International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning in Drug Discovery: A Review

Helly Paraskumar Rana[1]*, Tannir Anjali Kishorbhai*[2]*, Helly Sunil Shah*[3]*, Janki Tejas Patel*[4]

[1]Student, B.E. Computer Engineering, Sal College of Engineering, Ahmedabad, Gujarat, India
[2]Student, B.E. Computer Engineering, Sal College of Engineering, Ahmedabad, Gujarat, India
[3]Student, B.E. Computer Engineering, Sal College of Engineering, Ahmedabad, Gujarat, India
[4]Assistant Professor, Computer Engineering Department, Sal College of Engineering, Ahmedabad, Gujarat, India

ABSTRACT:

Drug discovery is a cornerstone of modern healthcare, driving the development of novel therapies to address complex and evolving diseases. However, the process remains time-consuming and resource-intensive, posing significant challenges to pharmaceutical innovation. Recent advances in artificial intelligence (AI) and machine learning (ML) present transformative opportunities to accelerate this pipeline, reduce costs, and improve overall efficiency. By leveraging vast biological datasets, advanced algorithms, and increasing computational power, AI has shown remarkable potential for optimizing target identification, molecular design, and drug repurposing. Among these approaches, Transformer-based models—originally developed for natural language processing—have emerged as particularly promising, owing to their ability to capture intricate patterns in sequential and structural biological data. Their applications are enabling breakthroughs in predicting molecular interactions, understanding protein folding, and generating candidate compounds. As the field advances, Transformer-driven drug discovery holds the potential to revolutionize biomedical research and significantly propel the development of precision medicine.

Keywords: Machine Learning, Drug Discovery, Deep Learning, Artificial Intelligence, De Novo Drug Design

1. Introduction

Drug research and development constitute a cornerstone of modern healthcare, significantly contributing to improved human health and well-being. However, the discovery and development of a new drug remain highly complex, costly, and time-intensive, with an average duration of more than ten years and an estimated cost of approximately USD 2.6 billion. Despite such high investment, the success rate of a small molecule advancing from phase I clinical trials to market approval remains below 10%, reflecting the considerable risk, financial burden, and uncertainty associated with drug development. Consequently, reducing costs while accelerating the pace of discovery has emerged as an urgent challenge for the pharmaceutical industry.

The rapid expansion of large-scale biomedical datasets presents unprecedented opportunities for computational approaches to transform drug discovery. Nonetheless, effectively mining, correlating, and interpreting vast data streams remains a significant scientific challenge. The advent of artificial intelligence (AI), supported by advanced algorithms and scalable computational resources, provides a powerful solution. Machine learning (ML), the most widely applied AI method, enables predictive modeling by learning from existing data through supervised, unsupervised, or reinforcement learning paradigms. Deep learning (DL), a subset of ML employing multi-layered artificial neural networks, offers particular advantages in handling complex, high-dimensional biomedical data. Such approaches have already revolutionized key stages of drug discovery, including target identification, de novo drug design, and drug repurposing. For instance, DL-based frameworks such as DeepDTA and DeepAffinity have improved the prediction of drug–target interaction affinities, accelerating candidate screening and prioritization. Collaborations between pharmaceutical giants like Sanofi, Merck, Takeda, and Genentech with AI-driven companies underscore the growing recognition of these technologies.

Most notably, Transformer-based models, such as GPT, BERT, and T5, originally designed for natural language processing, have catalyzed a paradigm shift. With their superior capacity to capture long-range dependencies, process sequences in parallel, and integrate multimodal data, these models are increasingly applied to drug-related biological sequences. Frameworks like TransDTI exemplify their potential, demonstrating superior performance in predicting novel drug–target interactions. Thus, ML—particularly Transformer-based architectures—offers unprecedented opportunities to revolutionize drug discovery pipelines. This paper explores the recent advances, opportunities, challenges, and future directions of ML, with a particular focus on the transformative role of Transformers in drug discovery.

2. Application of Machine Learning in Drug Design

2.1. Prediction of the Target Protein Structure :

Proteins play essential roles in diverse biological processes, and their dysfunction can result in abnormal cellular behavior and disease onset. For selective disease targeting, small-molecule compounds are typically designed based on the three-dimensional (3D) chemical environment surrounding ligand-binding sites of target proteins. Thus, accurate prediction of the 3D structure of target proteins is of great importance in structure-based drug discovery. Traditionally, homology modeling has been employed for this purpose, relying on known protein structures as templates. In comparison, machine learning (ML)-based approaches have shown greater promise in predicting protein structures with enhanced accuracy and efficiency. Notably, AlphaFold, a protein structure prediction system developed by DeepMind, utilizes deep neural networks (DNNs) and has achieved remarkable success in multiple international competitions. By analyzing amino acid distances and peptide bond angles, AlphaFold has demonstrated the ability to generate highly accurate 3D protein structures, marking a significant advance in the field and offering transformative potential for drug discovery. Nonetheless, proteins are dynamic and may adopt multiple conformations under different environmental conditions, which introduces additional complexity to structure prediction.

