International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Drug Generation Using Generative Models

Soham Bhure[1], Tanya Agrawal[2], Rutuja Walke[3], Mihir Tale[4], Ganesh Tarone[5]

[1]*Soham Bhure, Dept. of Computer Engineering, VIT Pune, Pune, India*
[2]*Tanya Agrawal, Dept. of Computer Engineering, VIT Pune, Pune, India*
[3]*Rutuja Walke, Dept. of Computer Engineering, VIT Pune, Pune, India*
[4]*Mihir Tale, Dept. of Computer Engineering, VIT Pune, Pune, India*
[5]*Ganesh Tarone, Dept. of Computer Engineering, VIT Pune, Pune, India*

Abstract - *Drug discovery is the first step in the drug development process to identify drug candidates for future research in clinical trials. The process of drug discovery and generation using traditional methods takes 12-14 years. The complexity of drug development has increased over the years. This paper aims at exploring various methods used to generate and validate drugs. The review of this paper starts with a brief history of how drugs are generated. We have also discussed some drug discovery processes. In this paper we have elaborated efficient deep learning methods used for drug discovery to generate simulated SMILE (Simplified Molecular Input Line Entry System) strings. Prominent use of Autoencoders, LSTM (Long Short-Term Memory), and GANs (Generative Adversarial Networks) to Generate drug molecules. Followed with a detailed discussion of Validating and Simulating these drug molecules. Followed by challenges and future scope of drug generation using deep learning models.*

Key Words: **Auto encoders, Drugs, FDA, Drug Approval Process, Generation, Generative Adversarial Networks, LSTM, Smiles.**

1. INTRODUCTION

Drug investigation is the process of discovering and/or designing medications in medicine, biotechnology, and pharmacology. Most medications have previously been found either by finding the active component in traditional treatments or by chance. Understanding how illness and infection are controlled at the molecular and physiological level, and then targeting specific entities based on this understanding, is a novel strategy.

The identification of candidates, synthesis, characterization, screening, and tests for therapeutic effectiveness are all part of the drug development process. Medical research has become much more complicated in the last 40 years, requiring preclinical studies, investigational new drug (IND) filings, and full clinical testing before receiving FDA marketing clearance. New drug applications (NDAs) and biologics license applications (BLAs) are usually thoroughly evaluated before being approved, and then drug performance is resubmitted to regulatory bodies for post-marketing research. After a comprehensive medical review, the main objective is to offer more efficient and safer therapies to patients as soon as feasible.

New drugs are developed through drug discovery. Drugs were formerly discovered mostly through discovering active components in traditional medications or by pure accident. Following that, traditional pharmacology was employed to go through chemical libraries containing small compounds, natural products, and plant extracts in order to locate those having therapeutic properties. Since the sequencing of human DNA, reverse pharmacology has used testing to find cures for existing ailments. Through the flow chart below, disease processes, molecular compound assays, current medicines with unforeseen side effects, and new technology promote drug development.

Medication discovery nowadays include hit screening, medicinal chemistry, and hit optimization to limit possible drug adverse effects (increasing affinity and selectivity). This stage of the medication development process also improves efficacy or potency, metabolic stability, and oral bioavailability. A gene or protein that has a substantial impact in illness is identified as a target. Therapeutic qualities are noted after they have been recognized. To validate targets, scientists employ disease associations, bioactive chemicals, cell-based models, protein interactions, signaling pathways analysis, and gene functional analysis, as well as in vitro genetic manipulation, antibodies, and chemical genomics.

