Drug Discovery

Sathiyan K[1], Santhosh P[2], Yathin S[3]
[1, 2, 3]Panimalar Engineering college

Abstract- *In the modern days, the advancements of information technology and related processing techniques have created a fertile base for progress in many scientific fields and industries. In the field of drug discovery and development, machine learning techniques have been used for the development of novel drug candidates. The methods for designing drug targets and novel drug discovery now routinely combine machine learning to enhance the efficiency, efficacy, and quality of developed outputs. The generation and incorporation of big data, through technologies such as high-throughput screening and high throughput computational analysis of databases used for both lead and target discovery, has increased the reliability of the machine learning and deep learning incorporated techniques. The use of these virtual screening and encompassing online information has also been highlighted in developing lead synthesis pathways.*

I. INTRODUCTION

At present, there are many diseases in the world for which the vaccine or the drug to cure those diseases is not found, and also for new diseases which can come in the future it takes Time and more Capital to find a cure for these diseases. The concept of DRUG DISCOVERY benefits society, particularly the aging society in a very significant manner. The general meaning of Drug Discovery is the process that underpins the entire pharmaceutical industry, encompassing the early stages of research from target discovery and validation, right through to the identification of a drug candidate or lead compound. In the future, Technology plays a vital role in drug discovery (ex: we can find the new biologically active compounds in computer-aided drug design). As per the reports, global drug discovery is expected to grow at the rate of approximately 12.2% over the next decade to reach approximately $160 billion by 2025. Your motive is to use the Machine Learning concept in order to find the cure for a disease in a short time and with less usage of Money. For example, drugs can reportedly take 12 years from the initial discovery stage to licensing approval, and the Association of the British Pharmaceutical Industry estimated the amount of investment to be at £1.15 billion per drug.

II. LITERATURE SURVEY

It is discovered that scientists across the world are in search of new drugs. About 2.6$million is the estimated price for developing the treatment. But the system fails because it includes money spent on nine out of ten therapies that fail between phase 1 trials and regulatory approval. Here, few leading biopharmaceutical companies believe that there is a solution for this downfall. Pfizer is using 1BM Watson (i.e.) a system that uses Machine Language, to power its search for immune-oncology drugs (used for cancer treatments). There is UK start-up Exscientia's Artificial Intelligence platform used to hunt for metabolic disease therapies. A company called Sanofi has signed a deal to use it. Following this, Roche subsidiary Genentech is using an AI system for GNS healthcare in Cambridge, to help Multinational Companies search for cancer treatments.[**1**].Many professionals have investigated this concept, For the past several decades, research in the molecular basis of human muscle aging has progressed. The major challenge is the development of accessible tissue-specific biomarkers of human muscle aging, which will evaluate the effectiveness of therapeutic interventions. Biomarkers could predict the functional capacity at some later age better than chronological age. Now, this is the method for tracking-related changes of human skeletal muscle. Here, they analyze the differential gene expression and pathway analysis of young and old tissue from healthy donors to pre-process the resulting data for a set of machine learning algorithms by Mamoshina, P. et al. [**2**]. Drug discovery and development is a long-term process, complex, and depend on several factors. Machine learning can improve discovery and decision making for well-specified questions with abundant, data, target validation, prognostic biomarkers, and analysis data in clinical trials. The lack of interpretability and repeatability of Machine learning generated results, which may limit their application. To tackle this situation, the application of machine learning can promote data-driven decision-making by itself, expeditious in the development process, and minimize the failure rate in drug discovery and development by Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham [**3**]. The process of discovery and development of drug-using manpower takes several years and much complicated. Finding the target element, clinical validation, completing the trial phases, and getting approval for the drug is the process. But novel target identification and validation can be done by various methods, tools, and data related to the target which will pre-validate a drug target in multiple views that will result in the new drug. This requires

