Grenze International Journal of Engineering and Technology, January Issue

Small Molecule based Drug Discovery using Machine Learning

Vasukumar Patel[1] and Dr. Shilpa Sonawani[2]
[1-2]M. Tech Data Science and Analytics, Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India
Email: vasupatel06102001@gmail.com, shilpa.sonawani@mitwpu.edu.in

***Abstract*—We outline a process for building particular kinds of neural networks made up of structures that are intimately related to the structure of the molecule being studied. A cellular neural network can be programmed to solve a special neural connection problem that represents a molecule. which a cellular neural network can be trained using. The concept involved converting chemical structures, such as peptides or tiny organic compounds, into a RNN-based self-learning environment. In order to find and create novel drugs to treat illnesses and enhance human health, the discipline of drug discovery is crucial. It entails a difficult, multidisciplinary process that combines several scientific and technological viewpoints. Machine learning techniques have become effective instruments in the drug discovery process, providing chances to speed up and enhance efficiency. For activities like virtual screening, predictive modeling of compound attributes, optimization of drug candidates, and discovery of new therapeutic targets, machine learning techniques can be used. By easing data processing, offering insights into intricate biological systems, and directing experimental decision-making, these methods have the potential to shorten the time and expense associated with the drug development process. They can aid in the selection of candidates that have a greater chance of success, minimizing the number of substances that must be synthesized and tested experimentally by doing this process we can get results much faster then traditional drug discovery methods.**

***Index Terms*— Recurrent Neural Network (RNN), Paralleling, Classification, Molecule, Quantitative Structure Activity Relation (QSAR), Connectivity, P-Value, Covid 19.**

I. INTRODUCTION

Machine Learning innovation in drug design is highly valued, since it facilitates the advancement of machine learning techniques and the collection of pharmacological [5] data. Drug discovery is a very long process that involves extensive experiments and different stages.It may take more than ten years and cost around billion dollars to discover a new drug. ML is becoming more and more crucial in converting medical data into research like reusable approaches, even though it doesn't rely on any hypothetical advancements. In the context of machine learning (ML), a variety of techniques are typically taken into consideration. These include Probabilistic Neural Networks (PNN), Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Random Forest, Naive Bayesian Classification (NBC), Multiple Linear Regression (MLR), Logistic Regression (LR), Linear Discriminant Analysis (LDA), etc [1].
Cells with distinct internal states and well-defined local interactions between adjacent cells, which control the

Grenze ID: 01.GIJET.11.1.538_1

dynamics of the entire system, characterise cellular systems [16]. Such systems have been constructed and studied in several instances thus far. Examples of discrete temporal cellular systems with discrete internal states are Wolfram's cellular automata. Introduced by Chua and Yang [16], the Cellular Neural Network (CNN) can be thought of as a vast array of simple continuous-time processing components with local connections.
This is where we used the concept of creating a so-called Molecular Graph Network (MGN) from the topology of a chemical compound originated. A molecule in an MGN can be thought of as a graph with every atom acting as a node and every chemical bond acting as an edge. The term "Graph Neural Network" was first used in the neural network field, where this idea was first developed [11].Cellular network topology is thus translated from a chemical structure as follows: each atom becomes a cell, each bond represents a local interaction between the cells, and the weights are determined by the types of elements and bonds. In this instance, the MGN functions as an Recurrent Neural Network(RNN). So We develop this method further by converting a peptide chain's chemical structure into a RNN's architecture.

