International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Artificial Intelligence-Based Approaches in Drug Discovery and Development: A Review

Nishchith Gosala J, Pooja K N
Department of Biotechnology, Sapthagiri College of Engineering, Bengaluru, Karnataka, India
Department of Master of Computer Application, Sidganga Institute of Technology, Tumkur, Karnataka, India
Email: nishchithgosala09@gmail.com, poojanatraj853@gmail.com

----------**********----------

Abstract:

Artificial intelligence (AI) is reshaping drug discovery by learning patterns from chemical, biological, and clinical data to propose targets, design molecules, and predict safety and efficacy. This review synthesizes recent advances across the drug discovery and development pipeline, emphasizing modern deep learning (DL) methods such as transformers and diffusion models for de novo design, graph neural networks for molecular property prediction, and protein structure prediction systems (e.g., AlphaFold2) that expand structure-enabled screening. We provide a pragmatic taxonomy of AI tasks, data modalities, and validation practices; highlight representative industrial and regulatory milestones; and discuss limitations including data bias, assay shift, uncertainty quantification, interpretability, and governance. Finally, we outline a risk-based framework for deploying AI in regulated settings and identify research directions—multimodal foundation models, active learning with laboratory feedback, and human-centered design—needed to convert algorithmic novelty into reproducible clinical impact.

Keywords— **Artificial intelligence, machine learning, deep learning, drug discovery, molecular generation, virtual screening, toxicity prediction, AlphaFold2, regulatory science**

----------**********----------

I. INTRODUCTION

Drug discovery is a high-risk, capital-intensive process spanning target identification, hit discovery, lead optimization, preclinical development, and clinical trials. Attrition is driven by insufficient efficacy, unexpected toxicity, poor pharmacokinetics, and suboptimal patient selection. AI offers a complementary paradigm in which models learn from large-scale datasets to prioritize hypotheses and accelerate design–make–test–analyze cycles. Recent progress is enabled by increased availability of multi-omics, imaging, and real-world clinical data; deep architectures capable of modeling sequences, graphs, and three-dimensional structures; and scalable computing. At the same time, deployment of AI for decisions affecting patient safety requires rigorous validation, uncertainty estimation, and governance.

II. MATERIALS AND METHODS (REVIEW METHODOLOGY)

Literature for this narrative review was identified from 2019–2025 using keyword combinations including AI drug discovery, graph neural network QSAR, transformer molecular generation, diffusion model drug design, protein structure prediction virtual screening, and AI regulatory decision making. Priority was given to peer-reviewed reviews, methodological papers with open benchmarks, and regulatory or policy documents. Inclusion criteria emphasized clear problem definition, reproducible experimental setups, and translational relevance.

III. AI ACROSS THE DRUG DISCOVERY PIPELINE

A. Target Identification and Validation

AI supports target discovery by integrating genomics, transcriptomics, proteomics, and phenotypic screening data to infer disease mechanisms. Common tasks include gene

prioritization, pathway activity estimation, and drug–target interaction prediction using graph-based and representation learning models. Knowledge graphs combine curated databases with embedding models to generate hypotheses, with success depending on leakage control and robust negative sampling.

B. *Hit Discovery: Virtual Screening and Binding Prediction*

Structure-based virtual screening ranks compounds for a target using docking and scoring. AI improves these workflows through learned scoring functions, pose re-ranking, and access to predicted protein structures. Ligand-based screening uses similarity metrics and quantitative structure–activity relationship models, where graph neural networks and transformers often outperform classical descriptors when trained with sufficient data.

C. *Lead Optimization*

Multi-objective Molecular Design Lead optimization balances potency with absorption, distribution, metabolism, excretion, toxicity, and developability. Generative AI addresses inverse design by proposing molecules that satisfy property constraints. Approaches include sequence-based language models, graph generators, and three-dimensional generative models, with diffusion models and reinforcement learning enabling controllable design under constraints.

