RESEARCH PAPER

Artificial Intelligence in Drug Discovery: Transforming Target Identification to Lead Optimization

Ranjan Banerjee[1], Thangaraj M[2], Rajath N Raju[3], Subham Ghosh[4], Jisu Das[5], Ramandeep Kaur[6], Krishna Raj J S[7]

[1]Assistant Professor, CSE Department, Brainware University, Barasat, Kolkata – 125, West Bengal, India, rbkpccst@gmail.com
[2]I.T.S Institute of Health & Allied Sciences Ghaziabad. thangarajmpt@gmail.com
[3]JSS College of Pharmacy, Mysuru-570015, rajath.1301@gmail.com
[4]Department Of Pharmaceutical Sciences, Rabindranath Tagore University, Hojai, Assam, subhamghosh8927@gmail.com
[5]Research scholar at Assam University, Silchar, djisu212@gmail.com
[6]Department of Electronics and Communication Engineering, School of Engineering and Technology, CGC University, Mohali – 140307, Mohali, Punjab, India ramandeep.j3335@cgcuniversity.in
[7]Associate Professor Community Medicine Department Malabar Medical College Hospital and Research centre, Calicut, Kerala Community Medicine, speaktodrkrishnaraj@gmail.com

ABSTRACT
Artificial intelligence (AI) is reshaping drug discovery by improving how teams prioritize biological hypotheses, triage chemical space, and optimize leads under multi-parameter constraints. Across the early pipeline—from target identification and validation to hit discovery and lead optimization—modern machine learning integrates human genetics, multi-omics, biomedical knowledge graphs, structure prediction, and molecular representation learning to support decision-making in iterative design–make–test–analyze (DMTA) cycles. At the same time, the field faces persistent barriers that determine real-world impact: biased and noisy labels, inflated retrospective benchmarks due to leakage and overly permissive data splits, limited prospective validation, and insufficient uncertainty calibration. This review synthesizes methods and best practices across the target-to-lead workflow, focusing on (i) data foundations and representations, (ii) AI for target discovery and evidence integration, (iii) structure-enabled screening and interaction modeling, and (iv) multi-objective, synthesis-aware lead optimization using predictive and generative models. We conclude with practical recommendations for building credible, auditable AI workflows aligned with emerging regulatory guidance and good AI practice principles.
Keywords: Artificial intelligence; machine learning; target identification; virtual screening; docking; generative models; lead optimization; retrosynthesis; ADMET; DMTA
How to cite this article: Banerjee R, Thangaraj M, Raju RN, Ghosh S, Das J, Kaur R, Krishna Raj JS. Artificial Intelligence in Drug Discovery: Transforming Target Identification to Lead Optimization. Int J Drug Deliv Technol. 2026;16(16s): 65-74. DOI: 10.25258/ijddt.16.16s.8

INTRODUCTION

Drug discovery remains characterized by high attrition and long cycle times, despite major advances in molecular biology, assay automation, and computation. A widely cited diagnosis of declining productivity ("Eroom's law") highlights that the number of new drugs approved per inflation-adjusted R&D dollar has decreased over decades, motivating approaches that improve early decision quality and reduce late-stage failure [1]. In parallel, analyses of development outcomes emphasize that failures are often driven by insufficient efficacy, safety liabilities, and inability to achieve adequate exposure, reinforcing the importance of robust target selection and developability-aware optimization [2].

Within this landscape, AI is best understood as a set of methods that convert heterogeneous data into actionable decisions at each stage of the target-to-lead workflow. A field-defining review described how machine learning can support drug discovery and development across target identification, screening, and optimization, while also

warning that success depends on data quality, appropriate validation, and integration into scientific workflows [3].

Since then, advances in biological data aggregation, structural modeling, and generative chemistry have accelerated progress, but have also surfaced recurring limitations: benchmark inflation from data leakage, weak generalization under distribution shift, and overconfident predictions in regions where training data are sparse.

A practical way to frame the promise of AI is through the DMTA loop that dominates medicinal chemistry operations. In this loop, scientists propose designs, synthesize candidates, test them in a panel of assays, and analyze the results to decide what to do next. AI contributes by (i) compressing evidence at scale (e.g., target evidence across genetics, pathways, and literature), (ii) providing fast surrogate predictions for expensive assays, (iii) proposing candidate sets that improve the trade-off between exploration and exploitation, and (iv) enabling decision policies that explicitly incorporate uncertainty. Importantly, these benefits are system-level: an AI tool that is only marginally better than a baseline predictor can still create large gains if it reliably reduces the number of failed experiments per iteration.