II. Related Work

Machine learning has been applied to drug discovery in a number of research and projects, showing its efficacy in different stages of the procedure [5][8][9]. For example, scientists have predicted chemical qualities like toxicity, solubility, and bioavailability using deep learning algorithms [4]. Large-scale datasets are utilized by these models to capture intricate interactions between biological activity and chemical structures, resulting in more precise predictions [15].
The main focus of machine learning applications in drug development has been virtual screening [1]. Research has shown that by examining molecular characteristics and forecasting their binding affinities to certain target proteins, prediction models are effective in locating possible therapeutic options.[11] By greatly accelerating the screening process, this method enables researchers to concentrate their efforts on the most promising molecules[1].
Gaining an understanding of the quantitative structure activity relation (QSAR) [14] is a crucial job in contemporary drug discovery. One can separate QSAR difficulties into learning and coding components. Regression approaches that link the biological activities with the physiochemical or structural aspects of those chemical substances under study could be used to address the learning component. In QSAR analysis, identifying the molecular descriptors that represent the key characteristics of the molecules is crucial. Alternative methods for doing the traditional QSAR avoid the challenge of determining and choosing a collection of molecular descriptors that are representative [18].

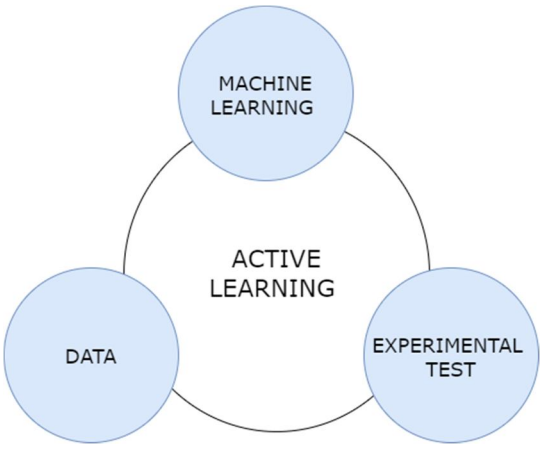

Fig.1 General Process

III. Methodology

The AI and ML-based prediction for the high throughput virtual screening is based on the information or data that is already accessible. The success rate of hit identification has increased significantly since the integration of AI and ML, with the provision of innovative medications being one of the major benefits. With numerous hidden layers, the RNN approach resembles a neural network and is widely utilized for its ability to learn arbitrary complex functions [1]. With enough data and processing time, it can learn as much as possible and produce

results that are extremely dependable. The various layers concealed in DL patterns give flexible access to learn arbitrarily complicated modules, which immediately provide relevant neurons and trained sets. [15] The RNN uses a gradient based optimization method and a backpropagation algorithm to enable neural network training and end-to-end differentiation. Additionally, layers, conventional architecture, and graph convolutional architecture are linked with feed-forward networks to advance the development of diverse domains and data kinds [1].

Machine learning-based drug discovery employs an organized, multidisciplinary method to quickly find putative medicinal molecules. The process usually starts with data gathering, when a variety of datasets pertaining to biological processes, chemical structures, and clinical results are acquired. Data cleansing and normalization are two pre-processing procedures that guarantee the dataset's consistency and quality. The next step is feature engineering, in which pertinent features corresponding to biological interactions or chemical structures are recovered [16].

The next step is to select a model, which includes selecting machine learning methods that are appropriate for the particular task at hand (e.g., regression or classification). The prepared dataset is used to train these algorithms, and cross-validation methods are used to evaluate the model's performance [6]. To improve the selected model's predicted accuracy, hyperparameter tuning is then used to optimize it.Virtual screening[7] is frequently used in the context of drug discovery to reduce the large chemical space and find possible therapeutic candidates. The probability that a chemical would display desired characteristics, like binding affinity to a target protein or bioavailability, can be predicted by machine learning models.

Furthermore, by giving priority to the synthesis and testing of compounds with the highest chance of success, the model's predictions can direct the design of experiments. Iterative feedback loops between model updates and experimental data improve the model's performance and help make the drug development process more effective [1].

In the end, applying machine learning to drug discovery speeds up the process of finding new therapeutic options, lowers expenses, and raises the success rate of introducing new medications to the market. Using a collaborative and dynamic approach, the methodology fosters a synergistic interaction between artificial intelligence and conventional drug discovery approaches by continuously validating and refining computer predictions through experimental validation.

