International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Machine Learning Approaches in Drug Discovery: A Brief Review

Sadhana Singh, Vishal Bharati
Dept of Computer Science,
Suyash Institute of Information Technology, U.P., India
sadhanasingh9651@gmail.com, vishalbharati161715@gmail.com

Abstract: Drug discovery is a complex, costly, and time-intensive process that traditionally relies on high-throughput screening, molecular synthesis, and extensive laboratory experimentation to identify effective therapeutic compounds. In recent years, machine learning (ML) has emerged as a transformative technology in pharmaceutical research by enabling faster, more accurate, and cost-efficient drug development. ML techniques can analyze large-scale biological and chemical datasets to predict molecular properties, identify potential drug candidates, optimize lead compounds, and evaluate drug-target interactions and ADMET properties. This review explores various machine learning approaches applied across different stages of the drug discovery pipeline, including supervised learning, deep learning, ensemble methods, and reinforcement learning. It also highlights recent advancements such as generative AI, graph neural networks, and explainable AI, while discussing existing challenges related to data quality, model interpretability, and regulatory acceptance. Overall, machine learning has significant potential to accelerate drug discovery, reduce development costs, and improve the efficiency of modern pharmaceutical research.

Keywords: Machine Learning, Drug Discovery, Virtual Screening, ADMET Prediction, Drug-Target Interaction

1. Introduction

Machine learning (ML) has emerged as a powerful and transformative technology in the field of drug discovery, significantly improving the speed, accuracy, and efficiency of identifying and developing new therapeutic compounds. Conventional drug discovery approaches are generally labor-intensive, expensive, and time-consuming, often requiring extensive laboratory experiments, clinical evaluations, and several years of research before a drug can successfully reach the market. In contrast, ML-driven methodologies offer intelligent, data-oriented solutions capable of processing and interpreting large-scale biological, chemical, and pharmacological datasets with greater precision. Various machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), are extensively utilized for applications such as virtual screening, hit and lead identification, molecular property prediction, and bioactivity analysis. Furthermore, ML-enhanced Quantitative Structure–Activity Relationship (QSAR) modeling and drug-target interaction (DTI) prediction techniques play a vital role in evaluating important pharmacokinetic parameters, including absorption, distribution, metabolism, excretion, and toxicity (ADMET). The integration of machine learning with bioinformatics, omics technologies, and natural language processing (NLP) has further strengthened biomarker discovery, target identification, and personalized medicine research. Although substantial progress has been achieved, several challenges such as insufficient data quality, limited model interpretability, overfitting issues, and poor cross-dataset generalization continue to affect the reliability of ML systems in pharmaceutical applications. To overcome these limitations, recent research has increasingly focused on explainable artificial intelligence (XAI), transfer learning, federated learning, and multi-task learning approaches to develop more transparent, robust, and trustworthy predictive models. Overall, machine learning is rapidly becoming an essential component of modern pharmaceutical research and drug development, offering tremendous potential to accelerate the discovery of safer, more efficient, and highly personalized therapeutic solutions for complex diseases.

2. Role of Machine Learning in Drug Discovery

ML algorithms excel at detecting patterns and making predictions from large datasets, making them ideal for the multi-dimensional data typical in drug discovery. Applications span various stages:

- Hit Identification: ML models assist in virtual screening by predicting bioactivity of molecules, helping prioritize compounds for synthesis and testing.
- Lead Optimization: Techniques like Quantitative Structure-Activity Relationship (QSAR) modeling predict chemical modifications that can improve efficacy and safety.
- Drug-Target Interaction (DTI) Prediction: ML methods identify interactions between small molecules and biological targets, crucial for efficacy and off-target effect prediction.
- ADMET Prediction: Algorithms forecast Absorption, Distribution, Metabolism, Excretion, and Toxicity properties, enabling early elimination of unsuitable candidates.

3. Common Machine Learning Techniques Used

- Support Vector Machines (SVM): Effective for classification tasks like identifying active vs. inactive compounds.
- Random Forest (RF): An ensemble learning method popular for its robustness in QSAR modeling.
- Artificial Neural Networks (ANN): Useful for modeling non-linear relationships in chemical and biological data.

