Received Date:12th February, 2025
Revision Date: 25th February, 2025
Accepted Date: 11th March,2025

Machine Learning in Modern Drug Discovery: Applications, Advances, and Future Horizons

Mausam Gurung[1*], Sujan Shrestha[2]
[1]Dept of Electronics, Communication and Information Engineering, Kathmandu Engineering College.
Email: mausaam.gurung593@gmail.com
[2]Assoc. Professor, Dept of Electronics, Communication and Information Engineering, Kathmandu Engineering College.
Email: sujan.shrestha@kecktm.edu.np

Abstract -The integration of machine learning (ML) into drug discovery is reshaping pharmaceutical development, addressing high costs exceeding $2 billion per drug, decade-long timelines, and success rates below 10%. This review categorizes ML applications into four domains: (1) Target Identification and Validation, where ML algorithms mine genomic and proteomic data to uncover novel therapeutic targets and validate their relevance; (2) Molecular Design and Screening, utilizing deep learning to design and optimize lead compounds, significantly reducing physical screening; (3) Predictive Toxicology and Pharmacokinetics, leveraging ensemble learning to predict safety profiles and ADMET properties with precision; and (4) Clinical Trial Optimization, employing patient stratification and real-time monitoring to streamline trials. Although ML offers transformative potential, challenges persist, including data quality inconsistencies, model interpretability, and regulatory hurdles. Emerging technologies like quantum computing, with its unparalleled molecular simulation capabilities, and generative AI, which excels in creating novel molecular entities, show promise in overcoming these obstacles. Drawing on recent advances, this review emphasizes the necessity of interdisciplinary collaboration among computational scientists, pharmacologists, and clinicians. Future directions include the adoption of federated learning for secure multi-institutional collaborations, explainable AI for greater transparency, and hybrid classical-quantum algorithms. Together, these innovations position ML as the cornerstone of next-generation drug discovery, accelerating timelines, reducing costs, and improving success rates.

Keywords - Machine Learning, Drug Discovery, Drug Development, Molecular Design, Clinical Trial Optimization, Generative AI, Explainable AI

Introduction

The discovery and development of new drugs is critical to addressing ever-evolving global health challenges, from combating infectious diseases to managing chronic conditions and rare genetic disorders. However, this process is one of the most intricate and resource intensive, typically spanning 10-15 years, requires an estimated investment of around USD 2.5 billion, and is characterized by a failure rate exceeding 90%, with only a fraction of drug candidates successfully reaching the market [1].The main obstacles in drug discovery and development are the mounting cost, risk, and time frame needed to develop new medicines. Fair pricing and accessibility are another unmet global challenge [2]. With such immense costs and hurdles, revolutionizing the drug development process through innovation has become critical.

Artificial intelligence (AI) and machine learning (ML) have emerged as transformative tools to address these challenges, offering the potential to reduce costs, improve efficiency, and increase the success rate of drug discovery [3] [4]. AI, a concept with roots dating back to the 1940's and formally coined in 1956 by John McCarthy, now drives advances in multiple areas of science, from discovering new drugs to analyzing medical images and modeling biological systems.

Transformer models [5] have revolutionized text generation, while AlphaFold [6] breakthrough in protein structure prediction has opened new frontiers in medical research. These advances demonstrate AI's potential to transform drug discovery and development [7].The impact of AI in pharmaceuticals is exemplified by the FDA's record-breaking approvals in 2020—40 new molecular entities (NMEs) and 13 biologics license applications (BLAs), the highest count in two decades [8].

Drug discovery and development employs three main

* *Corresponding Author*

computational approaches. Structure-based methods focus on analyzing the 3D structure of target proteins, enabling rational design of drugs that fit precisely into binding sites. Protein-ligand based approaches examine the interactions between drugs (ligands) and their protein targets, helping predict binding affinity and potential drug effectiveness [9]. System-based methods take a broader view, studying how drugs affect entire biological networks and pathways, providing insights into drug behavior at a cellular or organism level.

Computational [10] [11] Left some explanation of XAi and quantum parts. In this paper we also review the Explainable AI (XAI) which addresses the" black box" nature of traditional AI by making predictions transparent and interpretable, fostering trust among researchers and regulators. It helps elucidate how molecular features influence drug efficacy or toxicity, aiding informed decision-making. Quantum computing complements this by leveraging phenomena like superposition to solve complex problems, such as protein folding, with unparalleled efficiency. Together, XAI and quantum computing promise to revolutionize drug discovery by providing deeper insights and accelerating the development of novel therapies.

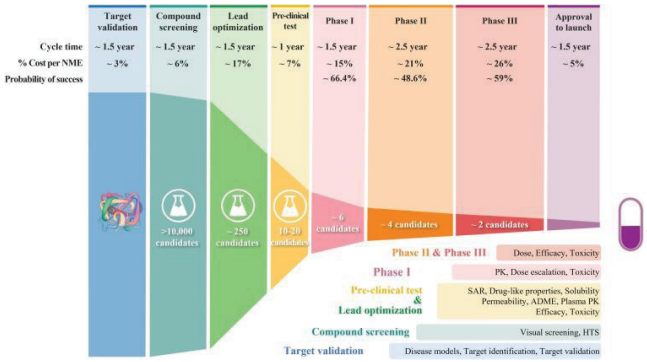

Figure 1: The process of drug discovery and development and the failure rate at each step. [12]

Drug Discovery and Development Process

The drug discovery and development process is a lengthy, costly, and multi-step journey [12] designed to bring safe and effective drugs to patients and following are the processes [13].

