ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2021; 7(4): 21-28
www.allresearchjournal.com
Received: 12-02-2021
Accepted: 21-03-2021

Ravi Manne
Biochemist, Chemtex Environmental Lab, Port Arthur, Texas, 77642, USA

Corresponding Author:
Ravi Manne
Biochemist, Chemtex Environmental Lab, Port Arthur, Texas, 77642, USA

Machine Learning Techniques in Drug Discovery and Development

Ravi Manne

DOI: https://doi.org/10.22271/allresearch.2021.v7.i4a.8455

Abstract
The advancement and progress in technology and related techniques have created an opportunity for progress in many scientific fields and various industries. Machine learning has become important tool for drug designs and discovery with the availability of bit data from large databases. IN this paper I analyze Machine Learning and Deep learning techniques which help Pharma industry in all stages of drug discovery which includes target validation, prognostic biomarkers, clinical trials.

Keywords: Drug discovery, Pharma industry, clinical trials

Introduction
Machine Learning methods and techniques and tools are used to help solve diagnostic and prognostic problems in wide medical domains. It is also being used to analyze the importance of clinical parameters and their combination for prognosis.
Advancements in technology in the medicinal field have accelerated development in medical field and also accuracy. Machine learning has influenced wide variety of task in modeling and cheminformatics like synthesis planning [1], toxicity prediction, and virtual screening. Artificial intelligence is being widely used in medicinal field and in drug and pharmaceutical field. Machine Learning is a subfield of AI and requires computational and mathematical theory. Machine learning is based on developing models from the exposure to training data. Machine learning now a days can be used with wide variety of data types, and methods, like imaging, protein structures, and instead of being restricted to certain data types previously like protein sequences and compounds. Use of machine learning for drug discovery has been growing continuously, which is yielding good results by using pattern recognition algorithms, discerning, mathematical relationships between empirical observations of small molecules and extrapolate them to predict chemical, biological and physical properties of novel compounds in contrast to the models which rely on explicit physical equations [2]. There are also certain limitations like need for huge amount of data, lack of interpretability etc. In comparison to physical model's machine learning techniques can also be easily used on big data sets without having the need for computational resources.,
There are totally seven phases in the process of drug discovery. Let us discuss about the phases.

1. Target Identification: Discovery: The first step of this process is target identification, it's not even about the drug, this phase is more responsible about understanding the targets. Target contains of misfold proteins, potential disease biomarkers and DNA mutations.
To develop the drug even though the possible way is to first identify targets and then work on the drug development, most of the times it's hard for human beings to identify all the possible combination of compounds. For most of the drugs this process takes for about 2 plus years.

2. Lead Discovery: pre-clinical: This is the second step of the cycle. IN this process we screen thousands of compounds which can interfere with the disease targets. In this step we can narrow down the potential compounds that can act on target. Usually, this cycle takes 1 to 2 years.

3. Medicinal chemistry: Pre-clinical: In this phase the narrowed compounds are further tested to analyze the interactions with the targets that caused the disease. Some analysis are carried by taking 3D configurations of compounds and interacting their interactions with disease targets. The results are taken from analysis and are further optimized towards the targets. This phase also takes around 1 to 2 years.

4. *In vitro* Studies: Pre-Clinical: compounds that are filtered until this stage are tested in cell system. *In vitro* studies is the phase where petri dish studies occur. In this phase the effectiveness of drug is tested by scrutinizing the compound that interferes with target.

5. *In vivo* Studies: Animal Studies: In this phase the compounds that pass through In Vitor phase are taken and tested on animals like rats or mouse. Compared to 2D *In vitro* cell structure models the results that are obtained in these animal studies are more representative. Due to the difference in the architecture of the cell model in animals, failure at his stage is also higher, and the results from *In vitro* may not correlate with *In vivo*.

