INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Review Article

Artificial Intelligence and Machine Learning Models in Early Drug Discovery: Current Paradigms, Challenges, and Future Prospects

Digambar Rahangdale, Aman Yadav, Puja Basule, Dr. Bhumeshkumar Wanjari

Shri Gurudev Institute of Pharmaceutical Education and Research, Gondia, Maharashtra, India 441601

ARTICLE INFO

Published: 11 Sept 2026
Keywords:
Artificial Intelligence, Deep Learning, Graph Neural Networks, Early Drug Discovery, De Novo Design, Virtual Screening, In Silico ADMET, Generative Chemistry.
DOI:
10.5281/zenodo.22702316

ABSTRACT

Traditional drug discovery is an arduous, high-risk process characterized by multi-billion-dollar R&D expenditures, protracted timelines exceeding a decade, and high early-stage attrition rates. The integration of Artificial Intelligence (AI) and Machine Learning (ML) has emerged as a disruptive paradigm across early discovery pipelines. This review provides an in-depth critical synthesis of computational architectures—ranging from classic Random Forest and Support Vector Machines to Deep Neural Networks, Graph Neural Networks (GNNs), Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and molecular Transformer models. We evaluate AI-enabled methodologies across four foundational pillars: biological target identification and validation, virtual screening, de novo molecular generation with multi-objective reinforcement learning, and in silico ADMET/toxicity de-risking. Furthermore, real-world case studies (such as the deep-learning-discovered antibiotic Halicin and GENTRL-designed DDR1 kinase inhibitors) are analyzed quantitatively. Finally, we address key translational roadblocks, including experimental negative-data scarcity, out-of-distribution generalization, synthetic accessibility constraints, and regulatory requirements for explainable AI (XAI).

INTRODUCTION

The pharmaceutical industry faces persistent economic and productivity challenges, commonly referred to as Eroom's Law, where the cost of bringing a novel molecular entity to market continues to escalate despite technological advances. A standard pre-clinical pipeline requires synthesizing and assaying thousands of chemical analogues to isolate a single viable clinical candidate, with early-stage attrition exceeding 90% due to poor pharmacokinetics, off-target toxicity, or suboptimal potency.

The advent of modern Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) methodologies represents a foundational shift from empirical trial-and-error screening to

***Corresponding Author:** Digambar Rahangdale
Address: *Shri Gurudev Institute of Pharmaceutical Education and Research, Gondia, Maharashtra, India 441601*
Email : digambarrahangdale2@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

rational, predictive in silico design. By leveraging massive multi-omics datasets, high-throughput crystallographic repositories, and curated chemical libraries, computational models are capable of mapping non-linear structure-activity relationships, exploring vast expanses of chemical space (~10^60 synthesizable drug-like molecules), and predicting biological interactions with high fidelity.

2. Core Methodologies and Algorithmic Frameworks

The computational toolset employed in modern molecular design spans supervised, unsupervised, and reinforcement learning modalities. Selecting the optimal architecture depends on the representation of the chemical entity—whether as 1D textual sequences (SMILES), 2D topological molecular graphs, or 3D stereochemical atomistic conformations.

Algorithm Class	Representative Architectures	Primary Drug Discovery Workflow	Key Advantage
Traditional ML	Random Forest (RF), SVM, XGBoost	2D QSAR, physicochemical classification, target triage	High baseline accuracy on small, tabular bioassay datasets
Deep Learning (DL)	Multilayer Perceptrons (MLP), Deep Multi-Task Nets	Cross-target affinity profiling, multi-assay prediction	Learns shared feature representations across thousands of targets
Graph Neural Networks	Graph Convolutional (GCN), MPNN, SchNet	Molecular property prediction, 3D binding affinity	Directly preserves topological atom-bond graph connectivity
Generative Models	VAE, GAN, Normalizing Flows	De novo scaffold generation, targeted molecular design	Smooth continuous latent representations of drug-like space
Transformers / LLMs	ChemBERTa, SMILES-Transformer, MolFormer	Sequence-to-sequence lead optimization, retrosynthesis	Captures long-range syntax and contextual chemical grammar

3. Applications Across Early Drug Discovery Pipelines

3.1 Target Identification and Validation

Target identification utilizes network biology, deep autoencoders, and knowledge graph embeddings (e.g., PrimeKG, Hetionet) to integrate multi-omics datasets (genomics, transcriptomics, proteomics, and disease phenomes). These architectures identify critical causal regulatory hubs in disease networks that standard genome-wide association studies (GWAS) often fail to detect due to complex epistasis. Furthermore, deep learning-based structural prediction tools (AlphaFold2, AlphaFold3, ESMFold) have resolved high-confidence 3D structural models for over 200 million proteins, enabling structure-based screening against historically uncharacterized or 'undruggable' targets such as membrane-bound G-protein coupled receptors (GPCRs) and transient allosteric pockets.

