The Applications of Machine Learning to Novel Drug Discovery

Siyu Lai

School of Architecture, Shandong University

larrylailai@gmail.com

Abstract. Drug discovery is a time-consuming, costly and often off-target discovery pipeline that nonetheless plays a crucial part in medical treatment field. In the past few decades, experimental assays remain the most reliable approach to screen compounds with huge cost. However, many artificial intelligence and machine learning algorithms have been implemented to modernize this field, such as through predicting molecular interactions or properties and analyzing biological data to identify potential drug targets, drug monitoring, and toxicity prediction. In summary, machine learning advancements provide critical support for logical drug design and discovery process, which could finally benefit all of humankind.

Keywords: Drug discoveries, Machine Learning, Data collection, Neural Network.

1. Introduction

The most vexing problem for drug discovery perhaps is the rising cost of finding new medicines. Reports said it cost average 800 million US dollars before producing a new drug, which added to 2.6 billion in 2013 [1]. Another problem is that FDA (Food and Drug Administration) places strict limitation on the new drug side-effect confirmation before it formally comes into market. As a result, although the total funding devoted to drug discovery is rising rapidly each year, the total number of new drugs that pass the FDA regulation continues to decline on an annual basis. [2]

Currently, a large quantity of biological big data is organized by scientists all over the world with open access. These databases provide abundant information among biological factors like existing drugs, cell lines, proteins and genes. Given the fact that using machine learning models on these datasets are much cost-effective than vitro or any other practical experiments, they have become understandably popular.

In this paper, my prime purpose is to guide people to easily and comprehensively understand how machine learning facilitates drug discovery process and the huge advantages that lie behind it.

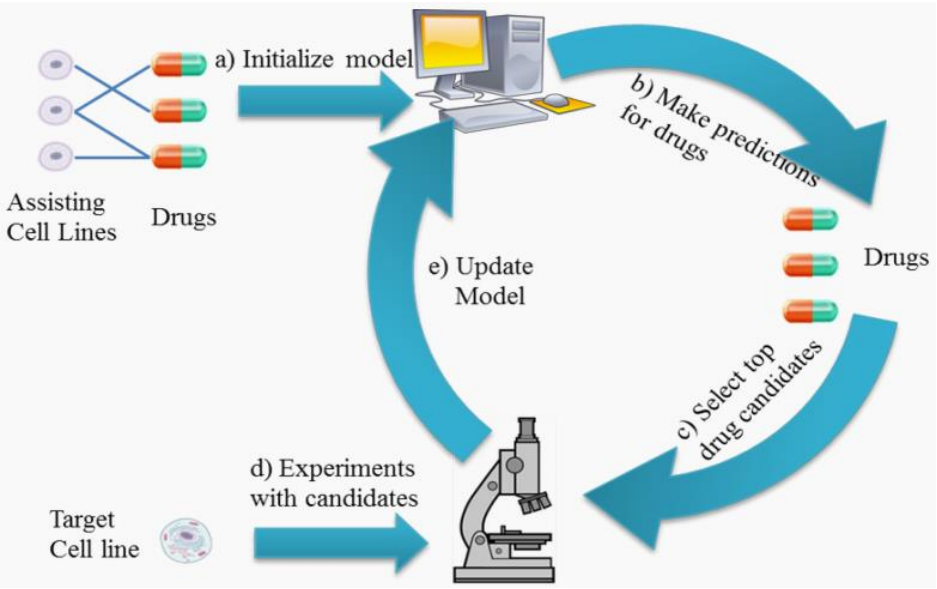

Figure 1. The work flow of computational drug discovery

2. Traditional Drug Design Process

Discovering new drugs is a complex and resource-intensive endeavor that relies on a combination of traditional methods and cutting-edge technologies. Traditional methods have played a fundamental role in drug discovery for centuries, providing a solid foundation upon which modern pharmaceutical research is built. In the following outline, I will present the ways in which traditional methods contribute to the creation of new drugs.