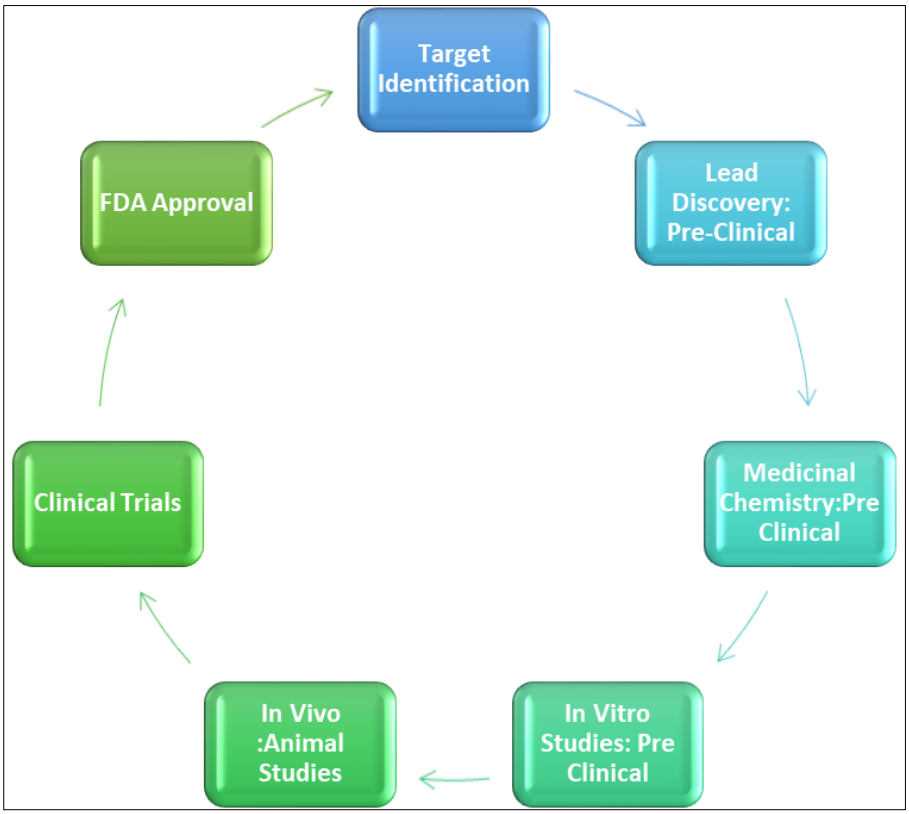


Fig 1: Phases in Drug Discovery process

6. Clinical Trails: The compound that has showed some promising features in the last phase are then proceeded to clinical trials. In this phase the trials are being done on human volunteers.

7. FDA Approval: The compound that passes all these phases is submitted to FDA for approval. As it passes the approval from FDA its available in the market for public to use.

The drugs which are approved in the year 2005 to 2006 took an average clinical development time of six and half years, and from 2008 to 2012 took an average time of 9.1 years [3]. In the later stage of clinical testing the rate of drug failure has increased, and this has always been a critical issue. Phase 2 and phase 3 Clinical trails that occurred between 1998 to 2008 has a failure rate of 54%. Safety concerns accounts to 17% in the rate of failure and lack of efficacy is another reason which accounts for the rest. Side effects and risk of death are also import reasons in phase 2 and phase3 drug failure [3, 4, 5]. As the failure of drug and the time consuming process, as it takes really long periods along with huge expenses can be frustrating, especially when we trails were not enough successful. Machine learning helps in this process where it learns from past experience and past data and helps eliminate some of the unknown factors and reduce human effort, expenses and time.

