FROM MACHINE LEARNING TO DEEP LEARNING:
"Progress in Machine Intelligence for Rational Drug Discovery"

Prof. Sridhar N K
Faculty: Electronics and
Communication Engineering Dept,
Sri Venkateshwara College of
Engineering, Bangalore
Bangalore, India
sridharnk@gmail.com

Archana P
Dept. of Electronics and
Communication Engineering
Sri Venkateshwara College of
Engineering, Bangalore
Bangalore, India
mail2archanajp@gmail.com

Abstract- Machine intelligence, which is normally presented as artificial intelligence, refers to the intelligence exhibited by computers. In the history of rational drug discovery, various machine intelligence approaches have been applied to guide traditional experiments, which are expensive and time-consuming. Over the past several decades, machine-learning tools, such as quantitative structure– activity relationship (QSAR) modeling, were developed that can identify potential biological active molecules from millions of candidate compounds quickly and cheaply. However, when drug discovery moved into the era of 'big' data, machine learning approaches evolved into deep learning approaches, which are a more powerful and efficient way to deal with the massive amounts of data generated from modern drug discovery approaches. Here, we summarize the history of machine learning and provide insight into recently developed deep learning approaches and their applications in rational drug discovery. We suggest that this evolution of machine intelligence now provides a guide for early-stage drug design and discovery in the current big data era.

I.INTRODUCTION

During Computational tools have been developed and applied to drug discovery as cost-effective alternatives to traditional experiment protocols. The accurate identification of new hits from large chemical libraries by computational models is desirable for the pharmaceutical industry because it can reduce the costs and time associated with experiments needed to obtain new drug candidates with optimized pharmacodynamics and pharmacokinetic (PK/PD) properties [1]. Virtual screening (VS), which is a standard computational approach, is widely used to guide rational drug discovery [2]. Historically, machine-learning Approaches, which are one of the most important components of machine intelligence, have been used to generate various QSAR models for VS over the past few decades [3]. The resulting models are based on molecular structures and target activities, such as physicochemical properties, therapeutic activities, and PK properties [4], which can vary in the different stages of drug discovery.

II. LITRARTURE SURVEY

In Silicon Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Frontiers in Chemistry. [1] The contributions of this paper are Machine learning algorithm is been used to predict the drug toxic or not and Analysis is made to find the toxic drugs. The drawback of this paper are Machine learning algorithm is been used to predict the drug toxic or not and Analysis is made to find the toxic drugs.

Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Molecular Pharmaceutics. [2] The contribution of this paper are Deep learning algorithm is been enhanced for the testing process and Many DL methods is been used for the drug discovery process. The drawback of this paper is Lack of miss classification from the Adaboost decision tree model.

Deep Learning in Drug Discovery. Molecular Informatics. [3] The contribution of this paper are Boltzman machine network technique is been used are Convolutional networks are used to get the information about the molecular. The drawback of this paper is very time consuming and as it is applied on a single drug on the group of drugs.

Survey of Machine Learning Techniques in Drug Discovery. [4] The contribution of this paper are Implemented in pharmacy industries and for the development of a drug delivery in the real world machine learning is been proposed. The drawback of this paper are Databases could be used widely for accurate results and can be implemented in online libraries, bio medical control etc.

Use of machine learning approaches for novel drug discovery. Expert Opinion on Drug Discovery, [5] the contribution of this paper are High hand based drug design, structure based studies are done in this paper and for novelty machine learning algorithm is been used. The drawback of this paper is several approximations need to be upgraded and Issues in molecular drug discovery.

III.QSAR MODELLING

The QSAR modeling procedure has been standardized across rational drug discovery processes [5]. Given the improvements in modeling approaches and the generation of descriptors, QSAR is widely applied at all stages of preclinical studies. The original hypothesis of QSAR ('similar compounds have similar activities') remains the foundation of all QSAR models developed so far. However, although different types of descriptor and different machine-learning methods used for QSAR modeling have their own pros and cons, the resulting models still suffer same issues, such as over fitting and active cliffs, which leads to the failure of predicting new compounds, especially those with chemical structures that different compared with those in the training sets used to develop QSAR models. Thus, new efforts are underway to make QSAR more applicable for drug discovery by integrating new modeling techniques. For example, currently the application of an applicability domain is a necessary step in QSAR modeling and

The use of combinatorial QSAR avoids the potential problems caused by using an individual approach [6].

