The Chitransh Academic & Research
International Journal of Multidisaciplinary Reseaech
An Online Peer Reviewed / Refereed Journal
Volume 1| Issue 04 |September 2025
ISSN: 3049-303X (Online)
Website: www.thechitranshacadmic.in

AI IN DRUG DISCOVERY: A Comprehensive Review

[1]Vinay Makhija,[2]Bhuvi Saini,[3]Dr. Sunil Kumar Shrivastav,[4]Dr. Kashish Parwani

ARTICLEDETAILS

Research Paper

Received: **30/08/2025**

Accepted: **10/09/2025**

Published: **30/09/2025**

Keywords:Data Resources, Molecular Representation Techniques, AI in pipeline, Virtual screening and hit

ABSTRACT

Traditional methods for finding new drugs include high-throughput screening, experimental tests, and intricate synthesis procedures—all of which are expensive, time-consuming, and prone to high attrition rates. By expediting several phases of the drug development pipeline, increasing prediction accuracy, and simplifying decision-making, the use of artificial intelligence (AI) has started to change this paradigm in recent years. Making use of large-scale multi-omics datasets, improvements in computational technology, and advanced algorithms like reinforcement learning, deep learning, and graph neural networks. Highlighting important data sources, molecular representation techniques, and computational frameworks, this paper summarizes current advancements in AI-assisted pharmaceutical research. With an emphasis on striking a balance between clinical translation success, interpretability, and predictive accuracy, we also look at recent uses, new developments, and enduring difficulties. AI-driven software platforms like AlphaFold2 and DeepChem, along with publicly available chemical and biological databases like ChEMBL, DrugBank, and the Protein Data Bank, are increasing the breadth and effectiveness of drug discovery initiatives. AI has proven useful in de novo molecular design, virtual screening, target identification, and the prediction of pharmacokinetic and toxicological profiles

INTRODUCTION-

New medicines development is a multi-phase, intricate process that usually takes more than ten years and requires investments of more than two billion US dollars. Despite this, the process frequently fails, especially in late-stage clinical trials. From target discovery to clinical validation, traditional approaches are frequently hindered by inefficiencies, low throughput, and rising expenses. In order to get beyond these restrictions, there is increasing interest in incorporating sophisticated computational tools.

Thanks to parallel developments in cloud computing, high-performance GPUs and TPUs, and the creation of massive, superior biomedical datasets, artificial intelligence has become a game-changer in drug discovery. From a variety of data modalities, like as genomic sequences, chemical structures, imaging data, and medical records, machine learning and deep learning algorithms are able to identify intricate, non-linear patterns. Notably, generative models make it easier to create new compounds with ideal features, while graph-based models allow for more complex description of molecular interactions.