- *Target Identification and Validation:* Researchers identify a biological target, such as a protein or gene, associated with a disease. The target's role in the disease process is validated using various techniques such as genetic studies or biochemical assays.
- *Lead Identification and Optimization:* Potential drug molecules (leads) are identified through high-throughput screening and computational methods. These leads are optimized for improved efficacy, safety, and bio- availability while minimizing toxicity.
- *Preclinical Studies:* Optimized leads undergo preclinical testing in cell and animal models to assess safety, pharmacokinetics (how the drug moves through the body), and efficacy.
- *Clinical trials:* Phase I: Tests the drug's safety and dosage in a small group of healthy volunteers. Phase II: Evaluates efficacy and side effects in a larger group of patients. Phase III: Confirms efficacy and monitors adverse reactions in an even larger patient population.
- *Regulatory Approval:* Data from preclinical and clinical studies are submitted to regulatory authorities, like the FDA, for review. If approved, the drug can be marketed.
- *Post-Marketing Surveillance:* After approval, ongoing monitoring ensures long-term safety and effectiveness.

This process, spanning 10–15 years, cost around US $ 2.5 billion, is resource-intensive with failure rate of more than 90 % with only a fraction of candidates reaching the market. [1]

Dataset and Database

The success of deep learning (DL) and machine learning (ML) in drug discovery hinge on diverse, high-quality datasets that enable robust model generalization. However significant challenges persist across the drug development pipeline, from early-stage target identification to clinical trials. This include limited access to comprehensive drug-target interaction data, patient privacy concerns that restrict data access, high cost associated with expert annotations, and data scarcity, particularly for rare conditions [14]. The field faces additional complexity due to the diverse nature of required datasets, spanning chemical structures (PubChem, ChEMBL),biological activity data (BindingDB, DrugBank), ADMET properties (ADMETLab 2.0), and clinical outcomes. Data quality and standardization issues further compound these challenges, particularly when integrating data from multiple sources.

To address these challenges the field has embraced FAIR (Findable, Accessible , Interoperable, and Reusable) data-guiding principles which provide a framework for the management of scientific research data [15] and offer a

systematic approach to tackling the challenges of reusing fast-growing, but frequently inaccessible and inconsistently annotated, research data resources [16]. The implementation of FAIR principles has improved dataset management through standardized metadata annotations, clear indexing, and compatible data formats (like SMILES and SDF), while establishing clear usage rights and quality descriptors [17].

Advanced learning techniques like few-shot and zero-shot learning enable drug discovery models to learn from minimal data by transferring knowledge from similar tasks, while transfer learning leverages pre-trained models to achieve high performance with limited samples [18]. Federated learning enables privacy-preserved collaboration between institutions, complemented by synthetic data generation through advanced generative model [19][20]. The integration of diverse data types, from quantum mechanical properties (QM9 dataset) to real-world evidence, combined with standardized reporting, has created a robust foundation for ML driven drug discovery while maintaining data integrity. List of Dataset for the drug discovery are given below. Table 1.

Table 1

Dataset for drug discovery

Dataset Type	Description	Example
Chemical Structure and Property Datasets	Contain information about chemical structures, molecular properties, and descriptors.	PubChem, ChEMBL, ZINC Database, ChemSpider
Biological Activity Datasets	Include bioassay results, binding affinities, IC50 values, EC50 values, and more.	BindingDB, PubChem, IUPHAR
Pharmacokinetic and ADMET Datasets	Focus on absorption, distribution, metabolism,-excretion, and toxicity (ADMET) data to evaluate drug likeness.	ADMETlab2.0, pkCSM, eTOX project
Genomic and Proteomic Datasets	Contain information on gene expression, protein sequences, and their interactions relevant to drug targets.	UniProt, GeneBank, TCGA
Clinical Trial Datasets	Provide clinical trial results, patient responses, and outcomes.	ClinicalTrials.gov, Vivli CLinical Trials Data sharing, FDA's Clinical Trial Data
Disease-Specific Datasets	Datasets focusing on specific diseases, including biomarkers, pathways, and epidemiological data.	DisGeNET, OMIM, TCGA
Drug-Target Interaction Datasets	Contain data on the interaction between small molecules and their biological targets.	DrugBank, STITCH, PharmGKB
Structural and Crystallographic Datasets	Include structural data of proteins and protein-ligand complexes.	Protein Data Bank (PDB), Binding-MOAD, Cambridge Structural Database (CSD)
Natural Product Datasets	Contain information on natural compounds and their pharmacological properties.	NPASS, Supernatural, TCMSP
Omics Datasets	Include large scale data for genomics, transcriptomics, proteomics, and metabolomics studies.	GEO, PRIDE, Metabo Lights
Toxicity Datasets	Provide data on drug safety, toxic doses, and adverse effects.	Tox21, ToxCast, OpenTox
Real-World Evidence (RWE) and Epidemiological Datasets	Include patient data, prescription records, and healthcare outcomes.	Truven Health Analytics, IQVIA, MIMIC
Image Datasets	Collaborative research projects to investigate the use of AI to make MRI scans faster.	fastMRI dataset, CT-PET dataset
Machine Learning Benchmark Datasets	Specifically designed for training and benchmarking drug discovery models.	MIMIC, DeepChem, QM9 Dataset

Methods

We use the keywords" Machine Learning", "Drug Discovery", "Drug Development", "deep learning", "drug design", "drug repurposing", "computational drug design" . We filter it on IEEE, PubMed, arxiv, biorxiv, Nature, Eelsvier and other publications based on web search. First section explains the background and section II is about drug discovery development process, section III about the different datasets that are present for the drug discovery and development, section V is about the role of machine learning and section VI is about the challenges and future direction of drug discovery and development and the last section of conclusion.

