. The synergy of machine learning and data produced in the course of high-throughput screening is highlighted. Data produced in HTS, which is a primary contributor of candidates, biological screening can be combined with cheminformatics techniques to significantly speed the verification of compound properties. Pharmaceutical companies synthesize and provide industrial applications. Data quality, chemical and in vitro testing costs, the increasingly greater number of compounds to examine, AI-driven systems for analyzing high-throughput screening data in HTS, calculation time, and computational workload are the most prevalent limitations during technology transfer. These obstacles appear to be incompressible in the foreseeable future. In addition, another disadvantage of any machine learning model is that it can only produce results based on training data. Consequently, inconsistencies and errors in the training data translated into the model may remain unchanged. The model's behavior can alter as a result. If the propulsive dataset is less reliable than when the model is deployed in a new setting, the model could give heterogeneous and therefore flawed results. Furthermore, creating a QSAR model based on ambient data and determining it can be challenging. Model validation requires representatives of the query population, but it is not always feasible to test compounds from HTS against specific targets, particularly novel ones.

3. Challenges in Analyzing High-Throughput Screening Data

Analyzing high-throughput screening data presents several challenges due to the complexity of the data. HTS generates a large amount of data, making data analysis an essential part of the process. When analyzing data obtained from HTS assays, several preprocessing methods are required to reduce data complexity. For example, the data needs to be cleaned and normalized to deal with missing data points and technical artifacts. Data analysis of HTS assays also remains complex due to the noise and variability present in biological data. Quality control is another crucial step in data preprocessing, as it can help identify any batch effects in the data that result from the experiment or computational analysis pipeline. After data comes from multiple sources or multiple assays, it is hard to collectively analyze the integrated data.

Moreover, there are no standardized bioinformatics tools and databases for preprocessing and analysis of data from HTS studies. This indicates a data governance gap in the data management systems of the institutes, which leads to problems in the systematic analysis of data. While some training studies and reviews on machine learning techniques for preprocessing and mining molecular and imaging data from one HTS assay or one screening study are present, HTS user communities have not framed any working guidelines for institutions on the regulatory aspects of setting up institutional data management systems. Although several standard operating procedures for data management for HTS are available, they center on the experimental and laboratory aspects and lack a bioinformatics training or setup of robust and scalable data processing pipeline perspective. It is necessary to develop new methods and tools that can be leveraged to enhance the quality of data based on machine learning techniques.

3.1. Data Preprocessing and Cleaning

Prior to analyzing high-throughput screening data, it is important to carefully preprocess and clean the raw data. This is a crucial first step, as the rigorous preparation of HTS data can ensure that analyses yield more accurate and reliable results. Often, raw HTS data contain missing values, outliers, and inconsistent patterns. Necessary measures to ensure the accuracy and credibility of downstream analyses are aimed at addressing such challenges and enhancing data reliability. The main goal of

preprocessing is to minimize inconsistencies and biases induced by such absent or erroneous information that could affect the quality of subsequent analyses.

Addressing missing values, outliers, batch effects, and other inconsistencies in the data forms a crucial part of good experimental design. Missing values in both independent variables and outcome measures need to be addressed, as they affect subsequent machine learning-based models. A statistical or random forest imputation is commonly used to infer missing values. Outliers and extreme concentrations of compounds need to be normalized or labeled as 'dilution' depending on the experimental protocols. Other methods that machine learning may use to exclude or down-weight such outliers must be used with caution, since they may result in discarding important information rather than noise. Batch effects due to instrumental or biological inconsistencies can be addressed using average, quartile, or user-specified regressions that aim to remove variable importance due to batch variability. Moreover, gaps in features, targets, protocols, or treatment conditions in HTS may cause the overall models trained from such data to be less accurate than those without gaps. Often, treatments may not correspond to common conditions due to the heterogeneity of experiments. Operations such as normalization, transformation, and scaling may be necessary in feature processing. These are important steps to ensure that data used during model training occurs on similar scales to facilitate interpretation. Insufficient normalization and transformation can affect the outcome of machine learning model interpretations. Normalization of feature values may be necessary to ensure that machine learning models do not bias towards features that have a larger average value, a phenomenon that is prominent in weighted regressions.

4. Case Studies and Success Stories

The aim of all these case studies is to provide readers with tangible proof regarding the effectiveness of AI-driven solutions in the high-throughput screening (HTS) process. Four representative stories of how AI can be used in the HTS process to obtain better hit compounds have been included in this section. These stories are not meant to be comprehensive but to provide a quick overview, an easy and comfortable read. We have grouped the papers according to the target for which AI was used: 1. TTR, focusing on seasoned researchers; 2. Ubiquitin-like conjugation systems; 3. Non-alcoholic steatohepatitis; 4. BRAFV600E mutant protein, also focusing on a start-up whose AI was

used. The outcome of these case studies is impressive. As with many of the AI papers, there can also be problems. In setting up collaborations, working on wetware and the difficulties in data sharing between companies and academic groups can be painstaking. AI can be an effective way to expand the chemical diversity of a company's HTS collection. In the three case studies presented, the process of introducing an AI-driven approach to compound collection was not only completed but also showed good promise. The importance of a collaborative effort between chemists, biologists, and computational scientists is the key to success. At the same time, in the case of merging platforms owned by different companies, anatomy and sharing data from scientists are the main prerequisites to pave the way for this new form of collaboration. This aims to provide an overview of the use of AI in the HTS process that drives the need for a new type of 'marriage' between big players in the pharma business and new start-ups.