IV. Model Workflow for Lead Molecule Prediction Utilizing Machine Learning Technique

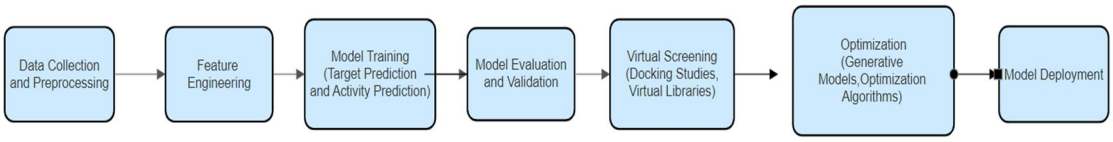

Fig.2 Model workflow for lead molecule prediction utilizing machine learning technique

AI innovation in drug design is highly valued, since it facilitates the advancement of machine learning techniques and the collection of pharmacological data. While AI doesn't rely on any theoretical developments, it is becoming more and more significant in converting medical data into research like reusable methods. In the area of machine learning (ML), a variety of techniques are often taken into consideration [1].

Data Collection and Preprocessing: This is the first stage in which pertinent information is acquired. In order to get the data ready for analysis, preprocessing may entail cleaning the data, dealing with missing values, scaling or normalizing features, and other operations SARS and MARS are similar to the Covid 19 (SARS-COV 2) that’s why we used drug compounds that used in research of SARS and MARS.

There is 3 CSV dataset for the Covid-19 dataset of carbon and hydrogen compound.

Master results table -Tables with the combinational compounds of carbon and hydrogen, along with their ratings and scores, make up this data.

Mers results - medications used in the MERS outbreak are included in this data.

Sars results - medicines used in the SARS outbreak are included in this data.

Feature Engineering: The preprocessed data is examined at this point in order to produce features for the model's training. To enhance model performance, feature engineering entails identifying the most pertinent qualities or producing new features from the available data based on features of the data and molecular structures we have used RNN.

Model Training (Target Prediction and Activity Prediction): In this instance, a predictive model is trained using the prepared features. One possible use for the model would be to forecast a certain target, such the activity of a possible medicinal molecule [4]. Depending on the nature of the problem, many algorithms and techniques can be employed for training; in this instance, we have used fig 3.
Model Evaluation and Validation: Metrics relevant to the task are used to assess the model's performance after training. Testing the model on a different dataset is a common step in validation to make sure it applies well to fresh, untested data [3].
Virtual Screening (Docking Studies, Virtual Libraries): Virtual screening is the process of searching databases of compounds using computational approaches to find compounds that are most likely to bind to a pharmacological target. This is done in the context of drug discovery. In order to forecast binding affinity, docking studies [7][8] model the interaction between the medication and the target [1].
Optimization (Generative Models, Optimization Algorithms): In this step, the model or the chemicals found by virtual screening are refined. New compounds can be designed using generative models, and for improved performance or efficacy, optimization algorithms can adjust the model's parameters or the compounds' chemical structures.
Model Deployment: Ultimately, the refined and verified model is implemented for real-world applications. This may entail incorporating the model into a drug discovery pipeline, enabling researchers to forecast the activity of novel molecules, or applying it to decision-making in a production setting.