IV. METHODOLOGY

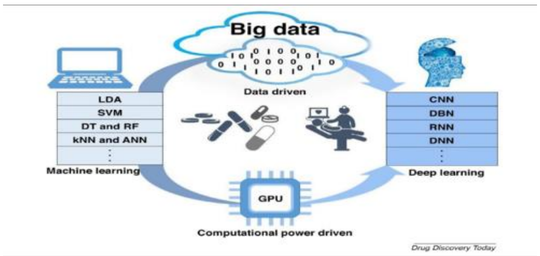

Figure 1: Block Diagram of Advances in machine intelligence for drug discovery

QSAR approaches used in drug discovery can be classified as linear and nonlinear techniques as follows:

- **Linear Discriminant Analysis (LDA):** Linear discriminant analysis (LDA), for example, introduced by Belhumeur in 1996 for pattern recognition and artificial intelligence, is a supervised machine-learning method that is suitable for dealing with small data sets. LDA is a classifier that considers a linear equation to maximize the between-class distance and minimize the within-class distance. LDA has been used to predict drug–drug interactions, identify new compounds, and detect adverse drug events, among others. Although LDA is a simple approach, the combination of LDA and novel descriptors is still considered a powerful modelling method. For example, Marrero et al. used a LDA algorithm combined with topologic, 3D-chiral, topographic, and geometric descriptors to predict the antifungal activity of drugs and yielded a higher accuracy compared with other nonlinear approaches.
- **Support Vector Machines (SVMS):** Support vector machines (SVMs) were proposed by Vapnik and colleagues for their ability to deal with high-dimensional variables in small data sets. For linear problems, the SVM model separates different categories by mapping points in space to maximize the margin between different classes of point. For nonlinear problems, SVMs use kernel mapping and transform nonlinear data sets into a high-dimension feature space for linear classification purposes. SVM has been widely applied for various modelling purposes in drug discovery. For example, Poorin-mohammad et al. combined the SVM approach with pseudo amino acid composition descriptors to classify anti-HIV peptides, with a prediction accuracy of 96.76%.
- **Decision Trees (DTS):** Decision trees (DTs) are a transparent and interpretable ma-chine-learning approach. Generally, there are two essential steps for the construction of decision trees: selecting attributes and pruning. First, molecule attributes are selected as a 'test' on a molecule (e.g., whether the partition coefficient of the molecule is >5). The selected attributes are viewed as internal nodes (in-clouding the root node and nonleaf nodes); the branch represents the outcome of the 'test' and the leaf node represents a classification label. Second, to avoid over fitting and to decrease the complexity of the tree, pruning algorithms are used to trim the generated tree. Recently, DTs were used to model absorption, distribution, and metabolism properties of drugs as well as their toxicity. For example, to

evaluate the toxicity of volatile organic compounds, Gupta and co-workers used DT forest and DT boost algorithms to model the sensory irritation potency of volatile organic compounds. The former algorithm combined DTs with the bagging technique, whereas the latter integrated DTs with a gradient boosting algorithm; both models showed improvement over standard DTs.