D. *ADMET and Toxicity Prediction*

AI models predict solubility, permeability, metabolic liabilities, cardiotoxicity risk, and organ-specific toxicity. Best practices include applicability-domain analysis, probability calibration, external validation, and prospective testing to address noisy and assay-dependent labels

E. *Preclinical Development and Translational Modeling*

Machine learning surrogates accelerate physiologically based pharmacokinetic modeling and dose optimization. Regulatory momentum to reduce animal testing has increased interest in computational models and new approach methodologies, including AI-enabled toxicology.

F. *Clinical Development*

Trial Design and Safety In clinical phases, AI assists with patient stratification, endpoint standardization, site selection, and pharmacovigilance. Imaging and histopathology models reduce inter-rater variability and are increasingly evaluated within regulatory qualification pathways.

IV. MODEL FAMILIES AND REPRESENTATIONS

Model performance depends on representation choices, including molecular strings, graphs, and three-dimensional conformers; protein sequences and structures; and multi-modal combinations of chemical, biological, and clinical data. Transformers, graph neural networks, and diffusion models provide complementary inductive biases across tasks.

V. VALIDATION, REPRODUCIBILITY, AND REGULATORY READINESS

Decision-relevant AI requires evaluation aligned to the intended context of use. Recommended practices include robust data governance, external validation on new targets and chemical series, uncertainty quantification, prospective wet-lab testing, and human oversight with audit trails. Regulators increasingly emphasize risk-based credibility assessment, transparency, and lifecycle management.

VI. INDUSTRIAL ADOPTION AND CASE STUDIES

Industrial adoption integrates end-to-end AI platforms into discovery workflows to shorten cycles. Prospective demonstrations show rapid identification of active compounds and progression of AI-designed candidates toward clinical evaluation. Partnerships between pharmaceutical companies and software providers reflect maturation of AI infrastructure under regulated development.

VII. CHALLENGES AND OPEN PROBLEMS

Key challenges include limited and biased datasets, distribution shift, objective misalignment, explainability requirements, data security, and

ethical and legal considerations such as privacy and intellectual property.

VIII. FUTURE DIRECTIONS

Future progress will likely be driven by multimodal foundation models, closed-loop autonomous experimentation, and hybrid physics–AI approaches. System-level integration, rigorous validation, and human-centered design will be critical for sustained impact.

IX. CONCLUSIONS

AI has expanded from traditional screening to a broad set of capabilities across the drug discovery and development lifecycle. Translational success depends on alignment with experimental reality, uncertainty-aware decision making, and risk-based governance frameworks.

ACKNOWLEDGMENT

The writer wishes to extend heartfelt thanks to the Department of Biotechnology at Sapthagiri College of Engineering in Bengaluru for the essential academic support and resources that facilitated this research. The writer also appreciates the advice and motivation received from the faculty during the development of this paper.

REFERENCES

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021.

[2] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1263–1272.

[4] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, and A. R. Leach, "ChEMBL: Towards direct deposition of bioassay data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D930–D940, Jan. 2019.

[5] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. Veselov, V. Aladinskiy, A. Aladinskaya, A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, and A. Volkov, "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," *Nature Biotechnology*, vol. 37, no. 9, pp. 1038–1040, Sept. 2019.

[6] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, 2018.

[7] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[8] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[9] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, "Molecular diffusion models for drug design," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 1–10, 2022.

[10] M. Elton, J. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular design—A review of the state of the art," *Molecular Systems Design & Engineering*, vol. 4, no. 4, pp. 828–849, 2019.

[11] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, and A. Palmer, "Analyzing learned molecular representations for property prediction," *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370–3388, 2019.

[12] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016.

[13] U.S. Food and Drug Administration, "Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products," Draft Guidance, Jan. 2025.

[14] European Medicines Agency, "Reflection paper on the use of artificial intelligence in the medicinal product lifecycle," EMA/149995/2024, Sept. 2024.

[15] A. Bender and R. C. Glen, "Molecular similarity: A key technique in molecular informatics," *Organic & Biomolecular Chemistry*, vol. 2, no. 22, pp. 3204–3218, 2004.

[16] J. Lyu, S. Wang, Y. Balius, I. Singh, R. Levit, Y. Moroz, M. J. O'Meara, and B. K. Shoichet, "Ultra-large library docking for discovering new chemotypes," *Nature*, vol. 566, no. 7743, pp. 224–229, Feb. 2019.