In drug discovery, the first and most crucial step is identifying appropriate targets related to the pathophysiology of diseases (such as genes or proteins), and then finding drugs or drug-like molecules that can interfere with these targets. Nowadays, we can achieve this by searching various biomedical databases such as NCBI GEO, The Cancer Genome Atlas (TCGA), and Arrayexpress, among others. Sometimes, even published literature can be used to identify targets. PubMed, for instance, is a database of various published biomedical literature, and mining this data can help identify targets for different diseases.

After identifying and validating suitable targets, the next step is to search for appropriate drugs or drug-like molecules that can interact with these targets and elicit the desired response. In the era of big data, we have access to vast chemical databases that can assist us in finding the perfect drugs targeting specific targets.

2.1. Synthesis and Medicinal Chemistry

Traditional medicinal chemistry plays a pivotal role in drug development. It involves the design and chemical synthesis of new compounds with specific therapeutic properties. Medicinal chemists modify existing molecules or create entirely new ones, aiming for improved efficacy, reduced side effects, and enhanced bio-availability. This process requires a deep understanding of chemical principles and structure-activity relationships that hitherto have largely depended on the knowledge of an experienced scientist.

2.2. Clinical Trials

Traditional experiments, involving in vitro and animal studies, is essential to assess the safety and efficacy of potential drug candidates. Animal models are used to understand a drug's pharmacology, toxicology, and potential side-effects before moving to human trials (empirical screening).

2.3. Post-Market Surveillance

Even after a drug is approved, traditional methods are employed to monitor its safety in the broader population. Adverse events and side effects are tracked, and appropriate actions are taken to safeguard public health.

Before computational datasets became involved, there were no systematic data sources for scientists to refer to. Furthermore, all the results would only have been available to a limited institution, which obviously dampened the production of new medical ideas.

3. Properties of Machine Learning

3.1. Simple Neural Networks

Neural networks are inspired by neuroscience and imitate the biological neural network (which consists of thousands of single neurons which are all connected with each other in different ways and approaches), while it was initially applied to drug discovery simply for analyzing datasets from patients with new algorithms. With the advanced technology in transistors, a single chip can contain over one billion transistors. The graphic processing unit (GPU) can perform complex computation, which accelerates the processing of the image. Finally, it can be used in practical image analysis and 3D virtual prediction.

However, if a fully connected neural network is used to process large-sized images, it has three obvious disadvantages: first, flattening the image into a vector will lose spatial information; second, having too many parameters leads to inefficiency and difficulties in training; finally, a large number of parameters quickly leads to network overfitting.

3.2. Convoluted Neural Network

Unlike conventional neural networks, the neurons in the layers of convoluted neural networks (CNNs) are arranged in three dimensions: width, height, and depth. The width and height are easily understandable since convolution itself is a two-dimensional operation. However, in convoluted neural networks, depth refers to the third dimension of the activation data volume, not the depth of the entire network. The depth of the entire network refers to the number of layers in the network.

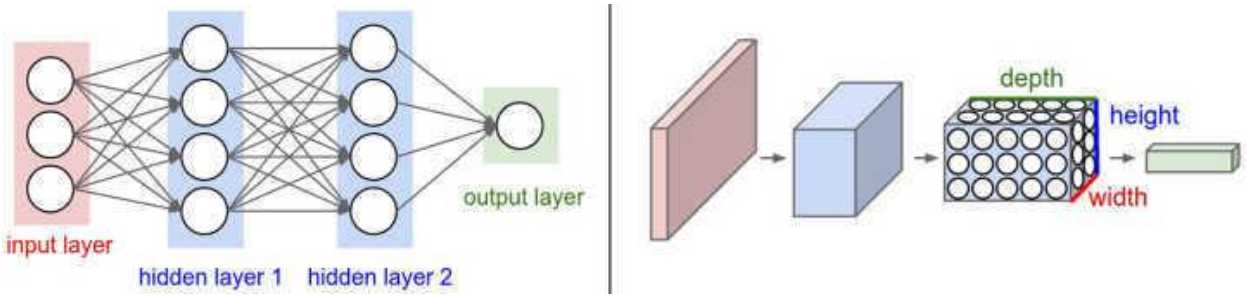

Figure 2. Comparing all-connected neural network with convoluted neural network

Essentially, convoluted neural networks are a regularized type of feed-forth neural network, which learning by itself via filter optimization. The function that is applied to the input values is determined by a vector of weights and a bias (typically real numbers). Learning consists of iteratively adjusting these biases and weights. The vectors of weights and biases are called filters and represent particular features of the input. [3]