Role of Machine Learning on different stages

5. 1 Machine Learning process based on data

Various learning paradigms offer distinct advantages in addressing the complex challenges of drug development, from molecular design to clinical trials. To address from data scarcity to privacy preserving, the following are the machine learning approaches.

- *Supervised learning* serves as the foundation for many

drug discovery applications, particularly in structure-activity relationship (SAR) modeling and binding affinity prediction. This approach relies on labeled datasets of known drug-target interactions and their outcomes to train models that can predict properties of novel compounds.

- *Semi-supervised learning* has emerged as a crucial paradigm in pharmaceutical research, addressing the common challenge of limited labeled data. This approach proves particularly valuable in scenarios where obtaining labeled data requires expensive wet-lab experiments. By leveraging a small set of labeled compounds alongside larger unlabeled datasets, semi-supervised learning enables more robust model training for tasks such as toxicity prediction and drug-target interaction analysis.
- *Unsupervised learning* plays a vital role in discovering hidden patterns within chemical spaces and biological data. This approach excels in analyzing large-scale molecular databases to identify natural clustering of compounds, potentially revealing new drug classes or unexpected therapeutic applications. Unsupervised learning comes into use when the data are not structured.
- The emergence of *self-supervised learning* has provided new avenues for understanding molecular structures and properties without explicit labels. This paradigm has proven especially valuable in pre-training models on vast chemical databases before fine-tuning them for specific drug discovery tasks.
- *Deep learning* has transformed the landscape of drug discovery, as highlighted by LeCun et al. (2015) in their seminal work. Their paper emphasizes how deep neural networks can learn hierarchical representations from complex chemical and biological data, enabling unprecedented accuracy in predicting drug properties and interactions. Deep learning architectures, particularly graph neural networks and attention mechanisms, have demonstrated remarkable success in modeling molecular structures and predicting drug-protein interactions.[21]
- *Transfer learning* has become increasingly important in drug discovery, allowing knowledge gained from one chemical space or therapeutic area to be applied to others. This approach is particularly valuable when dealing with rare diseases or novel drug targets where data may be limited. As noted by Cai et al., transfer learning enables more efficient model development by leveraging pre-trained models on larger chemical datasets.[22]
- *Zero-shot learning* represents an innovative approach in drug discovery where models can make predictions about entirely new classes of compounds or therapeutic targets without any prior training examples. This capability is particularly valuable when exploring novel chemical spaces or targeting newly discovered disease pathways where traditional training data may not exist. For instance, zero-shot learning enables models to predict potential drug-target interactions for emerging viral variants or previously unexplored protein families.
- *Few-shot learning* builds upon this concept by utilizing only a small number of examples to make accurate predictions about new chemical entities or biological targets. This paradigm has proven especially valuable in rare disease drug discovery, where limited data is available for model training. Few-shot learning enables rapid adaptation to new therapeutic targets with minimal experimental data, significantly accelerating the early stages of drug development.
- *Federated learning* represents a promising direction for pharmaceutical research, enabling collaborative model training while maintaining data privacy. This approach allows multiple research institutions or pharmaceutical companies to contribute to model development without sharing sensitive molecular or clinical data.
- *Active learning* strategies help optimize the expensive process of drug screening by intelligently selecting which compounds to test experimentally. This approach reduces the number of required experiments by identifying the most informative molecules for testing, thereby accelerating the drug discovery pipeline while minimizing costs.
- *Reinforcement learning* has found applications in de novo drug design, where models learn to generate novel molecular structures that optimize desired properties through an iterative process of exploration and exploitation.

5.2 Machine Learning Models

- *Random Forest:* An ensemble learning method that builds multiple decision trees and merges their outputs for more accurate and stable predictions.[23].

- *Support Vector Machine (SVM):* A supervised learning algorithm that finds the optimal hyperplane to classify data into different categories. [Cortes and Vapnik, 1995][24]
- *XGBoost:* A gradient-boosted decision tree algorithm known for its efficiency and performance in handling structured data. [Chen and Guestrin, 2016][25].
- *K-Nearest Neighbors (KNN):*A simple instance-based algorithm that classifies data points based on the majority class among its nearest neighbors. [Cover and Hart, 1967][26]
- *Neural Networks:*

Deep Neural Networks (DNN): A type of neural network with multiple layers between input and output, capable of learning hierarchical patterns. [McCulloch and Pitts, 1943] [27].

Convolutional Neural Networks (CNN): Designed for image data, CNNs use convolutional layers to capture spatial hierarchies. [LeCun et al., 1998][28].