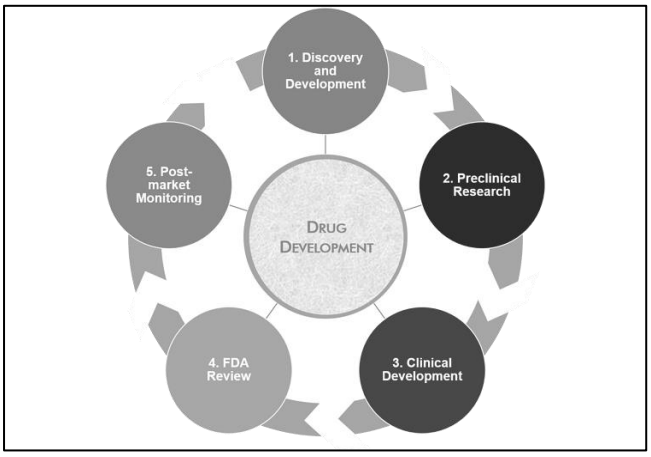

Figure 1: Drug Development Process

Artificial Intelligence techniques give a collection of tools that can aid in the discovery and decision-making process for well-defined problems involving a large amount of high-quality data. AI may be used at any level of the drug development process. Target validation, the development of prognostic biomarkers, and the analysis of digital pathology data in clinical trials are just a few examples. The context and technique of the applications varied, with some generating precise forecasts and insights. The lack of interpretability and repeatability of AI-generated outputs, which may restrict their applicability, is one of the most difficult aspects of implementing AI. In this paper, we will improve various generative models for the process of drug discovery.

2. LITERATURE REVIEW

It is critical for end-to-end learning that the input includes all of the chemical compound's latent feature information. There is a lot of chemical data available, such as weight, molecular formula, rings, atoms, and SMILES. SMILES shows the chemical structure as a line of ASCII characters among the numerous bits of data. C1CCCC1 and CC(=O)NC1=CC=C(O)C=C1 are the chemical symbols for cyclohexane and acetaminophen, respectively. SMILES can indicate atoms (such as carbon, nitrogen, and oxygen), bonds (such as single, double, and triple bonds), rings (such as open ring, close ring, and ring number), aromaticity, and branching.

Variational Auto-Encoders (Gomez-Bombarelli et al. 2016), Adversarial Auto-Encoders (Kadurin et al. 2017; Kadurin et al. 2017b), Recurrent Neural Networks and Reinforcement Learning (Jaques et al. 2017; Segler et al. 2017; Olivecrona et al. 2017) are examples of this approach, which are eventually combined with Sequential Gen (Guimaraes et al. 2017; Benjamin et al. 2017).

However, much of this research is still in the exploratory stage, with produced samples being evaluated solely visually or using metrics that aren't always relevant to the real drug development process. A thorough examination of the produced samples' internal chemical diversity would be very beneficial. Because drug candidates might fail in a variety of unanticipated ways later in the drug development pipeline, it's critical to generate a chemically varied stream of molecules. (Jaques et al. 2017, p. 8) claims that their Reinforcement Learning (RL) generative model produces simple molecules based on visual inspection. (Guimaraes et al. 2017, p.6, p.8), on the other hand, claims that their Objective-Reinforced Generative Adversarial Network (ORGAN) produces fewer repetitive and simple examples than reinforcement learning.

3. THEORY

Despite the availability of effective medicines and therapies in the pharmaceutical supply chain, new medications are desperately needed, particularly during the COVID-19 epidemic. Because the drug development life cycle is extensive and complex, requiring 10 to 15 years to generate a product, it is challenging to create and deliver treatments fast.

The drug development process is equally unclear and confronts its own set of obstacles, including difficulties identifying targets, a shrinking market for medication approval, and rising prices. Pharmaceutical firms join forces to break down the drug development life cycle and seek ways to speed things up while still producing safe and effective pharmaceuticals.

According to the FDA, the medication development process has five phases.

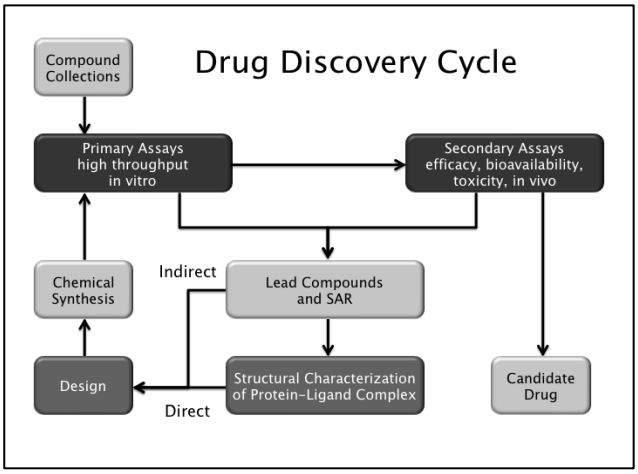

Figure 2: Drug Discovery Cycle

The discovery and development process is the initial phase. Researchers investigate fresh insights into a disease process in this stage, which allows them to build a solution to counteract the disease's symptoms. They then put molecular compounds to the test to see whether they have any anti-disease properties. Preclinical research, which includes in vitro and in vivo studies, is the second phase in the drug development process. Clinical research is the third phase. This is a term that relates to human research or trials. Researchers choose who is eligible to participate, how many individuals will be included in the study, and how long it will run.

