Review | Received 20 January 2026; Revised 23 February 2026; Accepted 16 March 2026; Published 11 May 2026
https://doi.org/10.55092/acr20260005

Artificial intelligence in oncology drug discovery: from target identification to therapeutic molecule generation

Jianxin Tang†, Jinhang Xu†, Wenqing Zhang†, Daohong Gong†, Qixing Huang, Xiaolong Cheng and Honglin Li*

Innovation Center for AI and Drug Discovery, School of Pharmacy, East China Normal University, Shanghai 200062, China

† These authors contributed equally to this work.

* Correspondence author; E-mail: hlli@hsc.ecnu.edu.cn.

Highlights:

- AI integrates multi-omics data to identify oncology targets and synthetic lethal pairs.
- Graph neural networks enable virtual screening of chemical libraries.
- Generative AI designs novel small molecules and protein binders for oncology targets.
- AI-driven ADMET prediction accelerates hit-to-candidate transition in drug discovery.
- Closed-loop systems advance autonomous AI-driven oncology drug discovery.

Abstract: The development of oncology therapeutics is currently impeded by exorbitant costs, protracted timelines, and high clinical attrition rates stemming from the inherent complexity of tumor biology. Artificial intelligence (AI) is transforming oncology drug discovery, shifting the paradigm from trial-and-error experimentation to one of data-driven rational design. In this paper, we review recent AI advances in four key areas. First, regarding target identification, we examine how multi-omics integration and deep learning uncover novel vulnerabilities, such as synthetic lethal pairs and immune checkpoints. Second, we analyze the evolution of virtual screening, moving from classical docking to graph neural networks that efficiently explore vast chemical spaces. Third, we highlight the shift toward generative molecular design, where AI models create de novo small molecules, protein binders, and nucleic acid therapeutics with tailored functional properties. Fourth, we discuss AI applications in preclinical evaluation for predicting toxicity and efficacy. Finally, we critically assess current challenges—including data standardization deficits and the “black box” nature of deep learning—and propose emerging strategies, such as automated design-make-test workflows, to bridge the gap between computational prediction and clinical reality.

Keywords: artificial intelligence; oncology drug discovery; target identification; generative design; virtual screening

1. Introduction

The development of effective cancer therapies faces formidable challenges. Despite massive investment, the pharmaceutical industry grapples with exceptionally low success rates; with the probability of a new oncology drug progressing from phase I to approval is remarkably low, estimated at approximately 3%–5% depending on the study period and methodology [1,2], substantially below the ~10% observed across all therapeutic areas [3]. This inefficiency stems from the inherent complexity of biological systems, the difficulty of predicting drug efficacy and safety at an early stage, and the unique obstacles presented by cancer itself: tumor heterogeneity, the rapid evolution of resistance, and a high proportion of undruggable targets [4–6]. The traditional drug discovery model, which relies heavily on expert intuition and serendipity-based high-throughput screening, has reached a bottleneck and requires a fundamental methodological shift. Figure 1 traces this evolution, illustrating key technological milestones from early phenotypic screening through the genomics era to the current AI-driven generative design paradigm.

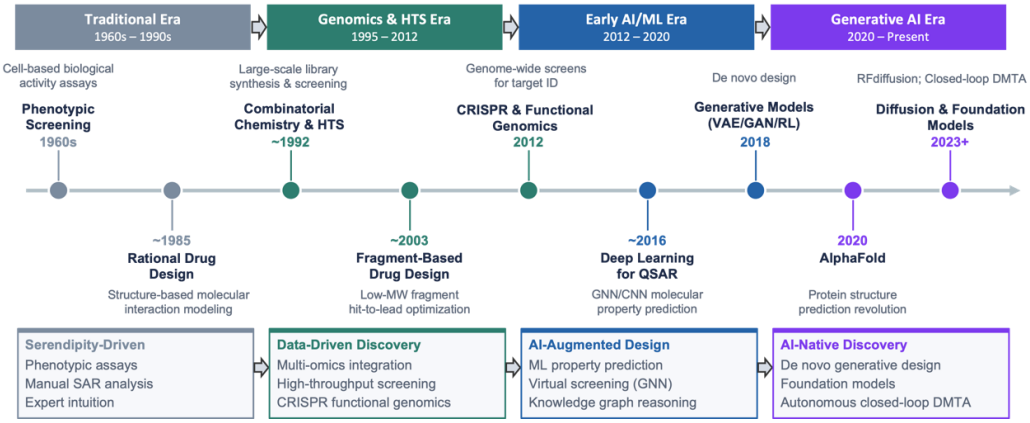

Figure 1. Historical evolution of target identification and small-molecule drug discovery, spanning from traditional phenotypic screening to the current generative AI era. Key milestones include combinatorial chemistry, CRISPR-based functional genomics, AlphaFold-driven structure prediction, and diffusion-based de novo molecular design. Bottom panels summarize the paradigm shift from serendipity-driven to AI-native autonomous discovery.

However, the convergence of exponential growth in biological data and breakthrough advances in computational power is positioning artificial intelligence (AI) as a pivotal driver for overcoming this impasse [7]. Unlike conventional computer-aided drug design (CADD) approaches which rely on physics-based simulations, modern AI leverages machine learning (ML) and deep learning (DL) to extract latent patterns from massive datasets, ranging from genomic sequences to unstructured scientific literature [8]. A particularly transformative development is the rise of generative AI. Powered by architectures such as Transformers and diffusion models, AI has evolved from merely screening existing libraries to the de novo design of novel molecular entities, significantly expanding the range of explorable chemical space. As summarized in Table 1, these AI-driven approaches offer marked improvements in speed, cost, and scalability over conventional methods, though the latter retain advantages in mechanistic interpretability, underscoring the complementary nature of both paradigms.