2.2. Prediction of Protein–Protein Interactions (PPIs) :

Proteins rarely function in isolation; rather, they cooperate with other proteins to establish intricate protein–protein interaction (PPI) networks. PPIs play indispensable roles in various biological processes, regulating protein activity, determining specificity, and creating novel binding sites for effector molecules. Consequently, elucidating and targeting PPIs offers opportunities to design innovative therapeutics that modulate complex biological functions.

ML-based methods for PPI prediction have been broadly categorized into structure-based and sequence-based approaches. Structure-based methods typically exploit protein structural similarities. For instance, IntPred, a random forest tool, predicts interface sites with strong accuracy across obligate and transient complexes (MCC = 0.370, accuracy = 0.811, specificity = 0.916, sensitivity = 0.411). Similarly, Struct2Graph, a graph attention network (GAT)-based model, identifies PPIs directly from protein chain structures with exceptional accuracy (0.9989 on balanced datasets, 0.9942 on unbalanced datasets). Sequence-based approaches, in contrast, leverage protein sequence data. DeepPPI, for example, applies deep neural networks to learn protein representations from descriptors and achieves excellent performance on *S. cerevisiae* data (accuracy = 0.925, precision = 0.9438, recall = 0.9056, specificity = 0.9449, MCC = 0.8508, AUC = 0.9743). Extensive testing further demonstrated its superiority across *S. cerevisiae*, *H. pylori*, and *H. sapiens* datasets. In addition, DELPHI, a deep ensemble model, has been introduced to predict PPI-binding sites using UniProt data. Despite these advances, structure-based approaches remain limited by the scarcity of experimentally resolved protein structures and their quality, giving sequence-based methods broader applicability.

2.3. Prediction of Drug–Target Interactions (DTIs) :

Most drugs exert therapeutic effects by interacting with specific biological molecules such as receptors, enzymes, or ion channels. Thus, accurate prediction of drug–target interactions (DTIs) is a crucial step in drug development. Traditional experimental methods, while accurate, are expensive and time-intensive. ML-based approaches now provide efficient alternatives, with a focus on three critical aspects: identifying drug-binding sites, estimating binding affinity, and determining binding poses within targets.

For binding site prediction, Cui introduced DeepC-SeqSite, a sequence-based deep CNN, which outperformed several state-of-the-art tools including COACH. Similarly, Zhou proposed AGAT-PPIS, an augmented GAT-based model that improved prediction accuracy by 8% over existing methods. In terms of binding affinity estimation, tools such as DEELIG and GraphDelta, based on machine and deep learning models, have proven highly effective. Moreover, Nguyen integrated random forest and CNN techniques to develop a scoring function that enhances docking pose evaluation across software platforms such as GOLD, GLIDE, and AutoDock Vina. These innovations underscore the growing utility of ML systems in predicting DTIs and facilitating the design of reliable drug candidates.

2.4. De Novo Drug Design :

De novo drug design refers to the computational creation of novel therapeutic molecules without relying on existing bioactive compounds or known structures. This approach enables the generation of molecules with defined biological properties tailored to specific diseases. Traditional *de novo* methods, such as fragment-based design, often produce compounds with poor pharmacokinetics, unfavorable metabolism, or impractical synthesis requirements. Thus, innovative approaches are in high demand to generate molecules meeting biological, pharmacological, and chemical suitability.

Recent advances in ML-based generative modeling, particularly autoencoder variants such as variational autoencoders (VAEs), adversarial autoencoders (AAEs), and generative adversarial networks (GANs), have transformed the field. For example, PaccMannRL applies a hybrid VAE with reinforcement learning to design novel anticancer molecules from transcriptomic data. druGAN, another AAE-based model, generates anticancer compounds with specific therapeutic properties. Additionally, MedGAN, which combines Wasserstein GANs with graph convolutional networks (GCNs), has been successfully used to generate quinoline-scaffold molecules, achieving high proportions of effective (25%), novel (93%), and unique (95%) compounds. To address challenges in chemical synthesis, Coley introduced the SCScore, which leverages reaction precedent knowledge with ML to evaluate synthetic complexity. Together, these advances highlight ML as a powerful driver in *de novo* drug design, revolutionizing the discovery and development of innovative therapeutics.