This review focuses on the early discovery pipeline—from target identification to lead optimization—organized around a decision-oriented question: how can AI measurably improve the quality and speed of choices made by discovery teams? We summarize methods, evaluation practices, and translation considerations, and we highlight what is mature enough for routine use versus what still requires careful validation.

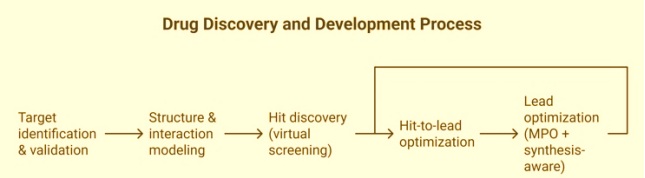

Figure 1. End-to-end AI-enabled drug discovery workflow from target identification to lead optimization, highlighting iterative Design–Make–Test–Analyze (DMTA) feedback.

2. Data foundations and molecular representations

Table 1. Representative AI applications across early drug discovery (target identification to lead optimization).

Stage	Typical inputs	AI tasks	Model families	Practical outputs
Target identification & validation	Genetics, multi-omics, pathways, literature, knowledge graphs	Target prioritization, mechanism inference, biomarker hypothesis	Graph learning, multimodal models, NLP/LLMs	Ranked targets with evidence, testable hypotheses
Structure & interaction modeling	Protein sequences/ structures, complexes, pockets	Structure prediction, pocket inference, pose generation, scoring	Protein language models, equivariant nets, diffusion models	Binding hypotheses, docking triage, mechanism insights
Hit discovery	Libraries, assay data, target structure	Virtual screening, docking/rescoring, hit expansion	GNN/transformers, ML rescoring, docking surrogates	Enriched hit lists, predicted poses
Hit-to-lead	SAR series, selectivity panels, early ADMET	Potency/selectivity prediction, prioritization, active learning	Multi-task DL, uncertainty estimation	Next-to-make list, reduced iterations
Lead optimization	MPO endpoints, synthesis constraints, PK/safety	Generative design, Pareto optimization, synthesis-aware filtering	Generative models, Bayesian optimization, retrosynthesis models	Pareto set of feasible optimized leads

AI performance in drug discovery is often limited less by model architecture than by the availability, quality, and governance of training data. Public bioactivity and annotation resources—including ChEMBL [4], DrugBank [5], PubChem [6], BindingDB [7], and UniProt [8]—

provide widely used starting points for model development and benchmarking, while project-specific screening and ADMET datasets (often proprietary) typically determine performance in deployment.

2.1 Data modalities across the pipeline. Target discovery draws on human genetics, functional genomics, pathway biology, and literature evidence. Screening and optimization rely on compound structures, assay outcomes, and protein/complex structures. Each modality has characteristic failure modes: genetics is sparse but often causal; literature is broad but biased toward fashionable targets; high-throughput assays can be noisy with systematic artifacts; and ADMET endpoints may vary by species, protocol, and lab conditions. These differences explain why "one model to rule them all" rarely works—models should be matched to decisions and supported by domain-specific quality controls.

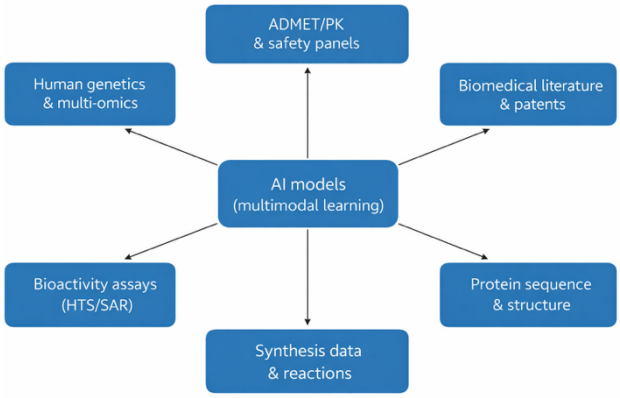

Figure 2. Core data modalities that feed AI models across the target-to-lead pipeline and motivate multimodal representation learning.

2.2 Molecular and biological representations. Small molecules can be represented as strings (e.g., SMILES), vectors (fingerprints), graphs (atoms and bonds), or 3D conformers. SMILES remains a standard encoding that supports both cheminformatics toolchains and modern language-model approaches [9]. However, generation from SMILES can produce invalid strings; SELFIES is a robust alternative that guarantees syntactic validity and helps stabilize generative workflows [10]. On the learning side, graph neural networks (GNNs) can encode chemical topology and local environments, while 3D representations support tasks that depend on geometry (binding poses, conformational strain, and steric clashes). For proteins, sequence embeddings can support functional prediction, while structures and complexes enable pocket-aware modeling and interaction reasoning.