AI integration has now permeated every critical stage of the oncology drug development. In the initial phase of target discovery, AI algorithms integrate multi-omics data to identify driver genes and reveal synthetic lethal interactions, providing precise intervention points for personalized medicine. In

the lead optimization and molecule design phase, generative models are engineering candidates that transcend the limitations of human intuition, including proteolysis-targeting chimeras (PROTACs) for protein degradation and novel binders for complex protein-protein interactions. Furthermore, in preclinical research, deep learning models are progressively refining toxicity prediction and efficacy assessment, thereby reducing reliance on costly animal models and mitigating safety risks at an earlier stage [9–13].

Table 1. Comparison of conventional and AI-driven approaches in oncology drug discovery.

Dimension	Conventional Approaches	AI-Driven Approaches
Target Identification		
Time/Efficiency	6–12 months	Weeks
Cost	High	Moderate
Accuracy/Hit Rate	Low	Improved
Scalability	Limited	High
Interpretability	High	Variable
Molecule Design and Screening		
Time/Efficiency	Months to years (high-throughput screening, HTS)	Days to weeks (virtual screening); hours (de novo design)
Cost	Very high	Moderate
Accuracy/Hit Rate	< 0.2%	Substantially improved
Scalability	~10^6 compounds physical	> 10^9 virtual; de novo unlimited
Interpretability	High	Variable

This review aims to provide a comprehensive framework for understanding the current state and future trajectory of AI in precision oncology. We systematically summarize advances in four core domains: target identification, virtual screening, generative molecular design, and preclinical evaluation. Beyond highlighting algorithmic achievements, we critically examine persistent technical bottlenecks, such as data quality disparities, model interpretability, and the translational gap between *in silico* predictions and wet-lab validation. By synthesizing these developments, we offer a perspective on how AI-driven workflows are evolving toward automated, closed-loop systems that promise to accelerate the delivery of next-generation cancer therapeutics.

2. AI-driven cancer drug design

Traditional drug development faces significant bottlenecks, including low efficiency in target discovery and high clinical attrition. The advent of AI offers a new paradigm to overcome these challenges. By integrating multi-modal data encompassing genomics, transcriptomics, proteomics, pathological imaging, and clinical information, AI can accelerate the generation of testable target hypotheses and support systematic evidence integration to minimize trial-and-error costs [14,15].

Currently, deep learning has been shown to extract certain molecular features from pathological images under specific conditions [16], simultaneously, graph neural networks (GNNs) can prioritize candidate targets by modeling protein-protein interaction networks [17], and knowledge graph approaches facilitate multi-source knowledge integration for drug repurposing [18]. From the perspective of the evidence hierarchy, AI-assisted target discovery typically progresses from

computational prediction to experimental validation and ultimately to clinical evaluation, frequently iterating across these stages. This chapter systematically reviews the methodological advances and translational cases of AI in tumor target identification and validation along this evidence-based chain.

2.1. Types and characteristics of oncology-related targets

The selection of oncology drug targets directly determines the efficacy and safety of therapeutic strategies. Based on the primary therapeutic axis and the mechanism of intervention, targets can be broadly classified into two major categories: (1) tumor-intrinsic targets, including driver oncogenes, synthetic lethal partners, DNA damage repair proteins, and metabolic/epigenetic regulators; and (2) immune-directed targets. The immune-directed category is further subdivided into immune checkpoint/modulation targets, encompassing immune checkpoints and other immunomodulatory factors, and antigen-based targets, comprising tumor neoantigens and tumor-associated antigens (TAAs). While both subcategories engage the immune system, they differ mechanistically: immune modulation acts by altering immunosuppressive signaling pathways, whereas antigen-based approaches rely on identifying and presenting tumor-specific antigens for T-cell recognition. Figure 2A provides an overview of these two major target categories and the corresponding AI-driven discovery methods operating at tissue, cell, and knowledge scales. Table 2 further details the specific targets within each category alongside their emerging therapeutic strategies, providing a reference framework for the subsections that follow.