Although first invented in the early 1980s, they only came into practical use until 2000 because of the high GPU requirements. Computational calculation can measure much more aspects of the cell about those cellular systems and see what it is correlated. As a result, there is an increasing interest in using computational, data science, informatics tools and AI for drug discovery. The initial impulse was to have computations simply manage large amounts of data, but there has been a shift towards machine learning.

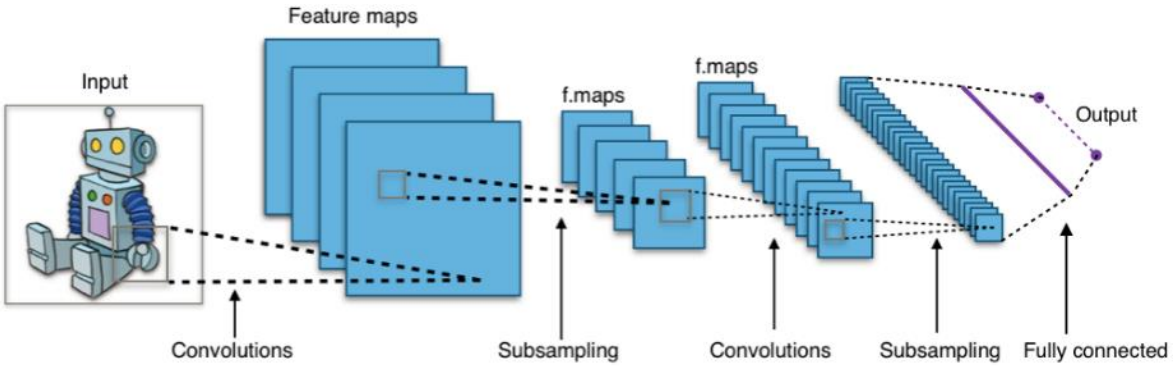

Figure 3. This figure depicts how input image transfer into feature maps and subsampling

3.2.1 Molecular Image Analysis

CNNs now are trained to analyze images of molecular structures, such as its chemical compounds. At first, we use known drugs to extract features that are relevant for drug design. This can aid in the prediction of a properties and interactions.

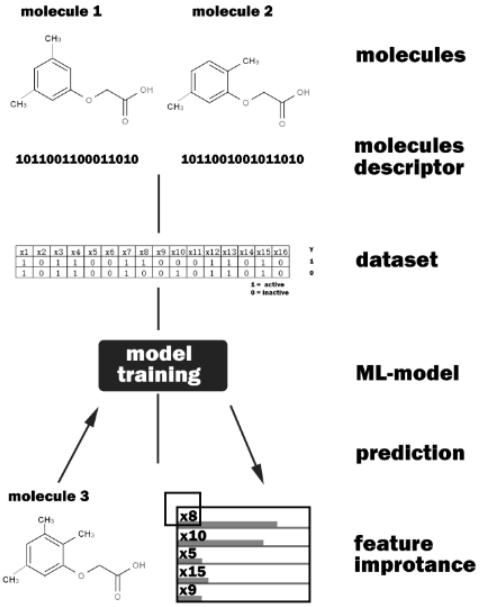

Figure 4. This figure depicts Quantitative Structure-Activity Relationship (QSAR) it is a technique that applies machine learning in order to learn the relationship between chemical structure and the biological activity of interest, which entails the collection of a data set of molecules. The chemical structure will be subjective to calculation of their molecular descriptors, which is binary representation. then all these molecule info are restored in the datasets which can be used to train the model so that it can predict the new molecules and enable the provision of insight into which features are important. Such information can be helpful for biologist and chemist in the design of future molecules. [4]

General workflow:

A. The data set of molecules to be used for the model building is preprocessed. Duplicate structures, doubtful biological values, etc. are eliminated.