Machine Learning Algorithms Used in Drug Delivery

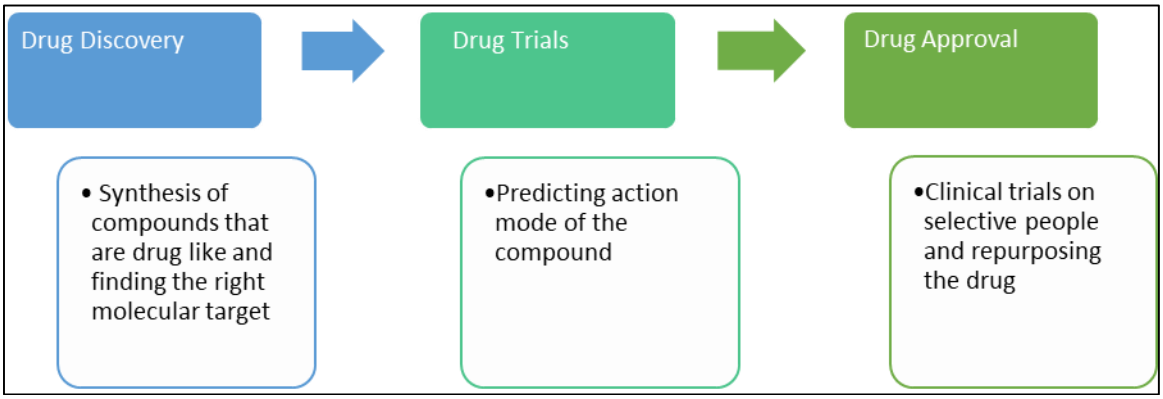


Fig 2: Machine learning flow in Drug Development

Machine learning is a branch of Artificial intelligence (AI). There are three types of categories in machine learning.
1. Supervised machine learning
2. Unsupervised machine learning
3. Reinforcement machine learning

Supervised machine learning: Supervised machine learning is defined by using labeled datasets for algorithm training that is used to classify the data or predict the outcomes accurately. When the input is fed into the model, it will adjust the weights using reinforcement learning process which we will discuss next, and that will ensure that model is perfectly fitted. Supervised learning helps many field and organizations solve variety of real-world problems.

Supervised learning can be separated into classification and regression problems.

Classification: Classification uses an algorithm to accurately classify and separate data into specific categories. It recognizes certain entities within the given dataset and tries to draw conclusions around those entities on how they should be labeled.

Regression: To understand the relation between dependent and independent variables regression is used.
Some commonly used algorithms for supervised learning are SVM, K-nearest neighbor (KNN), decision trees, Random Forest, Naive Bayesian classifier, linear regression, polynomial regression, logistical regression.

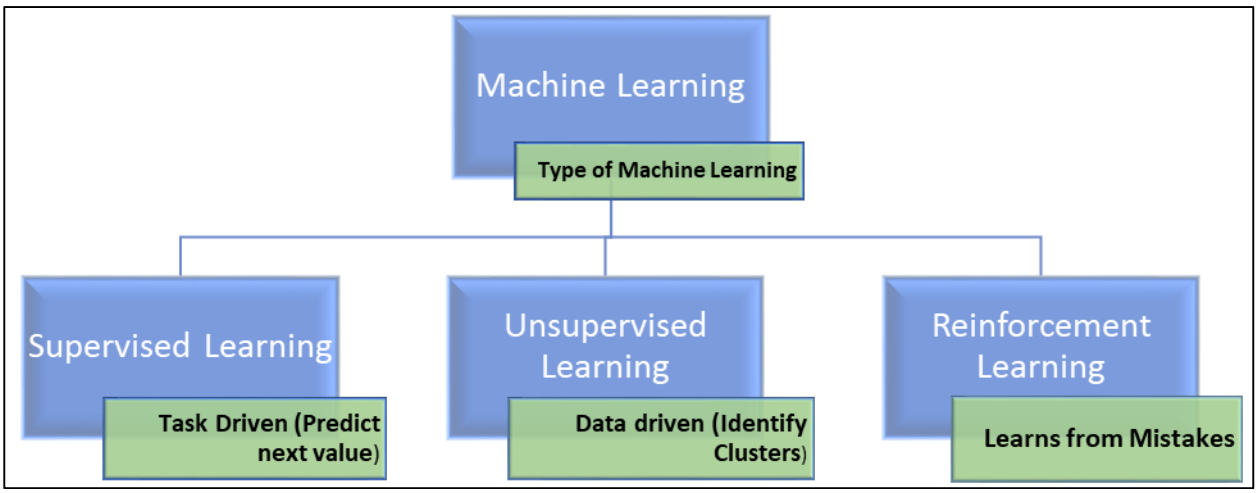


Fig 3: Types of Machine Learning

Unsupervised machine learning: Unsupervised learning is when you have the input variable, but the corresponding output variables are not there. The main aim of the unsupervised learning is to learn more about the data by understanding the distribution among the data. This can be further grouped into association and clustering.

Clustering: In a clustering problem you group the data depending on certain pattern or behavior.

Association: In association you want to discover rules which describes large part of your data.

Reinforcement machine learning: In reinforcement learning the is no pretrained data, the model decides what to perform to a given task. In supervised learning the training data will have the answer key, and the model is trained with the correct answer. In reinforcement learning, in the absence of training data, the model learns from its experience. There are two types in reinforcement learning. Positive and negative. Positive reinforcement learning is defined as the event that occurs due to a certain behavior, which increases the frequency and strength of the behavior. Advantages of positive reinforcement learning is it maximizes the performance. And sustain change for longer time period.