The FDA medication review is the process's fourth phase. A pharmaceutical business must first file a New Medicine Application, which is then reviewed by the FDA, which either approves or rejects the drug. Finally, an FDA advisory council weighs in with their thoughts. FDA post-market medication safety monitoring is the final phase in the medication development process. This is where the FDA evaluates reports of medication difficulties and decides whether to add cautions to dose advice, as well as further steps in the case of more significant difficulties.

The whole process from the discovery to the clinical trials takes around 12 years. This process can be too long and tedious. This paper compares various generative models to automate the very first step of the Drug Discovery Process. A generative model is a type of statistical model that is capable of producing new data instances. This model is commonly used to estimate probabilities, model data points, and differentiate between classes using these probabilities. Generic models may handle more complicated tasks than discriminative models since they generally depend on Bayes theorem. Unsupervised machine learning uses descriptive modelling to characterize phenomena in data, allowing computers to grasp the actual world.

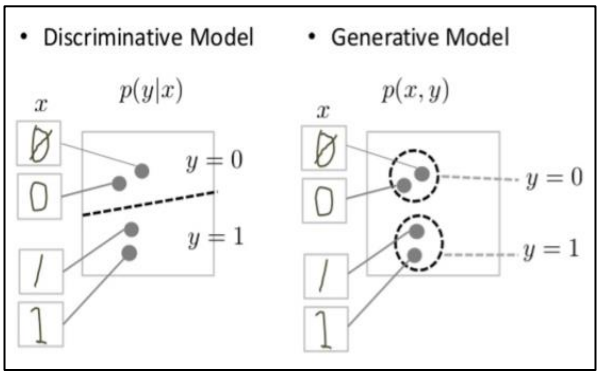

Figure 3: Generative Models Working

A generative model takes into account the data's distribution and informs you how likely a specific occurrence is. Because they can assign a probability to a succession of words, models that predict the next word in a series are often generative models. A discriminative model avoids the question of whether or not a particular event is likely, instead focusing on the likelihood of a label being applied to it.

In this paper, we explore Generative Models to generate Smiles structures. SMILES (Simplified Molecular Input Line Entry System) is a line format for inputting and expressing molecules and reactions (a typographical technique utilizing printable characters). SMILES includes the same information as a connection table with more connections.

The fact that SMILES is a verbal construct rather than a computer data structure makes it more helpful than a connection table. SMILES is a genuine language with just a few grammatical rules and a limited vocabulary (atom and bond symbols).

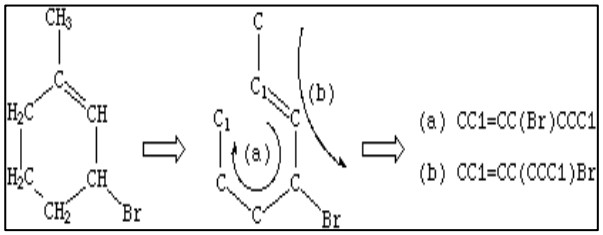

Figure 4: SMILES Representation

4. METHEDOLOGY

Various different methods were used for the generation of the SMILES Strings. The Aim being finding the method that gives the best and most accurate strings.

4.1 Long Short-Term Memory

The long short-term memory network (LSTM) is a recurrent neural network that can learn the order dependence of sequence prediction problems. This is a necessary behavior in complex problem areas such as machine translation and speech recognition. LSTM is a complex field of deep learning. It may be difficult to understand what LSTMs are and how they relate to domain terms such as bidirectional and sequence-to-sequence.

