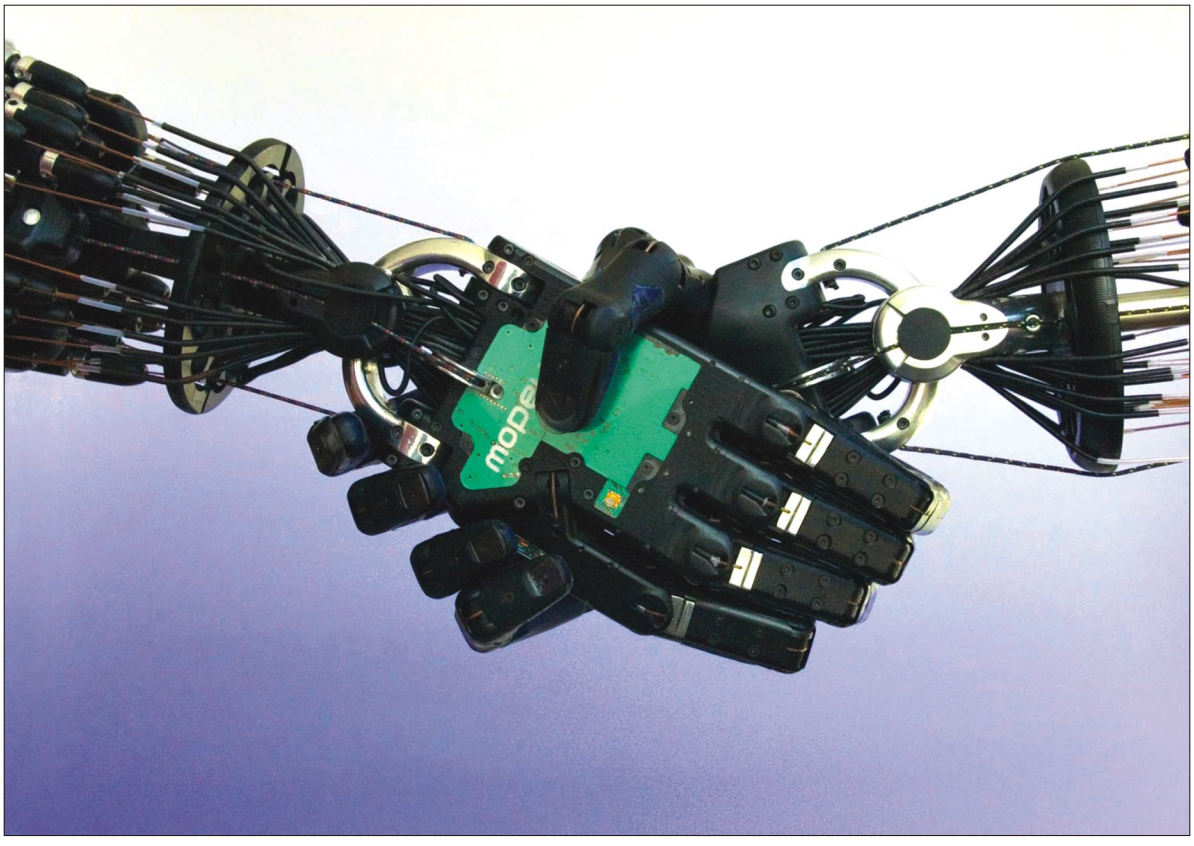


Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2027.1.14

Review Article
Artificial Intelligence in Drug Discovery: Progress, Prospects and Pitfalls

Tushar Gaikwad, Nachiket Atale and Ulhas Patil

Department of Microbiology, Institute of Science, Dr. Homi Bhabha State University, Mumbai, Maharashtra, India

Abstract

Artificial intelligence (AI) is rapidly reshaping drug discovery by accelerating key stages of the pipeline, reducing attrition and enabling exploration of complex chemical and biological spaces beyond the reach of traditional methods. This review provides a concise overview of eight major methodological domains: Structure-based virtual screening, QSAR modelling, ADMET prediction, generative molecular design, AlphaFold-integrated discovery, drug repurposing, retrosynthesis and the datasets and benchmarks that support these approaches. The evolution from classical machine learning methods to modern deep learning paradigms, including graph neural networks, transformer architectures, diffusion models and multimodal foundation models are aptly outlined in this study. Recent advances such as structure-aware affinity prediction, pocket-conditioned molecule generation, knowledge graph-based repurposing and synthesis-aware design are highlighted. Also, key challenges including benchmark bias, limited data accessibility, model interpretability and the gap between benchmark and real-world performance are critically evaluated. The review emphasizes the emerging solutions such as improved evaluation protocols, multimodal learning and integration with automated experimental platforms. AI is transitioning from a supportive tool to a central component of drug discovery. Despite persisting challenges, ongoing advancements in methodologies and infrastructure are establishing AI as a key catalyst for future therapeutic innovation.

Key words: Artificial intelligence, drug discovery, graph neural networks, molecular docking, QSAR, generative molecular design

Citation: Gaikwad, T., N. Atale and U. Patil, 2027. Artificial intelligence in drug discovery: Progress, prospects and pitfalls. J. Artif. Intel., 20: 1-14.

Corresponding Author: Ulhas K. Patil, Department of Microbiology, Institute of Science, Dr. Homi Bhabha State University, Mumbai, Maharashtra, 400032, India Tel: +91 9422260313

Copyright: © 2027 Gaikwad *et al.* This is an open access article distributed under the terms of the creative commons attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Developing a new drug from scratch remains one of the most complex and resource-intensive endeavors in modern science. The process is further compounded by critical ethical considerations and regulatory and legal challenges related to data privacy, algorithmic bias, intellectual property and patient safety. On average, bringing a single drug to market takes more than a decade and can exceed $2.6 billion when accounting for failed candidates[1]. Despite this investment, the overall success rate remains discouragingly low; Fewer than 10% of compounds entering clinical trials ultimately reach patients[2]. A significant portion of this attrition occurs during the early and preclinical stages, particularly in hit identification, lead optimisation and toxicity assessment. One of the fundamental challenges lies in the vastness of chemical space. Estimates suggest that the number of potential drug-like molecules ranges from 10^{23} to 10^{60}, yet only a tiny fraction can be experimentally screened. Even when promising candidates are identified, accurately predicting their pharmacokinetic and safety properties commonly denoted as ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) remains difficult. In fact, poor ADMET characteristics are responsible for approximately 40% of late-stage clinical failures[3].

In this regard, Artificial Intelligence (AI) has become a transformative force in drug discovery, enabling significant scientific advancements. Unlike conventional rule-based approaches, machine learning models can uncover complex and non-linear relationships between molecular structure and biological activity, often surpassing human intuition[5]. These capabilities are particularly valuable in navigating high-dimensional chemical space and accelerating decision-making across the drug discovery pipeline.

Early AI-based studies primarily employed classical machine learning methods, such as support vector machines and random forest algorithms, for predicting biological activity[4,5]. More recent AI approaches have further accelerated progress in activity prediction by integrating advanced computational and deep learning methodologies. Convolutional neural networks enabled richer molecular representations, while recurrent neural networks facilitated the generation of valid molecular structures using SMILES notation. More recently, Graph Neural Networks (GNNs) have become a dominant paradigm, offering a natural way to encode molecular topology and capture atom-level interactions[6,7].

Technological breakthroughs have further accelerated this transformation. Transformer architecture introduced self-attention mechanisms capable of modelling long-range dependencies within molecular graphs[8]. The release of AlphaFold2 revolutionized structural biology by providing high-accuracy protein structures at scale, enabling more reliable structure-based drug design[9,10]. At the same time, diffusion-based generative models have opened new avenues for de novo molecular design by directly generating three-dimensional ligand structures conditioned on protein binding sites[11-13].