3.2 Virtual Screening and Hit Identification

Virtual screening computationally triages millions to billions of chemical compounds to enrich active hit rates prior to physical assaying. While conventional high-throughput screening (HTS) yields hit rates typically between 0.01% and 0.14%, modern deep learning scoring functions and Directed Message Passing Neural Networks

(D-MPNN) achieve enrichment rates exceeding 15% to 50% in verified biological assays.

CASE STUDY 1: Deep Learning Discovery of Structurally Novel Antibiotics (Halicin & Abaucin)

Stokes et al. (2020) trained a Directed Message Passing Neural Network (D-MPNN) on 2,335 diverse molecules to predict growth inhibition against Escherichia coli. The model subsequently screened over 107 million chemical structures from the ZINC15 database and the Drug Repurposing Hub in under four days. Out of 99 computationally prioritized candidates physically tested in wet-lab assays, 51 demonstrated potent growth inhibition (empirical hit rate of 51.5% vs. ~0.05% historical HTS baseline). The lead candidate, Halicin (SU-3327), exhibited exceptional bactericidal activity against pan-resistant pathogens (including Clostridioides difficile and Acinetobacter baumannii) via dissipation of transmembrane electrochemical potential (proton-motive force), bypassing conventional enzymatic resistance mechanisms.

Deep Learning-Driven Discovery of Structurally Novel Antibacterial Therapeutics: The Halicin Paradigm

1. Problem Statement & Clinical Rationale

The clinical utility of conventional antibiotic classes is declining rapidly due to the proliferation of multidrug-resistant (MDR) and extensively drug-resistant (XDR) bacterial strains. The discovery pipeline for novel antibacterial chemotypes has been essentially stagnant for decades, largely because conventional discovery relies on screening natural product extracts or synthetically focused chemical libraries. Traditional HTS campaigns require massive physical infrastructure, incur substantial reagent costs, and yield empirical hit rates typically below 0.05%, with most active leads representing minor modifications of existing antibiotic classes subject to preexisting cross-resistance.

2. Neural Network Architecture and Training Pipeline

To overcome the limitations of fixed molecular fingerprint representations (e.g., Extended Connectivity Fingerprints / ECFP), the research team utilized a graph neural network architecture specifically tailored for molecular property prediction:

Algorithmic Framework: Directed Message Passing Neural Network (D-MPNN / Chemprop). Unlike standard Graph Convolutional Networks that pass messages along nodes (atoms), D-MPNN passes messages along directed bonds, preventing circular message propagation and retaining directional topological stereochemistry.

Training Cohort: The primary training dataset comprised 2,335 structurally diverse molecules, consisting of FDA-approved therapeutics and structurally varied natural products. Each compound was experimentally profiled at a fixed concentration (50 uM) for growth inhibition against wild-type Escherichia coli (strain BW25113).

Binarization Threshold: Molecules inhibiting growth by >=80% relative to control were classified as active hits, establishing a high-confidence binary training matrix.

3. Ultra-Large-Scale In Silico Virtual Screening

Following training and cross-validation (achieving an exceptional ROC-AUC of ~0.896 on blind test sets), the model was deployed across multiple large-scale libraries:

Broad Institute Drug Repurposing Hub: A library of ~6,000 clinical-stage and investigational molecules was computationally evaluated to identify potential repurposing candidates within hours.

ZINC15 In Silico Library: The model screened over 107 million commercially available small molecules. The entire multi-million-compound computational inference was completed in approximately 4 days on standard GPU workstations.

Dual-Filtering Optimization: Compounds were filtered not only for predicted antibacterial potency, but also through an in silico cytotoxicity model (trained on human HepG2 hepatic cell lines) and structural dissimilarity metrics (Tanimoto coefficient < 0.4 against canonical antibiotics) to enforce structural novelty and safety.