V. Model Implementation

Drug discovery artificial neural network (ANN) implementation is a methodical process with several steps. Gathering different datasets with information about chemical structures, biological activities, and clinical outcomes is the initial phase in the data collecting process. After that, in order to make sure the input data is appropriate for the neural network, a comprehensive data preprocessing step is essential. This step includes activities like data cleaning, normalization, and feature extraction. The following stage is model selection, in which the particulars of the problem at hand are taken into consideration while selecting the kind of neural network architecture. The prepared dataset is then used to train the neural network, which entails modifying the model's weights using optimization methods to reduce the discrepancy between expected and actual results [1].
By correctly forecasting the unknown data set, generalized machine learning techniques have been applied to choose the method that works best under difficult conditions. Numerous In order to avoid brute force sensitivity and optimize precisely, models are used to train on a single dataset by understanding the perspective in various model architectures. Models that are either supervised or unsupervised characterize machine learning approaches for classification. In supervised models, input and output variables are the mathematical connections between the variables in the dataset. It is too tough to comprehend linear regression between the known bi-variables while utilizing supervised learning models. that identify clusters based on inherent structure inside the input data and predict the relationship between the data points [17].
Data collection: Datasets that include details on biological processes, chemical structures, and clinical results that are pertinent to the task of finding new drugs. In these datasets we have chemical structure and connectivity and P-value of chemical compound. First the chemical structure of molecules is converted in to binary numbers.
Machine learning Model: Depending on the nature of the problem, so we selected RNN a feedforward, convolutional, or recurrent neural network architecture. In the input SARS and MARS this dataset is used because of the careerists of COVID-19 virus and using feature extraction and hidden layers we get output using that output we can do classification adding Master file which contains the carbon and hydrogen combinational compound and there scoring and rating. by using RNN model Recurrent neural networks (RNNs) are good at modelling sequential data, they are frequently used in machine learning drug discovery. Since molecules are made up of atoms arranged in particular sequences, molecular structures and interactions are by their very nature sequential in the context of drug discovery. Because RNNs have recurrent connections, they are able to recognise patterns and relationships within sequences, which makes them useful for tasks like activity prediction, property prediction, and molecule production. RNNs, for example, can produce new molecules in molecular generation tasks by sequentially predicting the next atom or bond based on previously formed molecular sections. Moreover, RNNs are able to process input sequences with variable lengths, which allows them to adapt to the flexibility needed in drug development, where molecules might differ in size and complexity. All things considered, RNNs are an attractive option for machine learning applications involving drug discovery due to the sequential nature of molecular data and their capacity to represent such sequences.

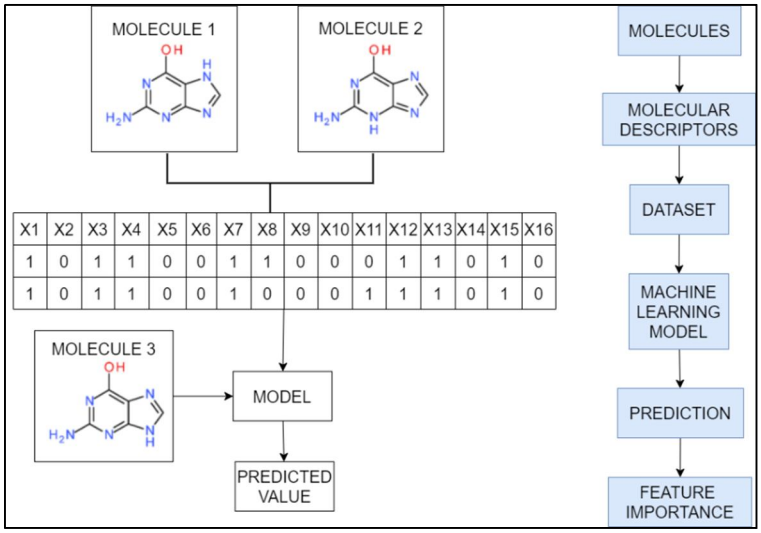

Fig.3 Model Overall Architecture

Predicted Value: Use the preprocessed dataset to train the chosen neural network. Utilizing optimization algorithms, modify the model's weights iteratively in order to reduce the discrepancy between expected and actual results. Results shown in predicted value of connectivity of molecular structure and used in covid 19 treatments.

In this Fig.4 we have the relation Bar-Chart that represents the SMILES (Simplified Molecular Input Line Entry System) vs. training, sources and generated Data it explains the connectivity range with first five chemical Compound based on that we have a range of probability value of trained and source and generated data this important feature we can extract after we can get connectivity range based on this rang for similar data or variant of same chemical compound.

