International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct-2026, PP 01-06

Artificial Intelligence For Drug Discovery: A Review Of Molecular Representation, Target Interaction, Generative Design And Explainable Models

Matkar Pallavi Rajendra[1]; Prof. R. V. Varhade[2]; Prof. M. C. Kolse[3]; Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Drug discovery is a complex, slow and expensive process that involves identifying biological targets, finding and optimizing compounds that act on them, and ensuring that the resulting candidates are effective and safe. Traditional approaches rely heavily on laboratory experiments and screening campaigns that consume years and large budgets, with high failure rates. Artificial Intelligence (AI) and Machine Learning (ML) are increasingly used to accelerate and improve this pipeline. This paper reviews AI methods across the stages of drug discovery: molecular representations such as SMILES strings, fingerprints and molecular graphs; graph neural networks and message passing for property prediction; deep models for drug–target interaction and binding affinity; prediction of toxicity and bioactivity; generative models and reinforcement learning for designing new molecules; protein structure prediction; drug repurposing; and explainable AI for interpreting predictions. Representative successes, including deep-learning-driven antibiotic discovery and rapid design of kinase inhibitors, are discussed alongside benchmark datasets. A layered AI-driven drug discovery architecture is proposed, with research gaps, objectives, a comparison of model families and a discussion of challenges such as data quality, bias, limited datasets and the need for experimental validation. The review concludes that AI can make drug discovery faster and more cost-effective when tightly coupled with laboratory testing.

Index Terms—Artificial intelligence, drug discovery, machine learning, graph neural networks, drug–target interaction, generative models, drug repurposing, explainable AI

I. Introduction

1.1 Motivation

Bringing a new medicine to patients typically takes more than a decade, and most candidates that enter clinical trials fail, often because of insufficient efficacy or unexpected toxicity. The early stages of discovery, in which researchers search for compounds that interact with a disease target, involve screening very large chemical libraries. Since the space of drug-like molecules is astronomically large, exhaustive experimental testing is impossible.

AI offers tools for navigating this space efficiently. Vamathevan *et al.* [1] reviewed ML applications across target identification, hit discovery, lead optimization and clinical development, and Schneider *et al.* [2] argued that AI is reshaping how medicinal chemists design molecules. Rather than replacing experiments, AI helps prioritize which experiments to run.

This paper reviews AI techniques in drug discovery, proposes an architecture that integrates them into a design–make–test–learn cycle and discusses the remaining challenges.

1.2 Objectives of the Review

The objectives of this review are:

- **O1:** To review AI and ML techniques across the stages of drug discovery.
- **O2:** To compare molecular representations and model families for property and interaction prediction.
- **O3:** To examine generative and reinforcement learning methods for molecular design.
- **O4:** To propose an AI-driven drug discovery architecture integrated with experimental validation.
- **O5:** To identify challenges and future directions, including explainability and data sharing.

II. Overview of Key Concepts

2.1 Stages of Drug Discovery

The pipeline begins with target identification and validation, continues with hit discovery through screening, lead optimization to improve potency and safety, preclinical testing and finally clinical trials. Fig. 1 maps AI techniques to these stages.