various aspects of knowledge and technologies. Based on the detailed specification and various comparisons the compatible elements are found by Byung-Cheol Kim; Sunghoon Kim [**4**].The process implementing using the Intel Open VINO toolkit for the identification of drugs. Using this technology one can identify the reactants which are added to the drugs and it automates the entire flow of the cycle. Using a custom object detection technique with the Intel OpenVINO toolkit the model can be trained using the R-CNN (Region-based Convolutional Neural Network) with the help of labelled drugs that also act as reactants. The total clinical trial process can be reduced by nearly 4-5 months; originally it takes nearly 10-11 months in general to complete this process. Thus we can make a stimulating drug to see the behaviour of the process and the implementation becomes much faster compared to the general process, implemented by Rise Biswas, Avirup Basu, Abhishek Nandy, Arkaprova Deb[**5**]. According to the report based on the Application of Machine learning in drug discovery by Jessica Vamathevan, Dominic Clerk, Drug Discovery is a long and complex process. Machine Learning helps in this process by improving discovery and decision making. Machine Learning is also applied in all stages like validation, identification of prognostic biomarkers[**6**].Usually it takes more than three years to discover a drug for a viral disease. It has to complete many stages to get a proper and effective drug as a output. To find a drug for a specific virus we need find a protein target which defense against the virus and stops that to grow(i.e. kills that virus from growing or getting replication). Binding of few more protein gives us the expected result, but sometime it is vain. For example, a drug which shows a good speedy recovery for covid patients is actually a drug for ebola. But the problem here is some of them had some side effects.This is the problem drug discovery faces.their goal is to make this man made long term process to automated short process.Artifical intelligence and automated chemistry used here to find a drug targets to the virus.Here machine learning is used to analyse and optimise to create the compound.the automated paltform tests the compound and proceeds it.From the results the promising compounds are found and tested with the Live samples**[7].**

III. PROPOSED SYSTEM

3.1 System architecture

Due to the lack of interpretability and repeatability of machine learning generated results may challenge the limit of their application. Systematic and comprehensive high-dimensional data still need to be generated. In this task, we are provided with the dataset of drug molecules in the form of SMILES and their binding affinity towards the disease. The molecules contain protein which is capable of replication and the transfer of the disease. One of the proteins is to be targeted and the drug is created which is capable of blocking the protein. The data has been generated using Protein-Ligand docking. Protein-ligand docking is a molecular modeling technique. The goal of protein-ligand docking is to predict the position and orientation of a ligand (a small molecule) when it is bound to a protein receptor or enzyme.