Negative reinforcement is, strengthening the behavior when a negative condition is stopped. Advantages of negative reinforcement learning is it increase the behavior of the model.

In Machine Learning the first step is to prepare the training dataset. Training dataset is the dataset on which the model will learn. Sometimes the data is labelled to call out features and the classification, and other times its unlabeled where the model will have to extract those features and do the classification on its own. Either way training data set needs to be prepared. And the next step would be choosing an algorithm to run through the data set. An algorithm is a set of statistical steps, and the type of algorithm depends on labeled or unlabeled data. Some types of algorithms used for labeled data are regression algorithms, decision trees, instance based algorithms. Some types of algorithms used for unlabeled data are clustering algorithms, association algorithms, neural networks. Next step would be training the algorithm by adjusting the weights, comparing the outputs it has generated with the actual results, and running the variables again until the algorithm returns the correct results most of the time [6].

Some of the Machine Learning algorithms thar are widely used in drug discovery are Support vector machine (SVM), Random Forest (RF), Decision trees, KNN, and Naïve Bayesian. Let us discuss about how these algorithms are applied in drug discovery.

Naïve Bayesian: It is a subset of supervised learning methods. It is classification techniques which is based on Bayes theory with an assumption of independence between predictors. It works on the assumption that in a given class the presence of a particular feature is unrelated to the presence of any other feature. Bayes theorems provide a way to calculate posterior probability P(c|x) from P(c), P(x) and P(x|c)

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

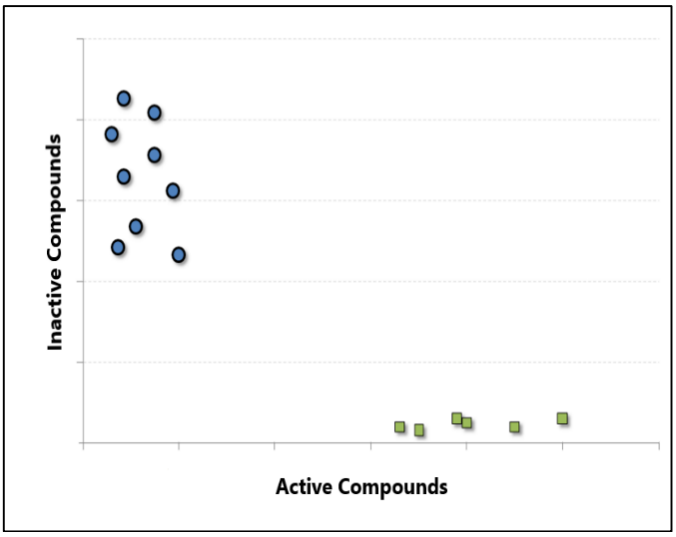
P(x|c) is the likelihood which is the probability of predictor given class.

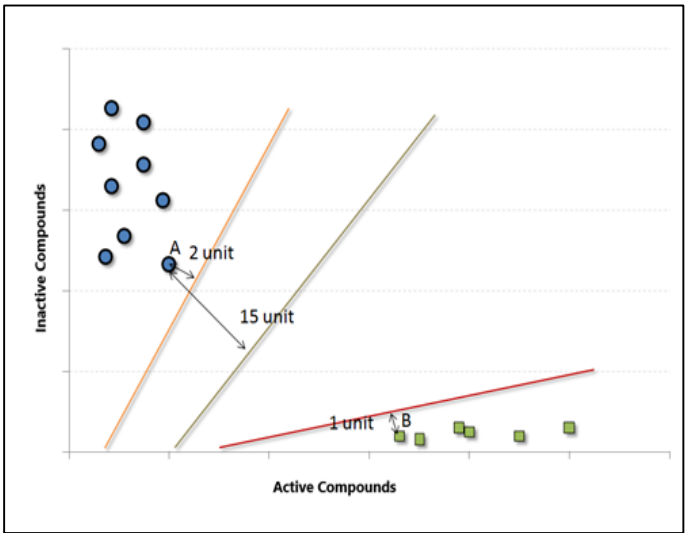
P(x) is the prior probability of predictor.