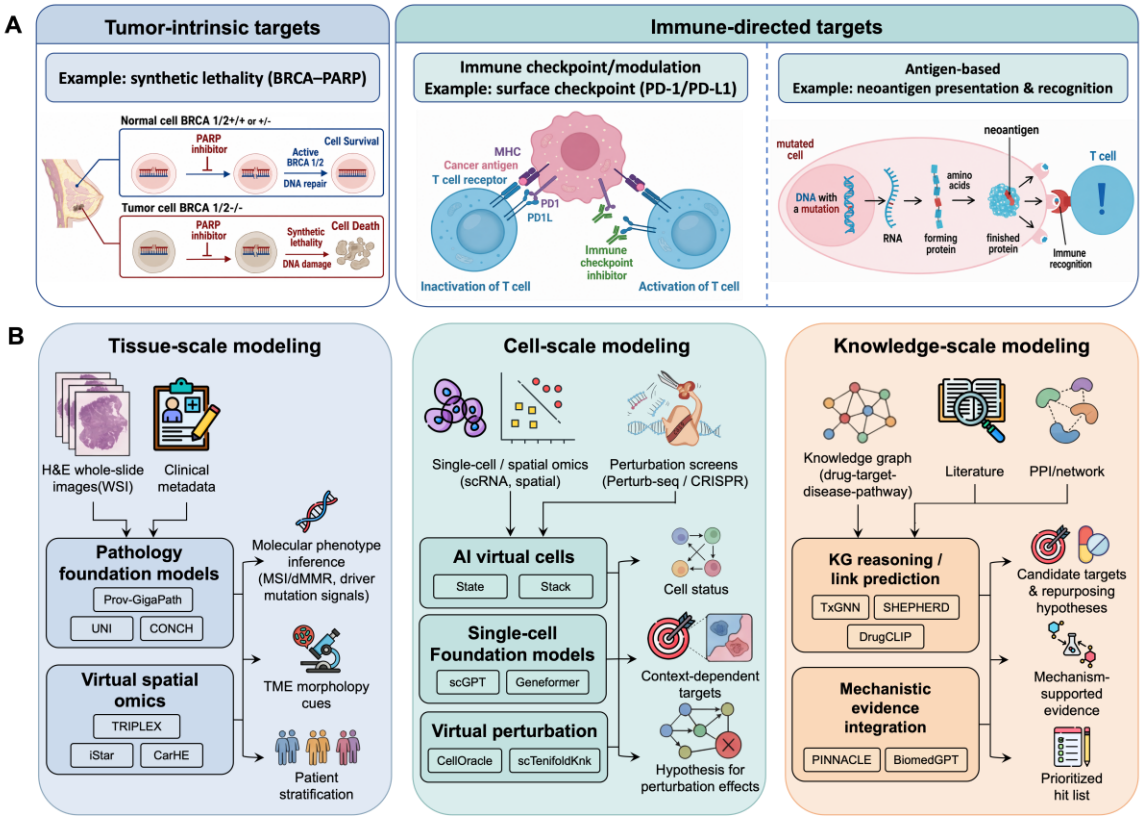

Figure 2. Tumor target classification and AI discovery frameworks. **(A)** Two major categories of tumor targets: tumor-intrinsic and immune-directed targets. Immune-directed targets are further subdivided into immune checkpoint/modulation targets and antigen-based targets; **(B)** Multi-scale AI methods for target discovery: tissue-scale pathology models, cell-scale single-cell models, and knowledge-scale graph reasoning.

Table 2. Classification of key oncology targets and emerging therapeutic strategies.

Targets	Sub-classification	Specific Target	Frontier Applications/Drug
Tumor-intrinsic	Driver Oncogenes	KRAS, EGFR, BRAF	Pan-KRAS inhibitors; G12D-specific non-covalent inhibitors; 4th-generation EGFR TKIs.
	Synthetic Lethality Partners	PARP1, Polθ, PRMT5, WRN	PARP1-selective inhibitors; Polθ inhibitors; MTA-cooperative PRMT5 inhibitors.
	Hard-to-drug Targets/Degradation	MYC, STAT3, Scaffold proteins	PROTACs/Molecular Glues; Transcription factor inhibitors.
Immune Modulation	Intracellular Signaling Regulators	DGKα/ζ, HPK1, SHP2	DGKα/ζ inhibitors; HPK1 inhibitors; SHP2 allosteric inhibitors.
	Next-generation Checkpoints	LAG-3, TIGIT, TIM-3	Bispecific antibodies; Fc-enabled anti-TIGIT.
	Innate/Myeloid Modulators	STING, NLRP3, CD40	NLRP3 inhibitors; CD40 agonistic antibodies.
Antigen-based	Tumor-Associated Antigens (TAAs)	B7-H3 (CD276), TROP2, Claudin 18.2, HER3	Next-generation ADCs; Tri-specific T-cell engagers; Radioligand therapies.
	Tumor Neoantigens	Shared Neoantigens (KRAS/TP53 mutations), Patient-specific SNVs	Personalized mRNA vaccines; Off-the-shelf TCR-T therapies.
TME & Emerging Modalities	Phagocytosis Checkpoints ("Don't Eat Me")	CD47/SIRPα, CD24	SIRPα-Fc fusion proteins; CD47/PD-L1 bispecific antibodies.
	Targeted Delivery Receptors	PSMA, Nectin-4, DLL3	Radio-conjugates; DLL3-targeted T-cell engagers.

2.1.1. Tumor-intrinsic targets

Tumor-intrinsic targets refer to molecules essential for tumor cell survival and proliferation, including driver oncogenes (such as EGFR and KRAS), synthetic lethal partners, DNA damage repair proteins, and metabolic/epigenetic regulators [19]. Among these, synthetic lethality strategies have emerged as a key focus for AI-assisted discovery due to their capacity to therapeutically exploit loss-of-function tumor suppressors. While tumor suppressors are inherently difficult to target directly, their synthetic lethal partners can be inhibited to achieve a similar therapeutic effect.

Synthetic lethality describes a genetic phenomenon wherein the simultaneous inactivation of two genes leads to cell death, while the inactivation of either gene alone is compatible with cell survival [20]. In cancer therapy, this principle is exploited to target tumors harboring loss-of-function mutations in undruggable tumor suppressor genes by inhibiting their corresponding synthetic lethal partners, thereby achieving the selective killing of tumor cells [21].

The sensitivity of BRCA1/2-deficient tumors to PARP inhibitors represents a landmark in the clinical translation of the synthetic lethality concept [22]. BRCA1/2 proteins participate in the homologous recombination (HR) repair of DNA double-strand breaks; their functional loss compromises HR and increases reliance on PARP-associated DNA damage responses during replication stress [23]. PARP inhibitors (including PARP trapping) exacerbate replication-associated lesions and promote the accumulation of lethal DNA damage specifically in BRCA-deficient cells [24]. Based on this mechanism, olaparib was approved by the FDA in 2014 as the first clinically approved therapy explicitly grounded in synthetic lethality rationale [25].

In the exploration of novel synthetic lethal relationships, the combination of MTAP deletion and MAT2A inhibition has attracted considerable attention [26]. The MTAP gene is located adjacent to CDKN2A in the 9p21 chromosomal region, with approximately 15% of human tumors, depending on

cancer type, harboring co-deletion of this region [27]. MTAP loss leads to the accumulation of methylthioadenosine (MTA), rendering cells metabolically dependent on the MAT2A–PRMT5 axis [28].

AI technologies have significantly accelerated the systematic discovery of synthetic lethal targets. Machine learning models trained on CRISPR screening data can now predict synthetic lethal interactions at a genome-wide scale [29]. Simultaneously, graph neural networks integrate protein-protein interaction networks and gene co-expression information to identify functionally complementary gene modules [30]. Notably, deep learning frameworks such as SLant and GCATSL have demonstrated superior predictive performance compared to traditional methods across various independent datasets [31,32].