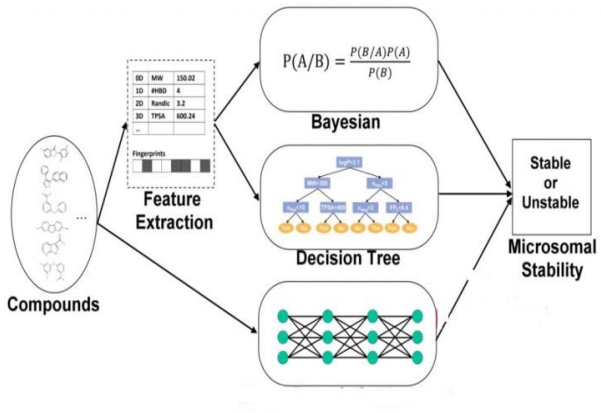


Fig 1.1Architecture Diagram

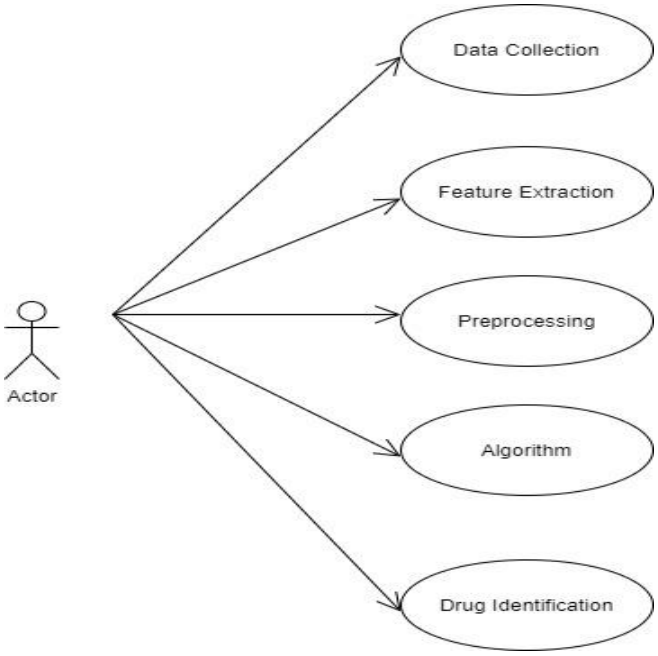


Fig 1.2 Use Case Diagram

3.2 Design description

3.2.1 Compounds and Feature extraction

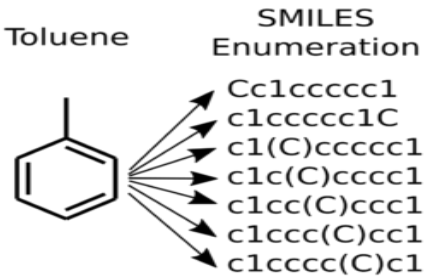


Fig 1.3 SMILES Representation

Coumpounds are the raw data which will denoted in SMILES(Simpliflied molecular input line entry system)describes the chemical structure. SMILES-based deep learning models are slowly emerging as an important research topic in cheminformatics.

3.2.2 Bayesian

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Fig 1.4 Bayesian equation

The Bayesian techniques are available for estimating the accuracy of various medical tests. It is a method of statistical inference in which Bayes theorem is used to update the probability for a hypothesis as more information becomes available.This piece of mathematics is used to find the probability that one might have or not have a disease.

3.2.3 Microsomal Stability

High throughput in vitro microsomal stability assays are widely used in drug discovery as an indicator for in vivo stability, which affects pharmacokinetics. drug discovery compounds are often not drug-like, are assessed with high throughput assays, and have many potential uncharacterized in vivo clearance mechanisms.

IV. IMPLEMENTATION AND RESULTS

4.1 Implemention

The project helps us in extracting the samples of combination from the dataset of the disease we collected. The major topics used while developing for the construction of the Project

1. Dataset (Represented by SMILES)
2. Protein-Ligand Docking
3. Morgan's Fingerprint
4. Mol2Vec

Module 1 – Dataset (Represented by SMILES)

SMILES-based deep learning models are slowly emerging as an important research topic in cheminformatics. In this study, we introduce SMILES Pair Encoding (SPE), a data-driven tokenization algorithm.

Module 2 –Protein-Ligand Docking

Protein–ligand docking is a molecular modelling technique. The goal of protein–ligand docking is to predict the position and orientation of a ligand (a small molecule) when it is bound to a protein receptor or enzyme.

Module 3 – Morgan's Fingerprint

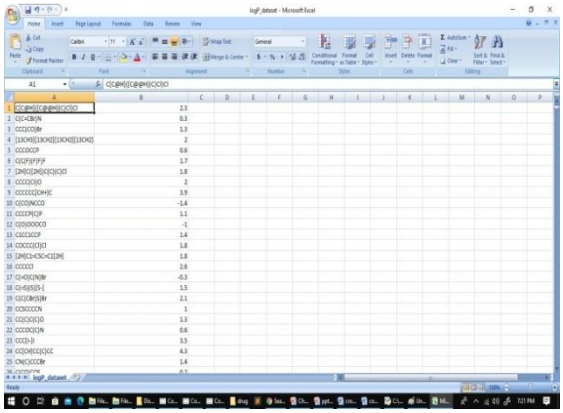
The Morgan fingerprint is basically a reimplementation of the extended connectivity fingerprint (ECFP). There is a paper describing it if you want more details but in essence you go through each atom of the molecule and obtain all possible paths through this atom with a specific radius.