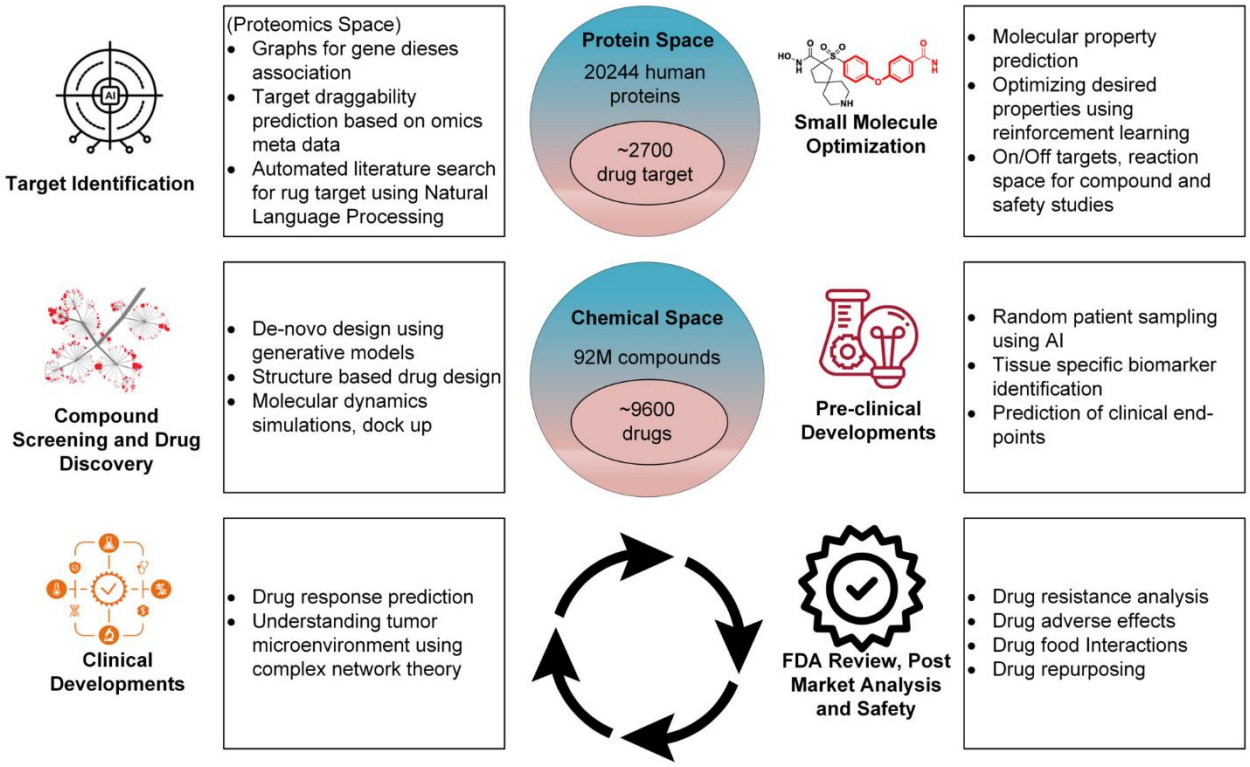
In order to take advantage of these capabilities, pharmaceutical corporations and AI-focused biotech businesses are working together more and more, which is producing intriguing clinical candidates and faster timelines. AI-assisted drug target identification, drug–target interaction prediction, molecular dynamics simulation, and logical lead compound optimization are recent achievements. Additionally, the design space for drug development has been greatly increased by the availability of open-access repositories like PubChem and ChEMBL as well as advances in protein structure prediction, such as AlphaFold2.

The quality and diversity of datasets, the interpretability of AI models, and the conversion of in silico predictions into in vivo efficacy are the main obstacles that still exist despite recent developments. The goal of this review is to present a thorough analysis of AI's function throughout the drug development spectrum, highlighting the technology's present advantages, disadvantages, and potential.

2. AI Techniques, Molecular Representation, and Data Resources

2.1 Data Resources for Drug Discovery Driven by AI

The availability of big, superior datasets encompassing chemical structures, biological targets, protein sequences, and experimental assay findings forms the basis of AI-assisted drug discovery. Public repositories make it possible to benchmark new algorithms and train predictive models:



Chemical Information Sources:

- ChEMBL—curated bioactivity information, such as pharmacological profiles and target interactions, for drug-like compounds.
- A huge library of chemical structures, characteristics, and biological test findings can be found on PubChem.
- DrugBank is an integrated database that links chemical, pharmacological, and pharmaceutical data about drugs and targets.

Organizational Databases:

- The Protein Data Bank (PDB) contains 3D structures of proteins, nucleic acids, and complex assemblies that have been determined through experimentation.
- The AlphaFold Protein Structure Database significantly increases structural coverage by

providing AI-predicted 3D models for a variety of proteins.

Datasets for benchmarking:

- Standardized datasets, such as ADMET, quantum chemistry, and bioactivity tasks, are used by MoleculeNet to assess molecular machine learning models.
- Atom3D is a benchmark for learning 3D molecule representations for a variety of biological prediction tasks.

Table 1: **Important Information Sources for AI in Drug Development**

- Description of the Category Database or Resource Main Application in AI Models
- Chemical ChEMBL curated information on bioactivity and binding Instruction Models of DTI, QSAR, and ADMET
- PubChem is a comprehensive chemical and bioassay database. Virtual screening and the creation of descriptors
- The DrugBank Data on drug-target pharmacology Repurposing drugs and identifying targets
- Database of Structural Proteins 3D architectures of macromolecules Docking and structure-based drug design
- AlphaFold Database AI-predicted architectures for proteins Identification of binding sites and target modeling
- Comparisons Typical molecular machine learning projects with MoleculeNet Model assessment and comparison
- Benchmarks for 3D molecular prediction using Atom3D Spatial interaction modeling and 3D graph learning.