Naïve Bayesian classifiers are used in cheminformatics to predict biological properties rather than physicochemical properties. This is applied in predicting the toxicity of the compound, protein target and bioactivity classification for drug-like molecule [7, 8, 9]. For the data sets that are retrieved Naïve base improve the accuracy. In the classification tools for biomedical data NB algorithm have shown great promise even though data is fille with unwanted data called noise. NB technique also proved to have shown important role in ligand-target interaction predictions, which is also a great step towards lead discovery. Author [10] has used NB techniques to classify active and inactive compounds, with possible activity as antagonists for estrogen receptors in breast cancer. NB has the ability to process large amounts of data by having unique tolerance toward noise, and researchers are quite utilizing this feature. This technique when combine with other tools like fingerprint -6, extended connectivity was able to collect excellent compounds. Author [11] has utilized this feature of NB combine d with SVM to identify the compounds that could act against the targets of human immunodeficiency virus type-1 and the hepatitis C virus generated from multiple QSAR models. A recent technique introduced in machine learning world is applying Bayes method to LBVS which is the Bayesian model averaging [12, 13], and is used by author [13] inn his paper. IN this paper author compared prediction of protein pyruvate kinase activity using dragon receptor for Bayesian models in average to SVM and ANN.

Support Vector Machine

Support Vector machine is a classification method. In this algorithm each data item is plotted as a point in n dimensional space, with value of each feature being the value of particular coordinate.





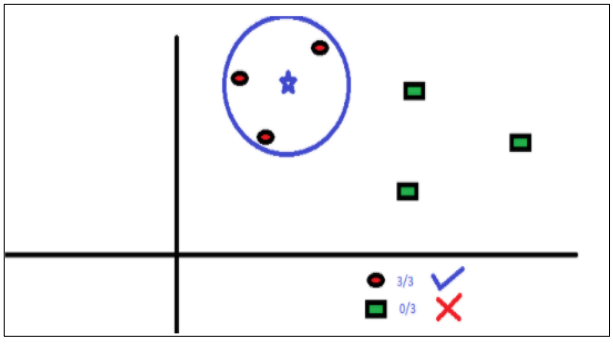
Let's take active and inactive compounds, we will first plot these in a two dimensional space where each point will have two coordinates, which are called support vectors. Now we should find a line which splits the data between two differently classified groups of data. This line will be farthest away from the closest point in each of the two groups. As shown in the above figure, black line is the one that splits data into two different classified groups. Use of

SVM plays an important role in drug discovery, due to its ability to distinguish between active and inactive compounds and also its ability to train the regression model. To determine the relationship between drug and ligand regression models are important [14, 15]. When we have a single protein of interest and several active compounds are screened against a it, SVM can be attributed in various scenarios. A wide number of new kernel functions have been introduced for SVMs, which includes ligand and kernels which captures different information for similarity assessment, like descriptor or graph similar and binding site or sequence similarity. These are expensive computationally and requires parameter determination. There is this kernel function which is parameter free. On the other hand, kernel functions that use 3D structure of compounds have been developed. As an example, pharmacophore kernel [16] will focus on three point pharmacophores in 3D space, outperforms SVM calculations on fingerprint representations of pharmacophores. Author [17] did research on searching novel c-Met tyrosine kinase inhibitors from eighteen million compounds with the application of docking calculations and 2 stage SVM. Compared to individual approaches this combined approach increased hit rates of active compounds as well as enrichment factors. Authors have identified top ranked hits of count 1000 and out of 5 selected hits eight of them tested active.
To enhance the prediction author [18] investigated drug target interactions and also integrated information obtained from published research of various source. To obtain the information on therapeutical and pharmacological effects of drug, chemical structures of drug, protein genomic information which is needed to characterize the drug interactions they have used kernel functions. By using the kernel function results displayed were of great potential.

Decision Tree: Decision tree is a type of supervised learning algorithm which is mostly used for classification problems. Depending on the set of decision rules decision trees are used to classify the data to make recommendations. Decision trees are used in drug filed for problems like, prediction of drug likeness, designing combinatorial libraries, generating compound profiling data etc. Decision trees are also used to predict ADME properties, like absorption, distribution, permeability and solubility of drugs, p-glycoprotein, metabolic stability and penetration. Decision tree models are simple and easy to understand, validate and interpret. Predictions of decision trees are known to suffer from high variance. Even with a small change in the data may lead to splits in the results. Due to the hierarchical nature this instability is caused. Along with that the decision tree structure is sensitive to small changes in data used for training. If the data set used for training is small the learning process will be affected. In contract huge data set may cause problems in decision trees too. So it is always recommended to used moderate size of data set, tree structure which is height balanced, with moderate number of levels. Decision tree performance is also dependable on sequence of splitting attributes selection. The splitting attributes need to be sort of according to importance or merit orders.

