Artificial Intelligence in Medicinal Chemistry: A Comprehensive Review

K. P. Beena[1]

[1]Associate Professor, Department of Pharmaceutical Chemistry, Sri Ramakrishna Institute of Paramedical Sciences, College of Pharmacy, Coimbatore- 641044, Tamil Nadu, India

Corresponding author: beenaakelesh12@gmail.com

Doi: https://doi.org/10.5281/zenodo.17947015
Received: 09 December 2025
Accepted: 22 December 2025

Abstract

Artificial intelligence (AI) has developed as a transformative force in medicinal chemistry, redesigning early drug discovery, hit identification, lead optimisation, and synthetic route planning. Effective exploration of chemical space and determination of key molecular properties has been enabled by advances in machine learning (ML), deep learning (DL), graph neural networks (GNNs), and generative models (such as variational autoencoders and transformer-based chemical language models). AI now plays a vital role across virtual screening, de-novo design, reaction prediction, retrosynthesis planning, ADMET profiling, and structure-based drug design. Remains challenging despite its rapid progress which includes data quality issues, model interpretability, experimental validation gaps, and integration into existing medicinal chemistry workflows. This review summarises the current AI methodologies, major applications, advantages, limitations and future opportunities in medicinal chemistry.

Key words: Artificial Intelligence, Medicinal Chemistry, Molecular properties

Introduction

Medicinal chemistry integrates chemical synthesis, structural biology, and pharmacological evaluation to design compounds with optimal potency, selectivity, and drug-like properties. Conventional discovery is expensive, sequential, and slow, typically requiring thousands of compounds to identify a single candidate. AI offers an accelerated, data-centric alternative by recognising patterns in chemical and biological data that are often too complex for manual analysis [1]. Driven by increased computational power, large curated datasets, and advances in deep learning architectures, AI has become a core component of modern drug design programs across academia and industry [2].

AI Methodologies in Medicinal Chemistry

Machine Learning and QSAR Models

Classical machine learning methods—random forests, support vector machines, logistic regression, and gradient boosting—continue to support ligand-based design. These approaches underpin QSAR modelling, predicting activity or property values based on molecular descriptors [3]. Although less flexible than deep learning, they often perform strongly on small datasets and provide interpretable relationships.

Deep Learning Approaches

Deep neural networks learn hierarchical chemical representations, allowing them to predict activity, physicochemical properties, and ADMET parameters more accurately [4]. Convolutional neural networks (CNNs) analyse 2D/3D molecular representations, whereas recurrent neural networks (RNNs) and transformer models interpret SMILES strings as chemical "language." Pretrained models enable transfer learning for low-resource tasks [5].

Graph Neural Networks (GNNs)

GNNs treat molecules as graphs of atoms (nodes) and bonds (edges). They excel at structure–property prediction because they naturally encode chemical connectivity [6]. Variants such as message-passing neural networks (MPNNs) and attention-based GNNs now represent the state-of-the-art in molecular property prediction, docking score refinement, and toxicity estimation.

Generative AI for De-Novo Design

Generative models create new molecular structures with desired properties [7]. Key classes include:

Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Normalizing flows, Transformer-based chemical language models. These models explore vast chemical spaces and optimize molecules using reinforcement learning or multi-objective scoring.

Reinforcement Learning (RL)

RL frameworks treat molecule generation as an optimization problem, refining molecules based on reward functions for potency, selectivity, logP, solubility, or synthetic accessibility [8]. RL accelerates lead optimization by automatically proposing improved analogs.

AI for Protein Structure and Target Modelling

Breakthroughs such as AlphaFold and RoseTTAFold have transformed structure-based drug design (SBDD) by predicting protein structures with near-experimental accuracy [9]. AI now enables virtual screening and docking on previously intractable targets, improving hit discovery.

Applications of AI in Medicinal Chemistry

Virtual Screening and Hit Identification

AI-driven virtual screening ranks compounds based on predicted affinity, reducing dependence on brute-force docking [10]. ML-based scoring functions learn from experimental binding data to outperform conventional docking in several benchmarks.

Applications include: Rapid triaging of million-compound libraries, Activity prediction for novel scaffolds, Ligand-based screening where structures are unavailable, Hybrid pipelines combining docking with ML-based rescoring enhance enrichment and reduce false positives.

De-Novo Design and Lead Optimisation

AI suggests novel molecules with desired properties while maintaining structural novelty. Common strategies include: Optimising lead compounds along potency/ADMET axes, Scaffold hopping by generating structurally diverse analogues, introducing subtle functional group modifications, Generative models integrated with RL enable multi-objective optimisation, a central challenge in medicinal chemistry.

Structure-Based Drug Design (SBDD)

With available or AI-predicted protein structures, AI guides: Binding pose prediction, Affinity estimation, Pocket property analysis, Fragment-based design, Machine-learned scoring functions now complement physics-based methods, offering improved accuracy and speed [11].

ADMET and Toxicity Prediction

Accurate ADMET prediction is crucial to avoiding late-stage failures. AI models can estimate:

Absorption and permeability, Cytochrome P450 interactions, Toxicity liabilities (cardiotoxicity, hepatotoxicity, mutagenicity), Blood–brain barrier penetration, Plasma protein binding and Multitask models leverage related ADMET endpoints to improve learning efficiency.