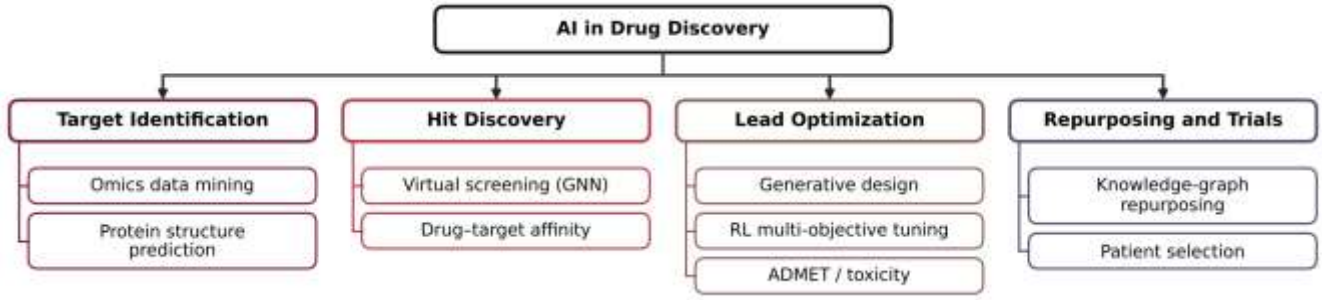


Fig. 1. AI techniques mapped to drug discovery stages.

2.2 Molecular Representations

Molecules can be represented as SMILES text strings [3], as fixed-length fingerprints such as extended-connectivity fingerprints [4] that encode substructures, or as graphs whose nodes are atoms and edges are bonds. The representation determines which ML models can be applied.

III. Review of Existing Work

Property prediction. Gilmer *et al.* [5] unified several graph-based models under the message passing neural network framework, in which atoms exchange information with neighbours over several steps before the graph is summarized, and achieved strong results on quantum-chemical properties. Yang *et al.* [6] analysed learned molecular representations and showed that a directed message passing network outperformed fingerprints on many industrial and public datasets. Wu *et al.* [7] introduced MoleculeNet, a benchmark collection that standardized evaluation of molecular ML.

Toxicity and interaction. Mayr *et al.* [8] developed DeepTox, a multitask deep network that won the Tox21 challenge for predicting toxic effects. Öztürk *et al.* [9] proposed DeepDTA, which uses convolutional networks on protein sequences and SMILES strings to predict drug–target binding affinity. Zitnik *et al.* [10] modelled side effects of drug combinations with graph convolutional networks on a multimodal graph of drugs and proteins.

Generative design. Gómez-Bombarelli *et al.* [11] used a variational autoencoder to map molecules into a continuous latent space where optimization can be performed and decoded back into new molecules. Popova *et al.* [12] combined generative and predictive networks with reinforcement learning to design molecules with desired properties. Zhavoronkov *et al.* [13] used a generative reinforcement learning system to design DDR1 kinase inhibitors, synthesizing and validating promising compounds within weeks.

Discovery and structure. Stokes *et al.* [14] trained a graph neural network to predict antibacterial activity and identified halicin, an antibiotic structurally different from known drugs. AlphaFold [15] and its successor for biomolecular complexes [16] predict protein structures with high accuracy, providing targets for structure-based design. Jiménez-Luna *et al.* [17] reviewed explainable AI methods that help chemists understand which molecular features drive predictions. Table I summarizes these works.

Ref.	Method	Key Findings	Limitations
[5]	Message passing neural networks	Unified graph models for molecules	Requires large labelled data
[6]	Directed MPNN (Chemprop)	Outperformed fingerprints on many datasets	Limited extrapolation
[8]	DeepTox multitask network	Top Tox21 toxicity prediction	Assay-specific endpoints
[9]	DeepDTA CNN on sequences and SMILES	Predicted binding affinity without 3-D data	Ignores structural detail
[11]	Variational autoencoder latent space	Continuous optimization of molecules	Many invalid decoded molecules
[12]	Deep RL for de novo design	Generated molecules with target properties	Synthesizability not guaranteed
[13]	Generative RL (GENTRL)	Rapid design of DDR1 inhibitors	Early-stage validation only
[14]	GNN antibacterial screening	Discovered halicin	Needs extensive follow-up testing
[15]	AlphaFold structure prediction	Near-experimental protein structures	Static, single conformations

TABLE I Summary of Literature Survey

IV. Identified Research Gaps

The following gaps are identified:

- **G1:** Public datasets are small, noisy and biased toward well-studied targets and chemical series.
- **G2:** Predictive models often fail outside their training chemical space, yet uncertainty is rarely quantified.
- **G3:** Many generated molecules are difficult or impossible to synthesize.
- **G4:** Explanations of predictions are not routinely provided to medicinal chemists.
- **G5:** Few AI-designed candidates have yet completed late-stage clinical trials, so real-world impact is still being established.