4. Pharmacological Profile of Lead Candidate: Halicin

Originally investigated as an inhibitor of c-Jun N-terminal kinase (JNK) for diabetic neuropathies, Halicin was predicted by the D-MPNN model to possess potent antibacterial efficacy despite bearing negligible structural similarity to any known antibiotic class.

Mechanism of Action (MOA): Biochemical assays confirmed that Halicin disrupts the electrochemical transmembrane potential (proton-motive force, delta-psi) by dissipating delta-pH across bacterial cytoplasmic membranes. This disruption selectively collapses ATP synthesis and metabolite transport without inducing rapid cell-wall lysis.

Spectrum of Activity: Halicin demonstrated exceptional in vitro and in vivo efficacy (nanomolar to low micromolar MICs) against critical pathogens, including multidrug-resistant Clostridioides difficile, pan-resistant Acinetobacter baumannii, and Mycobacterium tuberculosis.

Resistance Barrier: Due to its biophysical mechanism of targeting membrane potential rather than a single mutable protein binding pocket, E. coli exhibited zero measurable resistance acquisition across 30 days of continuous serial subculturing at sub-lethal concentrations (in stark contrast to ciprofloxacin, which induced rapid resistance within 3 days).

5. Key Takeaways and Translational Implications

1. High Predictive Accuracy from Modest Training Datasets: By training on only 2,335 carefully curated, high-quality data points, modern message-passing neural networks can learn highly generalizable representations of chemical bioactivity.

2. Expansion of Chemical Space: AI algorithms can effectively navigate structural topologies far removed from traditional pharmacophores, enabling the discovery of entirely novel functional scaffolds.

3. Drastic Reduction in Time and Capital: Pre-filtering multi-million-compound libraries computationally cuts initial hit identification timelines from years to days, fundamentally restructuring early pre-clinical drug discovery economics.

3.3 De Novo Molecule Design and Lead Optimization

Generative models (GANs, VAEs, Diffusion models) navigate continuous mathematical representations of chemical space to assemble

entirely novel scaffolds tailored to target binding pockets. Coupled with Reinforcement Learning (RL) and multi-objective Pareto optimization, these systems optimize binding affinity, Synthetic Accessibility Score (SAScore), Quantitative Estimate of Drug-likeness (QED), and solubility simultaneously.

CASE STUDY 2: Rapid De Novo Kinase Inhibitor Discovery via GENTRL

Zhavoronkov et al. (2019) utilized Generative Tensorial Reinforcement Learning (GENTRL) to design selective non-covalent inhibitors for Discoidin Domain Receptor 1 (DDR1), a kinase involved in fibrotic diseases. The complete workflow—from target formulation, de novo sampling of ~30,000 structures, synthesis prioritization, chemical synthesis of 6 lead candidates, to biological characterization—was accomplished in just 21 days (compared to traditional 2-3 year timelines). Four out of six synthesized compounds demonstrated potent nanomolar inhibition (IC50 = 10 to 21 nM), yielding a 66.7% active hit rate from synthesized leads, with one compound exhibiting favorable in vivo pharmacokinetic stability.

Rapid De Novo Molecular Design and Pre-Clinical Validation of DDR1 Kinase Inhibitors: The GENTRL Paradigm

1. Target Biology and Therapeutic Rationale

Discoidin Domain Receptor 1 (DDR1) is a transmembrane collagen-activated receptor tyrosine kinase critically involved in cell adhesion, proliferation, differentiation, and tissue remodeling. Aberrant activation and dysregulation of DDR1 signaling are strongly implicated in severe pathological conditions, notably idiopathic pulmonary fibrosis (IPF), renal fibrosis, and diverse malignant carcinomas. Although several non-selective multi-kinase inhibitors exhibit off-target DDR1 activity, developing structurally distinct, highly selective small-molecule DDR1 inhibitors with clean pharmacokinetic profiles remains an urgent therapeutic objective.

2. Generative Tensorial Reinforcement Learning (GENTRL) Framework

The GENTRL architecture combines Variational Autoencoders (VAEs), Reinforcement Learning (RL), and tensor-based continuous manifold representations to navigate drug-like chemical space toward targeted pharmacophoric profiles:

Continuous Latent Space Mapping: GENTRL encodes molecular structures into continuous low-dimensional latent vectors using tensor decompositions, allowing smooth probabilistic interpolation between bioactive chemotypes.