2.3 Benchmark datasets and what they measure. MoleculeNet has influenced common tasks, metrics, and split strategies for molecular property prediction [11]. Therapeutics Data Commons (TDC) extends this idea by curating additional tasks and standardized interfaces for drug discovery modeling [12]. These benchmarks are valuable for methodology comparison, but they often under-represent the realities of a project: dataset shift over time, series-level correlations, and multi-objective decisions. Therefore, benchmark performance should be treated as a screening signal rather than as decision-grade evidence.

2.4 Leakage, bias, and distribution shift. Random splits frequently leak close analogs into both training and test sets, inflating metrics. Scaffold-based splitting, commonly defined with Bemis–Murcko frameworks [13], reduces but does not eliminate this issue, because molecules with different scaffolds can remain close in chemical space. A large-scale analysis demonstrated that even scaffold splits can overestimate virtual screening performance and recommended more realistic splitting approaches for model selection and comparison [14]. In deployment, drift is common: the chemical series evolves, assay protocols change, and the model's effective domain shifts. Robust workflows therefore include monitoring and periodic re-training, together with uncertainty estimates that signal when predictions are likely extrapolative.

2.5 Practical evaluation checklist. For discovery decision support, evaluation should mirror intended use: time-based splits approximate forward deployment; target-based splits approximate transfer to new biology; and external validation across sources or laboratories is valuable when feasible. Reporting should include split definitions, leakage checks, calibration assessment, and error analysis by chemotype and endpoint. Finally, evaluation should connect to operational outcomes: fewer synthesis cycles to reach potency, fewer late-stage failures due to predictable developability issues, and higher hit rates for screened libraries.

3. AI for target identification and validation

Target identification and validation benefit most from methods that integrate heterogeneous evidence: genetics, pathway biology, perturbation screens, and literature knowledge. The Open Targets Platform aggregates disease–gene evidence and supports systematic prioritization, increasingly used for target selection and hypothesis framing [15]. However, target selection is not solely a ranking problem: it must also address tractability (can the target be modulated?), safety (is there a feasible therapeutic window?), and translational relevance (will modulation affect the right patient biology?). AI methods can contribute to each of these questions, but only if they produce interpretable evidence suitable for expert review.

3.1 Human genetics and functional genomics. Genetic association can strengthen causal hypotheses, but translating loci to actionable targets requires variant-to-gene linking and functional interpretation. Deep learning models for regulatory genomics, such as DeepSEA, can predict the functional impact of noncoding variation and help prioritize plausible regulatory mechanisms [16]. Models integrating long-range sequence interactions, such as Enformer, improve gene expression prediction from sequence and support hypotheses about regulatory control [17]. At the protein level, proteome-wide missense variant effect prediction (AlphaMissense) can help connect genotype to molecular phenotype, potentially flagging targets with strong intolerance patterns or identifying functionally important residues relevant to mechanism and resistance [18]. In practice, genetics-informed pipelines often combine AI scores with biological priors and orthogonal evidence (expression in relevant cell types, pathway coherence) before proceeding to experimental validation.

3.2 Knowledge graphs and network learning. Biomedical knowledge graphs (KGs) connect genes, diseases, drugs, phenotypes, and pathways, enabling link prediction and path-based reasoning for target discovery. Hetionet demonstrated that systematic integration of biomedical knowledge can prioritize repurposing hypotheses and reveal mechanistic connections across modalities [19]. More recent efforts build KGs tailored for target discovery (e.g., TarKG), reflecting a growing emphasis on target-centric evidence integration and benchmarking [20]. In target selection, KG-based methods can identify disease modules, highlight network proximity to known causal genes, and suggest biomarkers or combination strategies. Nonetheless, KGs inherit literature bias and often conflate association with causality; therefore they should be used as hypothesis generators with clear plans for experimental tests.

3.3 Transcriptomic perturbation signatures. The Connectivity Map established the idea of linking compounds, genes, and diseases via gene-expression signatures [21], and the L1000 platform enabled large-scale perturbational profiling and matching [22]. In target discovery, these resources support pathway inference and mechanism-of-action hypotheses, and they can suggest whether modulating a pathway moves a disease signature toward a healthier state. Key limitations include cell-context dependence (a signature in one cell line may not translate to a diseased tissue), confounding by toxicity signatures, and batch effects, all of which require careful normalization and control.

3.4 From prioritization to validation. The most useful AI outputs are not simply "a target list," but a package of evidence and a validation strategy: predicted biomarkers, relevant pathways, and experiments that can quickly de-risk causality and engagement. Validation often proceeds through (i) genetic perturbation and rescue experiments, (ii) orthogonal confirmation across model systems, and (iii) early safety hypotheses based on expression patterns and pathway roles. AI can help by ranking experiments by expected information gain, thereby reducing the number of iterations needed to establish whether a target is worth pursuing.