V. System Architecture

The architecture in Fig. 2 organizes AI tools around a closed design–make–test–learn loop.

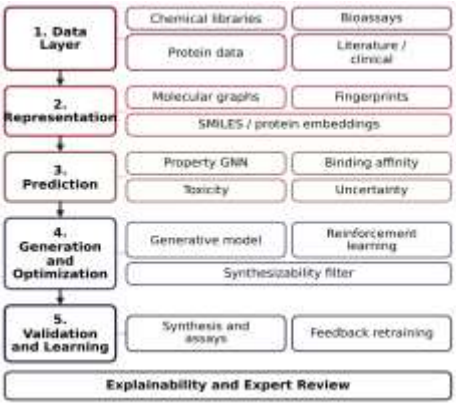


Fig. 2. Proposed AI-driven drug discovery architecture (design–make–test–learn).

5.1 Data Layer

Chemical libraries, bioassay results, protein sequences and structures, omics data, literature and clinical records are collected and standardized.

5.2 Representation Layer

Molecules are converted into graphs, fingerprints or SMILES embeddings, and proteins into sequence embeddings or predicted structures.

5.3 Prediction Layer

Graph neural networks predict activity, solubility, toxicity and other properties, while drug–target interaction models estimate binding affinity for the chosen target. Uncertainty estimates accompany every prediction.

5.4 Generation and Optimization Layer

Generative models propose new molecules, and reinforcement learning or Bayesian optimization steers generation toward compounds that score well on multiple objectives, including synthesizability.

5.5 Validation and Learning Layer

Top-ranked candidates are synthesized and tested in the laboratory. Results are fed back to retrain models, and explanation tools show chemists which substructures drive predictions.

VI. Implementation Methodology

The workflow is shown in Fig. 3. In a message passing network, the hidden state h_v of atom v is updated at every message-passing step as

$$h_v \leftarrow U(\, h_v\, ,\, \Sigma_\{w \in N(v)\}\, M(h_v, h_w, e_vw)\,) \qquad (1)$$

where N(v) are neighbouring atoms, e_vw the bond features, and M and U learned functions. Candidate molecules are ranked by a multi-objective score

$$R(m) = w_1 \cdot Act(m) + w_2 \cdot QED(m) - w_3 \cdot Tox(m) - w_4 \cdot SA(m) \qquad (2)$$

where Act is predicted activity, QED measures drug-likeness, Tox is predicted toxicity and SA estimates synthetic difficulty.

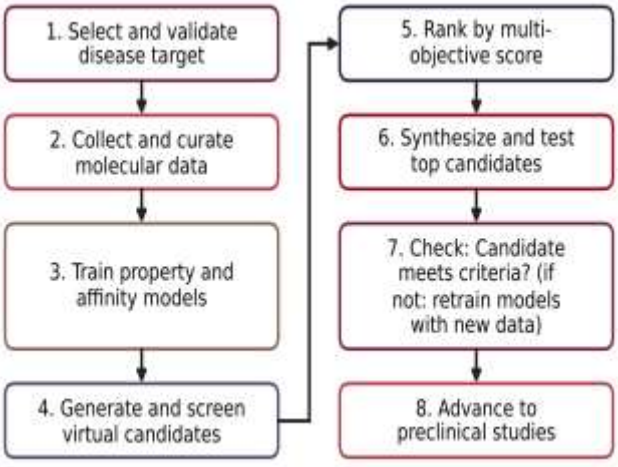


Fig. 3. AI-assisted drug discovery workflow.

VII. Performance Comparison

Table II compares model families. Fingerprint-based classical models are fast and robust with small data; graph neural networks learn richer representations; sequence models are convenient for binding affinity; and generative models enable new design but require careful filtering.