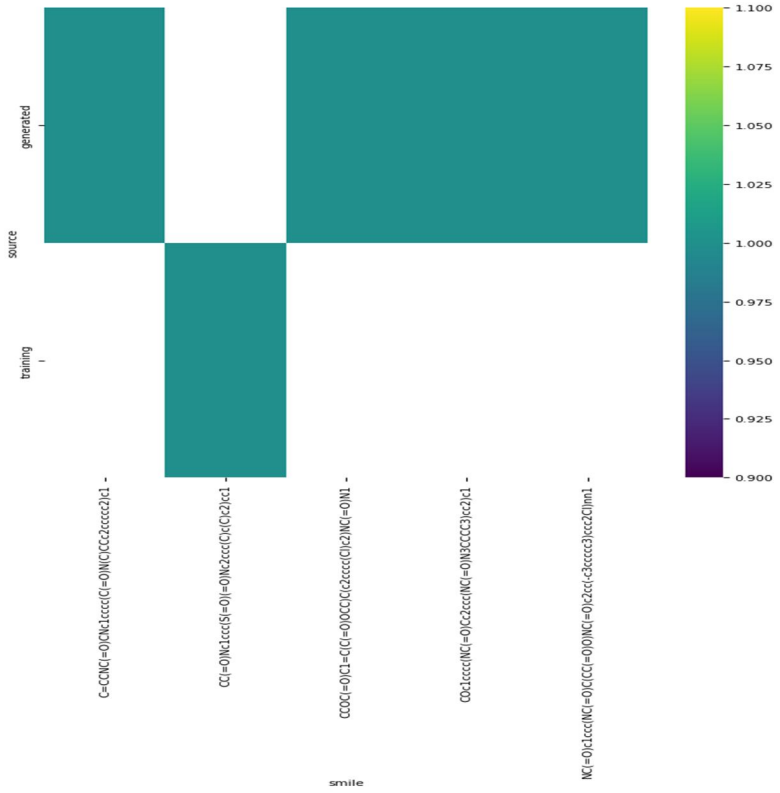

Fig.4 SMILE vs train and source Data

Considering all RNN Algorithm here I have used data comparison and classification using dataset of complete carbon hydrogen compound combination in covid-19 and the list of drugs which are used in SARS and MARS drug compound.

Historically, high-throughput screening of chemical compounds to find viable candidates has been a labor-intensive approach used in traditional small molecular-based drug development methodologies. These methods, although beneficial, are constrained by their high resource requirements and time-consuming nature. Recurrent Neural Networks (RNNs), on the other hand, have become a state-of-the-art technique in drug development, using deep learning to analysis intricate molecular interactions and forecast possible drug candidates with astounding accuracy. Because RNNs are so good at identifying sequential relationships in molecular structures, new substances with desired features can be created. Through the process of training the network on large datasets of biological activity and chemical structures, the network is able to learn complex patterns and relationships [6].

RNNs enable a more efficient and rapid drug development process, cutting down on the time and resources needed to find promising molecules. Consequently, RNN-based methods have demonstrated promise in both finding completely new molecules with therapeutic potential and improving already-existing drug candidates, so bringing about a radical change in the field of small molecular-based drug development.

VI. RESULTS

TABLE I. RESULT OF FIRST FIVE CARBON AND HYDROGEN COMBINATIONAL COMPOUND

id	Simplified Molecular Input Line Entry System (SMILE)	score
AAAA	COc1cccc(NC(=O)Cc2ccc(NC(=O)N3CCCC3)cc2)c1	99.9
AAAB	C=CCNC(=O)CNc1cccc(C(=O)N(C)CCc2ccccc2)c1	99.9
AAAC	CC(=O)Nc1ccc(S(=O)(=O)Nc2ccc(C)c(C)c2)cc1	99,9
AAAD	CCOC(=O)C1=C(C(=O)OCC)C(c2cccc(Cl)c2)NC(=O)N1	99.9
AAAE	NC(=O)c1ccc(NC(=O)C(CC(=O)O)NC(=O)c2cc(-c3cCC..	99.9

These are the outcomes and compounds that gives best reaction and connectivity so the drug can interact with bio-chemicals and use in testing. Here the id represents the specific notation. Similarities of SARS, MARS and covid 19 viruses and based on that we have used RNN and achieved results and get set of molecular compounds so we can use in covid 19 drug trial applying this method we can reduce over all time of new drug discovery.