4. AI-enabled structure and interaction modeling

Structure-enabled modeling provides a bridge from biological target hypotheses to chemical design, but must be framed with realistic expectations about protein flexibility and binding energetics. The most consistent value arises when structure-based AI narrows a search space, proposes plausible interaction hypotheses, and supports prioritization—rather than claiming precise affinity estimates in all settings.

4.1 Protein structure prediction and target modeling. AlphaFold2 produced highly accurate protein structure predictions for many targets and catalyzed broad adoption of predicted structures for hypothesis generation [23]. Complementary approaches such as RoseTTAFold [24] and ESMFold [25] expanded accessibility and throughput. AlphaFold3 extended modeling toward biomolecular interactions with a diffusion-based architecture, including complexes that may contain proteins, nucleic acids, ions, and small molecules [26]. In discovery practice, predicted structures can support pocket identification, mutation interpretation, and structure-guided library design. However, predicted structures usually represent one conformation (or a limited set) and may not capture ligand-induced rearrangements, allostery, or functionally relevant dynamics. For targets with multiple functional states, ensembles derived from experimental structures, homology, or simulations are often needed to avoid overfitting design decisions to a single receptor snapshot.

4.2 Docking, scoring, and binding prediction with ML. Docking remains a practical baseline for structure-based screening; AutoDock Vina is widely used due to speed and accessibility [27]. Learning-based scoring and rescoring aim to improve ranking and reduce false positives. Many models train on curated protein–ligand complex datasets such as PDBbind [28]. GNINA integrates deep learning with docking workflows to improve scoring and pose evaluation [29]. Drug–target affinity prediction models, including DeepDTA, learn from sequence and ligand representations to predict binding strength and are often used for triage and ranking when calibrated and evaluated for the intended context of use [30].

4.3 Diffusion and geometric methods for pose generation. Geometric deep learning approaches (e.g., EquiBind)

predict binding structures by modeling 3D constraints and symmetries [31], while diffusion-based methods such as DiffDock model pose generation as a probabilistic denoising process [32]. These approaches can complement docking by generating diverse pose hypotheses and by learning scoring functions that respect geometric invariances. More broadly, equivariant diffusion models are being explored for structure-based design and binding-aware generation [33]. For practical deployment, the relevant benchmark is prospective impact: do these methods increase hit rates at a fixed screening budget or reduce the number of design cycles needed to reach potency/selectivity targets?

4.4 Practical limitations and best practices. The gap between structural prediction and medicinal chemistry impact typically comes from: (i) protein flexibility and induced fit, (ii) uncertainty in protonation states and tautomeric forms, (iii) water networks and entropic effects, and (iv) assay-dependent readouts that may not correlate tightly with computed scores. Best practices therefore include calibration with known ligands or fragments, ensemble docking across plausible conformations, and consensus ranking from multiple scoring functions. In addition, it is often more reliable to use structure-based methods for *prioritization* rather than absolute affinity prediction: selecting a diverse set of plausible binders to test experimentally, learning from failures, and updating models accordingly.

5. Hit discovery and hit-to-lead optimization

Table 2. Evaluation and reporting checklist for decision-grade AI in drug discovery (recommended minimum).

Item	What to report	Why it matters
Split strategy	Scaffold/time/target splits aligned to context of use	Reduces leakage; matches deployment reality
Data provenance	Sources, curation rules, assay metadata, deduplication	Reproducibility and bias control
Baselines	Docking-only + classical QSAR/ML baselines	Ensures gains are meaningful
Uncertainty	Calibration, ensembles/conformal, abstention thresholds	Safer decisions; flags OOD cases
External validation	Independent dataset/lab or later campaign	Tests generalization
Prospective testing	New compounds/screen; pre-defined success metrics	Strongest utility evidence
Error analysis	Failures by scaffold/series/target class; noise sensitivity	Guides improvement; avoids blind spots
Synthesis feasibility	Retrosynthesis filter/constraints and route assumptions	Prevents unmakeable proposals

Hit discovery aims to identify chemically tractable starting points with measurable activity and acceptable early developability signals, while hit-to-lead refines potency and selectivity with early ADMET risk reduction. AI is widely used here because decisions are frequent, data accumulates rapidly, and even modest reductions in failed experiments can translate into meaningful time and cost savings.