- **K Nearest Neighbor (KNN):** The k nearest neighbor (kNN) is an unsupervised algorithm for classification and regression. In most cases, kNN is used for classifications that operate by counting the class of k nearest neighbors in the feature space. Thus, the kNN algorithm is one of the most simple and easy to perform of all machine-learning algorithms, and is normally integrated with other feature-selection algorithms. To identify antiviral drugs, Weidlich et al. applied kNN integrated with a simulated annealing method and RF for 679 drug-like molecules. Their results showed that this improved kNN model outperformed the RF models.
- **Artificial Neural Networks (ANN):** Artificial neural networks (ANN), which simulate brain function, are an attractive and powerful modeling approach widely used in recent drug discovery research. Based on their topological structure, ANN approaches can be classified into four types: forward, backward, random, and self-organized networks. Among these architectures, back propagation neural networks (BPNNs) are one of the most popular ANN methods. BPNN, proposed by Rumelhart and MeClelland, is a forward neural network with multi layered perception. It is a gradient-descendent method that minimizes the mean-square errors of the difference between the network outputs and the experimental data in the training set.

V. QSAR (QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP)

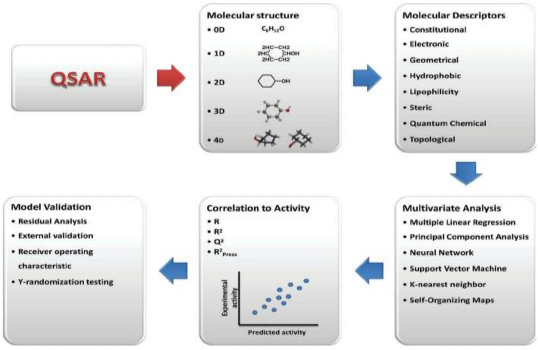

Figure 2: QSAR Structure

- Although QSAR approaches have been developed for decades, common issues remain that have not been solved by using any existing approaches. For example, most traditional QSAR studies have used a single modelling approach to develop a single model based on one type of descriptor. In other cases, when multiple QSAR models have been available, the model selection has always been performed based on statistics obtained from training sets (e.g., cross-validation). However, previous studies have shown that model performance based on training sets has a poor correlation with the external predictions of new compounds. Thus, traditional QSARs that aim to develop a single model and/or select a single model based on training set performance for prediction purpose are questionable.
- Compared with traditional QSAR modeling procedures (e.g., modeling by using one statistical tool and one type of descriptors), recent modeling studies in drug discovery have focused on predictions based on a combination of various types of model (by using different statistical tools and different types of chemical descriptor). Normally for a data set containing enormous and diverse compounds, an individual model would only cover part of its chemical and/or biological diversity. Consensus modeling based on a combinatorial QSAR (combi-QSAR) workflow take advantages of the output information obtained from various available individual models and fully explores the diverse chemical and/or biological information provided by a large training set. The combi-QSAR strategy has been applied to model various absorption, distribution, metabolism, and excretion (ADME) properties, the toxicity of drug molecules, and to select and design new drug candidates.

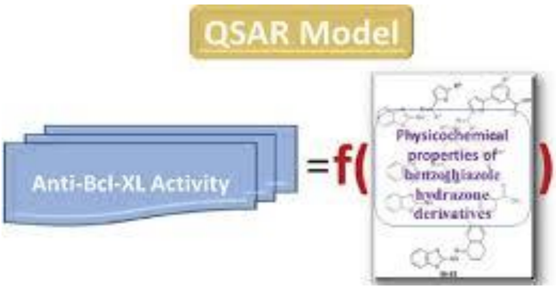
QSAR Model
Anti-Bcl-XL Activity
=f(
Physicochemical properties of benzothiazole hydrazone derivatives
)

Figure 3: QSAR model

- As a trade-off, combi-QSAR modeling is more time-consuming than the development of a single model. Another common issue related to QSAR modeling is the existence of ‘active cliffs’. QSAR modeling cannot deal with the situation whereby two compounds have similar structures but different activities, because it is against the basic hypothesis that 'similar compounds have similar activities'. In some early QSAR studies, additional physicochemical properties, such as partition coefficients (logP), water solubility, and melting point, were used successfully to augment computed chemical descriptors and improve the predictive power of QSAR models. These studies suggest that experimental results obtained from low-cost experimental testing can be used as extra biological descriptors in QSAR modeling to help resolve the 'active cliffs' issue.