Retrosynthesis and Reaction Prediction

AI-driven retrosynthesis tools propose synthetic routes and predict reaction outcomes. Template-based models use curated reaction rules, while template-free models learn directly from reaction examples. Applications include: Identifying shorter or more feasible routes, predicting yields, Estimating reaction likelihood. These tools are increasingly used to assess whether AI-generated molecules can be synthesized in practice.

Reaction Condition Optimisation

ML models also suggest optimal reagents, catalysts, and reaction conditions, accelerating medicinal chemistry assays and route development [12].

Data Sources and Computational Tools

Key Databases

Medicinal chemistry relies on large, high-quality datasets, including: ChEMBL – bioactivity data,

PubChem – chemical structures and assays, PDBbind – protein–ligand complexes, MoleculeNet – ML benchmark tasks, BindingDB, ZINC, Tox21, ADMET datasets. Data quality is critical; inconsistent assay conditions can reduce model performance.

Open-Source Tools and Platforms

Several widely used tools support AI-assisted design: RDKit – cheminformatics toolkit, AiZynthFinder, ASKCOS – retrosynthesis, DeepChem, ChemProp, DGL-LifeSci – deep learning libraries, AlphaFold – protein structure prediction, AutoDock-GPU + ML rescoring tools for docking optimisation. Commercial platforms integrate these capabilities with enterprise-level data pipelines.

Advantages of AI in Medicinal Chemistry

Accelerated hit discovery through rapid screening, Exploration of vast chemical space, including non-intuitive compounds, better property prediction, reducing attrition rates, Cost and time efficiency in prioritising compounds, Improved synthetic planning with AI-driven retrosynthesis., Integration of multi-modal data (chemical, structural, biological). These benefits allow chemists to focus efforts on promising candidates, reducing unnecessary synthesis and testing.

Challenges and Limitations

Data Quality and Bias

AI models depend heavily on training data quality. Missing, noisy, or inconsistent assay data propagate errors into predictions. Public datasets often contain biased chemical space dominated by specific scaffolds.

Model Interpretability

Deep learning models are often “black boxes.” Medicinal chemists need mechanistic rationale, not just predictions. Lack of interpretability hinders trust and adoption.

Synthetic Feasibility Issues

Generative models may produce molecules that are chemically valid but difficult or impossible to synthesize. Integrating retrosynthesis constraints is improving this area but remains a challenge.

Lack of Prospective Validation

Many AI successes are retrospective. Experimental validation lags behind computational advancements, creating a gap between theoretical performance and real-world utility.

Computational Costs and Expertise

Training large models requires hardware and skilled computational scientists, which may not be available in all institutions.

Regulatory and Ethical Concerns

Use of proprietary chemical data, model reproducibility, and intellectual property around AI-generated molecules present unresolved regulatory challenges.

Future Directions

Foundation Models for Chemistry

Large chemical language models trained on billions of molecules will allow zero-shot predictions, better generalisation, and more realistic de-novo design.

Integration of Structure and Generative Design

Combining AI-predicted protein structures with generative ligand models will enable target-guided molecular design, especially for challenging proteins.

Federated and Privacy-Preserving Learning

Pharmaceutical companies can train shared models without exposing proprietary data, enabling richer datasets and improved performance.

AI-Automation Coupling

Closed-loop systems that connect AI designs to automated synthesis and high-throughput screening (HTS) will significantly accelerate optimisation cycles.

Improved Explainable AI (XAI)

Models that provide mechanistic insight and identify key chemical features will enhance chemists' confidence in AI-generated predictions.

Multimodal Integration

Future models will integrate chemistry, omics, imaging, and clinical data, supporting more comprehensive drug design strategies.

Conclusion

AI is reshaping the landscape of medicinal chemistry by enabling rapid discovery, efficient optimization, and enhanced decision-making. Techniques ranging from classical ML to advanced generative models and GNNs provide powerful tools to explore chemical space and predict molecular behaviour. While challenges remain—particularly regarding data quality, interpretability, and prospective validation—the trajectory of AI-driven drug design is strongly positive. With continued methodological innovation and closer integration of computational and experimental workflows, AI is set to become an indispensable partner to medicinal chemists in driving next-generation therapeutics

References

1. Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity. *Bioinformatics, 26*(9), 1169–1175.
2. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today, 23*(6), 1241–1250.
3. Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Green, W. H., Barzilay, R., & Jensen, K. F. (2019). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science, 10*(2), 370–377.
4. Gaulton, A., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Research, 45*(D1), D945–D954.
5. Gilmer, J., Schoenholz, S., Riley, P., Vinyals, O., & Dahl, G. (2017). Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*.
6. Goh, G. B., Siegel, C., Vishnu, A., Hodas, N., & Baker, N. (2017). Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *ACS Central Science, 3*(8), 852–859.
7. Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.
8. Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics, 9*(48).
9. Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science, 361*(6400), 360–365.
10. Schwaller, P., et al. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science, 5*(9), 1572–1583.
11. Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature, 555*, 604–610.
12. Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics, 29*(6–7), 476–488.
13. Wu, Z., Ramsundar, B., Feinberg, E., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science, 9*, 513–530.