- Deep Learning: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown superior performance in image and sequence data analysis, respectively.
- Gradient Boosting Machines (GBM): Often outperform traditional ML models in predictive tasks related to bioactivity and ADMET properties.

4. Applications Across Drug Discovery Pipeline

- Virtual Screening: ML enables screening of millions of compounds against known biological targets with high precision.
- De Novo Drug Design: Generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are used to design novel compounds.
- Repurposing Existing Drugs: ML analyzes vast medical and molecular datasets to identify new indications for existing drugs.
- Predicting Clinical Trial Outcomes: ML models analyze patient data and trial parameters to forecast success probabilities.

5. Integration with Omics and Big Data ML integrates genomic, proteomic, transcriptomic, and metabolomic data to understand disease mechanisms and identify novel drug targets. Multi-omics data fusion allows more accurate disease modeling and personalized medicine.

6. Challenges in ML-Based Drug Discovery

- Data Quality and Availability: Incomplete, noisy, or biased datasets can reduce model accuracy.
- Interpretability: Deep learning models often act as black boxes, making it difficult to understand the rationale behind predictions.
- Generalizability: Models trained on specific datasets may perform poorly on external data.
- Regulatory and Ethical Concerns: Validation of ML-based predictions requires stringent regulatory oversight.

7. Related Work:

Aging studies are made on organisms such as Caenorhabditis elegans (C. elegans) worm, yeast, mice and (Buffenstein et al., 2008). Among them, C. elegans is highly used for conducting assessments on anti-aging and drug intervention due to its short lifespan, stereotypical development and small size (Kaletta and Hengartner, 2006) (Lucanic et al., 2013) (Carretero et al., 2017). There has been a decent increase in the number of works done on the identification of compounds that increase the lifespan. Latest research enabled us to develop compounds that enhance the longevity of caloric restriction (Fontana et al., 2010). In (Calvert et al., 2016), a pharmacological network is constructed by Ye et al. with the help of a connectivity map to identify drugs with overlapping gene expression profiles.

It helped to identify 60 compounds that improve longevity for C. elegans (Ye et al., 2014). Ranking of drug like compounds to modulate aging in C. elegans has been proposed in (Ziehm et al., 2017). This ranking method is based on genetic information, information on proteins associated with aging in an organism, 3D protein structure, homo-logy and sequence conversation between them and compound activity information. In (Ding et al., 2017) and (Carretero et al., 2015), a review on anti-aging and pharmacological compound classification on C. elegans lifespan has been discussed. The process of drug discovery includes several steps from selecting chemical compound candidates to clearing all drug requirement tests. It is also time consuming, expensive and labor-intensive. Recent advancements in machine learning is helping out in prediction of the chemical properties in drug discovery community. In (Zernov et al., 2003), Support Vector Machine is used for predicting the drug-likeness and agrochemical-likeness for large number of compounds. In (Putin et al., 2016), human chronological age was predicted using Deep Neural Network. In (Fabris et al., 2018), prediction of aging related genes was done using random forests.

Different possibilities of involvement of machine learning and artificial intelligence in longevity and anti-aging discoveries was discussed in (Batin et al., 2017). Machine learning was not greatly used for compound prediction on longevity. To the best of our knowledge, only a few studies tried to identify chemical compounds that effects longevity (Putin et al., 2016)(Barardo et al., 2017) (Fabris et al., 2017). Barardo et al. worked on classification of chemical compounds into two classes which are the ones that effect longevity and the ones that do not, using random forests (Barardo et al., 2017). They used random forest feature importance as a parameter to select the most important features among the two types of features which are chemical descriptors, gene ontology terms. Then, they implemented random forests for classification. Their results showed that impact of chemical descriptors was high compared to GO. However, using both chemical descriptors and Gene Ontology slightly improves the classification accuracy.