Recurrent Neural Networks (RNN): Suitable for sequential data, RNNs use loops to retain information over time. [Rumelhart et al., 1986][29].

Graph Neural Networks (GNN): Specialized in processing graph-structured data, GNNs capture relationships and interactions between entities. [Scarselli et al., 2009][30]. GNN review of 2017-2023 suggest review the application on drug target interaction, drug drug interaction, drug repurposing, drug representation ADMET prediction while Graph Convolution NN and its related optimization are currently the core algorithm in this fields.[31]

- *Transformer:* A model architecture leveraging self-attention mechanisms for sequence-to-sequence tasks, such as language modeling. [Vaswani et al., 2017] [5]. MVI ViT transformer model on early detection of Alzheimer disease found the SOTA model that has higher accuracy of 97.65, precision 96.98, recall 96.40, F1-score 96.69 and a computational overhead 00:12:45. However, the model requires more computation overhead, and the future direction is headed toward pruning the algorithm for simplification. On Albation test the result go down when MVI Vit was removed although other block removing loose less accuracy suggest that the transformer has the greater possibility on detection. [32]
- *Generative Models:*

Auto encoder: A neural network for unsupervised learning that encodes data into a latent representation and reconstructs it. [Hinton and Salakhutdinov, 2006] [33]

Generative Adversarial Networks (GANs): A generative model comprising a generator and a discriminator that compete to create realistic data. [Goodfellow et al., 2014] [34]

Diffusion Models: A class of generative models that iteratively denoise data samples to generate outputs. [Sohl-Dickstein et al., 2015][35].

5.3 Application of ML in drug discovery

The process of drug discovery represents one of the most challenging and time-intensive aspects of pharmaceutical development. However, machine learning (ML) and deep learning (DL) have emerged as powerful tools to address the inefficiencies and uncertainties of traditional methods [39]. As highlighted by Silva et al. [40], these technologies are transforming the industry through innovative models, tools, and databases that enhance various phases of drug development, from initial design to final synthesis which is present on table [2].

- *Target Identification:* **T**arget identification represents a crucial initial step in which ML algorithms excel in identifying key proteins, genes, and pathways involved in disease processes. Through the analysis of complex biological datasets and networks, these computational methods can predict potential drug targets and their interactions, significantly accelerating the early stages of drug development.
- *Virtual screening:* It's a computational technique that advanced quantitative structure activity relationship modeling and molecular docking simulations, leading to more accurate predictions of binding affinities and compound optimization [41].
- *Drug Design:* Powered by generative models including Vibrational auto encoders (VAE) and generative adversarial networks (GANs), enable the creation and optimization of novel molecular structure with desired properties while maintaining synthetic feasibility [42].
- *Drug Repurposing:* Network based ML models have benefited the efficient identification of new therapeutic uses for existing drugs. Through the analysis of biomedical literature and databases, technologies like

graph neural networks have become instrumental in discovering novel drug-target interactions and repositioning drugs for rare diseases [43].

- *Chemical synthesis:* ML/DL approaches have enhanced predictability and efficiency by forecasting reaction outcomes and optimizing conditions. These technologies analyze reaction databases to recommend efficient synthetic routes, reducing experimental trial-and-error and ensuring scalability in drug production, as emphasized by Silva et al. [40].

5.4 Explainable AI

Explainable AI (XAI) has emerged as a crucial component in modern artificial intelligence applications, especially in healthcare and pharmaceutical research. According to recent research, the implementation of explainable AI frameworks has become essential to ensure both scientific rigor and practical utility in medical applications [44]. The Five-Framework Approach to Explainable AI.

*a) Transparency***:** Transparency serves as the foundation of explainable AI in medical applications. As highlighted by Sendak et al. [45], transparency encompasses not just the technical aspects of AI models but also the comprehensive disclosure of methodologies, limitations, and potential biases. This framework ensures that both the scientific community and healthcare practitioners can understand and validate the AI decision-making process.

b) Reproducibility: The reproducibility framework, as emphasized by Nature's landmark study Benjamin et al.[46], focuses on the crucial ability to replicate AI models and their results across different contexts. This aspect is particularly vital in drug discovery, where consistent results across various research settings can significantly impact the development of new therapeutic approaches.

c) Effectiveness: The effectiveness framework examines the practical application and success rates of AI models in real-world scenarios. According to Cai et al. [47], effectiveness in medical AI applications is measured not just by raw performance metrics but by the quality of data utilized and the model's ability to generate actionable insights in clinical settings.

*d) Ethic***s**: The ethical framework, as outlined by Char et al. [48], addresses the crucial aspects of AI implementation in healthcare, including fairness, patient autonomy, and privacy considerations. This framework ensures that AI applications in medicine maintain high ethical standards while delivering valuable insights for drug discovery and development.

e) Engagement: The engagement framework, described by Richards et al. [49], emphasizes the importance of stakeholder involvement in AI development and implementation. This ranges from basic consultation to full partnership with healthcare providers and patients, ensuring that AI solutions meet real-world needs and expectations.

The integration of machine learning frameworks in drug discovery has revolutionized the field through MLops practices. As demonstrated by Wang et al. [50], combining this framework creates AI systems that are both powerful and trustworthy. This unified approach, highlighted by Chen et al. [51], ensures AI models meet dual objectives: achieving technical excellence while satisfying the rigorous demands of medical research and regulatory requirements. The result is a more efficient drug discovery pipeline supported by AI models that are not only high-performing but also transparent and interpretable enough to gain regulatory approval and clinical acceptance.