B. A series of molecular descriptors is calculated for this data set, producing a matrix of data with as many columns as molecules and as many rows as descriptors.

C. The data matrix is randomly divided into a training set and a test or validation set.

E. The model is built over the training set, using different techniques to produce a predictive model.

F. The model is validated by measuring its effectiveness on the test set.

Since 2013, public attention was drawn to a QSAR machine-learning challenge in drug discovery supported by Merck. A general task for machine learning is to uncover the relationship between the molecular descriptors used and the measured activity of the compounds to complete the model.

The features they extract from those process are the decisive component of model development pipelines.

3.2.2 Virtual Screening

Instead of filtering out the right drug candidates through endless experiments, CNNs can be set for virtual screening of compound libraries against specific drug targets. They can efficiently filter out compounds with low binding affinity, reducing the number of compounds that need to be actually tested. They may also detect subtle change in the cell that are produced by different compounds and thereby find out how compounds work.

Take the case of Deep Dock platform. Through this platform, the score classes (top or low scoring) of the remaining molecules are then inferred rather than explicitly calculated with actual docking. In the end, only the best-predicted scoring molecules remain to be conventionally docked, whereas unfavorable molecules are filtered out.

3.3. Nature Language Process

Nature Language Process (NLP) models can extract valuable insights from scientific literature, helping researchers stay updated on the latest discoveries and trends. By analyzing research paper and patents, they can select potential drug candidates and possible compound structure. It is usually applied in medicine (interpreting or summarizing electronic health records) and in document retrieval systems mainly for executing two processes: indexing and matching. In most modern systems, indexing is done by a vector space model through Two-Tower Networks, while matching is done using similarity or distance scores.

3.4. Deep Neural Networks

Deep neural networks are optimized convoluted neural networks, which add several hidden layers and are capable of calculating the non-linear features that capture extremely complicated data with each additional layer and which may be suitable for data mining in bio-medicine field because of their efficiency.

4. Application of the Machine Learning Process

In the process of drug discovery and development, the most daunting and frustrating step is finding suitable, biologically active drug molecules within the vast chemical space. What's even more disheartening is that nine out of ten drug molecules usually fail to pass the second stage of clinical trials and other regulatory approvals. The aforementioned challenges can be addressed by implementing tools and techniques based on artificial intelligence. Machine learning can be involved in every stage of the drug development process.

4.1. Primary and Second drug screening

In drug discovery, the screening of lead compounds is of paramount importance, and artificial intelligence plays a significant role in identifying new and potential lead compounds. In the chemical space, there are approximately 106 million chemical structures derived from various sources of research, including genomics, clinical and preclinical studies, in vivo analysis, and microarray analysis. Machine learning models such as reinforcement models, logistic models, regression models, and generative models are used to screen these chemical structures based on their activity at binding sites, structures, and target binding affinities.