Category Description	**Database or Resource**	Main Application in AI Models
Chemical	ChEMBL	Curated information on bioactivity and binding; Models of DTI, QSAR, and ADMET
Chemical	PubChem	Comprehensive chemical and bioassay database; Virtual screening and creation of descriptors
Drug-Target Pharmacology	DrugBank	Data on drug-target pharmacology; Repurposing drugs and identifying targets
Structural Proteins	Protein Data Bank (PDB)	3D architectures of macromolecules; Docking and structure-based drug design
Protein Structure Prediction	AlphaFold Database	AI-predicted architectures for proteins; Identification of binding sites and target modeling
Benchmark Datasets	MoleculeNet	Model assessment and comparison for typical molecular machine learning projects
Benchmark Datasets (3D)	Atom3D	Benchmarks for 3D molecular prediction; Spatial interaction modeling and 3D graph learning

2.2 Techniques for Molecular Representation

- The performance of AI models is largely dependent on an efficient molecular representation. Conventional cheminformatics relied on manually created descriptors like physicochemical property vectors or molecular fingerprints (ECFP, MACCS, etc.). Learned representations obtained directly from molecular graphs, sequences, or 3D coordinates are being used more and more in contemporary AI techniques:

- Graph-based Representations: Atoms are nodes and bonds are edges in graphs that represent molecules. By capturing connectivity and substructural patterns, Graph Neural Networks (GNNs) and message-passing algorithms improve molecular property and bioactivity predictions.

- Sequence-based Representations: RNNs, Transformers, and other natural language processing (NLP) architectures can be used for molecule production, property prediction, and retrosynthesis planning thanks to SMILES strings and InChI codes.

- 3D Structure-based Representations: Docking simulations, molecular dynamics, and structure-based drug design are made easier by the spatial coordinates of atoms, which are obtained from crystallographic data or structure prediction models.

2.3 AI-Based Drug Discovery Techniques

- A variety of supervised, unsupervised, and generative modeling techniques are used in AI drug discovery algorithms:

- Predictive modeling includes target prediction, ADMET property estimate, and quantitative structure–activity relationship (QSAR) modeling using deep neural networks, GNNs, and ensemble learning techniques.

- Generative design: de novo molecule creation with improved pharmacological profiles using reinforcement learning frameworks, generative adversarial networks (GANs), and variational autoencoders (VAEs).

- Structure Prediction: Target-based design is made possible even in the absence of experimental data thanks to AlphaFold2 and related deep learning pipelines for high-accuracy protein structure modeling.

- Virtual Screening & Docking: AI-powered scoring algorithms that combine pose estimation, chemical similarity, and binding affinity prediction for high-throughput ligand-receptor screening.