Information provided by features in a data set play a major role in classifying each instance into different classes. In many cases, a compromise on relevant and redundant features is observed in the data. Redundant features slow down the process of learning resulting in a decrease in accuracy. Feature selection helps us to identify and eliminate these redundant features helping us to improve the performance of the classification (Dash and Liu, 1997) (Xue et al., 2013). Exponential increase in the size of search space with the size of features makes feature selection much more challenging. Due to this, search methods such as greedy search or heuristic-based search have been proposed (Mao and Tsang, 2013). But these methods are prone to suffer from the local optima problem. Feature selection can be done using two approaches: 1) methods that are independent of a learning algorithm and do not consider classifier performance (Moradi and Rostami, 2015), and 2) wrapper methods which learns an algorithm in the feature selection process (Gasca et al., 2006) (Wang et al., 2008). Wrapper methods can enable us to achieve higher predictive accuracy (Dash and Liu, 1997).

Determining Drug-Target Interaction (DTI) is a major part in the area pharmaceutical sciences (Fakhraei et al., 2014). The process of drug discovery involves costs around $1.8 billion with a duration which may extend beyond 10 years (Ciociola et al., 2014). Thus, the drug-target interactions prediction helps in narrowing down the search area helping out the biologists. The main step in the process of drug discovery is to recognize the targets related to the drugs. These targets are mostly the proteins that can be drugged, and the ones related to diseases. The prediction of drug-target interactions aims at identifying the possible new targets for the existing drugs. It basically guides us through a proper experimentation process. There are many types of protein targets which include enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. These classes can modulate their function by interacting with various ligands. Thus, analyzing the genomic space produced by these classes of proteins, helps us to accurately estimate the possibility of an interaction. DTI can be used either for drug discovery or for re-positioning which is reusing available drugs for new targets. Approaches to predicting DTI can be classified into one of three categories: 1) Ligand-based, 2) Docking-based and 3) Chemogenomic approach. Prediction of DTI based on target proteins' ligand similarity is the Ligand-based approach (Keiser et al., 2007) (Keiser et al., 2009). 3D structure information of a target protein can help in estimating the likelihood of its possible interaction with certain drug. This is based on their binding capacity and strength (Cheng et al., 2007) (Morris et al., 2009). This method is the Docking-based approach. Lastly, if the chemical information of the drugs, genomic information from proteins and already identified DTIs are used, then this is the chemogenomic approach (Mousavian and Masoudi-Nejad, 2014) (Ding et al., 2013). A target with small number of binding ligands often leads to poor DTI predictions in a ligand-based method. This is a drawback in this method. Similarly, docking-based method is based on availability of 3D structures for target proteins and is also time consuming. These disadvantages made the chemogenomic approach more popular for identification of DTI in the recent times. This approach models the DTI problem as a machine learning problem and often builds a classifier which is trained by an available interaction data. This classifier is used to predict the unknown interactions (Ding et al., 2013).

Different techniques are employed in chemogenomic approach. Some of these include bipartite graph (Bleakley and Yamanishi, 2009)(Lu et al., 2017), recommendation systems (Alaimo et al., 2016) and supervised classification problem (Wen et al., 2017). But considering the data, we can see that there will be only a few positive interactions and the remaining possible interactions are unknown. For example, of the 35 million drug compound possible candidates, total number of positive drug interactions could just be 7000 (Bolton et al., 2008). Computational chemogenomic approaches can be classified into feature-based and similaritybased methods. Feature-based methods have features as inputs for a set of instances defined by a particular class label. The instances are generally the drugs and the features are the targets. The class label is a binary value indicating the presence of a possible interaction. Some of the feature-based methods like decision trees, random forests (Breiman, 2001) and support vector machines (Cristianini and Shawe-Taylor, 2000) are used for classification purposes. Generally, Random Forest and Support Vector Machine are used for the classification of drug-target interactions (Yu et al., 2012).

Similarity-based methods take the similarity matrices of drugs and targets as inputs including the interaction matrix which indicates the drug-target pairs that most likely interact with each other. These approaches assume the common features among the drugs and targets can determine the presence of an interaction between a drug and target. The main focus is on the structural similarity between the drugs and targets leaving out the remaining unknown attributes. This gives an opportunity for easy implementation of the similarity indices on the networks without any prior information. To implement the similarity-based algorithms, many similarity indices were used by (Lu et al., 2017). There are two ways to classify the similarity indices as local similarity indices and global similarity indices. The local similarity indices are node-dependent i.e. they require information related to neighbourhood of the network. On the other hand, global similarity indices are dependent on the path through the entire network topology (Lu et al., 2017). Use of the similarity indices outperform many random link predictors. A few examples of the similarity indices include Common Neighbours (CN), Jaccard Index, Preferential Attachment (PA) and Katz Index (Lu et al., 2017).