Although significant AI-based progress has been achieved, several critical challenges persist. The critical pitfall includes quality and bias of training datasets. Many widely used benchmarks suffer from analogue bias, where training and test sets contain highly similar compounds, leading to overly optimistic performance estimates[14-16]. Efforts to address these issues including scaffold-based data splitting and the development of more diverse datasets such as BigBind and MolData highlight the importance of robust evaluation practices[17,18]. Ultimately, the reliability of AI models is intrinsically tied to the quality of the data on which they are trained[19-21].

More recently, the rise of large-scale foundation models has introduced new models for molecular representation learning. Pretrained models, trained on vast chemical corpora, can be adapted to downstream tasks with minimal labelled data, offering improved generalization and efficiency[6,7,19,20]. Furthermore, multimodal approaches that combine chemical language models with graph-based and structural representations are beginning to demonstrate superior performance, suggesting a promising direction for future research[22,23].

To provide a comprehensive and methodological understanding of AI applications in pharmaceutical research, the objectives of this review includes (i) Examine major AI-driven areas in drug discovery, - like datasets and benchmarks, structure-based virtual screening, ligand-based QSAR and property prediction, ADMET and toxicity modelling, generative molecular design, AlphaFold-integrated structural workflows, drug repurposing and AI-assisted retrosynthesis, (ii) Assess recent algorithmic developments along with their practical limitations in preclinical applications, (iii) Highlight domains where AI has achieved measurable success in drug discovery, (iv) Identify challenges that hinder the translation of current AI performance into real-world applications and (v) Discuss future perspectives and unresolved issues that are likely to influence the evolving role of AI in the drug-discovery.

AI AUGMENTED DRUG DISCOVERY WORKFLOW

AI-augmented drug discovery pipeline needs the integration of machine learning methods across all stages, from target identification to preclinical candidate selection. AI is not merely augmenting traditional drug discovery workflows but fundamentally reshaping them. Its integration across all stages from early target discovery to preclinical evaluation positions as a central driver of innovation in the search for new therapeutics.

The modern AI-driven drug discovery workflow pipeline is divided into five key stages (Fig. 1). In the first stage, *target identification and validation*, multi-omics data sources such as genomics and proteomics are integrated using Graph Neural Networks (GNNs) and large language models to uncover biologically relevant targets. The second stage, *hit discovery*, leverages virtual screening and molecular docking, where deep learning models enhance both the speed and accuracy of identifying active compounds from large chemical libraries. The third stage, *hit-to-lead optimisation*, focuses on refining these initial hits into viable lead compounds using Quantitative Structure-Activity Relationship (QSAR) modelling and reinforcement learning for molecular optimisation. In the fourth stage, *lead optimisation*, AI models predict ADMET properties and generate novel compounds through generative techniques such as diffusion models and transformers. Finally, in the preclinical candidate selection stage, explainable AI (XAI) methods and knowledge graphs support decision-making, ensuring that selected candidates meet safety and efficacy requirements before advancing to clinical trials.

DATA FOUNDATIONS: DATASETS AND BENCHMARKS

Robust machine learning models are ultimately grounded in the quality and structure of the data on which they are trained. In AI-driven drug discovery, this dependence is particularly critical; The molecules a model encounters during training define not only what it can successfully predict, but also the boundaries of its generalization. This section examines the datasets that underpin modern molecular machine learning, the hidden biases they may introduce and emerging strategies to overcome data scarcity and improve evaluation fidelity.

Several large, publicly available datasets have enabled rapid progress in the field. Among them, ChEMBL is the most widely used, aggregating bioactivity measurements from medicinal chemistry literature across millions of compound target pairs. Its strength lies in its diversity and realism, though this comes at the cost of noise and experimental variability[4,5]. For structure-based modelling, PDBbind provides experimentally resolved protein ligand complexes paired with binding affinities, forming the backbone of many graph neural network (GNN)-based affinity prediction modelsasa QAASAZZ[24]. More recent efforts have focused on designing

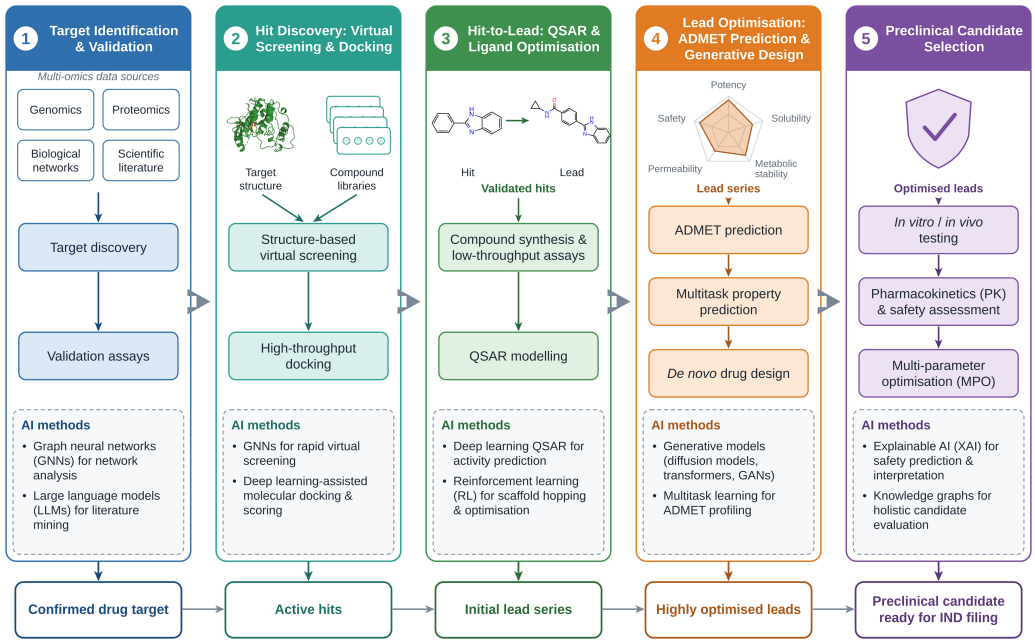


Fig. 1: AI-augmented drug discovery pipeline illustrating the integration of machine learning methods across all stages, from target identification to preclinical candidate selection

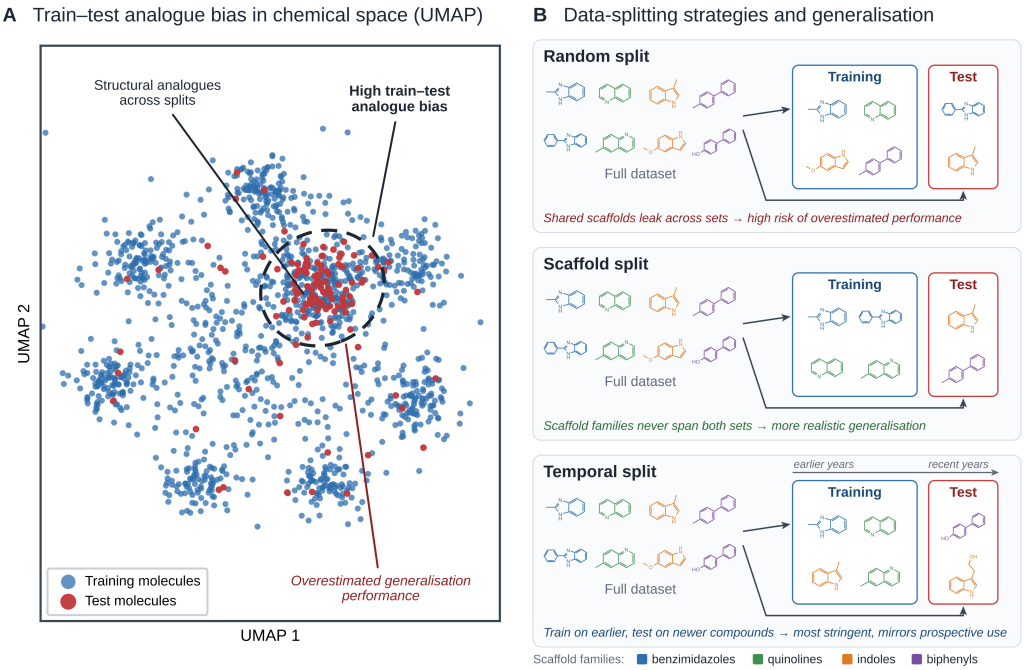


Fig. 2(a-b): Benchmark dataset biases and evaluation strategies. (a) UMAP projection illustrating analogue bias, where structurally similar molecules appear in both training and test sets and (b) Comparison of data splitting strategies: Random, scaffold-based and temporal splits, highlighting their impact on model generalisation