Beyond synthetic lethality, targeted protein degradation, encompassing PROTACs and molecular glues, has emerged as a powerful strategy for tackling traditionally undruggable targets such as MYC and STAT3. However, the rational design of PROTACs presents unique computational challenges: three molecular components (E3 ligase binder, target protein binder, and linker) must be simultaneously optimized, and the formation of a productive ternary complex is essential for efficient target degradation. AI methods are addressing these multi-component challenges at several levels. At the ternary complex prediction level, machine learning models such as DeepPROTACs [33] integrate structural features of E3 ligases, target proteins, and linker geometries with physicochemical descriptors to evaluate ternary complex compatibility and predict degradation potential. At the linker design level, generative models including PROTAC-INVENT [34] and Link-INVENT [35] employ reinforcement learning to explore chemical space for linkers with optimized length, flexibility, and cell permeability, enabling the automated generation of novel PROTACs candidates with desired pharmacological profiles. At the degradation outcome level, GNNs have been applied to predict degradation efficiency and target selectivity from molecular graphs, providing a computational framework to prioritize degrader candidates prior to costly experimental synthesis. For molecular glue discovery, which requires the identification of compounds that stabilize neo-substrate recruitment to E3 ligases, AI-driven virtual screening of E3 ligase–substrate protein interfaces are beginning to expand the accessible chemical space, although this subfield remains at an earlier stage of methodological maturity compared to PROTACs design. Databases such as PROTAC-DB [36] serve as critical data infrastructure for training and benchmarking these AI models.

2.1.2. Immune modulation targets

The clinical success of immune checkpoint inhibitors has ushered in a new era of cancer immunotherapy [37]. PD-1/PD-L1 inhibitors have been approved for multiple solid tumors, while CTLA-4 inhibitors have demonstrated durable efficacy in melanoma and other malignancies [38]. However, only approximately 20%–30% of patients achieve durable responses to existing immune checkpoint inhibitors [39]. Consequently, researchers are actively exploring novel immunomodulatory factors as next-generation therapeutic targets, a process that requires a deeper understanding of the endogenous signaling mechanisms in immune cells and their interactions with the tumor microenvironment [40].

Beyond classical cell surface checkpoints, intracellular signaling regulators have garnered attention as novel targets. Members of the diacylglycerol kinase (DGK) family, for instance, have been proposed as intracellular immune checkpoints [41]. DGKα and DGKζ negatively regulate T cell receptor (TCR) downstream signaling by metabolizing the diacylglycerol (DAG) signaling molecule [42]. Preclinical

studies have demonstrated that the inhibition of DGKα/ζ enhances T cell antitumor effector functions and exhibits synergistic effects with PD-1 blockade [43]. Based on this mechanism, Bristol Myers Squibb has developed the dual DGKα/ζ inhibitor BMS-986408, which has now entered clinical evaluation [44].

With regard to the balance between target druggability and safety, the upstream regulation of the CD47-SIRPα axis provides instructive insights. CD47, as a "don't eat me" signal molecule, is widely expressed on normal cells; consequently, direct blockade of the CD47-SIRPα interaction may lead to the unwanted clearance of normal cells such as erythrocytes [45]. QPCTL is a glutaminyl cyclase that catalyzes the *N*-terminal pyroglutamate modification of CD47, a modification essential for the high-affinity binding between CD47 and SIRPα [46]. Indirect modulation of CD47 function through QPCTL inhibition may theoretically achieve a more favorable therapeutic window [47].

AI technologies play multiple roles in the discovery of novel immune targets. Single-cell transcriptomic analysis combined with machine learning can identify the functional states and inhibitory markers of tumor-infiltrating immune cells [48]. AlphaFold-based protein structure prediction and molecular docking simulations facilitate the assessment of candidate target druggability [49]. Furthermore, methods integrating immune repertoire sequencing data with deep learning can predict T cell response characteristics following novel target intervention [50]. Recent studies have employed GNNs to identify immunosuppressive ligand-receptor pairs from cell communication networks within the tumor microenvironment, providing a computational framework for the systematic discovery of novel checkpoints [51].

2.1.3. Antigen-based targets

Antigen-based targets refer to tumor-associated molecules that can be recognized by the immune system or engineered immune cells, including mutation-derived neoantigens and TAAs. These serve as the molecular foundation for cancer vaccines, TCR-T, CAR-T, and other immunotherapeutic approaches [52].

Tumor neoantigens are tumor-specific antigenic epitopes generated by somatic mutations. Because they are absent from normal tissues, they possess high immunogenic potential [53]. Effective neoantigen presentation involves a complex multi-step cascade process: mutant proteins must be degraded by the proteasome to generate peptides, which are then transported into the endoplasmic reticulum via TAP transporters and bind to MHC class I molecules. The resulting pMHC complexes are presented on the cell surface and ultimately recognized by $CD8^+$ T cell TCRs to trigger immune responses [54]. Failure at any step results in loss of immunogenicity, and studies indicate that only approximately 1%–2% of nonsynonymous mutations can generate effectively immunogenic neoantigens [55]. Key factors influencing neoantigen immunogenicity include: binding affinity between mutant peptides and MHC molecules, the stability of pMHC complexes, the position of the mutation site within the peptide, physicochemical differences between mutant and wild-type amino acids, and the frequency of antigen-specific precursors in the T cell repertoire [56].

Among these determinants, T cell precursor frequency deserves particular attention because it is both critical for neoantigen immunogenicity and highly variable across patients, making it a major source of prediction uncertainty. AI methods are beginning to address this patient-specific variability through multiple complementary strategies. At the data integration level, coupling patient-derived TCR repertoire sequencing (TCR-seq) data with neoantigen prediction pipelines enables the estimation of pre-existing T cell reactivity against candidate epitopes. At the modeling level, transfer learning approaches leverage population-scale TCR repertoire databases such as VDJdb [57] and IEDB [58] to

learn generalizable binding patterns, which can then be fine-tuned on individual patient data to improve personalized predictions. More ambitiously, multi-modal frameworks that jointly model HLA genotype, peptide-MHC binding affinity, and TCR repertoire diversity are being developed to provide integrated immunogenicity scores that account for patient-specific immune contexts. Emerging computational tools for predicting TCR-pMHC binding specificity, including ERGO-II [59] and NetTCR [60], further support the prioritization of neoantigens based on individual immune repertoires, although achieving clinical-grade prediction accuracy remains an active area of investigation.