Drug-target link prediction is of high interest in the recent times. The application of machine learning in this challenging field of study makes it more promising and researchers have been trying to explore in depth in this field. (You et al., 2018) developed a lasso-based regularized linear classification method to predict the DTI overcoming the high-dimensional nature of the drug target data with huge number of features and small number of samples. A recent use of stacked auto encoder from the drug molecular structure and protein sequence to extract sufficient information was given by (Wang et al., 2018). DINES, a web server defined as drugtarget interaction network inference engine based on supervised analysis is used to predict unknown DTI for different biological data. It predicts with the help of machine learning methods integrating with the heterogeneous biological data in compatibility with the KEGG data (Yamanishi et al., 2014).

A framework was proposed by (Fakhraei et al., 2014) to work with a bipartite graph of drug-target interactions along with similarity measures. This used probabilistic soft logic to make predictions based on triad and tetrad structures. Enhancing the similarity measures to involve the non-structural information and to handle the possible missing interactions was done by (Shi et al., 2015). Matrix-factorization has been used in many research to identify new drug-target interactions (G□onen, 2012) (Eslami Manoochehri and Nourani, 2018) (Zheng et al., 2013). (Liu et al., 2016) used matrix-factorization method to focus on determining the probability of a drug-target interaction. In this method, the properties of drugs and targets were represented using their specific latent

vectors. A modification to this method which uses Bayesian optimization was developed by (Ban et al., 2017). It uses the neighborhood regularized logistic matrix factorization for DTI prediction. A multiple kernel link prediction on bipartite graphs was proposed by (Nascimento et al., 2016). In this method, relevant kernels are selected based on the weights which define the importance of a drug-target link. Bipartite Local Model (BLM) was extended to DTI prediction with hubness-aware regression in (Buza and Peska, 2017). It builds a projection-based ensemble of BLMs using similarity space of drugs and targets. In (Lan et al., 2016), the unknown links are treated as unlabeled samples. A majority voting method is used to decide on the label of these unlabeled samples using the method of Positive-unlabeled learning for drugtarget prediction (PUDT). This unlabeled data is divided into reliable negative samples and likely negative samples with the help of their similarity information. Weighted SVM is used to then classify this data. We can also have positive unlabeled data similar to negative unlabeled data. To deal with this data, two group of approaches were proposed. One identifies the reliable negatives among the data and use the positives and these negatives to build the model. The other group directly predicts the positive unlabeled data (Liu et al., 2017).

Typically, there is a lot of unbalanced data in the drug-target interaction database. Weighted profile can be used to determine the drug profiles with defines weights as similarities between drug-drug (Yamanishi et al., 2008). The nearest profile is an approach that predicts the interaction profile of new drugs. (van Laarhoven and Marchiori, 2013) developed a simple weighted nearest neighbor procedure for predicting the interacting pairs with high accuracy. A network based inference helps in building a predictive model similar to a network where the nodes are represented by drugs and targets. Interacting drugs and targets are represented as edges (Cheng et al., 2012).

8. Conclusion:

Machine learning is reshaping the landscape of drug discovery by enabling data-driven decision-making and reducing time-to-market for new therapeutics. While challenges remain, continued advances in algorithms, data integration, and computational power promise to make ML an integral component of the pharmaceutical R&D toolkit.