datasets explicitly tailored for machine learning. MolData restructures ChEMBL into disease and target-specific tasks, enabling consistent multi-task benchmarking[18]. BigBind, in contrast, constructs large-scale virtual screening benchmarks without relying on experimentally determined co-crystal structures, thereby better reflecting real-world scenarios where structural data are often unavailable[17]. Meanwhile, the widely used DUD-E dataset has been shown to suffer from significant analogue bias, leading to overly optimistic model performance[14,16].

The impact of dataset bias is illustrated in Fig. 2a. Ideally, the training and test sets should occupy distinct regions of chemical space. However, substantial overlap is observed, with clusters containing both training and test molecules. These clusters represent structurally similar analogues, indicating that the model is effectively evaluated on compounds that differ only marginally from those encountered during training. This phenomenon, known as train-test analogue bias, leads to inflated performance estimates and obscures the model's true capacity to generalize to novel chemical space. The Fig. 2b further illustrates how different data splitting strategies influence this effect. In a *random split*, molecules are assigned to training and test sets without regard to structural similarity, resulting in a high risk of leakage through shared scaffolds and consequently overestimated performance. In contrast, *scaffold-based splits* group molecules by their core chemical structures (e.g., Bemis-Murcko scaffolds) and ensure that structurally related compounds do not appear in both sets. This creates a more challenging and realistic evaluation of generalisation. Finally, *temporal splits* divide data based on time, training on older compounds and testing on newly discovered ones. This approach most closely mirrors real-world deployment, where models must predict properties for molecules that have not yet been synthesized and is therefore considered the most stringent benchmark.

Mitigating these biases requires careful dataset design and evaluation protocols. Scaffold-based and temporal splits are increasingly adopted as standard practice, although they often reveal significantly lower model performance than random splits. This discrepancy underscores the importance of realistic benchmarking in assessing true predictive capability. Beyond bias, data scarcity presents another major challenge. Even the largest public datasets cover only a small fraction of biologically relevant chemical space. For many targets, particularly those associated with rare diseases or emerging pathogens, only a limited number of experimentally characterized compounds may be available. To address this, researchers are exploring few-shot learning approaches, which enable models to generalize from a handful of labelled examples[19,21].

In parallel, semi-supervised learning techniques leverage the vast pool of unlabeled molecular data through strategies such as self-training and transfer learning, improving performance without requiring extensive experimental

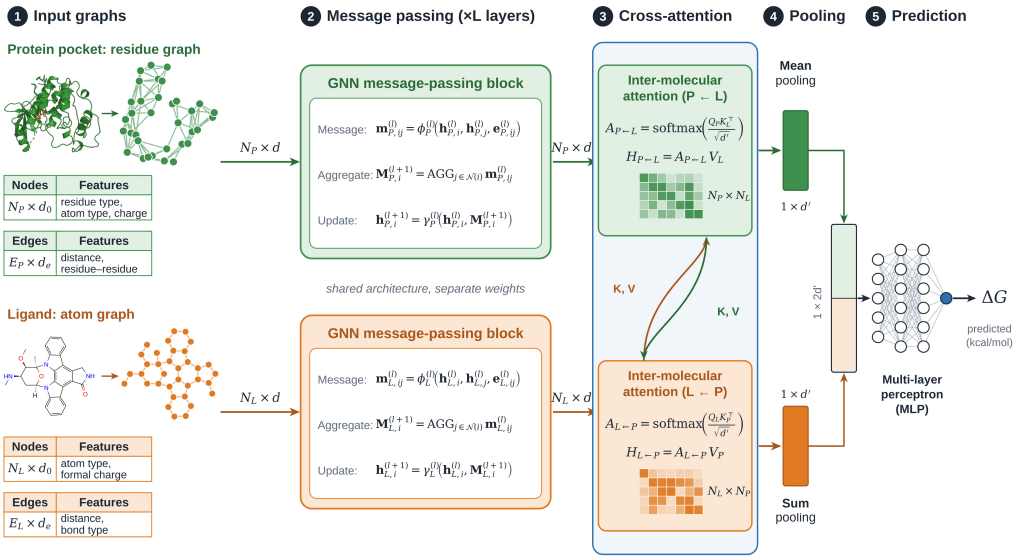


Fig. 3: Representative graph neural network architecture for protein ligand binding affinity prediction

annotation[20]. Furthermore, because biological activity is inherently multifaceted, modern approaches increasingly adopt multi-task and multi-label learning frameworks, allowing models to simultaneously predict multiple properties or endpoints[23,25].

Altogether, these developments highlight a central insight; in molecular machine learning, data quality, structure and evaluation strategy are as important as model architecture itself. As illustrated in Fig. 2, without careful consideration of dataset biases and splitting protocols, even the most sophisticated models risk learning shortcuts rather than meaningful chemical principles.

STRUCTURE-BASED VIRTUAL SCREENING AND MOLECULAR DOCKING

Virtual screening remains a cornerstone of early-stage drug discovery and the integration of artificial intelligence has significantly transformed its implementation. Traditional docking methods rely on physics-based scoring functions to estimate binding free energies; however, these approximations often lack accuracy and are susceptible to systematic biases. Compared with machine learning methods, emerging advanced approaches such as Graph Neural Networks (GNNs) have proven to be more efficient alternatives, enabling models to learn scoring functions directly from experimental and structural datasets.

GNN-based methods are especially well-suited for protein ligand affinity prediction, as they naturally represent both proteins and ligands as graphs. Through iterative message passing, these models capture complex structural and interaction patterns. As illustrated in Fig. 3, a typical architecture constructs separate graphs for the protein binding pocket (at the residue level) and the ligand (at the atom level), processes them through message-passing layers and then models their interactions via cross-attention mechanisms. The resulting representations are aggregated and used to predict binding affinity in an end-to-end manner.

Several recent models highlight the effectiveness of this approach. The PLANET model represents protein pockets as residue graphs and ligands as atom graphs, incorporating cross-attention to capture intermolecular interactions[24]. GNNSeq extends this framework by integrating sequence-based protein features, making it particularly useful when structural data are limited or unavailable[26].

Overall, these data-driven approaches have demonstrated clear advantages over classical docking methods. Across multiple studies, GNN-based models consistently reduce prediction errors by approximately 15-25% relative to traditional scoring functions, with the most significant improvements observed for complex or poorly characterised targets.

As compound libraries continue to grow, computational efficiency has become increasingly important. Methods such as KarmaDock leverage GNNs to predict docking poses in milliseconds rather than hours, enabling large-scale virtual screening[7]. In addition, consensus docking strategies where multiple scoring functions are combined through learned weighting have been shown to improve robustness and predictive performance[12].