The goal is to teach the model to recognize patterns in SMILES strings so that the output matches genuine molecules. LSTMs are the ideal network for producing new SMILES strings since we give the network text as data. We

have trained the RNN with two LSTM layers on a dataset of 100,000 SMILES strings. SMILES strings are made up of two sorts of characters: special characters like "/" or "=," and elemental symbols like "P," "Si," "Mg," and so on. We create a separate dictionary for each of the specified unique characters. Following the creation of our character mapping dictionaries, we normalized and translated every character in the SMILES string dataset into numbers. The input array X was then reshaped into a 3-dimensional array for [samples, time steps, physicochemical characteristics] using NumPy, which is the required input form for recurrent models. After training the model, the predicted output Y is one-hot encoded to produce new SMILES. When using integer representations like the ones we just generated, one-hot encoding removes the integer encoded value and replaces it with a new binary variable for each distinct integer value.

Because there might be delays of undetermined duration between critical occurrences in a time series, LSTM networks are well-suited to categorizing, processing, and generating predictions based on time series data. LSTMs were created to solve the problem of vanishing gradients that can occur while training standard RNNs. In many cases, LSTM has an advantage over RNNs, hidden

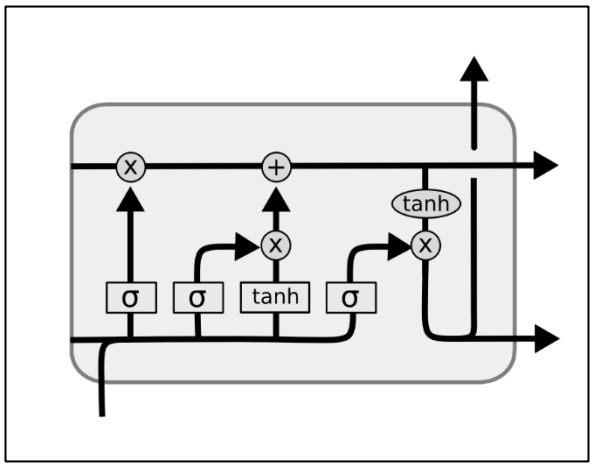

Figure 5: LSTM Architecture

Markov models, and other sequence learning approaches due to its relative insensitivity to gap length. A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. The three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time periods.

The final step of the approach was training of the model for which we used 10 epochs of 512 batch size which gave the optimal results. However, results giving better accuracy and more valid molecules were expected. Thus, other better approaches were explored.

4.2 Auto - Encoder

An Autoencoder is a neural network in which the output layer and the input layer have the same dimensions. Simply put, the number of output units in the output layer equals the number of input units in the input layer. Automatic encoders copy data from input to output without supervision, which is why they are sometimes called replicated neural networks.

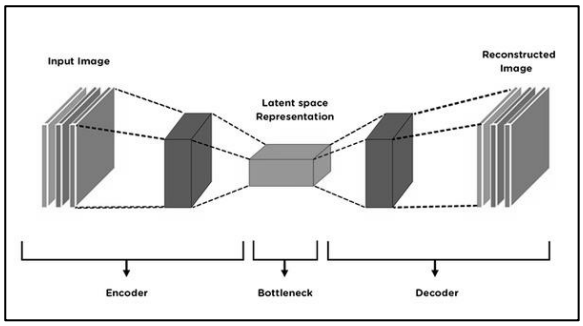

Figure 6: Auto- Encoder Architecture

Three major components of an auto-encoder are:

1. Encoder: The encoder is a connected feedforward neural network that shrinks the input picture into a dimensionality-reduced compressed representation and translates it into a latent dimensional space. The deformed representation of the actual picture is the compressed image.
2. Bottle-neck: This section of the network includes a reduced version of the decoder's input.
3. Decoder: The decoder and encoder are both forward networks with a topology identical to the encoder. The network is in charge of reassembling the input into the code's original size.

The input provided to the Encoder is a Correct/verified SMILE vector. And the output obtained from this encoder is a random size smaller vector. This Vector is also called latent space representation. Whereas the input taken by the decoder is the latent vector and the final output that is obtained from the decoder is a verified representation of a SMILE string.