Multi-Objective Reward Function: Reinforcement learning policy gradients guide the generator to optimize four concurrent objectives: (a) predicted binding affinity/selectivity against DDR1 kinase, (b) 3D pharmacophore alignment, (c) Synthetic Accessibility Score (SAScore <= 3.5), and (d) Quantitative Estimate of Drug-likeness (QED >= 0.6).

Data Pre-Training: The model was trained on general bioactive chemical libraries from ChEMBL and fine-tuned on targeted kinase structural datasets (including kinase-domain co-crystal structures from PDBbind).

3. De Novo Compound Generation and Lead Prioritization

The complete in silico pipeline generated approximately 30,000 unique drug-like structures within 21 days:

In Silico Filtering: From the 30,000 generated candidate molecules, 40 scaffolds with optimal predicted docking scores, selectivity against non-target kinases, and high synthetic tractability were shortlisted.

5. In Vitro and In Vivo Pharmacological Profiling

The four verified active compounds exhibited nanomolar inhibitory concentration (IC50) values against recombinant DDR1 kinase:

Enzymatic Potency: Lead compounds 1 and 2 demonstrated sub-nanomolar to nanomolar enzymatic potency (IC50 = 10.1 nM and 13.5 nM, respectively), effectively suppressing collagen-induced autophosphorylation in human lung fibroblasts.

Kinome Selectivity: Lead compounds were evaluated against a panel of human protein kinases (KinomeScan), confirming excellent selectivity for DDR1 over closely related kinases (such as DDR2, VEGFR2, and c-Kit).

In Vivo Pharmacokinetics (PK): When administered to C57BL/6 mice, lead compound 1 demonstrated favorable oral bioavailability, high plasma exposure (AUC), and acceptable terminal elimination half-life (T1/2 > 4.5 hours) without acute systemic toxicity.

6. Key Scientific & Translational Takeaways

1. Proof of Concept for De Novo Design: The study provided one of the first rigorous wet-lab validations that reinforcement learning and generative VAEs can design bioactive, synthesizable molecules completely from scratch.

2. Overcoming the Synthetic Bottleneck: By penalizing high SAScores directly within the loss function, GENTRL avoided the common pitfall of generating un-synthesizable molecular 'hallucinations'.

3. Compression of Discovery Timelines: Compressing the target-to-lead phase to under one month dramatically lowers pre-clinical burn rates and accelerates lead optimization cycles.

3.4 In Silico ADMET and Early Toxicity Profiling

Over 40% of drug candidate failures historically arise from unfavorable pharmacokinetics or unexpected organ toxicity. Machine learning models predict absorption, distribution, metabolism, excretion, and toxicity parameters early in the discovery funnel:

Aqueous Solubility and Permeability: Graph neural networks achieve a Mean Absolute Error (MAE) < 0.45 log units on standard benchmarks (FreeSolv, TDC), matching shake-flask experimental assay variance.

Cytochrome P450 (CYP450) Profiling: Multi-task classifiers predict inhibition across the 5 primary metabolic isozymes (CYP1A2, 2C9, 2C19, 2D6, 3A4) with ROC-AUC scores exceeding 0.88–0.93.

Cardiotoxicity (hERG Channel Blockade): Deep ensemble models identify QT-interval prolongation risk with >88% sensitivity and >85% specificity, reducing downstream automated patch-clamp testing load by >60%.

4. Benchmark Datasets and Computational Resources

Standardized benchmark platforms have been critical to the rigorous evaluation and reproducibility of molecular ML algorithms:

ChEMBL: Open bioactivity database containing >2.4 million distinct compound structures and >20 million quantitative binding/functional assay endpoints.

MoleculeNet: Standardized benchmark suite curated across quantum chemistry (QM9), physical chemistry (ESOL, FreeSolv), biophysics (PDBbind), and physiology (Tox21, ClinTox, BBBP).

Therapeutics Data Commons (TDC): Systematic framework for AI-driven therapeutics featuring curated datasets spanning target discovery, small-molecule design, and antibody engineering.

PDBbind: Comprehensive collection of experimentally determined 3D macromolecular crystal complexes with rigorously curated binding constants (Ki, Kd, IC50).