3. Machine Learning Methods to Drug Discovery

Artificial intelligence (AI) has emerged as a transformative force in drug discovery and design, offering innovative solutions to overcome the limitations of traditional methodologies. By enhancing machine learning (ML) approaches and integrating vast repositories of pharmacological data, AI facilitates the conversion of complex biomedical information into actionable insights. Unlike conventional strategies that rely heavily on incremental experimental improvements, AI holds the unique advantage of translating large-scale, heterogeneous datasets into predictive and reusable models, thereby accelerating the entire drug development pipeline.

A wide range of classical ML algorithms has been employed in drug design, including Random Forest (RF), Naïve Bayesian Classification (NBC), Multiple Linear Regression (MLR), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Probabilistic Neural Networks (PNN), Multi-Layer Perceptrons (MLP), and Support Vector Machines (SVM) (Lavecchia and Di Giovanni, 2013). These methods have proven valuable for classification, regression, and prediction tasks, such as virtual screening, quantitative structure–activity relationship (QSAR) modeling, and drug–target interaction analysis. However, their performance is often constrained by the need for manual feature engineering and limited capability in capturing nonlinear, high-dimensional patterns inherent in biomedical data.

Recent advancements have shifted attention toward deep learning (DL), a powerful extension of AI distinguished by its ability to perform automated feature extraction and generalization from complex datasets. Deep architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based models, have shown remarkable success in identifying intricate molecular structures, predicting protein–ligand binding affinities, and designing novel compounds. Unlike classical ML, DL scales efficiently with data volume and computational resources, uncovering hidden relationships within biological systems that were previously inaccessible.

4. Opportunities for Transformer-Based ML Models in Empowering Drug Discovery

The Transformer architecture, first introduced by Vaswani et al. in the seminal paper *"Attention is All You Need,"* represents a groundbreaking advancement in deep learning. By leveraging self-attention mechanisms, the Transformer enables efficient parallelization and excels at capturing long-range dependencies within data—capabilities that surpass traditional recurrent neural network (RNN) models. This innovation has significantly improved performance across numerous natural language processing and computer vision tasks, setting new benchmarks in several domains.

Given these advantages, the Transformer has emerged as a transformative paradigm within machine learning, particularly in the field of drug discovery. Its ability to model complex molecular interactions, predict chemical properties, and accelerate the identification of potential therapeutics underscores its growing significance in computational biology and cheminformatics. Consequently, the integration of Transformer-based models into drug discovery research offers a promising trajectory for developing more accurate predictive frameworks and expediting the overall drug development process.

4.1. Opportunity 1: Transformer Models Empowering PPI Identification :

Traditional machine learning approaches for protein–protein interaction (PPI) prediction commonly employ convolutional neural networks (CNNs) to extract low-dimensional representations from amino acid sequences. However, these methods often overlook the critical long-range dependencies that exist within protein sequences. Transformer architectures, with their powerful self-attention mechanisms, effectively capture these long-distance relationships, making them highly suitable for predicting protein interactions. For instance, Lin introduced *DeepHomo2.0*, a deep learning framework that integrates Transformer-derived features with monomer structural information and direct-coupling analysis. This hybrid design demonstrated exceptional performance, achieving accuracies exceeding 70% and 60% with experimental and predicted monomer structures, respectively, on the Protein Data Bank (PDB) test set—significantly outperforming traditional DCA-based and protein language model-based methods. Similarly, *AFTGAN*, proposed by Kang, merges Transformer and graph attention network (GAT) frameworks to enhance protein feature extraction and predict multi-type PPIs with superior precision. These findings highlight the transformative role of Transformer models in advancing the accuracy and depth of PPI prediction within computational biology.

4.2. Opportunity 2: Transformer Models Empowering DTI Identification :

Although deep learning models have substantially improved drug–target interaction (DTI) predictions, existing methods frequently provide limited molecular representations by focusing solely on SMILES, SMARTS, or molecular graphs. Transformers address this limitation by learning richer and more context-aware molecular embeddings, either independently or in combination with auxiliary AI methods. The *DeepMGT-DTI* model exemplifies this approach by integrating Transformer networks with multilayer graph representations to effectively capture intricate structural features of drugs. With an AUC of 90.24%, AUPR of 77.11%, F1 score of 79.31%, and accuracy of 85.15% on the DrugBank dataset, DeepMGT-DTI outperformed established baselines such as DeepDTA and TransformerCPI. Another influential model, *GSATDTA*, employs a triple-channel architecture combining graph–sequence attention with Transformer-based learning to achieve outstanding binding affinity predictions. Such developments illustrate the immense promise of Transformer-based architectures in enhancing DTI modeling and enabling more sophisticated drug discovery pipelines.