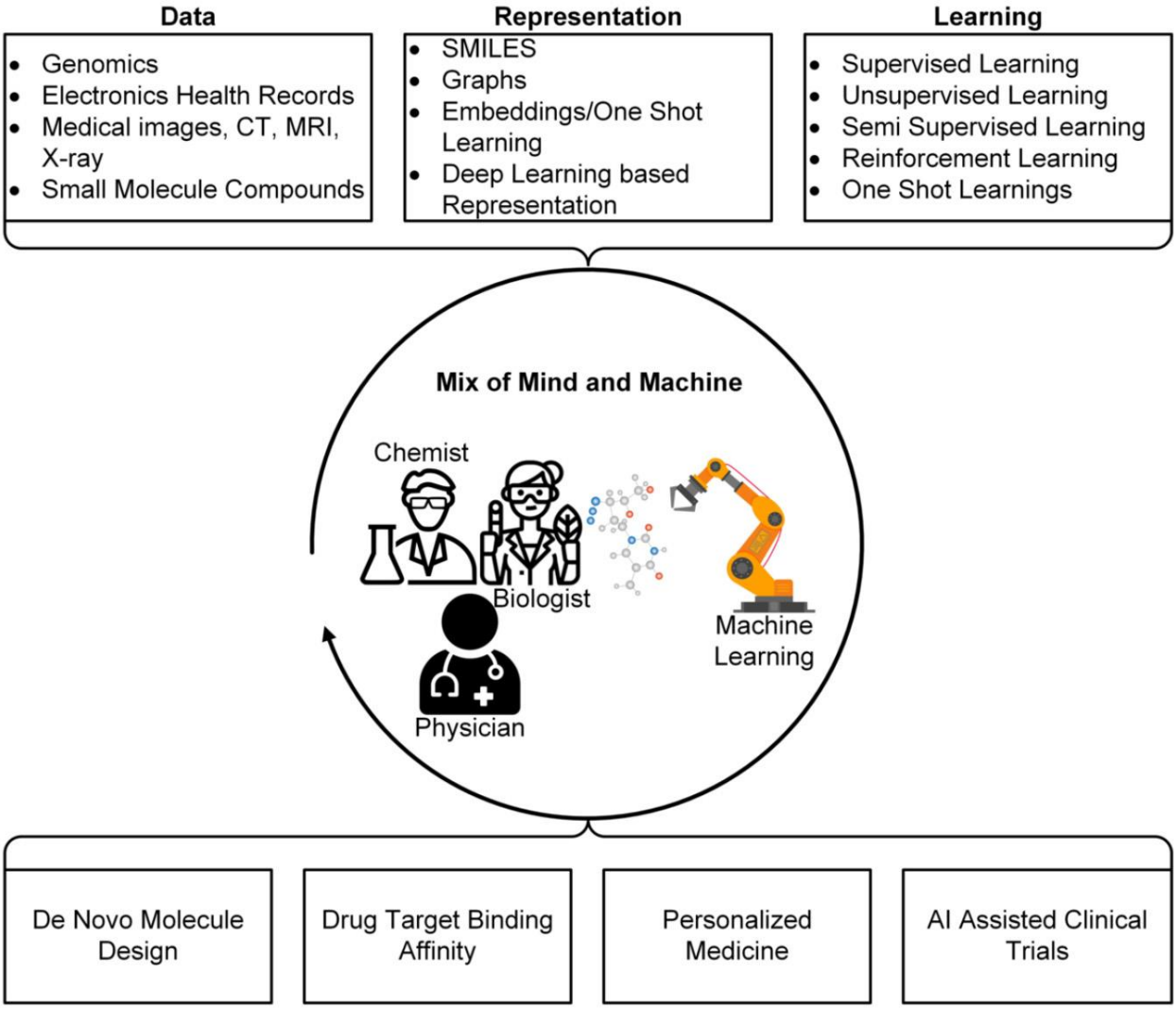
- Effectively, novel ligands customized to particular targets can be produced using diffusion models conditioned on protein pockets (3D-conditional generative models), which allows for inpainting, property tuning, and even negative design (avoidance of undesired interactions).

- Even in the lack of complete structural data, PharmaDiff, a diffusion model conditioned on 3D pharmacophoric characteristics, can direct molecule formation.

- It has demonstrated excellent performance in obtaining high docking scores and matching target pharmacophore patterns.

- Using AI-driven design, Insilico created ISM001-055 for idiopathic pulmonary fibrosis.

- Using diffusion models and LLM techniques through its firm ProPhet, AION Labs is advancing AI for molecular glue and small molecule targeting through subsidiary ventures with little biological data.



3. Using AI in the Pipeline for Drug Discovery

3.1 Identification and Validation of the Target

Rational drug discovery begins with the identification of a biological target, usually a protein, nucleic acid, or signaling system. AI facilitates this process by combining literature mining with multi-omics datasets (genomics, proteomics, and transcriptomics) to find new targets linked to disease. Prioritizing targets according to projected illness relevance, structural tractability, and safety profiles is aided by network analysis, causal inference techniques, and graph-based machine learning models. The ability of AlphaFold2 to predict structures has also made it possible to analyze proteins that have not yet been described, which has led to the development of new target hypotheses.