Beyond neoantigens, TAAs represent important targets for CAR-T, antibody–drug conjugates (ADCs), and bispecific antibodies. Unlike neoantigens, TAAs are not mutation-derived but rather normal proteins that are overexpressed or aberrantly expressed in tumor cells, such as CD19, BCMA, and HER2 [61]. CAR-T target selection requires balancing tumor specificity against normal tissue toxicity. AI methods can systematically evaluate candidate antigen expression profiles from single-cell and spatial transcriptomic data to identify target combinations with high tumor specificity and low off-target risk [62]. Furthermore, logic-gated CAR-T strategies enhance tumor specificity through the combinatorial recognition of multiple antigens, where AI can assist in optimizing the discriminatory power and safety window of these antigen combinations [63].

Regarding computational prediction, algorithms such as NetMHCpan-4.1 and BigMHC have optimized predictive performance by focusing on MHC binding affinity, antigen processing efficiency, and multi-dimensional feature integration [64,65]. Deep learning models including pMTnet and TITAN are attempting to address the more challenging task of pMHC-TCR binding prediction [66,67]. Whether neoantigens can become translatable therapeutic targets depends on multiple constraints including antigen processing and MHC presentation, pMHC stability, and TCR recognition. Therefore, the key to computational prediction lies in prioritizing the few candidate epitopes with genuine immunogenicity from numerous somatic mutations, which must ultimately be confirmed through experimental validation.

2.2. AI-driven target discovery methods

Building upon the biological characteristics of tumor-intrinsic and immune-directed targets (including immune checkpoint/modulation and antigen-based subtypes), the key role of AI methods lies in generating testable target hypotheses from multi-scale evidence and performing prioritization. The following sections elaborate on three hierarchical aspects: (i) virtual spatial omics and computational pathology at the tissue scale, (ii) AI virtual cells, single-cell foundation models, and virtual perturbation at the cellular scale, and (iii) knowledge graph reasoning with mechanistic evidence integration at the knowledge scale. As illustrated in Figure 2B, these multi-scale AI approaches integrate pathology foundation models, virtual spatial omics, AI virtual cells, single-cell foundation models, and knowledge graph reasoning to systematically prioritize candidate targets synthesized from diverse biological evidence.

2.2.1. Tissue-scale modeling: virtual spatial omics and computational pathology

At the tissue scale, AI methods leverage histopathological images to infer molecular and spatial information relevant to target discovery. Two main paradigms have emerged: virtual spatial omics, which estimates spatially resolved molecular profiles (e.g., spatial gene expression) from routine histology images—typically learned from paired spatial transcriptomic references and sometimes aided

by sparse spatial measurements—and molecular phenotype inference, which predicts molecular features such as mutation status and biomarkers directly from H&E whole-slide images (WSIs).

For virtual spatial omics, recent methods aim to predict spatially resolved gene expression from H&E images, potentially democratizing access to spatial molecular information. For instance, iStar integrates histology images with sparse spatial measurements to achieve super-resolution spatial gene expression prediction at near-single-cell resolution. TRIPLEX adopts a multi-resolution architecture to capture spatial and morphological context at different scales for improved gene expression prediction [68,69]. More recently, CarHE employs contrastive learning to align image features with spatial single-cell gene expressions, enabling the prediction of over 10,000 genes solely from H&E images during inference [70]. A comprehensive benchmark study evaluating multiple methods for predicting spatial gene expression from histology images has provided systematic guidance for model selection [71]. Collectively, these approaches provide spatially contextualized evidence for target discovery within the tumor microenvironment, while requiring careful external validation due to potential domain shifts across cohorts, staining protocols, and scanners.

For molecular phenotype inference, studies have demonstrated that WSIs contain morphological signals correlated with molecular phenotypes. Specifically, deep learning models can, under certain conditions, predict microsatellite instability (MSI), deficient mismatch repair (dMMR), and the mutation status of specific driver genes [72,73], thereby providing auxiliary cues for low-cost pre-screening and stratification prior to molecular testing.

The emergence of pathology foundation models has further enhanced the generalizability of this paradigm across different tasks. Slide-level foundation models, exemplified by Prov-GigaPath, learn transferable visual representations through large-scale self-supervised pretraining [74]. Universal pathology self-supervised models such as UNI are pretrained on massive H&E datasets, providing a robust feature backbone for diverse computational pathology tasks [75]. Vision-language models such as CONCH further incorporate image-text contrastive learning to align morphological representations with textual semantics, thereby enhancing cross-task adaptability [76]. These models provide a unified feature extraction framework for downstream tasks including tumor classification, tumor microenvironment phenotyping, and prognostic risk stratification, and serve as auxiliary evidence sources for target discovery and patient stratification.

2.2.2. Cell-scale modeling: virtual cells and single-cell foundation models

At the cellular scale, AI methods exploit single-cell data to identify cell type-specific targets and predict cellular responses to perturbations. Three interconnected paradigms have emerged: (i) AI virtual cells (AIVC), which aim to construct comprehensive “digital twins” of cells capable of simulating cellular behavior, state transitions, and responses to perturbations in a holistic manner; (ii) single-cell foundation models, which learn generalizable representations of gene expression across diverse cellular contexts; and (iii) virtual perturbation, which computationally simulates the effects of genetic or chemical interventions on cellular states.

For AI virtual cells, a landmark *Cell* perspective outlined a roadmap for building virtual cells capable of simulating cellular behavior at the molecular and structural levels [77]. The Arc Institute has released two complementary virtual cell models. State, the first-generation model, was trained on observational data from nearly 170 million cells and perturbational data from over 100 million cells

across 70 cell lines [78]. State operates across physical scales through two interlocking modules: the state embedding model that learns representations of individual cells, and the state transition model that predicts how perturbations shift cellular states. More recently, Stack introduced in-context learning capabilities to single-cell biology [79]. Trained on 149 million uniformly preprocessed human single cells, Stack leverages tabular attention to generate cell representations informed by cellular context, enabling the prediction of perturbation effects across entirely new biological conditions without requiring condition-specific perturbational data. These initiatives may ultimately enable comprehensive *in silico* simulation of target perturbation effects across diverse cellular contexts.