Table 2
Summary of tools used on drug discovery and development

Name	ML/ DL	Description
Prediction of the target protein structure		
TrRosetta Server	DNN	Predict 3D structures of proteins
AlphaFold	DNN	Predict 3D structures of proteins
ComplexQA	GNN	Predict protein complex structure
ProteinBERT	Transformer	Predict secondary structure
ESMfold	Transformer	Predict structure and function of proteins
TAPE	Transformer	Predict structure and function of proteins
ProtTrans	Transformer	Predict structure and function of proteins

Name	ML/ DL	Description
Predicting protein–protein interactions		
IntPred	RF	Predict PPI interface sites
eFindSite	SVM; NBC	Predict PPI interfaces
DELPHI	RNN; CNN	Predict PPI sites
PPISP-XGBoost	XGBoost	Predict PPI sites

HN-PPISP	CNN	Predict PPI sites
TAGPPI	GCN	Predict PPIs
Struct2Graph	GAT	Predict PPIs
DeepFE-PPI	DNN	Predict PPIs
SGPPI	GCN	Predict PPIs
DeepPPI	DNN	Predict PPIs
DL-PPI	GNN	Predict PPIs
DeepSG2PPI	CNN	Predict PPIs
MaTPIP	Transformer; CNN	Predict PPIs with explainable AI
ProtInteract	Autoencoder; CNN	Predict PPIs
MultiPPIMI	BAN (bilinear attention network)	Predict PPIs

Name	ML/ DL	Description
Predicting drug–target interactions		
MocFormer	Transformer	Predict DTI binding affinity
MFFDTA	Multimodal: GAT, CNN, GCNLSTM	Predict DTI affinity
DeepC-SeqSite	CNN	Predict DTI binding sites
DeepSurf	CNN; ResNet	Predict DTI binding sites
PrankWeb	RF	Predict DTI binding sites
PUResNet	ResNet	Predict DTI binding sites
AGAT-PPIS	GNN	Predict DTI binding sites
DeepDTA	CNN	Predict DTI binding affinity
SimBoost	GBM	Predict DTI binding affinity
DEELIG	CNN	Predict DTI binding affinity
DeepDTAF	CNN	Predict DTI binding affinity
GraphDelta	CNN, MPNN	Predict DTI binding affinity
PotentialNet	CNN	Predict DTI binding affinity
DeepAffinity	RNN, CNN	Predict DTI binding affinity

Name	ML/ DL	Description
De novo drug design		
ReLeaSE	RNN; RL	Conduct de novo drug design
ChemVAE	CNN; GRU, Autoencoder	Conduct de novo drug design
MolRNN	RNN, GGM (Graph generative model)	Conduct multi-objective de novo drug design
PaccMann (RL)	VAE	Generate compounds with anti-cancer drug properties
druGAN	AAE	Conduct de novo drug design
SCScore	CNN	Evaluate the molecular accessibility
UnCorrupt SMILES	RNN, VAE, GAN	Conduct de novo drug design
PETrans	Transfer learning	Conduct de novo drug design
FSM-DDTR	Transformer	Conduct de novo drug design
DNMG	GAN	Conduct de novo drug design
MedGAN	GAN	Design novel molecule

Name	ML/ DL	Description
Prediction of the ADME/T properties		
ADMETboost	XGBoost	Predict ADME/T properties
vNN	k-NN	Predict ADME/T properties
Interpretable-ADMET	CNN; GAT	Predict ADME/T properties
XGraphBoost	GNN	Predict ADME/T properties
DeepTox	DNN	Predict toxicity of compounds
LightBBB	LightGBM	Predict blood–brain barrier
Deep-B3	CNN	Predict blood–brain barrier
PredPS	GNN	Predict stability of compounds in human plasma

Name	ML/ DL	Description
Application of AI in drug repurposing		
deepDTnet	Autoencoder	Predict new targets of known drugs
NeoDTI	GCN	Predict new targets of known drugs
MBiRW	Birandom walk algorithm	Predict new indications of known drugs
GDRnet	GNN	Predict new indications of known drugs
deepDR	VAE	Predict new indications of known drugs
GIPAE	VAE	Predict new indications of known drugs
DrugRep-HeSiaGraph	Heterogeneous siamese neural	Predict new indications of known drugs
iEdgeDTA	GCNN	Predict DTI binding affinity
DeepPurpose	DNN, CNN	Predict DTI binding affinity
AI-DrugNet	DNN	Drug-target pair (DTP) network

Name	ML/ DL	Description
Others		
PyTrial	RNN, GNN, Transformer	A series of clinical trial task solutions and resources
Chemprop	MPNN	Molecular property prediction
DiffBoost	Diffusion Probabilistic Model	Data augmentation
Eghbali-Zarch et al,	Markov decision process	Predicting the adverse effect of drugs and medication plan.