3.2 Virtual Screening and Hit Identification

Once a target has been found, AI speeds up hit finding by using virtual screening in place of or in addition to conventional high-throughput screening. Millions of chemicals can be quickly scored against a target using deep learning models based on binding affinity data, weeding out possible hits before physical testing. Additionally, generative models enable the creation of completely original chemical scaffolds that are probably going to interact with the target. AI-based scoring combined with docking simulations in hybrid processes improves accuracy and lowers false positives.

3.3 Optimization of Leads

By striking a balance between potency, selectivity, pharmacokinetics, and safety, artificial intelligence approaches are being used more and more to optimize lead compounds. Bayesian optimization frameworks and multi-objective reinforcement learning can recommend chemical changes that increase binding affinity while lowering toxicity or metabolic instability. In order to find compounds that are both active and drug-like, these models frequently incorporate ADMET predictions into the optimization loop.

3.4 ADMET Forecast

The success of a drug candidate in the preclinical and clinical stages is largely determined by its ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) characteristics. Compared to conventional rule-based systems, AI-based QSAR models trained on extensive datasets like ChEMBL and Tox21 are more accurate in predicting these attributes. These forecasts lessen the possibility of expensive late-stage failures by allowing the early screening of undesirable candidates.

3.5 Repurposing Drugs

AI's capacity to integrate diverse datasets, such as chemical similarity networks, phenotypic screening findings, and real-world clinical data, is advantageous for drug repurposing, which is the process of finding new therapeutic uses for already-approved medications. Development timeframes can be greatly shortened by using machine learning models to identify subtle drug-disease connections. AI-assisted COVID-19 treatment candidate discovery from pre-approved drug libraries is one such example.

3.6 Design and Optimization of Clinical Trials

AI helps with clinical trial design in addition to preclinical research by identifying biomarkers for therapy response, refining trial protocols, and forecasting patient classification criteria. By simulating trial results and applying natural language processing to electronic health records and previous trial data, recruitment tactics are improved and trial efficiency is increased.

4. Issues, Restrictions, and Prospects for the Future

4.1 Availability and Quality of Data

Drug development AI models mostly rely on sizable, varied, and well-annotated datasets. However, experimental heterogeneity, imprecise labeling, and missing recordkeeping are common problems with publicly available chemical and biological datasets. Despite being more extensive, pharmaceutical companies' proprietary databases are rarely shared because of legal and competitive restrictions. Reproducibility and model generalizability may be hampered by this absence of publicly available, standardized data.

4.2 Model Interpretability and Trustworthiness

While deep learning and other complex AI models can achieve high predictive accuracy, their decision-making processes are often opaque. This “black box” nature poses challenges in regulatory approval and scientific acceptance. Without transparent reasoning, it becomes difficult for medicinal chemists and regulatory agencies to trust AI-generated predictions. Efforts to develop explainable AI (XAI) tools—such as attention maps, feature importance scores, and surrogate models—are essential to bridge this gap.

4.3 Benchmarking and Reproducibility

Numerous AI research in drug development show encouraging outcomes on small datasets but fall short in independent assessments. Inconsistent assessments among various research may result from the lack of generally recognized benchmarking standards. Standardized benchmarks are intended by projects like Atom3D and MoleculeNet, but wider use is required to guarantee fair comparisons and repeatable results.

4.4 Integration with Current Processes

Computational predictions and experimental validation pipelines must be in sync for AI to be integrated into conventional drug discovery processes. Bottlenecks may arise when AI systems and laboratory procedures have different timelines, output formats, and success measures. To guarantee seamless integration, hybrid teams made up of biologists, chemists, and computational scientists are essential.

4.5 Ethical and Regulatory Aspects

Regulations pertaining to algorithmic bias, patient privacy, and data provenance are brought up by the use of AI-driven drug discovery methods. Guidelines for the application of AI in pharmaceutical research are still being developed by regulatory bodies. Additionally, in order to prevent unintentional injury or unequal access to therapies, ethical concerns—like the possibility of bias in training datasets—must be addressed.