For single-cell foundation models, a growing family of Transformer-based methods has emerged to learn generalizable representations of gene expression. scGPT, pretrained on over 33 million cells, utilizes a transformer architecture and is pretrained with masked gene prediction-style objectives to model gene expression patterns [80]. Geneformer uses rank-based gene encoding to capture transcriptome-wide context from approximately 30 million cells [81]. Other notable models include scFoundation, which scales to 50 million cells by employing a read-depth recovery task for cross-dataset harmonization [82], and UCE (universal cell embeddings), which learns unified representations across species and sequencing platforms [83]. More recently, CellFM has pushed the scale further to 800 million parameters trained on 100 million human cells [84]. These models can be applied to downstream tasks including cell type annotation, batch integration, trajectory inference, and cross-dataset integration. However, recent systematic benchmarks have revealed that the zero-shot performance of these models does not consistently outperform simpler baselines such as highly variable gene selection, highlighting the necessity of rigorous evaluation and the need for further methodological development [85]. Several factors may account for this performance gap. First, single-cell datasets are characterized by high dropout rates, batch effects, and substantial technical noise. Large-scale pretraining may amplify rather than correct these artifacts when the training corpus lacks sufficient quality control. Second, foundation models pretrained with broad transcriptomic objectives do not necessarily learn representations that are optimally aligned with specific downstream tasks such as cell-type annotation or perturbation response prediction. Without dedicated fine-tuning, this task-model mismatch allows simpler, task-specific baselines, which often incorporate strong domain-specific inductive biases, to remain competitive. Third, the mapping from gene expression patterns to biological function is highly nonlinear and context-dependent, posing fundamental challenges for zero-shot generalization across diverse tissues, disease states, and experimental platforms. These observations do not diminish the long-term potential of foundation models but rather underscore that domain-specific adaptation strategies remain essential for translating model scale into consistent downstream performance gains.

For virtual perturbation, *in silico* perturbation represents a promising application direction for single-cell foundation models in target discovery, aiming to computationally simulate the effects of gene knockout or overexpression on cellular expression profiles [86]. However, systematic benchmark evaluations have shown that the advantages of current models over simple baselines in perturbation prediction tasks remain inconsistent [87]. Alternatively, methods based on gene regulatory networks, such as CellOracle and scTenifoldKnk, offer more mechanistically interpretable alternatives [88,89]. Virtual perturbation prediction is still in its early developmental stage, and target hypotheses generated by these approaches require rigorous validation through Perturb-seq or CRISPR screening experiments.

2.2.3. Knowledge-scale modeling: knowledge graphs and mechanistic evidence integration

At the knowledge scale, knowledge graphs provide a unified framework for integrating multi-source biological knowledge through the structured representation of relationships among entities such as drugs, targets, genes, diseases, and pathways [90]. Early methods such as graph embedding and link prediction could infer previously unreported drug-disease or target-disease associations [91]. In recent years, the fusion of GNNs with large language models (LLMs) has significantly enhanced both predictive performance and generalization capability.

TxGNN is a GNN-based foundation model designed for therapeutic drug indication prediction, covering over 17,000 diseases and demonstrating excellent performance in zero-shot prediction tasks, particularly for rare diseases lacking known therapeutic options [92]. SHEPHERD integrates knowledge graphs with patient phenotype data to support phenotype-driven rare disease diagnosis, facilitating causal gene prioritization and patient matching ("patients-like-me") for novel or heterogeneous presentations [93]. Lastly, DrugCLIP [94] has demonstrated the utility of contrastive learning for large-scale virtual screening by reformulating it as a dense retrieval task.

With respect to mechanistic evidence integration, PINNACLE constructs context-aware representations of protein-protein interactions, which are capable of distinguishing target functions across different cell types and tissue contexts [95]. Biomedical multimodal large models such as BiomedGPT further integrate literature, molecular structures, and omics data to provide interpretable mechanistic hypotheses for target-disease associations [96]. These methods, combined with transcriptomic reversal signatures and real-world data analysis, provide multi-dimensional supporting evidence from both functional and clinical perspectives.

3. Oncology drug discovery based on virtual screening

Once potential therapeutic targets have been identified and validated through the multi-scale approaches described above, the next critical task is to screen candidate molecules capable of effectively modulating these targets from vast compound libraries. Virtual screening (VS) constitutes a foundational technology in drug discovery (Figure 3), aiming to identify molecules with potential biological activity from massive compound libraries, thereby substantially reducing both the time and economic costs of experimental screening [97]. In oncology drug development, the complexity of target-associated signaling pathways and multidrug resistance mechanisms imposes elevated requirements on the accuracy of molecular activity assessment. Binding affinity, as a key metric quantifying the strength of small-molecule-target protein interactions, largely determines the hit rate of virtual screening through its prediction accuracy [98,99].

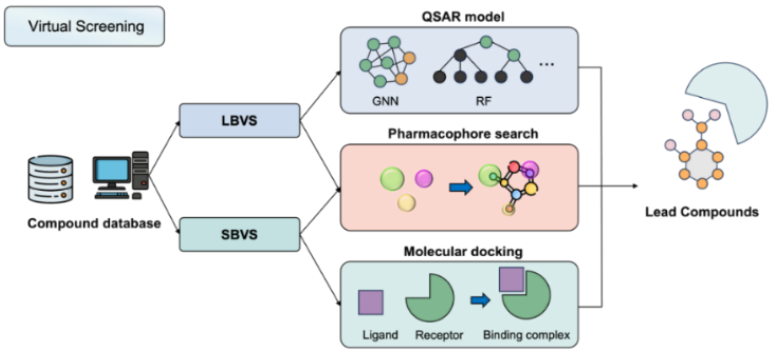

Figure 3. General workflow of AI-driven virtual screening.