- Over the past decade, the rapidly expanding HTS data sets available for large and diverse chemical libraries make it possible to extend the scope of conventional chemical descriptors in QSAR modeling to new hybrid descriptors, including both chemical descriptors and biological descriptors. Therefore, in recent drug discovery studies, models were generated based on new hybrid descriptors. For example, Kim et al. and Wang et al. showed that the oral bioavailability and blood–brain barrier (BBB) models can be improved by including biological descriptors of membrane transportations.
- In this new modeling strategy, the target properties of modeling are still biological activities in drug discovery, but the content and interpretation of 'descriptors' and the resulting models are different. This modeling focuses on the prediction of the same target property from different (chemical, biological, and genomic) characteristics of drugs and provides a unique opportunity to take advantage of both chemical and biological information relating to drug molecules.

VI. ADVANCES IN RATIONAL DRUG DISCOVERY RESULTING FROM DEEP LEARNING.

In the current era of big data and combined with the development of advanced screening protocols (e.g., HTS) and large chemical libraries, the amount of biological data is increasing dramatically. The availability of large data sets and their processing using graphics processing units (GPU) have promoted the development of new modeling approaches. In 2006, Hinton et al. introduced the deep belief networks that made it possible to construct nets with many hidden layers. This resulted in a new theory and caught the attention of many researchers and leading pharmaceutical companies. The concept of deep learning originated from the ANN approach, in which feed forward neural networks combined with many hidden layers are thought of as deep neural networks. Deep learning comprises simple but nonlinear processing units that each transform the representations or features at one level (starting with the raw input) into a representation at a higher, more representative level. Thus, the deep learning approach is a representation-learning method that results in learning multiple levels of representations from low- to high-level features. For example, to recognize images, deep learning networks can learn color information from raw pixel inputs in the first layer and then transform color information to edges of objects in the next layer. Without manually selecting the molecular descriptors, deep learning methods automatically select representations from raw, high-dimension, and heterogeneous data, which is exactly what big data modeling requires. Thus, this is likely to result in deep learning being widely used in various aspects of research, such as image recognition, speech recognition, video games, as well as model development in drug discovery. The most commonly used networks are convolutional neural networks (CNN), stacked auto encoders, deep belief networks (DBN), and restricted Boltzmann machines. As a relatively new approach, its applications in drug discovery can be summarized as follows: (i) new drug molecule identification; (ii) protein engineering; (iii) gene expression data analysis; and (iv) pharmacodynamics modeling.

- **New Drug Molecule Identification:** Identifying new drug candidates from large chemical libraries with computational models (e.g., VS) is an effective and feasible way to facilitate the drug discovery process. Generally, deep learning methods can also be used in this approach to perform VS [44–46]. For example, Pereira et al. introduced a novel deep learning-based VS method, called DeepVS [44]. They performed docking with 40 receptors and 2950 ligands, and compared the results with 95 316 decoys. The docking outputs were used to train deep CNN that could rank the list of ligands for each receptor. The results showed that DeepVS achieved the best performance reported for the VS of these 40 receptors. Similarly, deep learning can also be used to generate focused molecule libraries [47] or new molecular fingerprints and to model PK properties of potential drugs.
- **Protein Engineering:** Protein engineering involves developing and simulating proteins using computers. Recently, researchers used deep learning approaches to explore and discover protein structures and functions. To uncover protein functions, many efforts have been made to simulate interactions between proteins and other bio- logical molecules (e.g., DNA). For example, Hassanzadeh et al. used a recurrent convolutional network to predict the binding specificity of proteins to different DNA loci. They utilized data from in vitro high-throughput experiments to evaluate their

modeling. This modeling approach was shown to be the most accurate for detecting the binding preference between two proteins and individual DNA sub regions. Deep learning methods can also be used to predict biological functions of proteins directly from their raw 3D electron density and electrostatic potential fields.