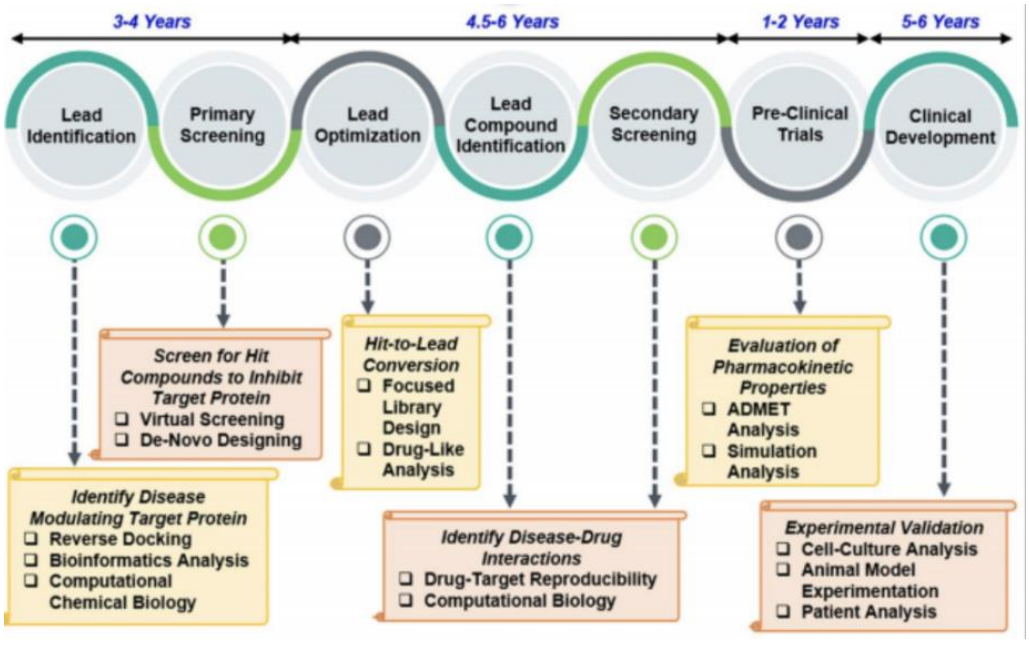

Figure 5. This figure depicts the process of primary and secondary screening

4.2. Structure-based and ligand-based virtual screening

Virtual screening is one of the essential methods in Computer-Aided Drug Design (CADD), which is an effective approach for screening promising therapeutic compounds from compound libraries. While being a crucial tool for high-throughput screening, it also presents challenges such as high costs and low accuracy. To apply machine learning to virtual screening, there should be a training dataset consisting of known active and inactive compounds. These training data are used to train the model using supervised learning techniques. The trained model is then validated, and if it is accurate enough, it is applied to a new dataset to screen compounds with the desired activity against the target. machine learning can accelerate the speed of virtual screening, making it more efficient and even reducing false positives in the process.

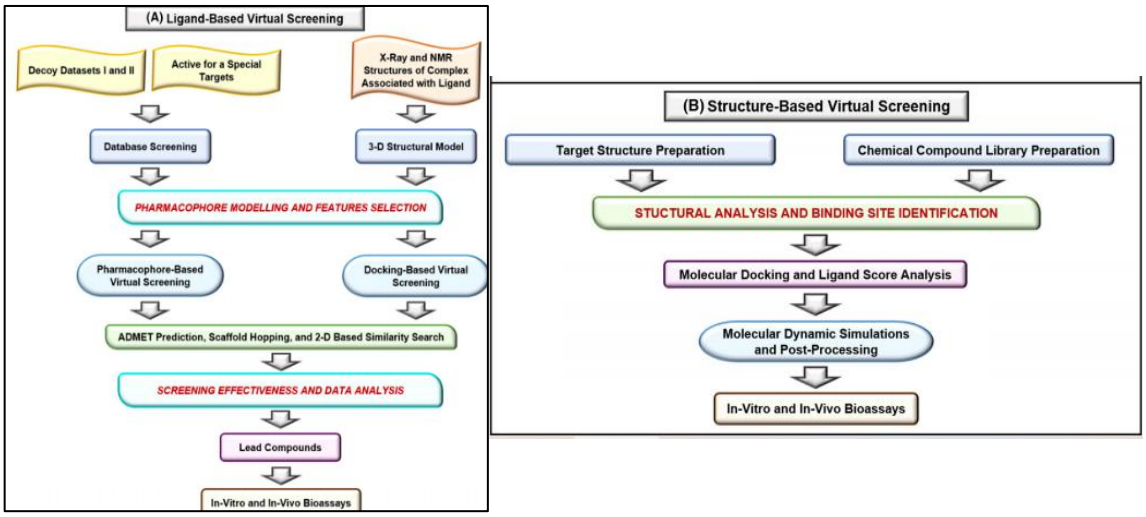

Figure 6. This figure depicts the process of Structure based and ligand base virtual screening [5]

