CASE STUDY on Application of Bioinformatics

Title of Paper:

Machine Learning Methods in Drug Discovery. [Lauv Patel, Tripti Shukla, Xiuzhen Huang, David W. Ussery and Shanzhi Wang]

Abstract:

Advancements in IT and related processing techniques have taken progress in scientific fields leaps and bounds ahead of the projected progress rate. This has now bled into Drug discovery, as Machine learning is now being used to enhance the efficiency, efficacy and quality of novel drug candidates. Big data incorporation techniques such as high-throughput screening and high-througput computational analysis of databases is being use to lead and target discoverabilty, this has in turn increased the reliability of the machine learning and deep learning techniques.

1. Introductions:

Advancements in computational science, Artificial Intelligence (AI) and Machine learning (ML) an essential component of AI, has accelerated drug discovery and development. ML models have been used in many promising technologies such as deep learning (DL) assisted self-driving cars, advanced speech recognition, support vertor machine-based smarter search engines, etc. Drug discovery has been based on a traditional approach that focuses on holistic treatment. The world's medical communities have started to used allopathic approach to medicine over the last few centuries which has increased disease fighting chances but the drug costs have also increased dramatically. The applications of ML and DL algorithms in drug discovery are not limited to a specific step, but for the whole process.

In this paper, the ML and DL algorithms that have been widely used are discussed.

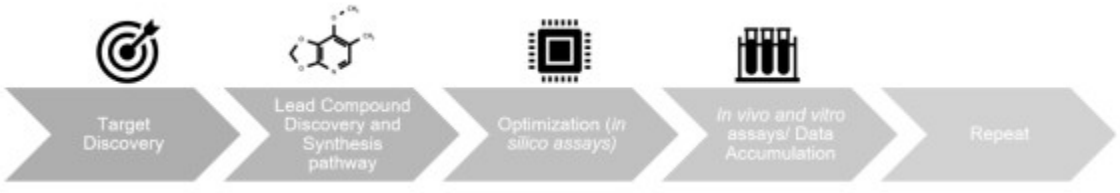

Fig1. General steps in drug discovery

2. ML Algorithms used in drug discovery:

Pharmaceutical organizations have extraordinarily profited from the use of different ML calculations in drug discovery. ML calculations have been utilized to foster modes for predicting chemical, biological and physical characteristics of compounds in drug dicovery. For instance, ML calculations have been utilized to foresee drug protein interactions, find drug viability, guarantee safety biomarkers and enhance the bioactivity of molecules. ML is used in Random Forest (RF), Naive Bayesian (NB) and support vector machine (SVM) methods, which we will be talking about later.

There are two fundamental kinds of ML algorithms : Supervised and unsupervised learning. Supervised learning gains from tests with realized marks to decide new examples. Unsupervised learning perceives

patterns in a bunch of tests without labels. Supervised and unsupervised learning can be combined into semi-supervised and reinforcement learning, where the two can be used for different datasets. The advancements of data analytics have succesfully attempted to describe and interpret the generated data. Using ML methods, like generalized linear models through NB, the issues of analysis and interpretation of data may be unburdened.

3. Random Forest (RF):

RF is widely used algorithm explicitly designed for large datasets with multiple feaures, as it simplifies by removing outliers, as well as classify and designate datasets based on relative features classified for the particular algorithm. The mathematical process of RF consists of several uncorrelated decision trees as an ensemble: each tree is responsible for determining one prediction. The tree that constitues the most votes is considered the best fit. Although false positives may happen in any statistical analysis, RF along with SVM and NB is considered to make the least amount of errors.

In drug discovery, RF is mainly used either for feature selection, classifier or regression. Some of the essential factors accompanying RF in drug discovery are: It expedites the training process, uses fewer parameters, imputes missing data and incorporates nonparametric data. Multivariate RFs specialize in limiting error by calculating several error estimates techniques within the system. RF was incorporated for the generation of the regression tree node and leaf nodes. RF algorithms have been implemented as a method of classification and regression in a quantitative structure-activity relationship (QSAR) modling used in lead discovery.