More recently, Large Language Models (LLMs) have been integrated into virtual screening workflows. Systems such as ChemCrow utilize LLMs to interpret docking outputs, propose molecular modifications and suggest synthetic pathways,

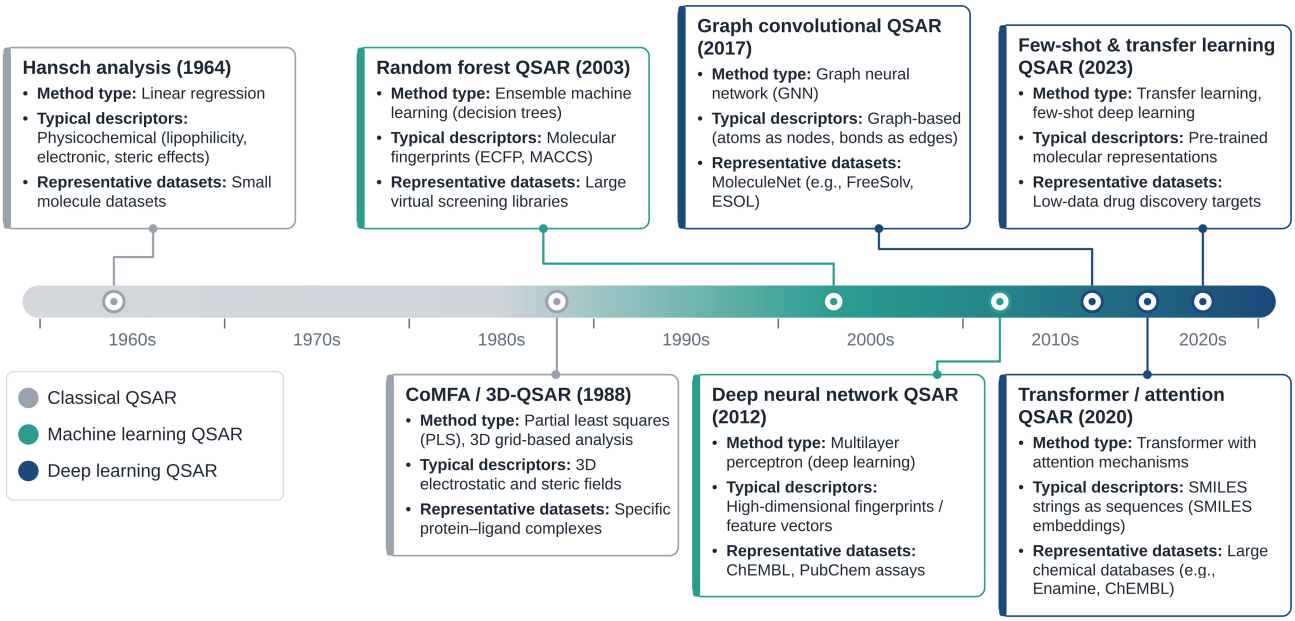


Fig. 4: Evolution of QSAR modeling from classical descriptor-based methods to modern deep learning approaches, highlighting key methodological milestones

enabling more automated and iterative discovery processes[27]. This emerging multimodal paradigm, which combines structural, numerical and textual information, is likely to play an increasingly important role in future drug discovery pipelines.

LIGAND-BASED METHODS: QSAR AND MOLECULAR PROPERTY REDUCTION

In scenarios where high-resolution structural information is unavailable, ligand-based approaches remain a fundamental component of drug discovery. Among these, Quantitative Structure Activity Relationship (QSAR) modelling has played a central role, evolving from simple descriptor-based models to modern deep learning frameworks capable of capturing complex molecular patterns.

As illustrated in Fig. 4, the development of QSAR methods can be broadly divided into several stages. Early approaches, originating in the 1960s, relied on manually engineered descriptors such as physicochemical properties (e.g., logP, electronic parameters) combined with linear regression models. These classical methods were later extended through three-dimensional QSAR techniques, incorporating steric and electrostatic interaction fields to better represent molecular behaviour.

The introduction of machine learning in the early 2000s marked a significant shift, with methods such as random forests and support vector machines enabling the modelling of non-linear relationships between molecular structure and activity. More recently, deep learning approaches particularly graph-based models have transformed QSAR by learning molecular representations directly from atomic connectivity, removing the need for handcrafted features[20].

Current advances in QSAR modelling increasingly leverage attention mechanisms and transformer-based architectures, which can capture long-range dependencies within molecular structures. These models extend QSAR beyond traditional activity prediction to a wide range of molecular properties, including solubility, permeability, metabolic stability and toxicity. However, many of these prediction tasks suffer from limited labelled data. To address this, recent work has focused on few-shot and transfer learning strategies, enabling models to generalize effectively even in low-data regimes[19].

ADMET AND TOXICITY PREDICTION

ADMET (absorption, distribution, metabolism, excretion and toxicity) properties play a decisive role in determining whether a candidate molecule will successfully progress

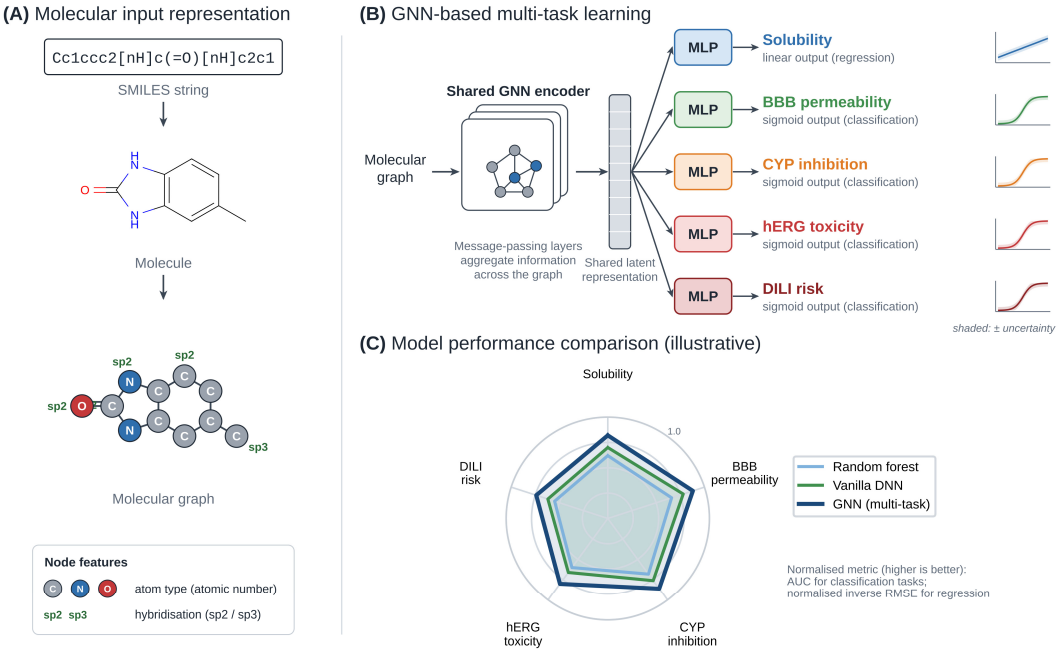


Fig. 5: Deep learning framework for multi-endpoint ADMET prediction. A shared graph neural network encoder extracts molecular features, which are then processed by task-specific output heads to predict multiple pharmacokinetic and toxicity endpoints

through clinical development. In practice, poor ADMET characteristics are responsible for a substantial proportion of late-stage failures, making reliable in silico prediction a critical challenge in drug discovery[3]. Recent advances in deep learning, particularly Graph Neural Networks (GNNs), have led to significant improvements in predictive performance across a range of ADMET benchmarks.