Random Forest: Random Forest is a collection of decision trees. In random forest we will have a collection of decision trees. Depending on attribute if we have to classify an object, each tree will vote for the class, and the decision forest chooses the classification which will have the most votes. Usually, a single decision tree does not provide high performance output. Pruning of the tree using cross validation or by model complexity parameters is a common process to limit high variance. RF models are proved to have improved the LBVS performance of individual Decision Trees. Random Forest have properties that improves the prediction of QSAR data. These properties are built in descriptor selection, high accuracy of prediction. Author [21] has published a method which was applied to mining estrogen receptors from a dataset of 57000 molecules and this method uses a different set of descriptors to build decision tree model which is accurate. False positive are possible with any algorithm, but Random Forest combined with SVM and NB can produce less amount of errors when compared to other algorithms. Since it has multiple decision trees errors there caused due to individual trees are minimized. IN drug discovery RFs are mainly used for classifiers, feature selection and regression. Some important factors which helps the use of RF in drug discovery are it expedites training process, imputes data that is missing, use few parameters, and incorporates non parametric data.

K Nearest Neighbor: It can be used for regression and classification problems. It is more used for classification problems. KNN works by storing all the available cases and classifies the new case by majority votes of its k neighbors. Distance function is used to measure the most common amongst its k nearest neighbors. These distance function can be Euclidean, Manhattan, Makowski, and Hamming distance. Sometime choosing K is challenging for performing KNN modeling. We can easily understand KNNN by applying it in our real lives. If you want to learn about a person, who you don't know, you need to ask their friends who are close to them.
Things to consider before selecting KNN:
KNN is computationally expensive
Variables should be normalized else higher range variables can bias it
Works on pre-processing stage more before going for KNN like an outlier, noise removal



To the local structure of the data K-NN algorithm is sensitive, and so it is ideal to calculate properties with strong locality, and so is the case with protein function prediction. K-NN approach don't have limitations, as only k neighbors are used to predict a new compound and this method is sensitive to noisy data.
If the training data is misclassified, this data could cause false prediction in the molecule. IN pharmacy world k-NN

is used for predicting anticonvulsants activity, d dopamine D1 antagonists, protein kinases inhibition, anticancer drugs, anti-inflammatory, and the activities of steroids.

Deep Learning: Deep learning is subfield of machine learning with algorithms inspired by the function and structure similar to brain called artificial neural network. Deep learning model will have an input layer, output layer and multiple hidden layers. The data is fed to the input layer, and it has to go through all the intermediate layer which are hidden layers, before the output is generated through output layer.

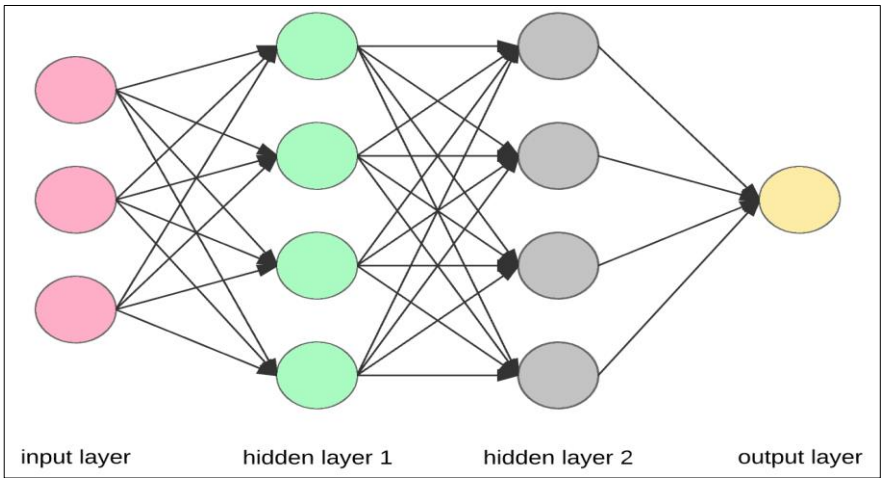


Fig 4: Deep Learning Model

Deep learning can be used in drug discovery in three different categories

- Prediction of drug properties
- De Novo Drug Design
- Dug -Target Interaction prediction

Drug Properties Prediction: Deep learning which is a sub field of machine learning is used to predict drug properties. The input to the algorithm is a drug and output is drug property like drug toxicity or solubility. Input: A drug. Output: 0-1 label which indicated whether a drug has certain properties or not [22].