4.6 Prospects for the Future

Multi-modal integration, which combines chemical, biological, imaging, and clinical data to produce more comprehensive predictive models, is where artificial intelligence in drug discovery is headed. Mechanistic understanding may be improved by advances in causal machine learning, which could make it easier to discern between correlation and causation. Large task-specific datasets may become less necessary when foundation models trained on extensive chemical and biological corpora become available. Additionally, open-source toolkits and cloud-based collaboration platforms will democratize access to state-of-the-art AI capabilities, encouraging creativity in startups, academia, and business. In the end, human–AI collaboration—where algorithms support but do not replace domain specialists in decision-making—is probably going to produce the most significant advancements.

4.7 AI in Drug Discovery Case Studies

Case Study 1: **DSP-1181 (Exscientia + Sumitomo Dainippon Pharma) was presented**.

An AI-driven platform was used to produce DSP-1181, a serotonin 5-HT1A receptor agonist, to treat obsessive-compulsive disorder (OCD). In contrast to the conventional multi-year timetable, the lead candidate impressively advanced from project commencement to clinical trial in just 12 months. Deep reinforcement learning was used by the AI system to navigate large chemical areas while adhering to several pharmacological restrictions. This accomplishment highlights how AI has the ability to significantly accelerate the early stages of drug discovery.

Case Study 2: The Fibrosis Candidate from Insilico Medicine (INS018_055)

The discovery of a new inhibitor that targets idiopathic pulmonary fibrosis was disclosed by Insilico Medicine in 2021. Within 18 months, the business took a small molecule from concept to preclinical validation using a generative chemistry platform and deep learning-based target discovery. The method suggested new chemical entities with desired ADMET profiles by combining generative models and multi-omics data.

Case Study 3: Baricitinib with BenevolentAI for COVID-19

BenevolentAI identified baricitinib, which was initially created for rheumatoid arthritis, as a possible COVID-19 treatment by utilizing its knowledge graph and AI-driven inference capabilities. Clinical trial data later confirmed the forecast, which was based on mapping the relationships between drugs and diseases and targets. As a result, emergency use authorization was granted in some areas. This demonstrates how AI can be used to repurpose medications in times of severe medical emergencies.

4.8 The Legal Environment for AI-Powered Drug Discovery

- AI is being used more and more in pharmaceutical situations by regulatory agencies including the European Medicines Agency (EMA) and the U.S. Food and Drug

Administration (FDA).

- Although direct AI applications in preclinical drug discovery are still less regulated, the FDA's "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan" provides initial recommendations for software validation.

- In order to guarantee reproducibility, regulators stress the importance of model transparency, dataset disclosure, and auditability.

- Committees that oversee ethics are starting to assess AI processes for possible bias, particularly in data pertaining to patients.

- Similar to the CONSORT principles in clinical research, future regulation is anticipated to mandate consistent reporting of AI model designs, training datasets, and validation processes.

4.9 Collaborative Models and Industry Trends

- Pharma-AI Collaborations: To speed up their pipelines, companies like AstraZeneca, GSK, and Sanofi have formed long-term partnerships with AI companies (such as BenevolentAI, Exscientia, and Insilico).

- Open Science Initiatives: By utilizing AI and freely sharing datasets and models, initiatives such as the COVID Moonshot and Open Targets are promoting quick innovation.

- Hybrid Discovery Models: An emerging trend is the creation of iterative "design–build–test–learn" cycles by fusing microfluidic high-throughput experimentation with AI-led hypothesis development.

- Investment Growth: Since 2018, a total of billions of USD have been raised by AI drug discovery businesses, with venture funding concentrating on precision medicine, generative chemistry, and AI-driven clinical trials.

4.10 AI Technology Developments for Drug Discovery:

Data Fusion and Multimodal Learning:

Integrating various biomedical data, such as chemical compounds, protein sequences, expression profiles, and clinical trial results, into cohesive AI models is a focus of recent study. By utilizing

complementary information, multimodal learning architectures—like transformers with modality-specific encoders—can produce predictions that are more reliable. For instance, a model that combines transcriptome profiles and 3D structural embeddings is able to predict off-target effects more accurately than a structure model.