Contemporary approaches progressively adopt multi-endpoint prediction frameworks, where several ADMET properties are modelled simultaneously. As illustrated in Figure 5, these architectures typically employ a shared molecular encoder, most often a GNN that transforms molecular graphs into latent representations. This shared embedding is then passed to multiple task-specific output heads, each dedicated to a particular endpoint. Such a design enables the model to capture common structural patterns across tasks while retaining the flexibility to specialize for individual properties.

Important ADMET endpoints include solubility, membrane permeability (notably blood brain barrier penetration), cytochrome P450 (CYP) inhibition, hERG-related cardiotoxicity and drug-induced liver injury (DILI) risk[28]. In the framework shown in Fig. 5, molecular inputs are first encoded from representations such as SMILES into graph structures, processed through stacked graph convolutional layers and aggregated into a shared latent space. Task-specific predictors, often implemented as multilayer perceptron's subsequently generate outputs tailored to regression or classification objectives depending on the endpoint.

Despite these advances, each ADMET property presents distinct modelling challenges. Solubility prediction is influenced by complex intermolecular interactions and crystallisation behavior, hERG toxicity depends on subtle structural motifs that are difficult to generalise; and DILI risk reflects multifactorial biological processes that extend beyond simple molecular features.

To address these challenges, recent research has gradually focused on uncertainty quantification, allowing models to provide confidence estimates alongside predictions. This capability is particularly important in practical applications, where decision-making must account not only for predicted values but also for their reliability[28].

GENERATIVE MOLECULAR DESIGN

Generative modelling has transformed molecular design by enabling the de novo creation of compounds with tailored properties. Instead of relying solely on screening pre-existing chemical libraries, these approaches can directly generate novel molecular structures while optimizing multiple objectives.

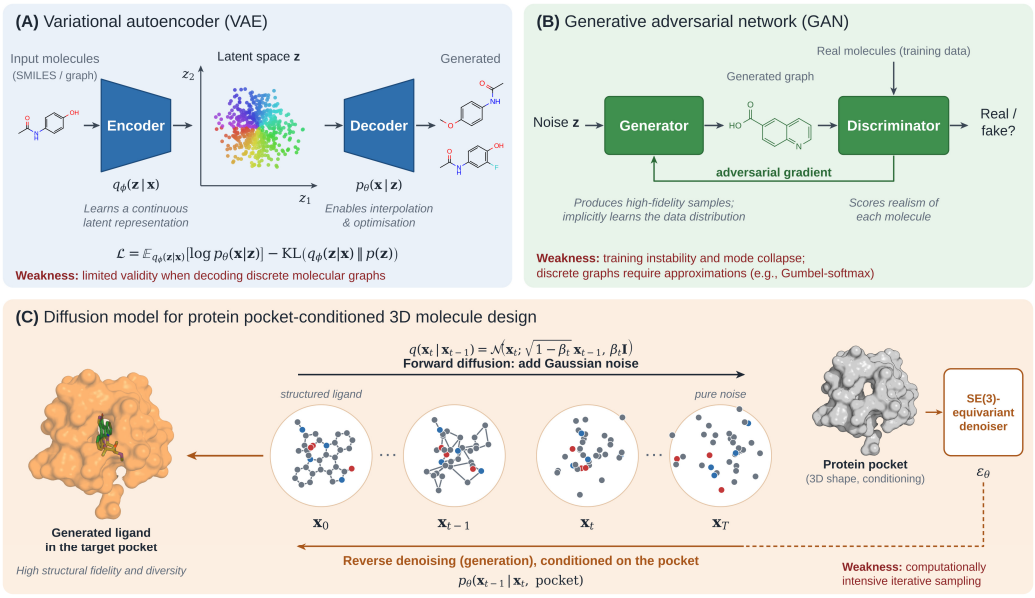


Fig. 6: Comparison of generative model architectures for molecular design, including variational autoencoders (VAEs), generative adversarial networks (GANs) and diffusion-based approaches

As illustrated in Fig. 6, modern generative molecular design is primarily based on three major architectures (i) Variational Autoencoders (VAEs), (ii) Generative Adversarial Networks (GANs) and (iii) Diffusion models.

The VAEs learn a continuous latent representation of molecules, allowing new compounds to be generated by sampling and decoding points in this latent space. GANs, in contrast, rely on an adversarial framework in which a generator produces candidate molecules while a discriminator evaluates their realism, guiding the generator toward more plausible outputs. Diffusion models, the most recent and rapidly advancing approach, operate by learning to reverse a gradual noise-adding process, reconstructing structured molecules from random noise.

Among these methods, diffusion models have demonstrated promise for structure-based drug design. By conditioning the generation process on protein binding pocket geometry, they enable the creation of molecules that are inherently compatible with specific biological targets. For example, DiffPROTACs focuses on generating proteolysis-targeting chimeras (PROTACs), which recruit cellular degradation machinery to selectively eliminate target proteins[29]. Similarly, PharmacoForge employs diffusion-based techniques to design pharmacophores spatial arrangements of functional groups critical for binding affinity[12].

A key challenge in practical drug design is the need to optimize multiple, often competing objectives, including potency, selectivity, pharmacokinetic properties and synthetic feasibility. Modern generative frameworks increasingly address this through multi-objective optimization strategies, such as reinforcement learning or conditional generation, which guide molecule generation toward desired property profiles. The integration of generative models with experimental feedback systems represents a significant shift in the drug discovery paradigm, progressively enabling closed-loop workflows that connect design, synthesis, testing and analysis in a continuous optimisation cycle.

ALPHAFOLD-INTEGRATED DRUG DISCOVERY

The introduction of AlphaFold2 in 2021 marked a breakthrough in structural biology, providing high-accuracy protein structures for millions of previously unresolved targets and significantly accelerating structure-based drug discovery efforts[9]. However, effectively incorporating these predicted structures into drug discovery pipelines introduces both new opportunities and important challenges[30,31].

As illustrated in Fig. 7, AlphaFold2 predictions are now routinely integrated into virtual screening workflows. The typical pipeline begins with a protein sequence input, which is processed by AlphaFold2 to generate a three-dimensional structure annotated with confidence scores (pLDDT). This is followed by binding site identification using specialized tools such as FPocket or SiteMap, virtual screening through molecular docking, selectivity analysis across homologous proteins and finally molecular dynamics (MD) simulations for structural refinement.