- **Gene Expression Data Analysis:** With the emergence of next-generation sequencing technology, massive amounts of heterogeneous genomics data can fit well with the requirements of deep learning methods. Thus, deep learning methods have been used in precision medicine development, sequence specification prediction, and genomics modeling for drug repurposing. For example, Aliper et al. Used transcriptional response data to predict the therapeutic categories of drugs. In their study, they used gene-level data of 26 420 drug perturbation samples belonging to 12 therapeutic categories across three cell lines. They integrated the gene expression profiles and pathway activation scores as new features into a deep neural network (DNN) modeling approach, which generated the highest classification accuracy compared with other traditional approaches. They also showed that DNN can accurately predict the category of drugs with different PK and PD conditions.

- **Pharmacodynamics Modelling:** PD modelling is vital to determine the interactions between drugs and their associated targets. Given the diversity of drug molecules and their targets, the potential drug–protein interactions are also complex and have many potential conformations. Recently, deep learning methods were used to predict the interactions of different complexes, such as drug–protein and homogenous complexes. In a recent report, Wen et al. Used DBN to predict drug– target interactions. To identify new drug–target interaction pairs, they used 2 146 240 drug–protein interaction pairs that contained approved drugs and targets without separating them into different classes.

VII. HOW MACHINE LEARNING HELPS IN DRUG DISCOVERY?

- The whole process of creating a new drug generates a lot of data.

- Machine learning offers an excellent opportunity to process chemical data and create outcomes that help us in drug development.

- Machine learning can help us process the data that has been collected over many years and some time decades in very little time.

- Machine learning to its full potential, then it could help the healthcare sector to generate $300 billion in revenue every year.

ADVANTAGES AND DISADVANTAGES OF DEEP LEARNING

ADVANTAGES OF DEEP LEARNING:

- **Conceptually simple:** A deep-learning model is just a chain of simple, continuous geometric transformations.

- **Non-linear**: Non-linear means that the output cannot be reproduced from a linear combination of the inputs.

- **Highly flexible:** Flexibility is a property of a statistical learning method. It is a measure of how much a fitted model can vary with a given train data. The more the flexible the model is, the better it can fit the train data.

- **Can be fine-tuned with more data:** Fine tuning is a process to take a network model that has already been trained for a given task, and make it perform a second similar task.

- **Excellent for pattern recognition:** Pattern recognition is the automated recognition of patterns and regularities in data. It has applications in statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

DISADVANTAGES OF DEEP LEARNING:

- **Hard to interrupt:** Interrupt is a signal emitted by hardware or software when a process or an event needs immediate attention. It alerts the processor to a high priority process requiring interruption of the current working process.

- **Theory not well understood:** Deep Learning Models are Complex Composition of Functions: This single fact is the reason why Deep Learning is powerful and hard. This complex composition introduces many numerical optimization challenges like Vanishing and Exploding Gradients.

- **Slow to train:** Fitting a neural network involves using a training dataset to update the model weights to create a good mapping of inputs to outputs. This training process is solved using an optimization algorithm that searches through a space of possible values for the neural network model weights for a set of weights that results in good performance on the training dataset. Hence it is a slower procedure to train the data sets.

- **May over fit:** Over fitting occurs when you achieve a good fit of your model on the training data, while it does not generalize well on new, unseen data. In other words, the model learned patterns specific to the training data, which are irrelevant in other data.

- **Data hungry:** In problems where data are limited, deep learning often is not an ideal solution. When deep learning algorithm doesn't have enough quality training data. It can fail spectacularly, such as mistaking a rifle for a helicopter, or humans for gorillas.

VIII. RESULTS AND CONCLUSION

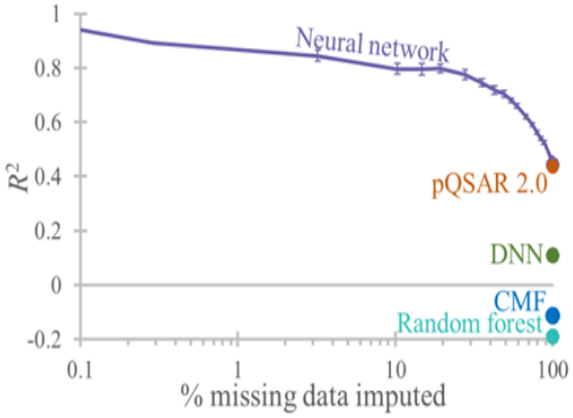

Figure 4: Initial model outcomes of drug discovery

- Figure represents the initial model outcomes of drug discovery using machine learning algorithms.