5. Critical Challenges and Translational Roadblocks

Despite rapid technological advances, translating computational hits into clinical development faces several non-trivial bottlenecks:

Data Scarcity and Positive Reporting Bias: The vast majority of published pharmacological literature reports only positive/active hits. The systematic absence of validated inactive assay data creates imbalanced training distributions and limits classifier precision.

Synthetic Accessibility & Feasibility: Generative architectures frequently propose highly complex polycyclic structures, strained bridgehead rings, or chemically unstable motifs that pose prohibitive synthetic barriers in organic synthesis labs.

The 'Black-Box' Problem & Model Interpretability: Regulatory agencies (US FDA, EMA) require clear mechanistic and structural rationales for drug approvals. Explainable AI (XAI) tools like Integrated Gradients, SHAP, and Attention-weights are increasingly necessary to validate that models rely on true pharmacophoric interactions rather than spurious dataset artifacts.

Out-of-Distribution (OOD) Generalization: Models frequently suffer significant performance degradation when applied to novel chemical scaffolds that lie outside the applicability domain of their training datasets.

6. Synthesis of Empirical Results and Performance Metrics

A comparative assessment of traditional empirical pipelines versus modern AI-augmented workflows reveals distinct structural advantages:

Discovery Metric	Traditional Empirical Workflow	AI-Augmented Paradigm
Screening Library Capacity	10^5 to 10^6 physical compounds	>10^9 virtual ultra-large libraries
Average Primary Hit Rate	0.01% – 0.14% in random HTS	1.5% – 51.5% in targeted ML screening
Hit-to-Lead Timeline	24 to 48 months	3 to 12 months
Early Toxicity Filtration	Late in vitro / in vivo assays	In silico multi-task ADMET pre-synthesis

CONCLUSION AND FUTURE DIRECTIONS

Artificial intelligence and machine learning have matured from speculative computational experiments to indispensable cornerstones of early pharmaceutical research. The convergence of generative molecular chemistry, high-accuracy 3D structural determination, and predictive multi-target ADMET profiling has compressed early-

stage hit-to-lead discovery cycles by orders of magnitude.

To realize the full clinical promise of AI-driven drug discovery, the field must continue advancing self-driving automated laboratories (cloud-based closed-loop robotic synthesis and validation), physics-informed neural networks that combine quantum mechanics with deep learning, and rigorous explainable AI frameworks aligned with regulatory guidelines. As these technologies mature, AI will continue to lower developmental costs and accelerate the delivery of life-saving therapeutics to patients worldwide.

REFERENCES

1. Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2(10), 573-584.
2. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.
3. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463-477.
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
5. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology, 37(9), 1038-1040.
6. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. Chemical Science, 9(2), 513-530.
7. Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., ... & Schneider, G. (2020). Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19(5), 353-364.
8. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., ... & Zitnik, M. (2021). Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery. Advances in Neural Information Processing Systems, 34, 188-201.
9. Rifkin, R. M., & Klambauer, G. (2022). Graph neural networks for molecular property prediction: A review. Journal of Chemical Information and Modeling, 62(11), 2600-2615.
10. Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., ... & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268-276.
11. Liu, X., Ye, K., van Vlijmen, H. W., Emmerich, M. T., & IJzerman, A. P. (2021). Drug target prediction using deep learning: Current status and future perspectives. Briefings in Bioinformatics, 22(4), bbaa344.
12. Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.
13. Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022). DiffDock: Diffusion steps, twists, and turns for molecular docking. arXiv preprint arXiv:2210.01776.
14. Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular representations for machine learning in drug discovery. WIREs Computational Molecular Science, 10(6), e1498.
15. Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A. R. (2019). ChEMBL: Towards

direct deposition of bioassay data. Nucleic Acids Research, 47(D1), D930-D940.
16. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.
17. Yang, K., Swanson, K., Jin, W., Gudbas-Rodriguez, C., et al. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370-3388.
18. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology, 37(9), 1038-1040.
19. Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268-276.

HOW TO CITE: Digambar Rahangdale, Aman Yadav, Puja Basule, Dr. Bhumeshkumar Wanjari, Artificial Intelligence and Machine Learning Models in Early Drug Discovery: Current Paradigms, Challenges, and Future Prospects, Int. J. of Pharm. Sci., 2026, Vol 4, Issue 9, 1119-1127. https://doi.org/10.5281/zenodo.22702316