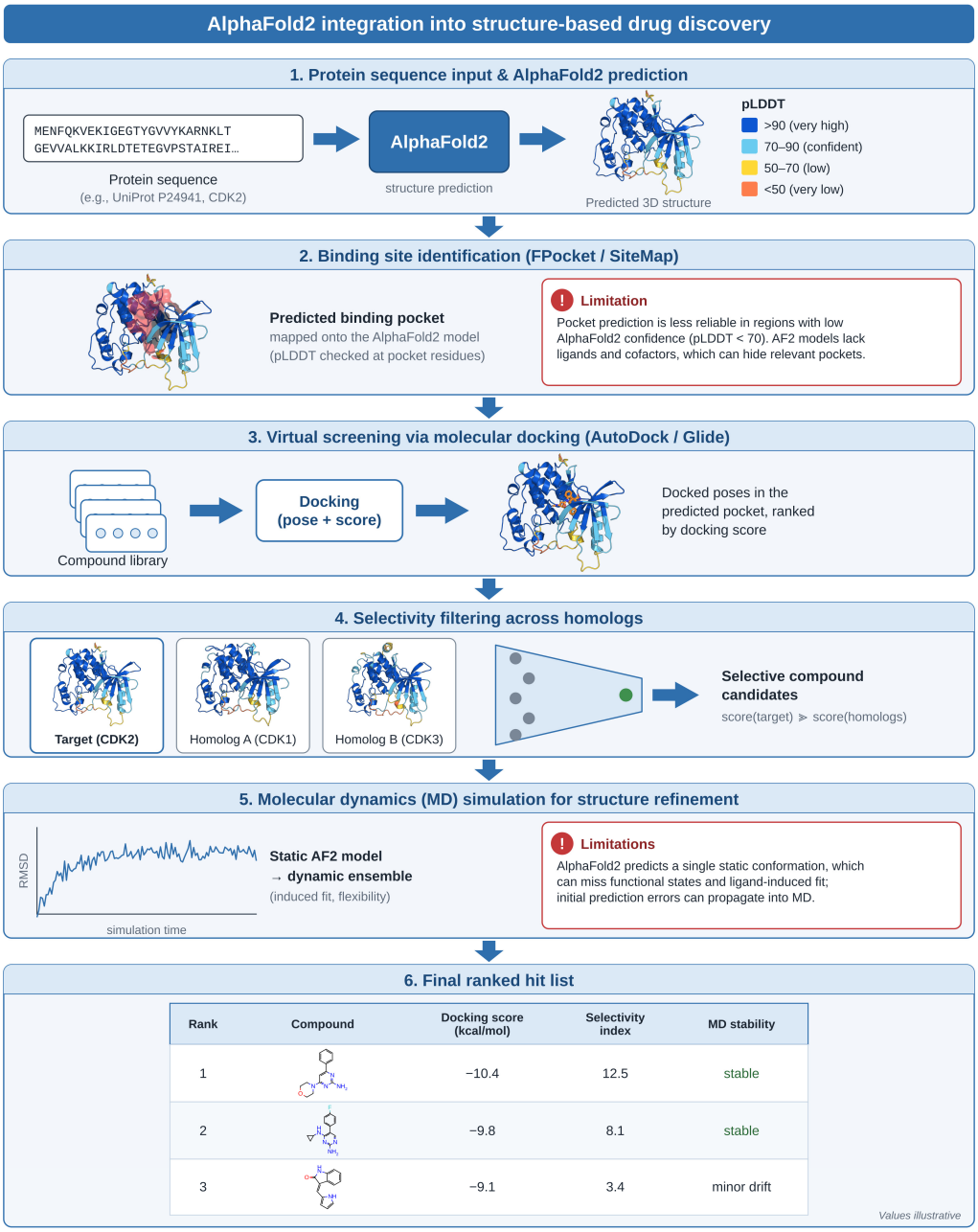


Fig. 7: Workflow integrating AlphaFold2-predicted protein structures into structure-based drug discovery, including binding site identification, virtual screening, selectivity filtering and molecular dynamics refinement

A key consideration when working with AlphaFold2 structures is the reliability of predicted regions. Binding site identification must account for pLDDT confidence scores, as low-confidence regions may not accurately represent the true protein geometry and can lead to misleading docking results. Furthermore, the absence of ligands or cofactors in AlphaFold models can complicate the identification of biologically relevant binding pockets.

Selectivity profiling is another important application enabled by AlphaFold2. By generating structural models for homologous proteins, researchers can evaluate whether candidate compounds preferentially bind to the intended target while avoiding closely related proteins, thereby reducing the risk of off-target effects.

Despite its strengths, AlphaFold2 produces static structural predictions, whereas proteins are inherently dynamic. Molecular dynamics simulations are therefore often

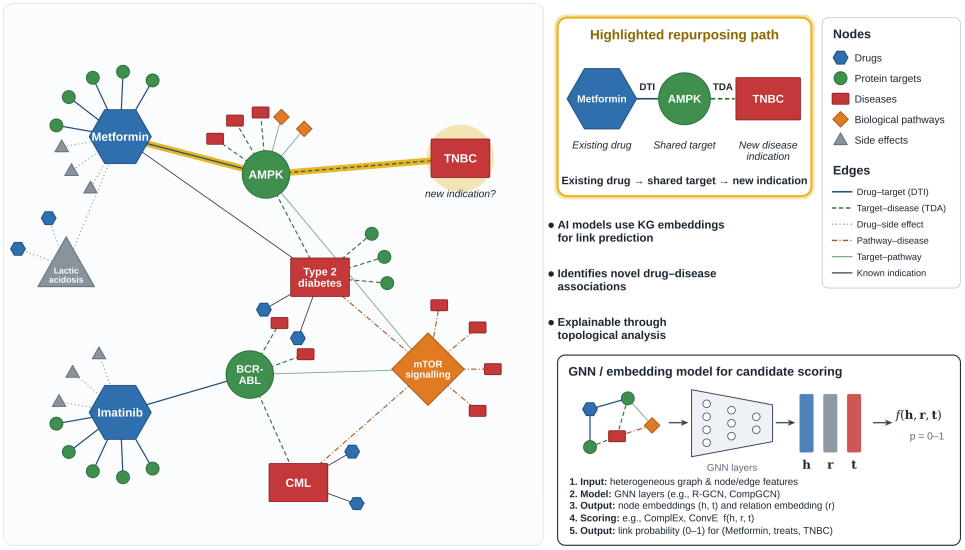


Fig. 8: Knowledge graph-based drug repurposing framework illustrating how heterogeneous biomedical relationships enable the identification of novel drug disease associations

employed to refine predicted structures and explore conformational flexibility, although this introduces additional computational cost. Moreover, AlphaFold2 performs best on single-domain proteins, while multi-domain systems and proteins undergoing large conformational changes remain more challenging to model accurately[30].

Nevertheless, AlphaFold-integrated workflows have already demonstrated practical impact. Several compounds identified through structure-based screening using AlphaFold-derived models have advanced to preclinical stages, underscoring the growing importance of this approach in modern drug discovery[31].

DRUG REPURPOSING WITH AI

Drug repurposing is the identification of new therapeutic applications for existing drugs. It provides a faster and more cost-effective alternative to traditional de novo drug development. By leveraging prior knowledge of safety and pharmacokinetics, repurposing strategies can significantly reduce both development time and risk. Recent advances in artificial intelligence have further expanded the scope and efficiency of these approaches[23,25].

A central paradigm in AI-driven repurposing is the use of knowledge graphs, which integrate diverse biomedical data sources into a unified network. These graphs typically encode relationships among drugs, protein targets, diseases, biological pathways and side effects. As illustrated in Fig. 8, such heterogeneous networks enable the modelling of complex biological interactions, where indirect connections between entities can reveal previously unrecognized therapeutic opportunities. Graph neural networks (GNNs) applied to knowledge graphs can learn embeddings for nodes and relationships, allowing models to predict missing links such as potential drug disease associations. In this framework, repurposing candidates are identified by tracing multi-hop relationships (e.g., drug → target → pathway→ disease), effectively uncovering mechanistic links that are not immediately apparent from direct associations alone[23]. The highlighted pathway in Fig. 8 exemplifies this process, where an existing drug connects to a new disease indication through a shared biological target.

The COVID-19 pandemic accelerated the adoption of AI-driven repurposing strategies. Knowledge graph-based methods were used to rapidly identify approved drugs with potential antiviral activity by linking drug-target profiles to SARS-CoV-2-related biological pathways[23,25]. AI-based repurposing has also shown promise for rare diseases, where limited patient populations make conventional drug development economically challenging. Network-based approaches can connect disease-associated genes to known drug targets through shared pathways, enabling the identification of viable therapeutic candidates[23,25].

In oncology, repurposing efforts have revealed that drugs originally developed for non-cancer indications can modulate cancer-relevant pathways, offering new treatment possibilities with established safety profiles. Overall, the integration of knowledge graphs and machine learning is transforming drug repurposing into a scalable, data-driven discovery paradigm.