4.1. Examples of AI-Driven Drug Discovery Platforms

To demonstrate some of the recent work that AI-driven drug discovery companies have been performing, we selected examples based on the latest publications, the approaches they employ, and the methodologies they use for the integration of AI and HTS data. The examples given below illustrate how each of the companies approaches the problem of identifying the most active compounds or influential biologically relevant targets in a more efficient way using state-of-the-art algorithms. Some of these approaches can even predict target binding by the incorporation of multiomics sequence data.

Insilico Medicine, a company founded in 2014, is using machine learning for the large-scale examination of urinary and other markers. They are able to identify suitable aging-related drug targets. Their AI-driven system has been tested by this group and, using machine learning techniques, they were able to provide cardiologists with the most promising drug candidates for atherosclerosis and related chronic conditions more quickly than traditional control methods searched for in databases. The platform was able to identify approximately 30 compound classes that could be developed into leading candidates after testing them with standard HTS assays, rapidly quantifying markers linked with decreased or increased activity. Furthermore, a small case study using patients with overlooked diseases was also presented. The company showcased a platform based on deep generative chemistry and reinforcement learning that enables the development of complex molecules. In this platform, a range of techniques were

used, including GAN architecture, Graph Inception Networks, de-convolutional networks, DDBSCAN, and two hybrid techniques called RL-Gru-Dense clustering and RL-Junct-DenselyCLU. The agents were trained to identify molecules that are both structurally novel and likely to be active across more than 30 disease and tissue targets. Further validation was undertaken where multiple receptors were assayed for every compound. Training was based on existing generated compounds as well as a function, an indicator of breast cancer clinical outcome, and a customized tailor immune phenotype multi-omics function for immune cell analysis. Finally, they have provided reports from the patients trialing this approach. The AI system proposed and trained was able to identify a number of compounds which could be used as leads for an oncogene and therapeutic target, with compounds also having hits. Results showed that compounds with dual activity might be good leads for combinatorial therapy against a member of the epidermal receptor family. On testing the platform, it was able to successfully identify the clinical target of the test compound and the relevant pharmacodynamic assay. The speed of identification was given as 40 minutes, comprising 20 minutes of computer processing time and 20 minutes of manual interpretation.

5. Future Directions and Emerging Trends

Continuous advances in machine learning will lead to future algorithms that are also increasingly powerful and efficient. Notably, machine learning-driven models that use the full assay context, such as plate maps, will become a well-established tool in computational biology for analyzing HCA measurements. Future research in this area will likely integrate methods from diverse disciplines, including biology, chemistry, and machine learning. Together with larger scientific consortia and more contributors in this field, these AI-driven models will replace current time- and money-consuming systems in the pharmaceutical industry. It is becoming increasingly likely that scientists and other impacted stakeholders will demand more transparency of AI models as they aim to understand the underlying mechanisms.

Regulatory authorities will need to provide guidelines for potential AI model integration while also monitoring developments in the field. Moreover, competitive market developments within the pharmaceutical sector will increasingly shape the development and application of AI-driven systems. In personalized medicine and

precision therapies, AI applications will find particularly promising opportunities for expanded clinical use. The broader application of advanced data-driven systems that can analyze more difficult high-throughput screening data, such as mass spectrometry imaging or label-free data. The development of MSI AI-based approaches is a nascent field. Emerging technologies, such as quantum computing, will likely further complement the technological framework for data analytics that can rival current drug discovery systems.

6. Conclusion

The motivation for this essay is a simple one - to draw attention to the transformative role that artificially intelligent systems can have in the context of high-throughput screening and drug discovery. The challenges that have already been highlighted when it comes to data generation are enormous, but they are dwarfed by the task of analyzing and translating the data into biological insights that can help in the drug discovery process. Given the increasingly large and complex harmonization and integration of data from a myriad of high-throughput technologies, there is good reason to believe that machine learning will increasingly become a weapon of choice for analyzing and interpreting such valuable data in the years to come. As AI becomes increasingly sophisticated and continues to evolve, drug discovery projects will be able to take advantage of more and more of the toolbox that such techniques offer. The end result will be to make research and development more efficient and our treatment of patients more effective. The fact is that this has already begun to happen and a solid pipeline of AI-driven tools for screening and drug discovery already exists. These tools represent the evolutionary start of integrating AI into drug discovery efforts. AI is not just a new tool for drug discovery: it is a way of transforming the industry to ultimately become more efficient. The practical application of these tools is in characterizing target and probe biology, where they can benefit high-throughput screening campaigns as well as in designing the preparation of molecules for hit to lead and lead optimization programs. However, they are just the first drops in an oncoming storm on the AI horizon. Fundamental evolutionary areas in AI to be developed further include the use of different types of - and importantly multiple - ‘omics profiling datasets in AI-driven systems, and also embedding other phenotypic and physiological data, including patient-derived samples. The future collaborative efforts between AI experts, leaders in the field of high-throughput screening, target biology, chemistry, and clinicians will be

key in this ongoing and very exciting transformative period in the drug discovery paradigm. Nonethless, research into the capabilities of all AI approaches and work flows combined across the community with the physical and empirically driven laboratory experience is still in its infancy, as is the classic drug screening area after decades of practical application and community research.