There are different ways to represent a drug:

- Molecular fingerprint
- Text based representation like Smiles, Selfish
- Graph structure like 2D or 3D graph

Molecular Fingerprint: Molecular fingerprint is one of the ways to represent a drug in the input pipeline of machine learning. The most common type would be binary digits, which can represent the absence or presence of particular substructures in the molecule. It is evident that encoding a molecule as vector is not process which can be reversed. It's not possible to reconstruct the molecule form the fingerprint, which means there is a loss of information that is lost during this operation.

Smiles code: This is another way of representing a molecule is by encoding structure as text. It is a process in which graphical structure data is converted to textual content, and this text is used in the pipeline for machine learning. SMILES (Simplified Molecular -Input Line Entry System) is the most popular representation. Once the conversion is done we can use other algorithms like NPL to process the drug and to predict the properties, chemical interaction and side effects.

Graphical Structured Data: In this approach graph data is used directly as an input to the deep learning pipeline, for example compound can be considered as graph, in where atoms are vertices, and edges are chemical bonds.

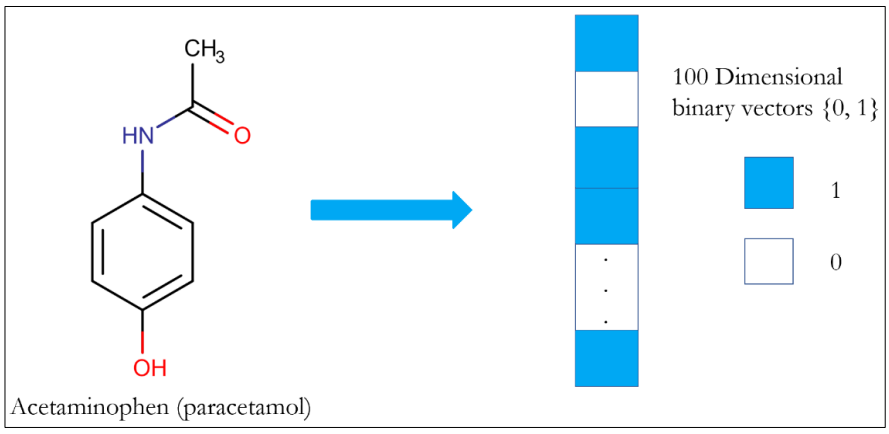


Fig 5: Representing a molecule in a binary vector

De Novo Drug Design: This way of approach is used when we want to design a compound to have certain specific properties for example, if we want to design a compound which can bind to a particular protein modify some pathways, and not interact with others, which can some have certain physical property like solubility range. This can be achieved by the use of this toolkit.

Drug Target Interaction Prediction: Proteins are important in any creature and are responsible for critical functionality in any being. Proteins functions is dependent on 3-Dimensional structure. By changing this structure, the functionality of protein can be changed, and this is an important factor in drug discovery. Most of the drugs are designed to bind to the specific protein. So this is the important factor in drug delivery is to determine if the protein can bind to the drug or not. This is called drug target interaction prediction.

Author [23] has proposed a framework which is based on deep learning for drug target interaction prediction. These deep learning frameworks which are used for DTI prediction take both compounds and protein information as input. As mentioned above, numerous ways are used to represent a compound and proteins can have different representations as well. Author [24] proposed a method with Convolutional Neural Network for Ligand Protein scoring. They used 3D representation instead of text based representation. Author [24] suggested a method to act on viral proteins of 2019 nCoV and used molecule transformer Drug Target Interaction.

Discussion

Recent developments in machine learning and deep learning techniques have provided many opportunities in the field of drug discovery and development. For a lot of issues across the industry we can expect to see applications and solutions. The data becoming bigger will also help improve machine learning techniques. Machine learning and deep learning has been used in many fields including pharmacy, medicine [25, 26], agriculture, cars, etc. Some studies proved that using machine learning techniques have outperformed compared to traditional methods in many of the subfields in medicine. ML algorithms along with DL methods, have enabled utilizing AI in the industry in our day-to-day life. Impact of machine learning methods in all the areas including healthcare, speech recognition, NPL, computer vision are being felt.