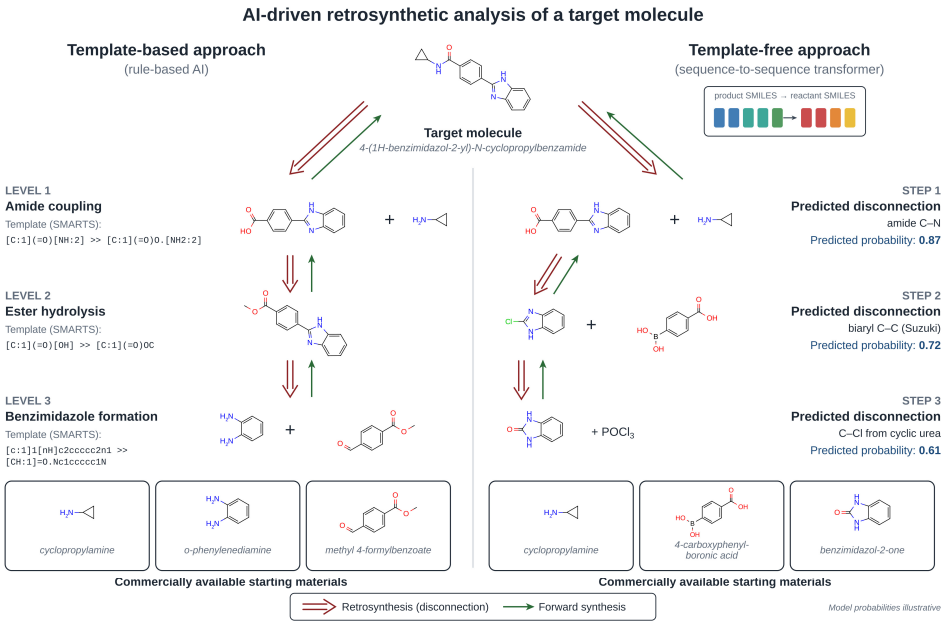


Fig. 9: AI-driven retrosynthesis illustrating template-based and template-free strategies for decomposing target molecules into synthesizable precursors

RETROSYNTHESIS AND REACTION PREDICTION

AI-driven retrosynthesis, the task of decomposing target molecules into simpler, synthetically accessible precursors has significantly advanced modern synthetic planning. Traditionally, retrosynthetic analysis depended heavily on the expertise and intuition of chemists. In contrast, AI-based approaches formalize this process, enabling systematic exploration of possible synthetic routes and often identifying strategies that may not be immediately apparent to human experts[32]. As illustrated in Fig. 9, two primary approaches dominate AI-driven retrosynthesis: Template-based and template-free methods. Template-based approaches extract reaction rules from curated reaction databases and apply these learned templates to new target molecules, ensuring chemically valid transformations grounded in known chemistry[32].

In contrast, template-free methods typically based on sequence-to-sequence transformer architectures learn to directly predict precursor structures from product molecules without relying on predefined reaction templates. These models treat chemical representations such as SMILES as sequences, enabling flexible and generalizable prediction of reaction outcomes[33]. The retrosynthetic tree shown in Fig. 9 highlights how both approaches can be used to iteratively deconstruct complex molecules into feasible starting materials.

The integration of retrosynthesis with generative molecular design has emerged as a particularly powerful direction. Rather than generating candidate molecules without regard to synthetic feasibility, modern frameworks incorporate retrosynthetic constraints during the design process, favouring compounds that are both biologically promising and practically synthesizable[13]. An important emerging model is the coupling of retrosynthetic analysis with hit-to-lead optimisation, allowing synthetic accessibility to be evaluated alongside biological activity early in the design cycle. This integrated approach reduces the risk of investing resources in compounds that are theoretically attractive but difficult or impractical to synthesise, thereby improving the efficiency of the overall drug discovery pipeline[32,33].

CHALLENGES AND FUTURE OUTLOOK

Despite the rapid advances seen in recent years, AI-driven drug discovery still faces several fundamental challenges before it can achieve reliable and widespread real-world impact; a major concern is the lack of reproducibility. As in many areas of computational science, apparent improvements in model performance are not always matched by genuine progress in practical utility. Inflated benchmark results, often caused by overly permissive evaluation settings, can create a misleading impression of model capability. Analogue bias in train test splitting allows models to perform well on

compounds that are structurally very similar to those seen during training, while failing on truly novel chemotypes. Addressing this problem will require stricter benchmarking standards, more realistic evaluation protocols and broader adoption of prospective validation[15,19,20].

A second major challenge is interpretability. Many high-performing deep learning models function as black boxes, which limits their acceptance in regulatory and high-stakes decision-making contexts. This issue is especially important for applications such as toxicity prediction, where understanding why a model makes a particular prediction can be as important as the prediction itself. Although current interpretability techniques including Shapley values, attention visualization and concept-based methods offer partial insight, they still fall short of the mechanistic transparency often required by regulators and domain experts[6,9,20].

Limited access to high-quality proprietary data presents another obstacle. Pharmaceutical companies possess extensive experimental datasets that could substantially improve model training and validation, yet these data are seldom shared because of intellectual property and competitive constraints. As a result, many published models are developed using only public data-sets, which may not fully capture the complexity and diversity of industrial discovery settings. This contributes to a persistent generalization gap between benchmark performance and real-world effectiveness[15,19,20].

Three converging trends are likely to shape the next phase of AI-driven drug discovery. First, foundation models trained on large-scale, multimodal molecular and biological data are expected to become increasingly important, providing general-purpose representations that can be adapted to a wide range of downstream tasks with limited additional data[6,19,20]. Second, future models will likely become more physics-aware and dynamics-informed, incorporating structural symmetry, conformational flexibility and mechanistic constraints rather than relying solely on static representations. Third, the continued integration of AI with autonomous experimental platforms may progressively close the design make test analyze loop, giving rise to increasingly self-optimizing discovery systems capable of iterative compound generation, synthesis, testing and refinement with reduced human intervention[12,13].

CONCLUSION

Artificial intelligence has evolved from a supportive technology into a core element of the modern drug-discovery pipeline. Across multiple domains highlighted in this review, including virtual screening, QSAR modelling, ADMET prediction, generative molecular design, AlphaFold-assisted screening, drug repurposing and retrosynthesis, AI has enabled faster decision-making, enhanced predictive accuracy and more efficient exploration of chemical and biological space.

At the same time, current evidence supports a perspective of cautious optimism rather than unquestioned enthusiasm. The disparity between models that achieve strong benchmark performance and those that consistently provide practical preclinical utility has gradually decreased, particularly after 2020, due to improvements in algorithms, datasets and evaluation strategies. Significant methodological progress has been made through the development of more realistic benchmarks, rigorous validation frameworks and increasingly integrated drug-discovery workflows. Despite these advances, important challenges persist, especially with respect to reproducibility, interpretability, data availability and the ability of models to generalize reliably beyond curated benchmark datasets.

Overall, the future of AI in drug discovery is unlikely to involve replacing medicinal chemists, pharmacologists, or structural biologists; instead, AI is expected to function as an essential collaborative tool. By accelerating research workflows, lowering attrition rates and uncovering novel opportunities that may otherwise remain undetected, AI can substantially improve the speed, efficiency and productivity of drug discovery. Nevertheless, human expertise will continue to play a critical role in transforming computational predictions into clinically successful therapeutics, while AI increasingly supports this process by reducing development time, minimizing costs and improving the likelihood of success.

SIGNIFICANCE STATEMENT

This review highlights the transformative role of Artificial Intelligence (AI) in reshaping modern drug discovery pipelines. By critically evaluating deep learning methodologies, from virtual screening to retrosynthesis, it bridges the gap between theoretical benchmark success and practical preclinical application. The study provides essential insights for overcoming dataset biases and interpretability challenges, guiding future therapeutic innovation.