3.1. Traditional computer-aided drug design methods

Early virtual screening primarily relied on classical CADD methods, such as molecular docking and pharmacophore modeling. Molecular docking predicts binding affinity by simulating the spatial orientation and interaction patterns of ligands within target protein binding pockets, while pharmacophore modeling extracts key three-dimensional spatial and chemical features from known active molecules as screening criteria. These methods have served vital roles in drug discovery over the past several decades but exhibit notable limitations.

First, constrained by computational efficiency and available resources, traditional virtual screening is typically limited to compound libraries of 10^5 to 10^7 magnitude, struggling to effectively cover the vast chemical space. Second, traditional empirical or semi-empirical scoring functions rely on simplified energy terms describing hydrogen bonds, hydrophobic interactions, and electrostatic effects; however, they struggle to adequately capture solvent effects and entropic contributions, leading to systematic errors in affinity prediction [100–102]. Moreover, most methods offer limited treatment of receptor flexibility and conformational variation, a problem particularly pronounced in highly dynamic oncology targets such as kinases, often resulting in substantially reduced enrichment rates [103].

Third, in ligand-based scenarios, classical methods rely on combinations of hand-crafted descriptors with traditional machine learning algorithms (such as random forests or support vector machines) [104]. While these models can capture structural correlations associated with biological activity, they exhibit limited expressiveness in handling complex non-linear relationships and demonstrate reduced generalization capacity when applied to small-sample or imbalanced datasets. Finally, regarding post-screening validation, traditional computational validation centers on molecular dynamics (MD) simulations and endpoint free energy calculations (such as MM-GBSA/MM-PBSA). These methods partially account for conformational dynamics and solvation but are computationally demanding and sensitive to initial binding poses and parameterization, thereby limiting their scalability [105].

These longstanding challenges have motivated the development of modern AI approaches aimed at more faithfully capturing thermodynamic and dynamic effects. Physics-informed GNNs such as PIGNet2 [106] integrate physical energy components within deep learning architectures, improving the physical consistency of binding affinity scoring and virtual screening. Machine learning extensions of implicit solvation models are designed to better account for solvent-related energetic contributions [107]. In addition, models trained on large-scale molecular dynamics datasets (e.g., MISATO and PLAS-20k) leverage conformational ensembles and atomic trajectories to implicitly capture dynamic fluctuations and solvent-mediated interactions that are difficult to represent in traditional static scoring frameworks [108,109].

3.2. AI-empowered screening and closed-loop validation

The emergence of deep learning has propelled a paradigm shift in virtual screening methodologies. Models exemplified by GNNs and equivariant neural networks can directly extract richer spatial and topological features from molecular structures, substantially improving binding affinity prediction performance. Within this framework, virtual screening has evolved from qualitative, similarity-based screening to quantitative processes oriented toward binding strength. Concurrently, improvements in computational efficiency have enabled virtual screening to scale to ultra-large libraries containing 10^{10} to 10^{15} compounds [110], achieving a leap from searching limited candidate sets to exploring entire chemical spaces.

In structure-based virtual screening (SBVS), deep learning is increasingly being integrated to address geometric and ranking challenges. Early studies employed convolutional neural networks (CNNs) such as GNINA to combine deep learning with physics-based conformational sampling [111–113]. To address geometric limitations, E(3)-equivariant graph neural networks (e.g., EquiBind, TankBind, and EquiScore) have emerged, explicitly modeling geometric relationships between atomic coordinates while respecting SE(3) symmetry constraints, thereby achieving more precise spatial characterization of interactions [114–116]. Addressing the challenge of fine-grained affinity ranking among closely related ligands for the same target, the pairwise binding comparison network (PBCNet) employs physics-informed graph attention mechanisms to directly model relative binding strength [117]. Furthermore, contrastive learning and joint embedding strategies (e.g., DrugCLIP [94]) allow for efficient pre-filtering by constructing shared representation spaces for protein pockets and small molecules, effectively bypassing exhaustive docking. Additionally, diffusion-based generative methods (such as DiffDock and SurfDock) offer an alternative paradigm for SBVS by directly predicting binding poses without relying on traditional docking searches [118–120].

In ligand-based virtual screening (LBVS) [121], deep learning has fundamentally reshaped the paradigm, particularly when reliable target structures are unavailable. GNN-based models (e.g., D-MPNN and Chemprop) learn structural representations directly from molecular graphs, eliminating tedious feature engineering and consistently outperforming traditional QSAR in capturing complex structure-activity patterns [122]. Similar strategies have been extended to infectious diseases [123,124] as well as oncology and neurological targets [125], yielding hit rates markedly higher than traditional high-throughput screening. As applications broaden, targeted innovations have emerged to address emerging challenges [126]: AdaptToR [127] employs adaptive anchor selection to enhance multi-target prediction, while eMOSAIC [128] combines multimodal representations with outlier detection for robust uncertainty estimation. The latter effectively addresses out-of-distribution (OOD) generalization issues, a common bottleneck in ligand-based discovery.

To bridge the gap between computational prediction and experimental reality [129], AI integration has substantially improved validation efficiency. Machine learning accelerates MD and free energy calculations through active learning-based prioritization [130,131], while dynamics-aware models (e.g., Dynaformer [132] and ProtMD [133]) explicitly incorporate conformational ensembles or pose perturbations to simulate receptor flexibility. Neural network potentials (NNPs) provide high accuracy at lower cost, enabling broader dynamics assessment [134,135]. Ultimately, closed-loop strategies integrating screening, experimental feedback (e.g., SPR, ITC, and functional assays) [136,137], and iterative model optimization are receiving increasing attention, positioning virtual screening to assume a more central role in precision oncology [138–141]. As standardized benchmarks and prospective validation frameworks mature, virtual screening is well-positioned to transform the landscape of precision oncology drug discovery.