References

[1] Collins FS, Varmus H. A new initiative on precision medicine. New England J Med 2015;372(9):793–5.
[2] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.
[3] Romond EH, Perez EA, Bryant J, Suman VJ, Geyer Jr CE, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, et al. Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. N Engl J Med 2005;353(16):1673–84.
[4] Blanco JL, Porto-Pazos AB, Pazos A, Fernandez-Lozano C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. Sci Rep 2018;8(1):1–11.
[5] Munteanu CR, Fernández-Blanco E, Seoane JA, Izquierdo-Novo P, Angel Rodriguez-Fernandez J, Maria Prieto-Gonzalez J, Rabunal JR, Pazos A. Drug discovery and design for complex diseases through qsar computational methods. Current Pharmaceutical Des 2010;16(24):2640–55.
[6] García I, Munteanu CR, Fall Y, Gómez G, Uriarte E, González-Díaz H. Qsar and complex network study of the chiral hmgr inhibitor structural diversity. Bioorganic Med Chem 2009;17(1):165–75.
[7] Liu Y, Tang S, Fernandez-Lozano C, Munteanu CR, Pazos A, Yu Y-Z, Tan Z, González-Díaz H. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. Expert Syst Appl 2017;72:306–16.
[8] Riera-Fernández P, Munteanu CR, Dorado J, Martin-Romalde R, Duardo- Sanchez A, Gonzalez-Diaz H. From chemical graphs in computer-aided drug design to general markov-galvez indices of drug-target, proteome, drugparasitic disease, technological, and social-legal networks. Current Computeraided Drug Des 2011;7(4):315–37.
[9] Shirvani P, Fassihi A. Molecular modelling study on pyrrolo [2, 3-b] pyridine derivatives as c-met kinase inhibitors, a combined approach using molecular docking, 3d-qsar modelling and molecular dynamics simulation. Mol Simul 2020:1–16.
[10] B. Suay-Garcia, J.I. Bueso-Bordils, A. Falcó, M.T. Pérez-Gracia, G. Antón-Fos, P. Alemán-López, Quantitative structure–activity relationship methods in the discovery and development of antibacterials, Wiley Interdisciplinary Reviews: Computational Molecular Science e1472.
[11] Fernandez-Lozano C, Gestal M, Munteanu CR, Dorado J, Pazos A. A methodology for the design of experiments in computational intelligence with multiple regression models. PeerJ 2016;4:e2721.
[12] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic acids research 46 (D1) (2018) D1074–D1082..
[13] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.
[14] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, et al. Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1): D1100–7.
[15] Sterling T, Irwin JJ. Zinc 15–ligand discovery for everyone. J Chem Inform Modeling 2015;55(11):2324–37.
[16] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
[17] Gramatica P, Sangion A. A historical excursus on the statistical validation parameters for qsar models: a clarification concerning metrics and terminology. J Chem Inform Modeling 2016;56(6):1127–31.
[18] Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, Zhang J, Chan L, Cao R. Survey of machine

learning techniques in drug discovery. Current Drug Metabolism 2019;20(3):185–93.
[19] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery 2019;18 (6):463–77.
[20] Gramatica P. Principles of qsar modeling: comments and suggestions from personal experience. Int J Quantitative Structure-Property Relationships (IJQSPR) 2020;5(3):61–97.
[21] Cronin MT, Schultz TW. Pitfalls in qsar. J Mol Struct (Thoechem) 2003;622(1– 2):39–51.
[22] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110(18):5959–67.
[23] Valdés-Martiní JR, Marrero-Ponce Y, García-Jacas CR, Martinez-Mayorga K, Barigye SJ, d'Almeida YSV, Pérez-Giménez F, Morell CA, et al. Qubils-mas, open source multi-platform software for atom-and bond-based topological (2d) and chiral (2.5 d) algebraic molecular descriptors computations. J Cheminformatics 2017;9(1):1–26.
[24] Fourches D, Ash J. 4d-quantitative structure–activity relationship modeling: making a comeback. Expert Opinion Drug Discovery 2019;14 (12):1227–35.
[25] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6), 1241-1250.
[26] Vamathevan, J., Clark, D., Czodrowski, P., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18, 463–477.
[27] Zhavoronkov, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37, 1038–1040.
[28] Anurag et. al., "Load Forecasting by using ANFIS", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020
[29] Raghawend, Anurag, "Detect Skin Defects by Modern Image Segmentation Approach, Volume 20, Issue 1, 2020
[30] Ekins, S., & Puhl, A. C. (2017). Exploiting machine learning for end-to-end drug discovery and development. *Trends in Pharmacological Sciences*, 38(5), 478-494.