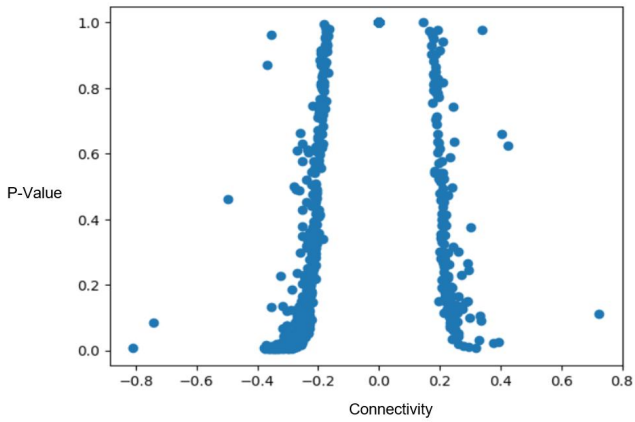

Fig.5 SARS drug P-Value and Connectivity graph

This Graph shows the SARS drug P-Value and connectivity using this data graph and insights we can get similar virus drug for outcome like SARS-CoV-2 virus. Also, we can make specific classification of particular molecular compounds and use as in clinical trial. The proposed RNN model's accuracy in classifying Drug compound from given data. These are the four outcomes that gives best outcome.

TABLE II. RESULT OF CHEMICAL COMPOUND AND THEIR CONNECTIVITY AND P-VALUE

	Name	Connectivity	P-Value
0	valproic acid	-0.376625	0.003243
1	rifabutin	-0.340380	0.003250
2	levomepromazine	-0.329905	0.003363
3	promazine	-0.338460	0.003409
4	Prestwick-559	-0.332285	0.003454
5	maprotiline	-0.345995	0.003484
6	nicergoline	-0.297975	0.003541
7	GW-8510	-0.365775	0.003643
8	lasalocid	-0.287965	0.003653
9	pyrvinium	-0.312920	0.003693

This result gives insight which drugs may give best outcome agents COVID-19. These results have a same range of connectivity and probability value by classifying the cluster we have gotten a 10 drugs that we can use as a medical trails use of this drugs can help to treat symptoms of covid 19.
When it comes to COVID-19 small molecular-based drug development, the use of Recurrent Neural Networks (RNNs) has greatly advanced the field of study. RNNs have been important in examining chemical structures and identifying promising therapeutic candidates due to their capacity to analyze sequential data and recognize complex patterns. Through the utilization of sophisticated computational models, scientists have accelerated the process of identifying chemicals that demonstrate encouraging antiviral characteristics against the SARS-CoV-2 virus so by using these methods we can expedite process of drug clinical trials much faster compared to regular traditional method.

VII. CONCLUSION

A well-constructed RNN, which corresponds to Molecular Graph Network architectures, is highly beneficial in the field of cheminformatics, namely in the construction of molecules with certain features. In this work, we demonstrate its utility. When designing peptides. There are currently more instances of these applications in the field of molecular drug design in this case we achieved most promising drugs for Covid 19 that we can use in clinical trials.
The pharmaceutical industry can benefit greatly from the application of AI's ML and DL techniques to comprehend the chemical structure and activity relationship of lead molecules from a variety of pharmaceutical data sets. Because of their superior data mining capabilities, AI-based techniques have been used in computer-assisted medication development recently. Nevertheless, there are certain drawbacks to this approach. For instance, the effectiveness of both machine learning and deep learning techniques is directly impacted by the volume of data in data mining technologies. The fact that the successful development of deep learning techniques offers a potential solution to these issues, one of the key drawbacks is the lack of clarity surrounding the deep learning models' workings. Furthermore, neural models of formation are used to modify many parameters; nevertheless, only a small number of useful models [1].
The potential applications of artificial intelligence (AI) and machine learning (ML) in small molecular-based drug development are enormous and have great promise [4]. Through the expedited identification of promising drug candidates, the optimization of their attributes, and the streamlining of the whole research and development pipeline, machine learning and artificial intelligence (AI) technologies are transforming the drug discovery process [11]. ML algorithms can uncover novel molecular structures, anticipate possible drug-target interactions, and optimize compound attributes for increased safety and efficacy because to their enormous dataset analysis capabilities. This shortens the time frame for drug research, lowers expenses, and raises the possibility that novel treatments will be successfully introduced to the market [1].
By combining AI with small molecular-based drug development, novel approaches to challenging biological problems may be found, opening the door to more specialized and focused therapies. The combination of artificial intelligence, machine learning [2], and drug discovery has enormous potential to revolutionize the pharmaceutical [5] sector and enhance the quality of healthcare worldwide as technology develops.