5.1 Ligand-based virtual screening and QSAR. Modern QSAR increasingly uses representation learning rather than fixed descriptors. Analyses of directed message passing networks (Chemprop/D-MPNN) illustrate how learned representations can improve performance across endpoints and provide practical tooling widely adopted in medicinal chemistry [34]. Attention-based GNNs such as AttentiveFP further show how attention mechanisms can highlight substructures associated with activity or properties, supporting qualitative interpretation and SAR reasoning [35]. Molecular language models, exemplified by ChemBERTa, extend self-supervised pretraining to chemistry and can improve data efficiency when fine-tuned on project assays [36]. In practice, ligand-based models are most useful when paired with careful split strategies (e.g., time or series splits) and with uncertainty estimates that indicate when the model is extrapolating beyond known chemical space.

5.2 Structure-based virtual screening as triage. In many workflows, docking generates poses and baseline scores; ML-based rescoring improves ranking; and uncertainty-aware methods prioritize candidates most likely to yield experimentally confirmed hits. Screening success should be measured by early enrichment (how many actives appear in the top-ranked fraction) and hit rate (fraction confirmed), rather than by global metrics alone. For highly imbalanced screening problems, precision–recall curves and top-k enrichment can be more informative than ROC-AUC. Operationally, the goal is to reduce the number of compounds that must be synthesized or purchased to obtain a fixed number of confirmed hits.

5.3 Hit expansion and SAR navigation. After initial hits, teams explore analogs to map structure–activity relationships (SAR) and identify tractable vectors for

optimization. Predictive models can propose substitutions likely to improve potency while maintaining physicochemical balance, and can help prioritize which regions of a molecule to explore next. The most useful models in this phase typically optimize for ranking within a chemical series and for identifying uncertainty when proposing more radical modifications. Incorporating constraints such as available building blocks, synthetic complexity, and project timelines improves practical relevance.

5.4 Early developability: ADMET prediction and multi-task modeling. Many programs fail due to pharmacokinetics or safety liabilities; therefore early integration of ADMET prediction can reduce wasted cycles. ADMETlab 3.0 provides a recently updated resource for ADMET prediction and illustrates community movement toward standardized, integrated pipelines [37]. In project settings, multi-task models that jointly predict potency, selectivity, and key ADMET endpoints can support balanced decisions, but require careful handling of missing labels and assay heterogeneity. A pragmatic approach is to treat ADMET predictions as risk flags and triage criteria rather than as precise numerical forecasts.

5.5 Uncertainty and active learning in DMTA cycles. The highest impact of AI in hit-to-lead is often experimental prioritization. Active learning selects compounds that balance expected improvement with uncertainty to learn efficiently, particularly when expensive assays (e.g., safety panels) are bottlenecks. Calibrated uncertainty helps avoid overconfident extrapolation as chemotypes shift over time and enables explicit policies (e.g., abstain or request new data) when predictions fall outside the domain of applicability.

6. Lead optimization: generative design and multi-objective optimization

Lead optimization is a multi-objective search problem constrained by synthetic feasibility and project timelines. AI contributes through predictive modeling, generative proposal, and search strategies that optimize across endpoints. The central shift is from single-endpoint QSAR toward integrated decision support that jointly considers potency, selectivity, ADMET, and makeability.

6.1 Generative design methods. Generative modeling for inverse molecular design was highlighted as a path toward proposal-driven discovery, with caveats about constraints and evaluation [38]. Reinforcement-learning approaches demonstrated goal-directed de novo design by directly optimizing scoring functions [39]. Continuous latent-space approaches (e.g., variational autoencoders) enable interpolation and optimization in embedded spaces and can support local exploration around a lead series [40]. Structure-aware graph generators such as junction-tree VAEs improve chemical validity by composing molecules from substructures and enforcing realistic graph structure [41]. MolGAN illustrates implicit graph generation [42], and policy-based graph generation methods (e.g., GCPN) frame molecule construction as a sequential decision process [43].

6.2 Multi-parameter optimization (MPO) and Pareto trade-offs. Lead optimization rarely has a single “best” molecule; it has trade-offs. A practical MPO setup defines target ranges (e.g., potency threshold, clearance below a limit, acceptable solubility) and uses desirability functions or constrained optimization to rank proposals. Pareto front thinking is useful: if a candidate improves potency but worsens clearance, it may still be valuable depending on program priorities. AI models support MPO by providing fast multi-endpoint predictors and by proposing candidate sets that reflect desired trade-offs while maintaining diversity.

6.3 Benchmarking generative design and avoiding ‘gaming’. Benchmarks such as GuacaMol [44] and MOSES [45] provide standardized evaluation for validity, novelty, and distribution matching. However, these metrics can be optimized in ways that do not yield medicinally useful compounds. Therefore, practical evaluation should include: synthetic feasibility, stability and reactivity heuristics, novelty relative to internal IP constraints, and predicted risk flags (e.g., liabilities). In many organizations, the most effective use of generative models is to propose a *shortlist* for expert review, rather than to automate final selection.