4.3 Generative Adversarial Network

GANs, or generative adversarial networks, are a type of artificial intelligence. Networks is a deep learning-based method to generative modelling. It is an unsupervised learning job in machine learning that is used for automatic

drug discovery and then learning the regularities and patterns in the input data so that the model may produce or output fresh verified grin strings.

GANS is a simple approach to train a new dataset of validated medicines by framing the dataset as a problem with two sub-models: one is the Generator model, which we have trained to produce validated drug modules, and the other is the Modeler model, which we have trained to generate validated drug modules. In a contradictory, zero-sum game, the two models are trained simultaneously until the discriminator model is tricked roughly half of the time, indicating that the generator model is providing plausible instances.

The architecture of how GANs combine actual and synthetic data to create the desired outputs is depicted in the diagram below.

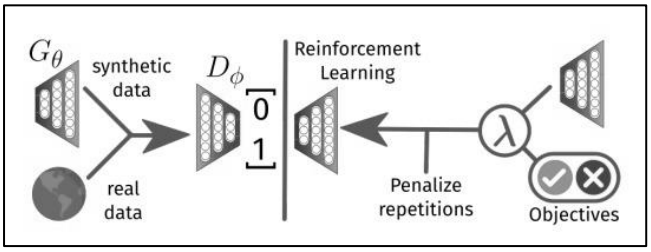

Figure 7: GANs Architecture

A unique computational approach for designing pharmacological compounds with the required characteristics has been implemented. Deep learning and reinforcement learning methods are used in the model. The model combines reinforcement learning with two distinct deep neural networks: a generator and a discriminator. These two algorithms are trained independently but combined to create unique targeted chemical libraries. Our model employs a simple molecular input line entry mechanism to represent molecules as strings (SMILES).

The generative model is employed in this system to produce new chemically viable compounds, acting as an agent, while the discriminative model (predicting the characteristics of the new compound) is also important, and the agent's behavior is evaluated to assign a number to each formed molecule. The generative model is trained to maximize the predicted reward, which is a function of the numerical characteristics provided by the predictive model. The generative model and discriminator model are trained using supervised learning techniques in the first step of the process.

The second step uses the RL method to train both models together to bias the creation of new chemical structures toward those with the required physical and/or biological characteristics. In a nutshell, we employed the GAN model combined with RL to create chemical compounds with certain physical qualities like stability, originality, and synthesizability.

5. VALIDATION

After the generation of the smiles strings by various models, the next step included the Validation. There is no guarantee that the resulting Smiles strings will be valid and represents a reasonable molecular structure.

To make a SMILES string, we convert a SMILES string into an alternate syntax that keeps the atom order but changes the parenthesis and ring closure symbols. SMILES are not chemically interpreted during the transformation process; instead, the syntax is string processed. As a result, it makes assumptions about the shape of the SMILES string that may or may not be satisfied if the SMILES string was created using a typical cheminformatics toolkit. The algorithm, for example, assumes that the SMILES were constructed by traversing the chemical network in depth first. The SMILES string CC(C1)CCCC1 cannot be encoded, for example, since this assumption is not satisfied.

We substitute the two ring closure symbols in a combination with a singular one at the latter symbol's location when using SMILES to alter ring closure symbols. Why not just replace the first symbol with something else? The reason for this is due to the tree structure of a SMILES string; there is only one path from the root to each node in the tree, but the path may split numerous times after that. While the atom that occurred five bonds earlier (along the path leading to this atom) can be referred to, the atom that appears five bonds later may have several possibilities.

Validity is basically checking the molecule for Syntactic and Semantic Errors. For this, the library Rdkit has been used. After passing the smiles string via the model, we get the structure of the Smile generated. If it is not a valid molecule, we will get an error.

6. SIMULATION

Docking is a method in molecular modelling that predicts the preferred orientation of one molecule to another when they are linked together to create a stable complex. Using scoring functions, for example, knowledge of the preferred orientation may be used to predict the strength of connection or binding affinity between two molecules. In signal transduction, the interactions between physiologically relevant molecules such as proteins, peptides, nucleic acids, and lipids are crucial. Furthermore, the sort of signal produced may be influenced by the relative direction of the two interacting partners. As a result, docking is beneficial for forecasting the signal's intensity and kind.