REFERENCES

[1] Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries by Xingjin Zhang, Runchuan Li , Yongpeng Liu, Shan Gao, Hongpo Zhang, Shengya Shen,Zongmin Wang, (2021) Springers: Molecular Diversity. https://doi.org/10.1007/s11030-021-10326-z

[2] Sanal MG, Paul K, Kumar S et al (2019) Artifcial intelligence and deep learning: the future of medicine and medical practice. J Assoc Physicians .https://doi.org/10.1007/ s10916-019-1419-x

[3] India. Sousa MJ, Pesqueira AM, Lemos C et al (2019) Decision-making based on big data analytics for people management in healthcare organizations. J Med Syst 43:290.

[4] Vamathevan J, Clark D, Czodrowski P et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477.

[5] Mohs RC, Greig NH (2017) Drug discovery and development: role of basic biological research. Alzheimers Dement (N Y) 3:651–657. https://doi.org/10.1016/j.trci.2017.10.005

[6] Paul D, Sanap G, Shenoy S et al (2021) Artifcial intelligence in drug discovery and development. Drug Discov Today 26:80–93. https://doi.org/10.1016/j.drudis.2020.10.010

[7] Abagyan R, Totrov M, Kuznetsov D. Icm–a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem. 1994;15(5):488–506.

[8] Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews. Comput Molec Sci. 2015;5(6):405–424.

[9] Alpaydin E. Introduction to machine learning. Cambridge: MIT press; 2020.

[10] Ament SA, Pearl JR, Cantle JP, Bragg RM, Skene PJ, Coffey SR, Bergey DE, Wheeler VC, MacDonald ME, Baliga NS, et al. Transcriptional regulatory networks underlying gene expression changes in huntington's disease. Mol Syst Biol. 2018;14(3):e7435. doi: 10.15252/msb.20167435.

[11] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, Thegraph neural network model, IEEE Transactions on Neural Networks,Vol. 20, 2009, pp 61–80.

[12] An H, Li M, Gao J, Zhang Z, Ma S, Chen Y. Incorporation of biomolecules in metal-organic frameworks for advanced applications. Coord Chem Rev. 2019;384:90–106.

[13] Bryce RA, Hillier IH (2014) Quantum chemical approaches: semiempirical molecular orbital and hybrid quantum mechanical/molecular mechanical techniques. Curr Pharm Des 20:3293– 3302.

[14] Aguiar-Pulido V, Gestal M, Cruz-Monteagudo M et al (2013) Evolutionary computation and QSAR research. Curr Comput Aided Drug Des 9:206–225.

[15] Souza PCT, Thallmair S, Confitti P et al (2020) Protein-ligand binding with the coarse-grained Martini model. Nat Commun 11:371.

[16] Machine learning for small molecule drug discovery in academia and industry https://doi.org/10.1016/j.ailsci.2022.100056.

[17] L. Chua, L. Yang, Cellular neural networks: Theory, IEEE Trans. On Circuits and Systems 35, 1988, 1257–1272.

[18] A. Goulon, T. Picot, A. Duprat, G. Dreyfus, Predicting activities without computing descriptors: Graph machines for QSAR, SAR and QSAR in Environmental Research, Vol. 18, 2007, pp. 141–153.