Model	Input	Strength	Limitation
Random Forest / XGBoost	Fingerprints	Robust, small data	Fixed features
Graph neural network	Molecular graph	Rich representation	Data hungry
CNN / Transformer	SMILES, sequences	Easy input handling	Ignores 3-D shape
VAE / generative model	Latent space	Novel molecules	Validity, synthesis
Reinforcement learning	Reward function	Goal-directed design	Reward hacking

TABLE II Comparison of Model Families

VIII. Open Issues and Future Directions

Table III summarizes key challenges. Future directions include foundation models trained on large chemical and biological corpora, automated laboratories that close the loop between AI and experiments, federated learning across pharmaceutical companies to share knowledge without sharing proprietary data, and AI for clinical-trial design and patient selection.

Challenge	Mitigation
Poor data quality	Curation, standardized assays
Limited datasets	Transfer learning; multitask models
Model bias	Diverse chemical space; applicability domain
Interpretability	Attribution of substructures (XAI)
Validation	Tight loop with laboratory experiments

TABLE III Challenges and Mitigation

IX. Conclusion

AI is changing how new medicines are found. The reviewed literature shows that graph neural networks predict molecular properties accurately, deep models estimate drug–target interactions, generative models and reinforcement learning propose novel compounds, and protein structure prediction opens new targets. Notable successes such as AI-assisted antibiotic discovery demonstrate real value. Data quality, generalization, synthesizability and interpretability remain significant challenges, and experimental validation is indispensable. Used within a design–make–test–learn cycle, AI can make drug discovery faster, more efficient and more cost-effective.

9.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

References

1. J. Vamathevan et al., “Applications of machine learning in drug discovery and development,” Nat. Rev. Drug Discov., vol. 18, no. 6, pp. 463–477, 2019.

2. P. Schneider et al., “Rethinking drug design in the artificial intelligence era,” Nat. Rev. Drug Discov., vol. 19, no. 5, pp. 353–364, 2020.
3. D. Weininger, “SMILES, a chemical language and information system,” J. Chem. Inf. Comput. Sci., vol. 28, no. 1, pp. 31–36, 1988.
4. D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010.
5. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in Proc. ICML, 2017, pp. 1263–1272.
6. K. Yang et al., “Analyzing learned molecular representations for property prediction,” J. Chem. Inf. Model., vol. 59, no. 8, pp. 3370–3388, 2019.
7. Z. Wu et al., “MoleculeNet: A benchmark for molecular machine learning,” Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.
8. A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, “DeepTox: Toxicity prediction using deep learning,” Front. Environ. Sci., vol. 3, Art. no. 80, 2016.
9. H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: Deep drug–target binding affinity prediction,” Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.
10. M. Zitnik, M. Agrawal, and J. Leskovec, “Modeling polypharmacy side effects with graph convolutional networks,” Bioinformatics, vol. 34, no. 13, pp. i457–i466, 2018.
11. R. Gómez-Bombarelli et al., “Automatic chemical design using a data-driven continuous representation of molecules,” ACS Cent. Sci., vol. 4, no. 2, pp. 268–276, 2018.
12. M. Popova, O. Isayev, and A. Tropsha, “Deep reinforcement learning for de novo drug design,” Sci. Adv., vol. 4, no. 7, Art. no. eaap7885, 2018.
13. A. Zhavoronkov et al., “Deep learning enables rapid identification of potent DDR1 kinase inhibitors,” Nat. Biotechnol., vol. 37, no. 9, pp. 1038–1040, 2019.
14. J. M. Stokes et al., “A deep learning approach to antibiotic discovery,” Cell, vol. 180, no. 4, pp. 688–702, 2020.
15. J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, pp. 583–589, 2021.
16. J. Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” Nature, vol. 630, pp. 493–500, 2024.
17. J. Jiménez-Luna, F. Grisoni, and G. Schneider, “Drug discovery with explainable artificial intelligence,” Nat. Mach. Intell., vol. 2, no. 10, pp. 573–584, 2020.