Due to its capacity to anticipate the binding-conformation of small molecule ligands to the proper target binding site, molecular docking is one of the most often utilized strategies in structure-based drug design. The initial condition for performing a docking screen is the structure of the protein of interest. The structure is usually determined using a biophysical approach like x-ray crystallography, NMR spectroscopy, or cryo-EM, although it can also come through homology modelling. A docking tool uses this protein structure and a database of possible ligands as inputs. A docking program's success is determined by two factors: the search algorithm and the scoring mechanism.

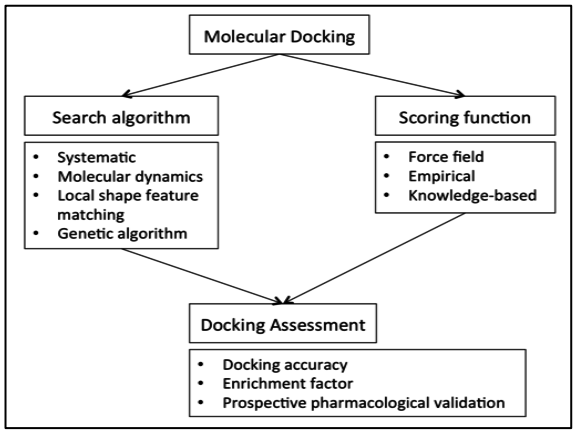

Figure 8: Molecular Docking Flowchart

In theory, the search space is made up of all potential protein and ligand orientations and conformations. However, with current computational resources, exhaustively exploring the search space—which would entail enumerating all possible molecule distortions as well as all possible rotational and translational orientations of the ligand relative to the protein at a given level of granularity—is impossible. Most docking systems take into consideration the ligand's whole conformational space, and a few even try to represent a flexible protein receptor.

The ligand and the receptor have been subjected to a range of conformational search techniques. These are some of them:

1. Torsion searches on rotatable bonds might be systematic or stochastic.
2. Simulations of molecular dynamics.
3. Genetic algorithms are used to "evolve" new low-energy conformations, with the fitness function used to choose individuals for the next iteration being the score of each position.

Docking algorithms yield a vast number of possible ligand poses, some of which can be eliminated right away owing to protein conflicts. The remaining ligands are ranked using a scoring system that takes a posture as input and outputs a number indicating the chance that the posture reflects a good binding interaction. The majority of scoring functions are molecular mechanics force fields based on physics that estimate the energy of a posture within the binding site. An additive equation may be built to represent the various contributions to binding:

$$\Delta G_{bind} = \Delta G_{solvent} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{rot} + \Delta G_{t/t} + \Delta G_{vib}$$

Figure 9: Binding Equation

Solvent effects, conformational changes in the protein and ligand, free energy owing to protein-ligand interactions, internal rotations, ligand and receptor association energy to create a single complex, and free energy owing to changes in vibrational modes are among the components. A stable system with a low (negative) energy suggests a potential binding contact. Alternative techniques include limitations based on known critical protein-ligand interactions or knowledge-based potentials generated from interactions found in vast databases of protein-ligand structures into modified scoring systems. In the Docking Process we predict the binding affinity between ligand and protein and the structure of protein ligand complex. The use of OpenBabel software was made that converts our Smile format into a pdb format which will be used further. The file downloaded from here is loaded on PyRx software for simulation.

The main protein taken as a sample is SARS 2 Covid 19 strand. After running the simulation, docking is achieved where the ligand tries various positions it can attach with

the virus. The binding affinity for various positions are tested thus to give the best interlocking location. The best locations are displayed as a result in a sequential manner and the most optimum location can thus be found.

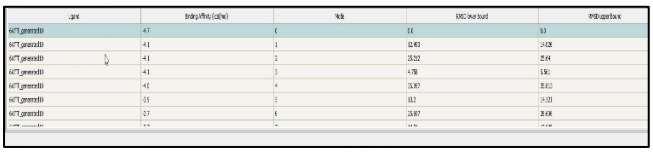

Figure 10: PyRx Software

We use PyMol software can be used for further visualization and clear understanding of the process. All the possible docking locations, the molecule can attach with are displayed here.