6.4 Synthesis-aware design and retrosynthesis planning. Neural and hybrid systems for synthesis planning have advanced rapidly. Planning chemical syntheses with deep neural networks and symbolic search showed that learned one-step models can be combined with search to propose plausible routes [46]. The Molecular Transformer achieved strong performance for reaction prediction with uncertainty calibration, supporting more reliable reaction planning [47]. AiZynthFinder provides an open-source retrosynthesis planner that is widely used to assess makeability and propose routes [48], and ASKCOS has evolved into an open-source synthesis planning suite integrating multiple one-step models and search strategies [49]. Integrating retrosynthesis feasibility as a filter (or objective) helps ensure that AI-proposed leads can be made within reasonable time and cost.

6.5 Closed-loop optimization in DMTA. The highest leverage emerges when models are updated as new assay data arrive and when active learning selects the next experiments. Closed-loop systems aim to reduce the number of syntheses needed to reach a multi-objective target profile and to increase the probability that a final lead meets developability gates. Success metrics should therefore

include cycle time reduction, fewer syntheses per milestone, and improved probability of meeting potency/ADMET/selectivity thresholds, not only improved retrospective accuracy.

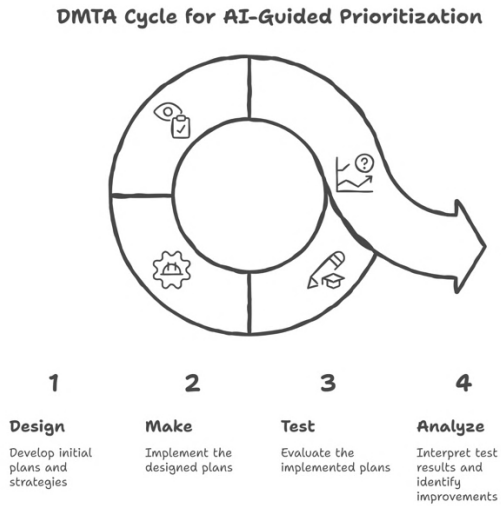

Figure 3. AI-enhanced Design–Make–Test–Analyze (DMTA) loop with uncertainty-aware prioritization to reduce iterations and improve lead quality.

6.6 Common pitfalls and mitigation. Several recurring pitfalls limit the usefulness of AI in lead optimization. First, models may optimize toward "easy" proxies (e.g., predicted potency) while silently degrading properties that are not modeled or poorly labeled. Second, generators can drift toward synthetically complex or unstable structures unless constrained by retrosynthesis feasibility and medicinal chemistry heuristics. Third, optimization can collapse diversity, producing many near-duplicates that add little information gain. Mitigation strategies include enforcing diversity constraints, explicitly modeling key liabilities, and using human-in-the-loop triage with clear acceptance criteria. Finally, teams should measure success in operational terms: how many cycles were saved, how quickly the Pareto frontier moved, and whether the final lead met pre-specified developability gates.

7. Translation, governance, and regulatory considerations

Table 3. Illustrative prospective or translational examples of AI contributing to drug discovery and development.

Example	AI contribution	Evidence type
Deep learning–guided antibiotic discovery (halicin)	Model-driven screening/prioritization from large chemical space	Prospective experimental validation
Generative AI–discovered TNIK inhibitor (rentosertib/ISM001-055)	Generative design + experimental progression reported in clinical study	Clinical/translational report
Regulatory-qualified AI DDT for MASH (AIM-NASH)	AI standardization of pathology readouts to support trials	Regulatory tool qualification

Translation of AI from retrospective benchmarks to robust discovery impact depends on credible evaluation, reproducibility, and governance aligned with context of use.

7.1 Evidence hierarchy: retrospective to prospective. Retrospective benchmarks are useful for iteration and for comparing methods under controlled settings, but prospective validation—new compounds prioritized by the model and tested in experiments—provides stronger evidence of practical value. The risk of inflated performance from permissive splits and residual similarity in scaffold splits reinforces the need for realistic evaluation designs and cautious interpretation of high benchmark scores [14]. A pragmatic evidence hierarchy for discovery teams is: (i) internal time-split evaluation within a program, (ii) external or cross-source validation, (iii) pseudo-prospective testing on held-out campaigns, and (iv) fully prospective studies where the model drives selection and experimental confirmation. For screening tasks, reporting should include enrichment at low false-positive rates and hit rates at realistic budgets; for optimization tasks, reporting should connect model use to operational outcomes such as fewer syntheses to reach a target profile or fewer late-stage failures due to predictable developability issues.