- The main machine learning algorithms such as neural networks, QSAR, DNN, and RF have been used.

- The data collaborated in above figure shows the probability of QSAR is maximum.

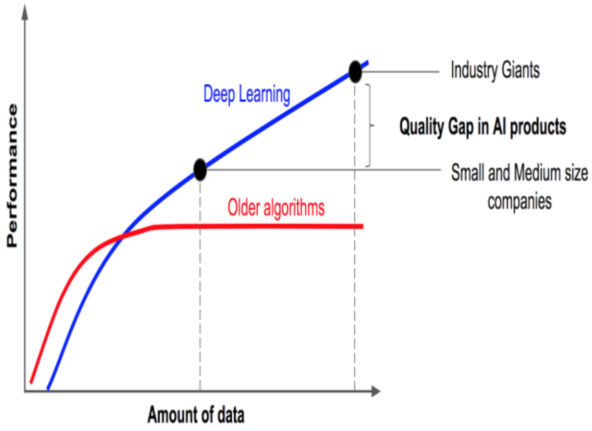

Figure 5: Demonstrates the amount of data versus the performance rate.

- The figure demonstrates the amount of data versus the performance rate.

- The red line in the plot represents the older machine learning algorithms whose performance was low.

- Whereas, observe the blue line which is been switched from machine learning to deep learning that enhanced the performance and lead to massive drug discoveries.

- This deep learning algorithm is more accurate and the results are approximate.

FUTURE SCOPE

- Overall, as a newly developed machine intelligence technique, deep learning has demonstrated the potential for use in the new big data era of drug discovery.

- With more data becoming available and new approaches being developed.

- Deep learning methods will become a major computer-aided drug design (CADD) approach in the near future.

- Artificial Intelligence and machine learning in particular, present the pharmaceutical industry with a real opportunity to do R&D industries.

- So that it can operate more efficiently and substantially improve success at the early stages of drug development

CONCLUSION

Machine intelligence has been applied in the drug discovery field for decades. The development of deep learning methods is driven by the accumulation of massive amounts of biomedical data and the powerful parallel computing capacity of GPUs. Deep learning methods can deal with complex tasks based on large, heterogeneous, and high-dimensional data sets without the need for human input. These methods have been shown to be useful in many practical and commercial applications, including drug discovery studies.

REFERENCES

[1] Yang, H., Sun, L., Li, W., Liu, G., & Tang, Y. (2018). In Silicon Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Frontiers in Chemistry, 6. doi:10.3389/fchem.2018.00030

[2] Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Molecular Pharmaceutics, 14(12), 4462–4475. doi:10.1021/acs.molpharmaceut.7b00578

[3] Gawehn, E., Hiss, J. A., & Schneider, G. (2015). Deep Learning in Drug Discovery. Molecular Informatics, 35(1), 3–14. doi:10.1002/minf.201501008

[4] Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., … Cao, R. (2018). Survey of Machine Learning Techniques in Drug Discovery. Current Drug Metabolism, 19. doi:10.2174/1389200219666180820112457

[5] Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. Expert Opinion on Drug Discovery, 11(3), 225–239. doi:10.1517/17460441.2016.1146250

[6] Danishuddin, M. and Khan, A.U. (2015) Structure based virtual screening to discover putative drug candidates: necessary considerations and successful case studies. Methods 71, 135–145

[7]Wang, T. et al. (2015) Quantitative structure-activity relationship: promising advances in drug discovery platforms. Expert Opin. Drug Dis. 10, 1283–1300

[8] Mistry, P. et al. (2016) Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. Soft. Comput. 20, 2967– 2979

.