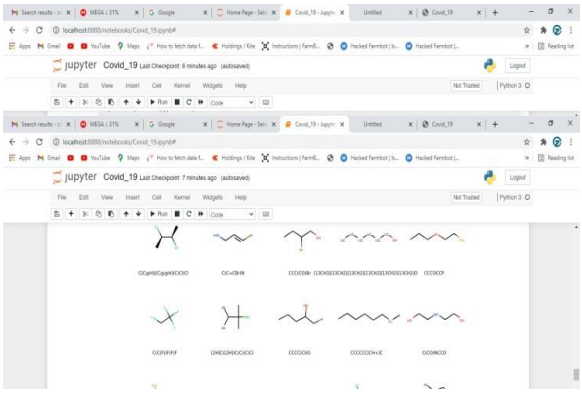
Modeule 4 – Mol2Vec

Mol2vec is an unsupervised pre-training method to generate an information rich representation of molecular substructures. Since it is an unsupervised method, it does not require labeled data as input and can leverage from larger amounts like the here employed 19.9 million compounds.

4.2 Results

The screen shots of the project is shown below:





V. CONCLUSION

We found that Drug Discovery is a long-term process by clinical experts. It will have it's pre-clinical &amp,clinical phases, and drug approvals from IND(Investigational New Drug) and NDA(New drug Approval). Nevertheless, this long-term process can be diminished using Machine learning. The time consuming will be reduced, the limited success rate will raise, investment can be shortened and we can avoid ethical challenges. Despite the advantage in technology and understanding of biological systems, drug discovery is still a long process with a low rate of new therapeutic discovery. Data indicates the new targets, as opposed to established targets, are prone. Although combinatorial approaches have provided a new and effective way to discover drug leads, there still exists an abundance of natural organisms that have not been screened for potential new leads. Therefore, we feel the search for new drugs from natural sources (bioprospecting) should be continued even with the advent of combinatorial methods to drug discovery.

REFERENCES

[1] 1.N. Fleming, "How computer science is dynamic drug discovery", Nature News, vol. 557, no. 7706, May 2018.

[2] Mamoshina, P. et al. Machine learning on human muscle transcriptomic information for biomarker discovery and tissue-specific drug target identification. Front. Genet. 9, 242 (2018).

[3] Jessica Vamathevan, St. Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee. Applications of machine learning in drug discovery and development. PMID: 30976107

[4] Byung-Cheol Kim; Sunghoon Kim. Assemblable project formulation for drug target discovery.IEEE International Conference on Bioinformatics and Biomedicine Workshops.IEEE954-955 (2012).

[5] Risab Biswas, Avirup Basu, Abhishek Nandy, Arkaprova woman. Drug Discovery and Drug Identification mistreatment AI.10.1109/Indo-TaiwanICAN48429.2020.9181309(2020)

[6] Jessica Vamathevan, St. Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee. Application of Machine learning in Drug Discovery. PMID: 30976107(2019)

[7] M. Scudellari, "Researchers ar dissipated that AI and automation will cut drug discovery from 5 years to 6 months: Automating antivirals," in IEEE Spectrum, vol. 57, no. 10, pp. 44-49, Oct. 2020, doi: 10.1109/MSPEC.2020.9205548.

[8] P. Szymanski, M. Markowicz and E. Mikiciuk-Olasik, "Adaptation of high-throughput screening in drug discovery-toxicological screening tests", Int J metric weight unit Sci., vol. 13, no. 1, pp. 427-452, 2012.

[9] M. Benhenda, "ChemGAN challenge for drug discovery: will AI reproduce natural chemical diversity?", arXiv preprint, 2017.

[10] H. Chen and Z. Zhang, "Prediction of Drug-Disease Associations for Drug positioning Through Drug-miRNA-Disease Heterogeneous Network," in IEEE Access, vol. 6, pp. 45281-45287, 2018, doi: 10.1109/ACCESS.2018.2860632.

[11] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. the increase of deep learning in drug discovery. Drug Discov. Today 23, 1241–1250 (2018).

[12] Hinton, G. Deep learning — a technology with the potential to remodel health care. JAMA 320, 1101–1102 (2018).

[13] Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of trial success rates and connected parameters. Biostatistics “https://doi.org/10.1093/biostatistics/kxx069” (2018).

[14] eon, J. et al. a scientific approach to spot novel antineoplastic targets mistreatment machine learning, matter style, and high-throughput screening. Genome Med. 6, 57 (2014).

[15] J.P. Hughes et al., "Principles of early drug discovery", British Journal of materia medica, vol. 162, no. 6, pp. 1239-1249, 2011.