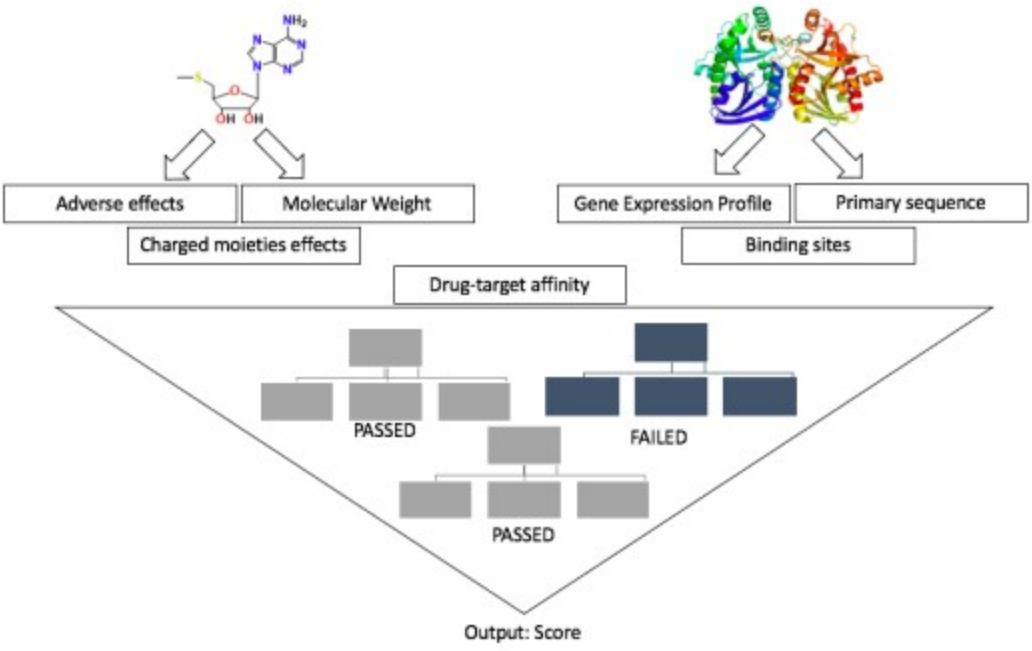

Fig2. Schematic view of drag development using random forest (RF)

4. Naive Bayesian (NB):

NB algorithms are a subset of supervised learning methods that have become an essential tool used in predictive modeling classification. Standard algorithms work to classify features of datasets, and depending on the input characteristics, factor correlation and dimensionality of the data. NB techniques could also serve important roles in predicting ligand-target interactions, which could be a massive step forward in lead discovery. In a recent study NB along with SVM was utilized in combination to predict possible compounds that could be active against targets of human immunodeficieny virus type-1 and hepatitis C virus generated from multiple QSAR models.

Use of NB in combination with other systems and techniques has shown to be very useful in drug discovery process.

5. Support Vector Machine (SVM):

SVMs are supervised machine learning algorithms used in drug discovery to separate classes of compounds based on the feature selector by deriving a hyperplane. SVM uses the similarities between classes to formulate infinite numbers of the hyperplane. SVM is curcial to drug discovery because of its capability of distinguishing between active and inactive compounds, ranking compounds from each database or trainnig regression model. Regression models are vital in determining the relationship between the frug and ligand, as it employs a query for datasets to predict. In case of drug discovery SVM can rank compounds from different databases based on the probability of being active for any computational screening. The process could be manipulated by training the algorithm using various descriptors for feature selectors such as 2D fingerprints and target protein.

In case of drag-target interaction, it is specifically designed for integratng ligands and proteins of interest information as an essential component for SVM modeling. Kernel functions were used to incorporate information on drug pharmacological and therapeutic effects, drug chemical stuctures and protein genomic information to characterize the drug-target interactions. The results from all the different sources were promising but the kernal function for prediction showed the most potential. SVM is maining used to predict drugs that could have multiple bioactivities.

6. Limitations:

ML algorithms are trained with inputted data, this is the major limitation of ML algorithms. Even though ML has been around for quite some time now, the biological pathways/targets being discovered are still new. Information for a particular protein of interest may be limited. Not all data is gathered from wet-lab, computer generated prediction is utilized, this leads to inaccuracy in the training data and futher more the results. Even though the algorithms discussed above have a higher threshold for minimizing erros there are still some categorial cerrors from the training datasets.

In algortithm based predictions there is always concern for overfitting or underfitting. Overfitting is when the model consists of lower quality information/technique but generates higher quality performance. It occurs when the model picks up unusual features during the training, resulting in a negative impact on the model. Underfitting models fail to recognize the data sets, underlying trend and generalize the new data inputted. Cross validation is an often-used techniques used to estimate the accuracy of the ML algorithms' models, by using independent data sets to infer the models.

7. Deep learning (DL) Methods:

Deep learning are considered one of the cutting-edge areas of development and study in almost all scientific and techonological fields. DL has allowed resolving many challenges faced by standard ML algorithms, including image recognition and speech recognition. The basis of DL is often imlpicated in NN systems, where they are used to create systems that have the capability to complete complex data recognition, interpretation, and generation. The main subsets of artifical NNs used in current drug discovery are deep neural networks (DNNs), recurrent neural networks (RNNs) and convolution neural networks (CNNs).