Foundation and Self-Supervised Models:

Drug discovery is adopting self-supervised learning, which has shown great effectiveness in natural language processing. The model is able to learn general chemical principles through extensive pretraining on billions of molecules, which can subsequently be refined for particular applications such as retrosynthesis or ADMET prediction. Foundation models with promising cross-task transfer capabilities are MolBERT and ChemBERTa.

Co-designing Proteins and Ligands using Generative AI

Generative AI is now being used in protein engineering, which goes beyond small-molecule design to create enzymes or antibodies that are optimized for stability, binding affinity, or manufacturing. New technologies have the ability to completely change the design of biologics by co-generating a ligand and a protein target that interact ideally.

Integration of Quantum Computing

Although they are still in their infancy, quantum algorithms have the ability to replicate molecular interactions with previously unheard-of precision. Companies like Polaris Quantum Biotech are exploring with hybrid quantum-classical pipelines for applications like energy minimization and molecular docking.

4.11 Difficulties in Bringing AI from the Bench to the Bedside

Clinical Acceptance and Confidence:

- Adoption in clinical settings is conservative even when AI models produce intriguing preclinical prospects. Before clinicians to trust AI suggestions for therapeutic decision-making, they need clear rationale, strong validation, and data from randomized controlled trials.

- Drug discovery is still less uniform in this area, even though regulatory clearance for AI-assisted medical devices has started. Regulators will probably demand that AI models go through formal qualification procedures with established metrics and validation datasets, much like diagnostic tools do.

Integration with Evidence from the Real World (RWE)

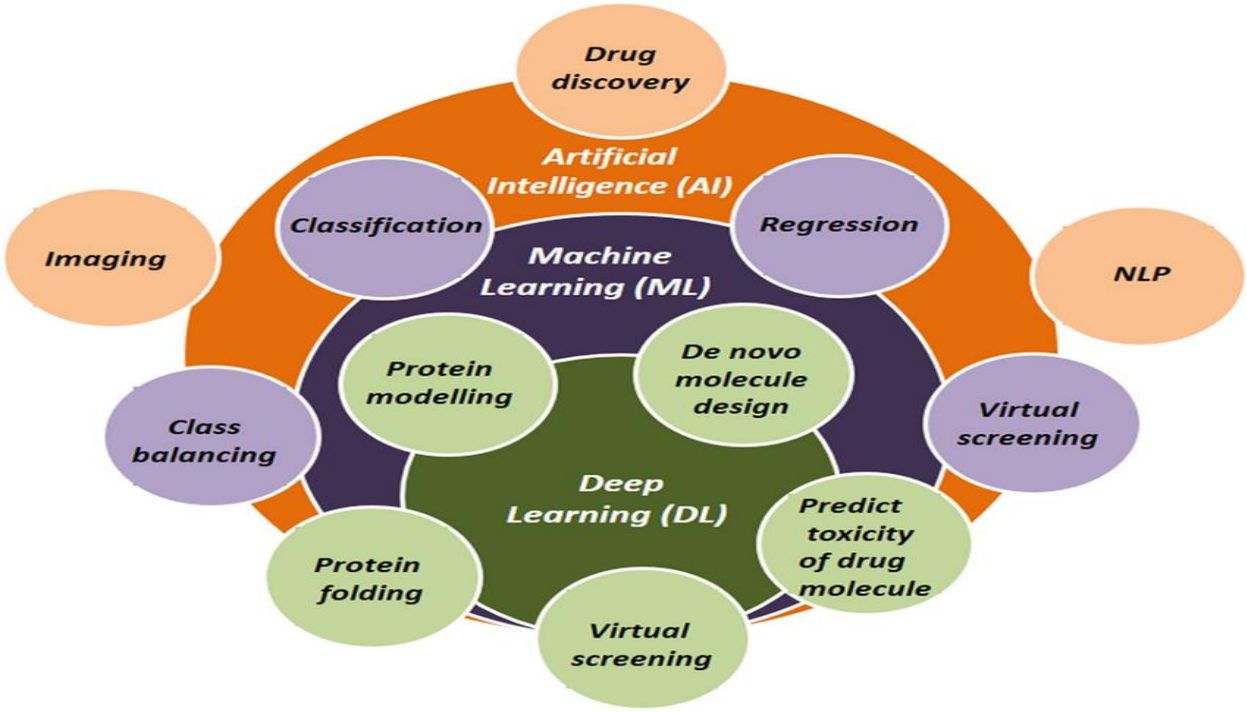
One important gap is connecting post-marketing surveillance data with preclinical projections. AI has the potential to establish a feedback loop that enhances drug safety and future forecasts by connecting preclinical safety estimates with actual adverse event monitoring.