Note: DNN, deep neural network; RNN, recurrent neural network; RF, random forest; CNN, convolutional neural network; XGBoost, eXtreme gradient boosting; GCN, graph convolutional network; GAT, graph attention network; SVM, support vector machine; NBC, naive Bayes classifier; ResNet, residual network; GBM, gradient boosting machines; RL, reinforcement learning; GRU, gated recurrent unit; VAE, variational autoencoder; AAE, adaptive adversarial autoencoder; GNN, graph neural networks; k-NN, k-nearest neighbor; LightGBM, light gradient boosting machine; NTN, neural tensor network; GAN, generative adversarial network; GCNN, graph convolutional neural network.

5.5 Quantum Computing

Quantum computing represents a transformative approach to drug discovery by leveraging quantum mechanical principles like superposition and entanglement through qubits, offering exponentially faster computational capabilities than traditional systems. This technology enables researchers to simulate complex molecular interactions, predict protein folding mechanisms, and analyze drug-target interactions with unprecedented accuracy and speed. As demonstrated by Fujitsu in collaboration with Toray Industries Inc., quantum-inspired computing methods, specifically combining FEP (Free Energy Perturbation) and QM/MM with Digital Annealed technology, have successfully reduced early-stage drug development time from 15 months to just seven weeks in COVID-19 [52] and dengue fever research, enabling rapid evaluation of billions of molecular compounds for toxicity, synthesizability, and biological activity.

The most significant promises of quantum computing in drug discovery lie in several key areas: molecular simulation and modeling for more accurate drug-target interaction predictions, enhanced drug optimization through quantum algorithms like the Variational Quantum Eigensolver (VQE), improved machine learning capabilities via Quantum Support Vector Machines (QSVM) for better drug property prediction, and secure data management through Quantum Key Distribution (QKD)[53]. While practical implementation challenges remain, including the need for more powerful quantum computers and refined algorithms, the technology's potential to dramatically accelerate drug development while reducing costs makes it a crucial frontier in pharmaceutical research.

Challenges and Future Direction

Despite the transformative potential of machine learning in drug discovery, several significant obstacles impede its full implementation. The complexity of ML models, particularly deep learning architectures, creates a critical challenge in interpretability, making it difficult for researchers to validate and trust model predictions. Access to high-quality datasets remains limited, while computational resource demands pose substantial barriers to widespread adoption. Additionally, concerns about AI-generated predictions' safety and accountability create hesitation among stakeholders. The integration of quantum computing, while promising, introduces new complexities in algorithm development and infrastructure requirements [7].

The future of ML-driven drug discovery holds remarkable promise through several emerging architectures and approaches. Transformer models, with their sophisticated attention mechanisms, demonstrate exceptional potential in understanding molecular structures and predicting drug-target interactions [40]. Their ability to process long-range dependencies makes them particularly valuable for analyzing complex biological sequences and protein structures. Similarly, GNNs excel at capturing molecular geometry and chemical interactions, offering powerful tools for predicting molecular properties and drug-target binding [31].

The convergence of these advanced architectures with quantum computing represents a transformative opportunity, potentially revolutionizing molecular modeling and drug design capabilities beyond current classical computing limitations [54]. This integration could dramatically accelerate discovery timelines and enhance pre- diction accuracy, particularly in simulating quantum mechanical properties of molecules.

Strategic initiatives are advancing to address current limitations. Development of explainable AI [44] methodologies aims to demystify model predictions, while cross-sector collaborations strengthen the foundation for ethical and effective ML implementation [55]. Hybrid

approaches combining classical ML with quantum computing represent a frontier in computational drug discovery. Furthermore, the integration of multi-modal data sources, including genomic and clinical information, promises to enhance model robustness and reliability.

The establishment of comprehensive regulatory frameworks will play a crucial role in shaping the field's evolution. These guidelines must balance innovation with ethical considerations, ensuring responsible AI implementation while maintaining scientific progress. By embracing emerging technologies and maintaining rigorous standards, the industry can maximize ML potential while mitigating associated risks.

Conclusion

The integration of machine learning (ML) into drug discovery represents a transformative approach to address long-standing challenges in pharmaceutical research. This review demonstrates that ML technologies offer unprecedented potential to revolutionize the drug development process by addressing critical inefficiencies, including historically high costs, long timelines, and low success rates. Using advanced computational approaches such as deep learning, graph neural networks, and generative models, researchers can now accelerate target identification, virtual screening, drug design, and repurposing with remarkable efficiency.

The emergence of explainable AI (XAI) frameworks further enhances the credibility of these computational methods by providing transparency and interpretability, which are crucial to gain regulatory acceptance and scientific trust. Moreover, the convergence of ML with emerging technologies like quantum computing presents an extraordinary frontier, promising to dramatically reduce drug development timelines and enhance predictive capabilities. However, significant challenges remain, including data quality limitations, model interpretability concerns, and computational resource constraints. The future of drug discovery lies in continued interdisciplinary collaboration, strategic technological integration, and the development of robust regulatory frameworks that balance innovation with rigorous scientific and ethical standards. As machine learning continues to evolve, it holds the potential to transform pharmaceutical research, ultimately accelerating the development of more effective, personalized therapeutic solutions and addressing complex global health challenges.

References

[1] Olivier Wouters, Martin Mckee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*, 323:844, 03 2020.

[2] Koippallil Gopalakrishnan Aghila Rani, Mohamad A. Hamad, Dana M. Zaher, Scott McN Sieburth, Navid Madani, and Taleb H. Al-Tel. Drug development post covid-19 pandemic: toward a better system to meet current and future global health challenges. *Expert Opinion on Drug Discovery*, 16(4):365–371, 2021. PMID: 33356641.