REFERENCES

1. DiMasi, J.A., H.G. Grabowski and R.W. Hansen, 2016. Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ., 47: 20-33.

2. Wong, C.H., K.W. Siah and A.W. Lo, 2019. Estimation of clinical trial success rates and related parameters. Biostatistics, 20: 273-286.
3. Waring, M.J., J. Arrowsmith, A.R. Leach, P.D. Leeson and S. Mandrell *et al*., 2015. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat. Rev. Drug Discovery, 14: 475-486.
4. Korotcov, A., V. Tkachenko, D.P. Russo and S. Ekins, 2017. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol. Pharmaceutics, 14: 4462-4475.
5. Lane, T.R., D.H. Foil, E. Minerali, F. Urbina, K.M. Zorn and S. Ekins, 2021. Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. Mol. Pharmaceutics, 18: 403-415.
6. Zhang, X.C., C.K. Wu, Z.J. Yang, Z.X. Wu and J.C. Yi *et al*., 2021. MG-BERT: Leveraging unsupervised atomic representation learning for molecular property prediction. Briefings Bioinf., Vol. 22. 10.1093/bib/bbab152.
7. Zhang, X., O. Zhang, C. Shen, W. Qu and S. Chen *et al*., 2023. Efficient and accurate large library ligand docking with KarmaDock. Nat. Comput. Sci., 3: 789-804.
8. Yin, Y., H.Y.I. Lam, Y. Mu, H.Y. Li and A.W.K. Kong, 2024. Advancing bioactivity prediction through molecular docking and self-attention. IEEE J. Biomed. Health. Inf., 28: 7599-7610.
9. Jumper, J., R. Evans, A. Pritzel, T. Green and M. Figurnov *et al*., 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596: 583-589.
10. Atz, K., L. Cotos, C. Isert, M. Håkansson and D. Focht *et al*., 2024. Prospective de novo drug design with deep interactome learning. Nat. Commun., Vol. 15. 10.1038/s41467-024-47613-w.
11. Kadan, A., K. Ryczko, E. Lloyd, A. Roitberg and T. Yamazaki, 2025. Guided multi-objective generative AI to enhance structure-based drug design. Chem. Sci., 16: 13196-13210.
12. Flynn, E.L., R. Shah, I. Dunn, R. Aggarwal and D.R. Koes, 2025. PharmacoForge: Pharmacophore generation with diffusion models. Front. Bioinf., Vol. 5. 10.3389/fbinf.2025.1628800.
13. Nippa, D.F., K. Atz, Y. Stenzhorn, A.T. Müller and A. Tosstorff *et al*., 2025. Expediting hit-to-lead progression in drug discovery through reaction prediction and multi-dimensional optimization. Nat. Commun., Vol. 16. 10.1038/s41467-025-66324-4.
14. Chen, L., A. Cruz, S. Ramsey, C.J. Dickson and J.S. Duca *et al*., 2019. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS ONE, Vol. 14. 10.1371/journal.pone.0220113.
15. Graber, D., P. Stockinger, F. Meyer, S. Mishra, C. Horn and R. Buller, 2025. Resolving data bias improves generalization in binding affinity prediction. Nat. Mach. Intell., 7: 1713-1725.
16. Chen, Y., D. Luo, S. Chen, T. Hou, C. Huang and W. Xue, 2025. CNSGT: Generative transformer for de novo drug design targeting the central nervous system. J. Chem. Inf. Model., 65: 10054-10067.
17. Brocidiacono, M., P. Francoeur, R. Aggarwal, K.I. Popov, D.R. Koes and A. Tropsha, 2024. BigBind: Learning from nonstructural data for structure-based virtual screening. J. Chem. Inf. Model., 64: 2488-2495.
18. Arshadi, A.K., M. Salem, A. Firouzbakht and J.S. Yuan, 2022. MolData, a molecular benchmark for disease and target based machine learning. J. Cheminf., Vol. 14. 10.1186/s13321-022-00590-y.
19. Vella, D. and J.P. Ebejer, 2023. Few-shot learning for low-data drug discovery. J. Chem. Inf. Model., 63: 27-42.
20. Liu, Y., H. Lim and L. Xie, 2022. Exploration of chemical space with partial labeled noisy student self-training and self-supervised graph embedding. BMC Bioinf., Vol. 23. 10.1186/s12859-022-04681-3.
21. Chen, Y., J. Huang, C. Liu, S. Zhang and X. Li *et al*., 2026. DualPG-DTA: A large language model-powered graph neural network framework for enhanced drug-target affinity prediction and discovery of novel CDK9 inhibitors exhibiting *in vivo* anti-leukemia activity. Adv. Sci., Vol. 13. 10.1002/advs.202513099.
22. Berreziga, R., M. Brahimi, K. Kraim and H. Azzoune, 2025. Combining GCN structural learning with LLM chemical knowledge for enhanced virtual screening. J. Chem. Inf. Model., 65: 11510-11520.
23. Abdullah, Z. Fatima, M.A. Ather, L. Chanona-Hernandez and J.L.O. Rodríguez, 2026. Hybrid dual-context prompted cross-attention framework with language model guidance for multi-label prediction of human off-target ligand-protein interactions. Int. J. Mol. Sci., Vol. 27. 10.3390/ijms27021126.
24. Zhang, X., H. Gao, H. Wang, Z. Chen and Z. Zhang *et al*., 2024. PLANET: A multi-objective graph neural network model for protein-ligand binding affinity prediction. J. Chem. Inf. Model., 64: 2205-2220.
25. Wu, Y., K. Li, M. Li, X. Pu and Y. Guo, 2023. Attention mechanism-based graph neural network model for effective activity prediction of SARS-CoV-2 main protease inhibitors: Application to drug repurposing as potential COVID-19 therapy. J. Chem. Inf. Model., 63: 7011-7031.
26. Dandibhotla, S., M. Samudrala, A. Kaneriya and S. Dakshanamurthy, 2025. GNNSeq: A sequence-based graph neural network for predicting protein-ligand binding affinity. Pharmaceuticals, Vol. 18. 10.3390/ph18030329.
27. Bran, A.M., S. Cox, O. Schilter, C. Baldassari, A.D. White and P. Schwaller, 2024. Augmenting large language models with chemistry tools. Nat. Mach. Intell., 6: 525-535.
28. Jiang, D., T. Lei, Z. Wang, C. Shen, D. Cao and T. Hou, 2020. ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning. J. Cheminf., Vol. 12. 10.1186/s13321-020-00421-y.
29. Li, F., Q. Hu, Y. Zhou, H. Yang and F. Bai, 2024. DiffPROTACs is a deep learning-based generator for proteolysis targeting chimeras. Briefings Bioinf., Vol. 25. 10.1093/bib/bbae358.

30. Baselious, F., S. Hilscher, D. Robaa, C. Barinka, M. Schutkowski and W. Sippl, 2024. Comparative structure-based virtual screening utilizing optimized AlphaFold model identifies selective HDAC11 inhibitor. Int. J. Mol. Sci., Vol. 25. 10.3390/ijms25021358.
31. Bernatavicius, A., M. Šícho, A.P.A. Janssen, A.K. Hassen, M. Preuss and G.J.P. van Westen, 2024. AlphaFold meets de novo drug design: Leveraging structural protein information in multitarget molecular generative models. J. Chem. Inf. Model., 64: 8113-8122.
32. Baylon, J.L., N.A. Cilfone, J.R. Gulcher and T.W. Chittenden, 2019. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. J. Chem. Inf. Model., 59: 673-688.
33. Westerlund, A.M., S.M. Koki, S. Kancharla, A. Tibo and L. Saigiridharan *et al*., 2024. Do chemformers dream of organic matter? Evaluating a transformer model for multistep retrosynthesis. J. Chem. Inf. Model., 64: 3021-3033.