The utilization of specific NNs from variations that exist in the subset is dependent on multiple factors. DNNs, a specific type of feed forward neural networks, function with signular path data flow from the input layer through the hidden layers reaching an output layer. A generative DNN can create novel chemical compounds from existing libraries and training sets; while predictive DNN can predict the chemical attributes of the novel compounds. QSAR models are currently being used to find the correlation between these compounds' chemical structure and activity.

There are always some error sources and imprecision over the multiplicity of studies conducted using these AI algorithms. It has been found the NNs face a few deficiencies in comparison to other ML algorithms in their application of QSAR studies. QSAr studies is te most advanced form of DL based AI in current drug discovery and development. It has allowed researchers to take 2D chemical structures and determine physicochemical descriptors related to the molecule's acticity. 3D QSAr has allowed further inquiry of geometric structure impacting ligan-targeting interactions. NNs face a few deficiencies in comparison to other ML algorithms in their applications of QSAR studies. The presence of excess descriptors is a major problem, this causes redundancy in NN and eventual cloggin of outputs. These issues have been alleviated using more specific feature selection algorithms to get a smaller number of higher quality descriptors.

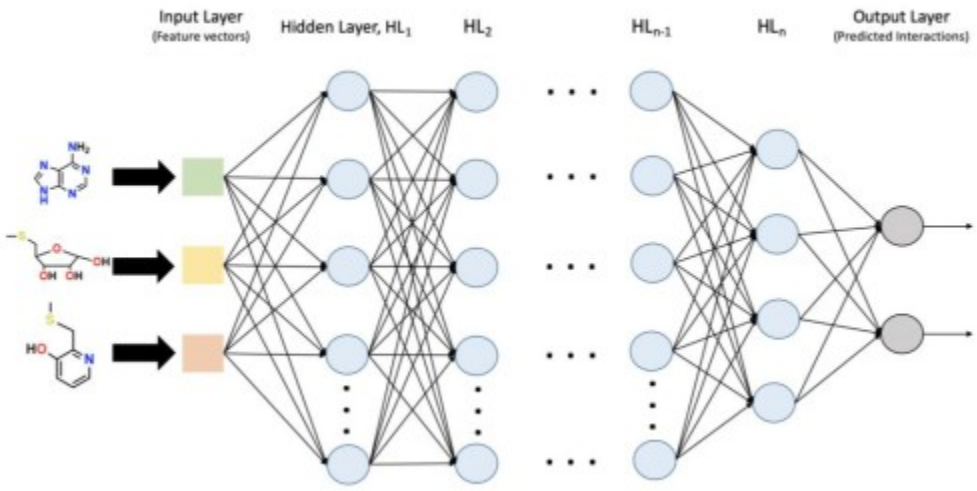

Fig3. General Scheme of deep neural network DNN and recurrent neural network RNN

RNNs are used in descriptive simplified molecular-input line-entry system (SMILES) nomeclatures in much of the algorithms regarding de novo deug design and discovery. The subset RNN-type long short-

term memory have become a reliable, standardized method for generating novel chemical structures. RNNs are unqire in their ability to use neuros connected in the same hidden layer tof orm a fucntioning cycle of processing inputes and outputs compared to DNNs and feed forward neural networks which have no connections within the same layer and only push outputs. These generative RNNs have shown promising results in the generation of sensible, structurally, correct and feasibile, novel SMILE strcutures that were not included in the original SMILE training sets.

CNNS are a subset og DNNs that take inputs, assign weights to specific parts of the input then build the ability to differentiate the data. While traditional DNNs are limited in their ability to function correctly on higher-deminsional datasets, CNNs server as a gleaming solution to tackling this issue with their ability to preserve input dimensionality. Training required by CNN model is significantly less than that of DNN and RNN. All these advantages has allowed CNN to become a very important and sought after learning algorithm for image recognition, surpassing other standard ML algorithms. Combination of there DL techniques such as CNNs, have been very successful in identifying gene mutations and disease targets.

8. Conclusion:

ML based techniques seek to revitalize the development of drugs. These methods are based on separate applications in target discovery, lead compound discovery, synthesis, protein-ligand interactions, etc. ML applications are paving they way for algorithm-enhanced data query, analysis, anad generation. Even though applications of ML have been used in drug development for a while now there is still a long way to go. Not to take away from the achievements though, ML has accelerated the pace of drug discovery and also increased accuracy up to 90% compared to drug discovery in the 60s. Owing to more precise algorithms, more powerful supercomputers, substantial preivate and public investment in the field, these applications are becoming more intelligent, cost-effective and time efficient while coosting efficacy.

9. Reference

1. **Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine Learning Methods in Drug Discovery.** ***Molecules*****, *25*(22), 5277. https://doi.org/10.3390/molecules25225277**