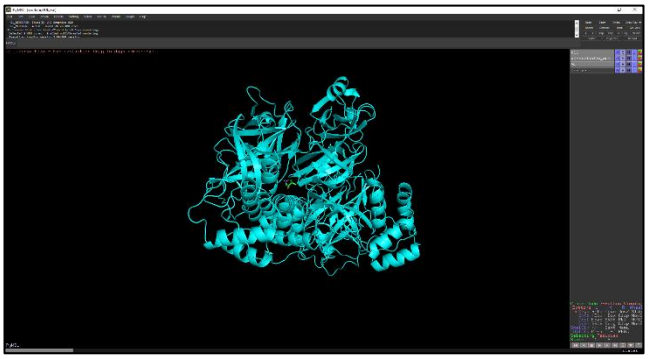

Figure 11: Simulation using PyMol

In this manner, various molecules can be tested for their affinities with the main drug and the whole drug generation process can be automated to a certain extent thus reducing the time significantly.

7. CONCLUSIONS

To summarize, in this paper we have suggested an alternate approach to the creation of new medicines. We have tried to generate new drugs with the help of a huge 'SMILES' database and different generative models like LSTMs, GANs and Auto-Encoders. All the generated drugs were validated for a feasible chemical formula and structure. Finally, the valid drugs were simulated on PyMol and PyRx to check if they can be docked well on the target compounds.

REFERENCES

[1] Mohs RC, Greig NH (2017) Drug discovery and development: role of basic biological research. Alzheimers Dement Transl Res Clin Intervent 3(4):651–657

[2] Scannell JW, Blanckley A, Boldon H, Warrington B (2012) Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 11(3):191–200

[3] .Lionta E, Spyrou G, Vassilatis D, Cournia Z (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem 14(16):1923–1938

[4] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4(2):90–98

[5] Babak Alipanahi, Andrew Delong, Mathew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specicities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology 33, 8 (2015), 831–838.

[6] Christof Angermueller, Tanel Parnamaa, Leopold Parts, and Oliver Stegle. 2016. ¨ Deep learning for computational biology. Molecular systems biology 12, 7 (2016), 878.

[7] W. Patrick Walters & Regina Barzilay (2021) Critical assessment of AI in drug discovery, Expert Opinion on Drug Discovery, DOI: 10.1080/17460441.2021.1915982

[8] Petko Bogdanov and Ambuj K Singh. 2010. Molecular function prediction using neighborhood features. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 7, 2 (2010), 208–217.

[9] Leo Breiman. 2001. Random forests. Machine learning 45, 1 (2001), 5–32.

[10] Lei Chen, Tao Huang, Jian Zhang, Ming-Yue Zheng, Kai-Yan Feng, Yu-Dong Cai, and Kuo-Chen Chou. 2013. Predicting drugs side effects based on chemical - chemical interactions

[11] Lei Chen, Jing Lu, Jian Zhang, Kai-Rui Feng, Ming-Yue Zheng, and Yu-Dong Cai. 2013. Predicting chemical toxicity eects based on chemical-chemical interactions. PLoS One 8, 2 (2013), e56517.

[12] Lei Chen, Jing Yang, Mingyue Zheng, Xiangyin Kong, Tao Huang, and Yu-Dong Cai. 2015. The use of chemical-chemical interaction and chemical structure to identify new candidate chemicals related to lung cancer. PloS one 10, 6 (2015), e0128696.

[13] Francois Chollet. 2015. Keras: Deep Learning library for eano and TensorFlow. https://github.com/fchollet/keras. (2015).

[14] Ryszard Czerminski, Abdelaziz Yasri, and David Hartsough. 2001. Use of support ´ vector machine in pattern classification: Application to QSAR studies. Molecular Informatics 20, 3 (2001), 227–240.

[15] George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. 2014. Multi-task neural networks for QSAR predictions. arXiv preprint arXiv:1406.1231 (2014).

BIOGRAPHIES

Soham Bhure
Student, Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Tanya Agrawal
Student, Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Rutuja Walke
Student, Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Mihir Tale
Student, Computer Engineering
Vishwakarma Institute of Technology
Pune, India

Ganesh Tarone
Student, Computer Engineering
Vishwakarma Institute of Technology
Pune, India