4.3. Opportunity 3: Transformer Models Empowering De Novo Drug Design :

Existing generative models for *de novo* drug design predominantly focus on database-driven virtual screening or unconditional molecule generation, often neglecting protein target specificity. Transformer models address this gap by incorporating protein target information into the molecular generation process, enabling the design of biologically relevant compounds. *AlphaDrug* represents a pioneering example, employing a modified Transformer combined with Monte Carlo Tree Search (MCTS) and docking-based scoring to generate target-specific molecules. It surpasses conventional models such as LiGANN, SBMolGen, and SBDD-3D in several metrics, including docking score, uniqueness, QED, logP, synthetic accessibility, and NP-likeness. Furthermore, Transformer-derived architectures inspired by GPT—such as *cMolGPT*—have successfully adapted language generation principles to molecular generation. Fine-tuned for specific protein targets, cMolGPT generates compounds whose chemical spaces align closely with those of experimentally validated molecules, establishing new standards for AI-guided drug design.

4.4. Opportunity 4: Transformer Models Empowering Molecular Property Prediction :

A major obstacle in molecular property prediction is the scarcity of labeled data. Transformer-based self-supervised learning models, particularly those inspired by BERT, have emerged as promising solutions to leverage vast unlabeled datasets for effective pre-training. *K-BERT*, a Transformer approach tailored for chemical data, captures SMILES semantics in a manner analogous to human chemists. Demonstrating superior results in 8 of 15 benchmark tasks, K-BERT achieved an average AUC of 0.806, outperforming models such as XGBoost-MACCS, XGBoost-ECFP4, HRGCN+, and AttentiveFP. Similarly, *SMILES-BERT*, proposed by Wang, adopts a two-stage pre-training and fine-tuning paradigm, incorporating both labeled and unlabeled molecular data. It achieved high accuracies on datasets such as LogP (0.9154), PM2 (0.7589), and PCBA-686978 (0.8784), outperforming competitive graph-based and sequence-based frameworks. These innovations underscore the capability of Transformer-based architectures to enhance molecular property prediction and streamline the identification of promising drug candidates.

4.5. Opportunity 5: Transformer Models Empowering Chemical Synthesis :

Earlier sequence-based methods for reaction prediction often relied on recurrent neural networks (RNNs) with single-head attention, treating reactants and reagents separately and requiring explicit atom mapping. This approach limited interpretability and predictive flexibility. In contrast, Transformer-based architectures—particularly the *Molecular Transformer*—have revolutionized chemical synthesis modeling by capturing relational dependencies among reaction components through multi-head attention. This model demonstrated remarkable accuracy in forward reaction prediction and reaction condition estimation. Building upon this foundation, Schwaller proposed an enhanced Molecular Transformer framework integrated with a hypergraph exploration algorithm for automated retrosynthetic pathway prediction. This advanced model extends beyond reactant prediction to identify reagents for each retrosynthetic step, offering superior complexity handling compared to prior machine learning approaches. Collectively, Transformer-based reaction models signify a major advancement toward fully automated, interpretable, and efficient chemical synthesis prediction systems.

5. Challenges of ML-Based Models in Drug Discovery

Machine learning (ML)-based models have demonstrated outstanding capabilities in analyzing and extracting meaningful representations from high-dimensional, complex datasets, driving substantial progress across multiple stages of drug discovery. Despite these advancements, several challenges continue to impede their full potential and practical adoption in the field.

First, the success of ML models is profoundly influenced by the availability and quantity of high-quality training data. In general, larger datasets tend to yield more accurate and reliable models; however, the scarcity of labeled biomedical data often leads to overfitting and limited model generalizability. This data limitation remains one of the most critical bottlenecks in ML-driven drug discovery. To address this, the application of transfer learning has emerged as a promising strategy, enabling models trained on related tasks to adapt effectively to new but similar challenges. In parallel, researchers are shifting toward the use of carefully curated, smaller datasets that emphasize data quality over quantity. This approach supports the extraction of meaningful insights from limited data, improving both the precision and applicability of ML models to complex biological phenomena.