5. In Summary:

- With tools and frameworks that greatly speed up and optimize the process from target selection to clinical development, artificial intelligence has become a game-changing force in drug discovery. The exploration of large chemical spaces, increased predictive accuracy, and more rational design of small molecules and biologics have all been made possible by developments in deep learning, graph neural networks, generative modeling, and multimodal learning. Public datasets, top-notch structural resources, and AI-powered software platforms have all been combined to create previously unheard-of chances for cooperation and innovation between regulatory agencies, business, and academia.

- Notwithstanding its achievements, the area continues to encounter difficulties with reproducibility, data quality, model interpretability, and integrating computational predictions with experimental procedures. Furthermore, the rate and extent of AI use in drug research will be influenced by changing legal frameworks and ethical considerations. Standardized benchmarking procedures, the creation of transparent, explicable systems that can close the gap between in silico predictions and clinical realities, and synergistic human–AI collaboration are likely to be key components of future advancement as the technology develops. AI will continue to transform early-stage drug discovery and assist in providing patients with safer, more effective treatments at a never-before-seen speed if these obstacles are removed.
- clinical development, artificial intelligence has become a game-changing force in drug discovery. The exploration of large chemical spaces, increased predictive accuracy, and more rational design of small molecules and biologics have all been made possible by developments in deep learning, graph neural networks, generative modeling, and multimodal learning. Public datasets, top-notch structural resources, and AI-powered software platforms have all been combined to create previously unheard-of chances for cooperation and innovation between regulatory agencies, business, and academia.

- Notwithstanding its achievements, the area continues to encounter difficulties with reproducibility, data quality, model interpretability, and integrating computational predictions with experimental procedures. Furthermore, the rate and extent of AI use in drug research will be influenced by changing legal frameworks and ethical considerations. Standardized benchmarking procedures, the creation of transparent, explicable systems that can close the gap between in silico predictions and clinical realities, and synergistic human–AI collaboration are likely to be key components of future advancement as the technology develops. AI will continue to transform early-stage drug discovery and assist in providing patients with safer, more effective treatments at a never-before-seen speed if these obstacles are removed.



Conclusion :

Drug development is an expensive, time-consuming, and dangerous process that has historically relied on iterative chemical optimization and experimental screening. This paradigm is being redefined by artificial intelligence (AI), which makes data-driven decision-making possible, improves predictive accuracy, and speeds up procedures throughout the discovery pipeline. AI systems can discover new drug targets, create new molecular entities, optimize lead compounds, and more accurately predict absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties by utilizing sophisticated algorithms like graph neural networks, generative models, reinforcement learning, and multimodal learning.

The basis for training and validating such models is provided by important resources like as the Protein Data Bank, DrugBank, PubChem, ChEMBL, and AlphaFold. This paper summarizes the latest developments in AI-driven drug discovery, looks at existing constraints, explores potential future approaches, and provides example case studies. Data quality, model interpretability, legal constraints, and new prospects in foundation models, multimodal integration, and collaboration

frameworks between humans and AI are given special attention. AI has the ability to lower attrition rates, accelerate development schedules, and provide patients throughout the world with more effective therapies with sustained innovation and responsible application.

References :

- Wikipedia.com
- Written sources
  - The Hindu
  - The Times Of India

- https://pmc.ncbi.nlm.nih.gov/articles/PMC7577280/

- https://iucrc.nsf.gov/centers/center-for-data-driven-drug-development-and-treatment-assessment-data/

- https://www.sciencedirect.com/science/article/pii/S2095177925000656

- https://pubmed.ncbi.nlm.nih.gov/35755871/

- https://pubmed.ncbi.nlm.nih.gov/34734228/

- https://arxiv.org/abs/2506.06915