[3] Veer Patel and Manan Shah. Artificial intelligence and machine learning in drug discovery and development. *Intelligent Medicine*, 2(3):134–140, 2022.

[4] Catrin Hasselgren and Tudor I. Oprea. Artificial intelligence for drug discovery: Are we there yet? *Annual Review of Pharmacology and Toxicology*, 64(Volume 64, 2024):527–550, 2024.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin ˚A½A¨dek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon.

[7] A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[8] Alexandre Blanco-Gonz´alez, Alfonso Cabez´on, Alejandro Seco-Gonz´alez, Daniel Conde-Torres, Paula Antelo-Riveiro, A´ngel Pin˜eiro, and Rebeca Garcia-Fandino. The role of ai in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6):891, June 2023.

[9] Usman Shareef, Aisha Altaf, Maghfoor Ahmed, Nosheen Akhtar, Mohammed S. Almuhayawi, Soad K. Al Jaouni, Samy Selim, Mohamed A. Abdelgawad, and Mohammed K. Nagshabandi. A comprehensive review of discovery and development of drugs discovered from 2020–2022. *Saudi Pharmaceutical Journal*, 2023.

[10] Divya Vemula, Perka Jayasurya, Varthiya Sushmitha, Yethirajula Naveen Kumar, Vasundhra Bhandari, Divya Vemula, Perka Jayasurya, Varthiya Sushmitha, Yethirajula Naveen Kumar, and Vasundhra Bhan- dari. Cadd, ai and ml in drug discovery: A comprehensive review. *European Journal of Pharmaceutical Sciences*, 2022.

[11] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W. Lowe. Computational methods in drug discovery. *Pharmacological Reviews*, 66(1):334–395, 2014.

[12] Nalini Schaduangrat, Samuel Lampa, Saw Simeon, Matthew Paul Gleeson, Ola Spjuth, and Chanin Nan- tasenamat. Towards reproducible computational drug discovery. *Journal of Cheminformatics*, 2020.

[13] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90 *Acta Pharmaceutica Sinica B*, 12(7):3049– 3062, 2022.

[14] Amol B Deore, Jayprabha R Dhumane, Rushikesh Wagh, and Rushikesh Sonawane. The stages of drug dis- covery and development process. *Asian Journal of Pharmaceutical Research and Development*, 7(6):62–67, Dec. 2019.

[15] Amit Gangwal, Azim Ansari, Iqrar Ahmad, Abul Kalam Azad, and Wan Mohd Azizi Wan Sulaiman. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biology and Medicine*, 179:108734, 2024.

[16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip E. Bourne, Jil- dau Bouwman, Anthony J. Brookes, Tim Clark, Merc`e Crosas, Ingrid Dillo, Olivier Dumon, Scott C. Edmunds, Chris T. A. Evelo, Richard Finkers, Alejandra N. Gonz´alez-Beltr´an, Alasdair J. G. Gray, Paul Groth, Carole A. Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob W. W. Hooft, Tobias Kuhn, Ruben G. Kok, Joost N. Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel Laerte Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene C. van Schaik, Susanna-Assunta Sansone, Erik Anthony Schultes, Thierry Sengstag, Ted Slater, George O. Strawn, Morris A. Swertz, Mark Thomp- son, Johan van der Lei, Erik M. van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katy Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

[17] Ebtisam Alharbi, Yojana Gadiya, David Henderson, Andrea Zaliani, Alejandra Delfin-Rossaro, Anne Cambon-Thomsen, Manfred Kohler, Gesa Witt, Danielle Welter, Nick Juty, Caroline Jay, Ola Engkvist, Carole Goble, Dorothy S. Reilly, Venkata Satagopam, Vassilios Ioannidis, Wei Gu, and Philip Gribbon. Selection of data sets for fairification in drug discovery and development: Which, why, and how? *Drug Discovery Today*, 27(8):2080–2085, 2022.

[18] Jens M. Kelm, Marc Ferrer, Martin-Immanuel Bittner, and Madhu Lal-Nag. Data standards in drug discovery: A long way to go. *Drug Discovery Today*, 29(2):103879, 2024.

[19] Ana M. Barrag´an-Montero, U. Javaid, G. Valdes, D. Nguyen, P. Desbordes, B. Macq, S. Willems, Liesbeth Vandewinckele, M. Holmstr¨om, F. L¨ofman, S. Michiels, K. Souris, E. Sterpin, and J. Lee. Artificial intelli- gence and machine learning for medical imaging: A technology review. *Physica medica (Testo stampato)*, 2021.

[20] Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 2024.

[21] Dac Thai Nguyen, Trung Thanh Nguyen, Huu Tien Nguyen, Thanh Trung Nguyen, Huy Hieu Pham, Thanh Hung Nguyen, Thao Nguyen Truong, and Phi Le Nguyen. Ct to pet translation: A large-scale dataset and domain-knowledge-guided diffusion approach, 2024.

[22] Yu-Xuan Tang. Deep learning in drug discovery: applications and limitations. *Frontiers in Computing and Intelligent Systems*, 2023.

[23] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63(16):8683–8694, 2020. PMID: 32672961.

[24] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[26] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[27] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13:21–27, 1967.