4.3. QSAR modeling and drug retargeting

In drug design and development, understanding the relationship between chemical structure, physicochemical properties, and biological activity is crucial. Quantitative Structure-Activity Relationship (QSAR) modeling is a computational approach used to establish quantitative mathematical models between chemical structure and biological activity. Traditional QSAR models are broadly categorized into two types: regression models (such as Gaussian Processes, GPs) and classification models. [6]

4.4. Peptide synthesis and small molecule design

Peptides are biologically active short chains composed of approximately 2-50 amino acids. Due to their ability to cross cellular barriers and reach desired targets, they are increasingly being used in therapies. In recent years, researchers have utilized the advantages of artificial intelligence to discover new peptides. For instance, Yan in 2020 developed Deep-AmPEP30, a DL-based platform for the identification of short anti-microbial peptides (AMPs). Deep-AmPEP30 is a CNN-driven tool that predicts short AMPs from DNA sequence data. [7]

4.5. Predicting bioactive agents and monitoring of drug release

Various online tools have been developed recently to analyze drug release and assess the feasibility of selected biologically active compounds as carriers. The most commonly used method is the pharmacophore evaluation based on chemical features. To study ligand-based chemical properties, various successful experiments have been established using the CATALYST program. Additionally, researchers can utilize artificial intelligence to identify biologically active compounds for specific targets related to diseases. For instance, Wu et al. integrated DL and RF methods to devise WDL-RF for determining bioactivity of G protein-coupled receptors (GPCRs) targeting ligands. Likewise,

Cichonska et al. [8] developed pairwiseMKL, a multiple kernel learning-based method, for determining the bioactivity of compounds. [9]

5. Summary

Despite these advanced applications of machine learning technology, the currently success rate of drugs across the process of discovery and testing is still very low, at about 10%. Now, there is increasing interest in actually placing the AI or the computational system at the heart of the discovery process. These methods range from very low-level physics-based interaction of molecules with their actions, models of cells, analysis of patient data, to the integration of all these aspects.

At present, the major challenge for the pharmaceutical industry while developing a new drug is its increased costs and reduced efficiency. However, machine learning approaches and recent developments in Deep learning come with great opportunities to reduce this cost, increase efficiency, and save time during the drug discovery and development process. [10]

Just as Rohan Gupta said, professor in Department of Biotechnology in Delhi Technological University has said, "Though there are some unavoidable barriers and numerous amounts of work must be done to incorporate AI tools into the drug discovery cycle, there is no doubt that in the near future AI will bring revolutionary changes in the drug discovery and development process."

References

[1] A. Mullard, "New drugs cost us [dollar] 2.6 billion to develop," Nature Reviews Drug Discovery, vol. 13, no. 12, pp. 877–877, 2014.

[2] L. Fricker, "Drug discovery over the past thirty years: Why are not there more new drugs?" Einstein Journal of Biology and Medicine, vol. 29, no. 1, pp. 61–65, 2016.

[3] Gupta, R., Srivastava, D., Sahu, M. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25, 1315–1360 (2021).

[4] Carpio, L.E., Sanz, Y., Gozalbes, R. et al. Computational strategies for the discovery of biological functions of health foods, nutraceuticals and cosmeceuticals: a review. Mol Divers 25, 1425–1438 (2021).

[5] Gupta, R., Srivastava, D., Sahu, M. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25, 1315–1360 (2021).

[6] Gupta, R., Srivastava, D., Sahu, M. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25, 1315–1360 (2021).

[7] Yan J, Bhadra P, Li A et al (2020) Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. Mol Ther-Nucleic Acids.

[8] Wu J, Zhang Q, Wu W et al (2018) WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest. Bioinformatics.

[9] Cichonska A, Pahikkala T, Szedmak S et al (2018) Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics 34: i509–i518.

[10] Gupta, R., Srivastava, D., Sahu, M. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25, 1315–1360 (2021).