7.2 Reproducibility and reporting. Minimum reporting should include data provenance, curation steps, split protocols, baseline comparators, and error analysis. For deployed models, monitoring for drift and recalibration are essential as assay protocols and chemical space evolve. Practical teams often implement model cards and data cards, define approval processes for training data inclusion, and log decision outcomes (what was proposed, what was synthesized, what succeeded/failed) to learn whether the model is improving operations rather than simply improving retrospective metrics.

7.3 Regulatory credibility and good AI practice. Regulators have begun to articulate expectations for AI models used to generate evidence intended to support regulatory decisions. The U.S. FDA issued a draft guidance on the use of AI to support regulatory decision-making for drugs and biologics,

proposing a risk-based credibility assessment framework tied to a defined context of use [50]. EMA and FDA jointly published guiding principles for good AI practice in drug development (January 2026), emphasizing human-centric design, data governance, clear context of use, risk-based performance assessment, and life-cycle management [51]. EMA also published a reflection paper on AI in the medicinal product lifecycle, highlighting governance and validation considerations [52]. Even when AI is used only for early discovery decisions, these principles are helpful because project narratives increasingly incorporate model-based analyses when selecting candidates and justifying development strategies.

7.4 Practical governance for discovery teams. Governance-ready AI workflows include documentation, access control for datasets, versioned training pipelines, and clearly defined decision policies (including uncertainty thresholds and human review). Governance is also cultural: teams must agree on when the model is trusted, when it must be challenged, and how to learn from failures without overfitting to the most recent outcomes. A concise governance checklist is: define the decision the model supports; define the acceptable error and risk; validate on realistic splits; calibrate uncertainty; monitor drift; and keep a complete audit trail of inputs, model versions, and decisions.

8. Future directions and conclusion

Several trends are likely to shape the next phase of AI-enabled discovery. First, multimodal models that unify text, sequences, structures, and chemistry will enable richer evidence integration and smoother tool orchestration across the pipeline. The key constraint will be trust: these models must be coupled to retrieval and provenance so that scientific claims can be traced to underlying evidence rather than generated as unsupported text. Second, geometry-aware generation and binding-aware design (including diffusion-based approaches) may improve the physical plausibility of proposed molecules and poses, but must be evaluated prospectively to demonstrate that they increase hit rates or reduce DMTA cycles. Third, closed-loop experimentation and laboratory automation will shift emphasis from static model accuracy to system-level learning efficiency and robustness—how quickly a system learns a useful SAR model from new experiments and how reliably it avoids unproductive synthesis.

Real-world traction is increasingly visible in prospective demonstrations. Deep learning–guided screening identified a structurally novel antibiotic candidate (halicin) with broad activity in murine models, illustrating prospective enrichment beyond conventional heuristics [53]. More recently, a generative-AI discovered TNIK inhibitor (rentosertib; formerly ISM001-055) was reported in a randomized phase 2a trial in idiopathic pulmonary fibrosis, providing a rare example of an AI-enabled discovery narrative extending into clinical evaluation [54]. These examples should be interpreted cautiously—success is multi-factorial and not attributable to a single algorithm—but they illustrate the direction of travel: AI is most persuasive when coupled to disciplined experimental validation, transparent reporting, and synthesis-aware constraints.

In conclusion, AI is transforming target-to-lead discovery by improving evidence integration for target selection, accelerating hit identification, and enabling multi-objective, synthesis-aware lead optimization. The most durable impact will come from integrating AI into DMTA systems with realistic evaluation, calibrated uncertainty, and governance aligned to context of use.

References

1. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191–200. doi:10.1038/nrd3681.

2. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B. 2022;12(7):3049–3062.

3. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477. doi:10.1038/s41573-019-0024-5.

4. Zdrazil B, Felix E, Hunter F, et al. The ChEMBL Database in 2023/2024: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D1180–D1192. doi:10.1093/nar/gkad1004.

5. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265–D1275. doi:10.1093/nar/gkad976.

6. Kim S, Chen J, Cheng T, et al. PubChem 2025 update. Nucleic Acids Res. 2025;53(D1):D1516–D1525. doi:10.1093/nar/gkae1059.

7. Liu T, Lin Y, Wen X, et al. BindingDB in 2024/2025: a FAIR knowledgebase of protein–small molecule binding data. Nucleic Acids Res. 2025;53(D1):D1633–D1644. doi:10.1093/nar/gkae1075.

8. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–D531. doi:10.1093/nar/gkac1052.

9. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–36. doi:10.1021/ci00057a005.