Second, the quality of data plays an equally crucial role in determining the predictive performance of ML models. Data derived from public drug databases often exhibit inconsistencies caused by variations in experimental protocols, measurement techniques, and assay conditions, leading to difficulties in ensuring comparability across datasets. Implementing rigorous data preprocessing techniques—such as noise filtering, outlier detection, and normalization—can significantly enhance data reliability. Statistical methods such as Z-score analysis, box plots, and iterative deletion are effective for identifying and eliminating noisy or erroneous entries. Moreover, employing cross-validation techniques allows researchers to evaluate the generalization ability of models across diverse datasets, ensuring robust predictive performance on unseen data.

Third, the increasing diversity and complexity of ML architectures introduce challenges in model selection and optimization for specific tasks within drug discovery. Selecting the most suitable model requires careful consideration of multiple factors, including problem complexity, dataset characteristics, and computational constraints. Once a model is selected, hyperparameter tuning becomes critical for optimizing its predictive performance. While automated hyperparameter optimization methods can streamline the process, they often remain computationally intensive and partially reliant on human intervention, which can lead to suboptimal parameter configurations. To mitigate this, cross-validation remains an essential practice for objective model evaluation. Establishing clear performance metrics—such as accuracy, precision, recall, F1 score, AUC, and AUPR—further enhances the transparency and rigor of model comparison across different tasks.

Fourth, a major limitation of ML, especially in deep learning (DL) models, lies in their lack of interpretability. Unlike traditional statistical models with transparent decision-making processes, DL models function as complex networks filled with nonlinear transformations, making it challenging to understand how specific outcomes are derived. This "black-box" nature often hinders their acceptance in biomedical research, where mechanistic interpretability is essential. Recent advancements in explainable AI techniques, including Activation Maximization, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP), contribute valuable tools for visualizing and elucidating the inner workings of ML models. These methods enhance transparency by identifying the most influential features contributing to a model's decision, thereby supporting greater trust and insight into underlying biological mechanisms.

In conclusion, while ML-based methodologies have revolutionized drug discovery by accelerating data analysis and hypothesis generation, their widespread application still demands significant improvements in data quality, model interpretability, and methodological standardization. Addressing these challenges through innovative solutions, robust validation frameworks, and interdisciplinary collaboration will be crucial to unlocking the full transformative potential of ML in modern drug discovery.

6. Conclusion

The continuous advancement of drug research and development plays a critical role in addressing the global demand for effective disease treatments, providing safer, more efficient, and patient-friendly therapeutic options. Compared with traditional drug discovery strategies, machine learning (ML)-based approaches offer numerous advantages, including reduced costs and development time, improved prediction accuracy, and enhanced safety assessments. By enabling data-driven insights and predictive modeling, ML bridges the gap between early-stage discovery and clinical drug efficacy, making it increasingly vital to both the pharmaceutical industry and academic research.

Notably, the emergence of large language models such as ChatGPT has intensified research interest in applying Transformer architectures—particularly their self-attention mechanisms—to accelerate and optimize various stages of the drug discovery pipeline. These advancements have created new opportunities for tackling complex challenges in molecule design, protein–ligand interaction prediction, and target identification, establishing a new paradigm in computational drug discovery.

Nonetheless, ML-based models still face persistent challenges that hinder their full integration into real-world drug development. The generation of false positives and false negatives remains a significant issue, leading to erroneous predictions and inefficient allocation of laboratory resources. Therefore, rigorous *in vitro*, *in vivo*, and clinical validation studies remain indispensable to confirm the reliability and practical applicability of ML predictions. Future research should prioritize enhancing the quality and consistency of training data, improving the interpretability of ML algorithms to ensure transparency in decision-making, and promoting the synergistic integration of computational intelligence with expert biological and chemical knowledge. Such interdisciplinary collaboration will be essential to maximize the effectiveness, reliability, and safety of ML-driven drug discovery.

REFERENCES

1. https://www.mdpi.com/1420-3049/29/4/903
2. https://share.google/wzBWJjYR8cjjKwQDU
3. https://www.sciencedirect.com/science/article/pii/S2667102621001066
4. 4..https://share.google/nol6cE7oShPBbOXVR
5. https://share.google/cVY3McTFuD50cMk2B
6. https://share.google/wNgJcuI7A1XcVoXtP
7. https://ijrpr.com/uploads/V6ISSUE4/IJRPR42841.pdf
8. https://ijrpr.com/uploads/V6ISSUE2/IJRPR39121.pdf