[28] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 52(1):99–115, 1990.

[29] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back- propagating errors. *Nature*, 323:533–536, 1986.

[31] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[32] Rufan Yao, Zhenhua Shen, Xinyi Xu, Guixia Ling, Rongwu Xiang, Tingyan Song, Fei Zhai, and Yuxuan Zhai. Knowledge mapping of graph neural networks for drug discovery: a bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15, 2024.

[33] Junde Chen, Yun Wang, Adnan Zeb, M.D. Suzauddola, and Yuxin Wen. Multimodal mixing convolutional neural network and transformer for alzheimer's disease recognition. *Expert Systems with Applications*, 259:125321, 2025.

[34] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[35] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[36] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[37] R. Bellman. *Dynamic Programming*. Dover Books on Computer Science. Dover Publications, 2013.

[38] Maryam Eghbali-Zarch, Reza Tavakkoli-Moghaddam, Fatemeh Esfahanian, Amir Azaron, and Moham- mad Mehdi Sepehri. A markov decision process for modeling adverse drug reactions in medication treatment of type 2 diabetes. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 233:793 – 811, 2019.

[39] Andrew J. Schaefer, Matthew D. Bailey, Steven M. Shechter, and Mark S. Roberts. Modeling medical treatment using markov decision processes. In Margaret L. Brandeau, Fran¸cois Sainfort, and William P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 593–612. Springer US, Boston, MA, 2004.

[40] Yuyuan Wu, Lijing Ma, Xinyi Li, Jingpeng Yang, Xinyu Rao, Yiru Hu, Jingyi Xi, Lin Tao, Jianjun Wang, Lailing Du, Gongxing Chen, and Shuiping Liu. The role of artificial intelligence in drug screening, drug design, and clinical trials. *Frontiers in Pharmacology*, 15:1459954, November 2024.

[41] Xin Qi, Yuanchun Zhao, Zhuang Qi, Siyu Hou, and Jiajia Chen. Machine learning empowering drug discovery: Applications, opportunities and challenges. *Molecules*, 2024.

[42] Neeraj Kumar and Vishal Acharya. Advances in machine intelligence-driven virtual screening approaches for big-data. *Medicinal Research Reviews*, 44:939 – 974, 2023.

[43] Irini Doytchinova. Drug design—past, present, future. *Molecules*, 27(5), 2022.

[44] Xiaoqin Pan, Xuan Lin, Dongsheng Cao, Xiangxiang Zeng, Philip S. Yu, Lifang He, Ruth Nussinov, and Feixiong Cheng. Deep learning for drug repurposing: methods, databases, and applications, 2022.

[45] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system. *Sensors*, 22(20), 2022.

[46] Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3):491–497, 11 2019.

[47] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Thakkar Shrad- dha, Rebecca Kusko, Susanna-Assunta Sansone, Weida Tong, Russ D. Wolfinger, Christopher E. Mason, Wendell Jones, Joaquin Dopazo, Cesare Furlanello, Levi Waldron, Bo Wang, Chris McIntosh, Anna Golden- berg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J. W. L. Aerts. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, October 2020.

[48] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410–414, May 2019.

[49] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar Ossorio, Sonoo Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25, 08 2019.

[50] Sebastian Vollmer, Bilal A. Mateen, Gergo Bohner, Franz J Kir´aly, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S. L. McAllister, Puja Myles, David Granger, Mark Birse, Richard Branson, Karel GM Moons, Gary S Collins, John P. A. Ioannidis, Chris Holmes, and Harry Hemingway. Machine learning and ai research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness, 2018.

[51] Faiza Khan Khattak, Vallijah Subasri, Amrit Krishnan, Elham Dolatabadi, Deval Pandya, Laleh Seyyed- Kalantari, and Frank Rudzicz. Mlhops: Machine learning for healthcare operations, 2023.

[52] Norah L Crossnohere, Mohamed Elsaid, Jonathan Paskett, Seuli Bose-Brill, and John F P Bridges. Guide- lines for artificial intelligence in medicine: Literature review and content analysis of frameworks. *J Med Internet Res*, 24(8):e36823, Aug 2022.

[53] Fujitsu. Disrupting and accelerating drug discovery for faster and more accurate lead identification. Tech- nical report, Fujitsu, 2020. Accessed: 2025-01-18.

[54] Sai Krishna Kandula, Nagaveni Katam, Pranav Reddy Kangari, Adithya Hijmal, Rakesh Gurrala, and Mohammed Mahmoud. Quantum computing potentials for drug discovery. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1467–1473, 2023.

[55] Raffaele Santagati, Alan Aspuru-Guzik, Ryan Babbush, Matthias Degroote, Leticia Gonz´alez, Elica Kyo- seva, Nikolaj Moll, Markus Oppel, Robert M. Parrish, Nicholas C. Rubin, Michael Streif, Christofer S. Tautermann, Horst Weiss, Nathan Wiebe, and Clemens Utschig-Utschig. Drug design on quantum com- puters. *Nature Physics*, 20(4):549–557, March 2024.

[56] Andrew M. Davis, Ola Engkvist, Rebecca J. Fairclough, Isabella Feierberg, Adrian Freeman, and Preeti Iyer. Public-private partnerships: Compound and data sharing in drug discovery and development. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 26(5):604–619, 2021. PMID: 33586501.