10. Krenn M, Häse F, Nigam A, et al. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. Mach Learn Sci Technol. 2020;1:045024 (preprint 2019).
11. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513–530. doi:10.1039/C7SC02664A.
12. Huang K, Fu T, Gao W, et al. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. arXiv. 2021. doi:10.48550/arXiv.2102.09548.
13. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem. 1996;39(15):2887–2893. doi:10.1021/jm9602928.
14. Guo Q, Hernandez-Hernandez S, Ballester PJ. Scaffold splits overestimate virtual screening performance. arXiv. 2024. arXiv:2406.00873.
15. Ochoa D, Hercules A, Carmona M, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023;51(D1):D1353–D1359. doi:10.1093/nar/gkac1046.
16. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–934. doi:10.1038/nmeth.3547.
17. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021. doi:10.1038/s41592-021-01252-x.
18. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492. doi:10.1126/science.adg7492.
19. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6:e26726. doi:10.7554/eLife.26726.
20. Zhou C, et al. TarKG: a comprehensive biomedical knowledge graph for target discovery. Bioinformatics. 2024;40(10):btae598. doi:10.1093/bioinformatics/btae598.
21. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–1935. doi:10.1126/science.1132939.
22. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–1452.e17. doi:10.1016/j.cell.2017.10.049.
23. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589. doi:10.1038/s41586-021-03819-2.
24. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021. doi:10.1126/science.abj8754.
25. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023. doi:10.1126/science.ade2574.
26. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500. doi:10.1038/s41586-024-07487-w.
27. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–461. doi:10.1002/jcc.21334.
28. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015;31(3):405–412. doi:10.1093/bioinformatics/btu626.
29. McNutt AT, Francoeur P, Aggarwal R, et al. GNINA 1.0: molecular docking with deep learning. J Cheminform. 2021;13:43. doi:10.1186/s13321-021-00522-2.
30. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821–i829. doi:10.1093/bioinformatics/bty593.
31. Stärk H, Ganea O-E, Pattanaik L, Barzilay R, Jaakkola T. EquiBind: geometric deep learning for drug binding structure prediction. arXiv. 2022. doi:10.48550/arXiv.2202.05146.
32. Corso G, Stärk H, Jing B, et al. DiffDock: diffusion steps, twists, and turns for molecular docking. arXiv. 2022/2023. doi:10.48550/arXiv.2210.01776.
33. Schneuing A, et al. Structure-based drug design with equivariant diffusion models. Nat Comput Sci. 2024. doi:10.1038/s43588-024-00737-x.
34. Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370–3388. doi:10.1021/acs.jcim.9b00237.
35. Xiong Z, Wang D, Liu X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2019;62(22):10082–10089. doi:10.1021/acs.jmedchem.9b00959.
36. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv. 2020. doi:10.48550/arXiv.2010.09885.
37. Xiong G, Wu Z, Yi J, et al. ADMETlab 3.0: a comprehensive platform for accurate ADMET prediction.

Nucleic Acids Res. 2024;52(W1):W422–W431. doi:10.1093/nar/gkae222.
38. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. Science. 2018;361(6400):360–365. doi:10.1126/science.aat2663.
39. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de novo design through deep reinforcement learning. J Cheminform. 2017;9:48. doi:10.1186/s13321-017-0235-x.
40. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–276. doi:10.1021/acscentsci.7b00572.
41. Jin W, Barzilay R, Jaakkola T. Junction Tree Variational Autoencoder for molecular graph generation. arXiv. 2018. doi:10.48550/arXiv.1802.04364.
42. De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs. arXiv. 2018. doi:10.48550/arXiv.1805.11973.
43. You J, Liu B, Ying Z, Pande V, Leskovec J. Graph convolutional policy network for goal-directed molecular graph generation. arXiv. 2018. doi:10.48550/arXiv.1806.02473.
44. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096–1108. doi:10.1021/acs.jcim.8b00839.
45. Polykovskiy D, Zhebrak A, Vetrov D, et al. Molecular Sets (MOSES): a benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644.
46. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018;555:604–610. doi:10.1038/nature25978.
47. Schwaller P, Laino T, Gaudin T, et al. Molecular Transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572–1583. doi:10.1021/acscentsci.9b00576.
48. Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum EJ. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform. 2020;12:70. doi:10.1186/s13321-020-00472-1.
49. Tu Z, et al. ASKCOS: open-source, data-driven synthesis planning. (PubMed: 40397546). 2025.
50. FDA. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Draft guidance. 6 Jan 2025.
51. FDA & EMA. Guiding principles of good AI practice in drug development. January 2026.
52. EMA. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. 2024.
53. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702.e13. doi:10.1016/j.cell.2020.01.021.
54. Xu Z, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. Nat Med. 2025;31(8):2602–2610. doi:10.1038/s41591-025-03743-2.

.