References

1. Segler MH, Waller MP. Modelling chemical reasoning to predict and invent reactions. Chem. - Eur. J 2017;23:6118-61.
2. Yu-Chen Lo, Stefano Rensi E, Wen Torng, Russ Altman B. Machine learning in chemoinformatics and drug discovery, Drug Discovery Today 2018;23(8):1538-1546. ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2018.05.010.
3. Schuhmacher O, Gassmann M. Hinder Changing R&D models in research based pharmaceutical companies J Transl Med 2016;14 (1):105.
4. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of investigational drugs in late-stage clinical development and publication of trial results JAMA Internal Med 2016;176(12):1826-1833.
5. Lowe D. The Latest on Drug Failure and Approval Rates. Available at: blogs.sciencemag.org/pipeline/archives/2019/05/09/the-latest-on-drug-failure-and-approval-rates. Accessed [September 16, 2019].
6. https://www.ibm.com/cloud/learn/machine-learning
7. Lowe R, *et al*. Predicting the mechanism of phospholipidosis. J. Cheminformatics 2012;4:2.
8. Koutsoukas A, *et al*. In silico target predictions: defining a benchmarking dataset and comparison of performance of the multiclass Naı¨ve Bayes and ParzenRosenblatt Window. J. Chem. Inf. Model 2013;53:1957-1966.
9. Nigsch F, *et al*. Ligand-target prediction using Winnow and naı¨ve Bayesian algorithms and the implications of overall performance statistics. J. Chem. Inf. Model 2008;48:2313-2325.
10. Pang X, Fu W, Wang J, Kang D, Xu L, Zhao Y, *et al*. Identification of Estrogen Receptor α Antagonists from Natural Products via *In vitro* and In Silico Approaches. Oxid. Med. Cell. Longev 2018, 6040149. [Google Scholar] [CrossRef]
11. Wei Y, Li W, Du T, Hong Z, Lin J. Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method. Int. J Mol. Sci 2019;20:3572. [Google Scholar] [CrossRef]
12. Wasserman L. Bayesian model selection and model averaging. J Math. Psychol 2000;44:92-107.
13. Angelopoulos N, *et al*. Bayesian model averaging for ligand discovery. J Chem. Inf. Model. 2009;49:1547-1557.
14. Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM. Advances with support vector machines for novel drug discovery. Expert Opin. Drug Discov 2019;14:23-33. [Google Scholar] [CrossRef] [PubMed]
15. Lima AN, Philot EA, Trossini GH, Scott LP, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin. Drug Discov 2016;11:225-239. [Google Scholar] [CrossRef] [PubMed]
16. Mahe P, *et al*. The pharmacophore kernel for virtual screening with support vector machines. J. Chem. Inf. Model 2006;46:2003-2014.
17. Xie, Q-Q, *et al*. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met. Eur. J. Med. Chem 2011;46:3675-3680.
18. Wang YC, Zhang CH, Deng NY, Wang Y. Kernel-based data fusion improves the drug-protein interaction prediction. Comput. Biol. Chem 2011;35:353-362. [Google Scholar] [CrossRef]
19. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Molecules 2020;25(22):5277.
20. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug discovery today 2015;20(3):318-331.
21. Tong WD, *et al*. Decision forest: combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci 2003;43:525-531.

22. https://towardsdatascience.com/review-deep-learning-in-drug-discovery-f4c89e3321e1
23. Feng Q, Dueva E, Cherkasov A, Ester M. Padme: A deep learning-based framework for drug-target interaction prediction. arXiv preprint arXiv:1807.09741 2018.
24. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–ligand scoring with convolutional neural networks. Journal of chemical information and modeling 2017;57(4):942-957.
25. Ravi Manne, Snigdha Kantheti, Sneha Kantheti. Classification of Skin cancer using deep learning, Convolutional Neural Networks - Opportunities and vulnerabilities- A systematic Review, International Journal for Modern Trends in Science and Technology, ISSN: 2455-3778 2020;06(11):101-108. https://doi.org/10.46501/IJMTST061118
26. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/