3.3. From in silico hits to bioactive leads

AI-driven virtual screening has transcended theoretical utility to deliver experimentally validated lead compounds for complex oncology targets. Recent advances demonstrate the successful translation of computational predictions into bioactive molecules across four critical dimensions: kinase resistance, protein–protein interactions (PPIs), synthetic lethality, and drug repurposing.

In addressing kinase resistance, AI workflows have successfully tackled the EGFR C797S mutation, a major bottleneck for third-generation non-small cell lung cancer (NSCLC) therapies. By integrating machine learning-based screening with allosteric pocket analysis, researchers identified fourth-generation inhibitor candidates effective against L858R/T790M/C797S triple mutants [142,143]. These AI-derived hits demonstrated significant cytotoxicity in resistant cell lines and favorable pharmacokinetic profiles, underscoring the capability of computational models to navigate restrictive conformational spaces where traditional design often fails.

Regarding hard-to-drug PPI interfaces, AI has facilitated a transition from monoclonal antibodies to oral bioavailable small molecules. For the PD-1/PD-L1 checkpoint, pharmacophore-guided screening identified non-peptide small molecules capable of physically blocking the interaction surface [144]. Experimental validation confirmed that these compounds restore T-cell function and induce immune responses *in vivo*, validating AI's potential to "chemicalize" large biological targets.

In the realm of synthetic lethality, AI is accelerating the discovery of inhibitors for DNA damage response targets such as DNA polymerase theta (Polθ, POLQ). Consensus machine learning models and generative platforms (e.g., Chemistry42) have streamlined the screening of millions of compounds, yielding nanomolar-potency helicase inhibitors with novel scaffolds (such as 3-hydroxymethyl-azetidine derivatives) [145,146]. These compounds exhibit selective lethality in BRCA-deficient tumor models, showcasing the powerful synergy between predictive screening and generative design.

Finally, in system-level drug repurposing, AI platforms integrating multi-omics data identified (Z)-endoxifen as a potential therapeutic candidate for glioblastoma multiforme, with *in vitro* validation demonstrating its anti-proliferative and pro-apoptotic effects and known blood–brain barrier permeability further supporting its relevance to central nervous system tumors [147]. Collectively, these applications confirm that AI virtual screening has matured into a robust engine capable of delivering structurally novel, thermodynamically stable, and biologically active leads for precision oncology.

4. Generative artificial intelligence for de novo oncology drug design

Traditional computational workflows in drug discovery have long relied on enumerative search paradigms, where the discovery potential is strictly confined to the finite size and diversity of existing compound libraries. When confronting the high heterogeneity of tumor targets and the dynamic remodeling of binding interfaces driven by resistance mutations, such passive selection strategies often prove insufficient. Generative AI has fundamentally transformed this landscape by introducing a diverse arsenal of deep generative architectures, including variational autoencoders (VAEs), generative adversarial networks (GANs), normalizing flows, Transformers, and diffusion models, each offering distinct advantages for molecular design (Figure 4). This marks a paradigm shift from the screening of discrete candidate sets to the goal-driven, de novo creation of novel molecular entities. By learning the high-dimensional probability distributions of sequence, structure, and physicochemical properties embedded in massive datasets, generative models facilitate conditional sampling within continuous latent spaces. This capability enables the de novo design of therapeutics with tailored functional constraints across multiple molecular scales—ranging from small molecules to peptides, protein binders, and nucleic acids—thereby providing a systematic technical avenue to address undruggable targets and overcome rapid resistance evolution [148].

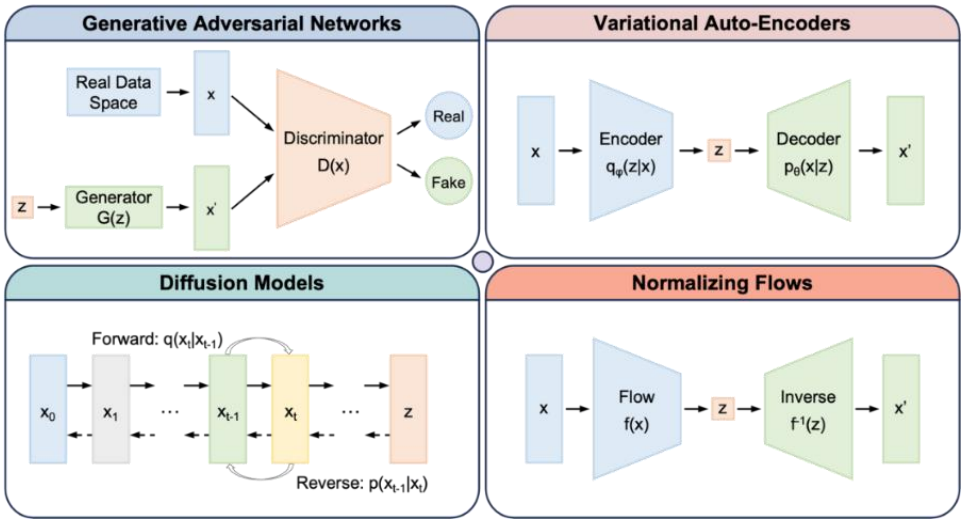

Figure 4. Schematic overview of generative artificial intelligence models. Representative architectures of generative models commonly used in drug design, including generative adversarial networks, variational auto-encoders, diffusion models, and normalizing flows.

4.1. Evolution of core generative model architectures

The core principle of generative models lies in learning the statistical patterns of molecules across sequence, structure, and physicochemical property spaces, thereby enabling conditional generation within latent spaces (Figure 4). Early molecular generation research primarily employed sequence modeling frameworks such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), utilizing simplified molecular input line entry system (SMILES) representations or primary sequences as input for autoregressive generation. These early sequence-based approaches laid the groundwork for molecular generation and demonstrated the feasibility of reinforcement learning-guided optimization. However, their inherent limitations in capturing global molecular topology motivated the development of more expressive architectures. VAEs and GANs were developed to map discrete molecular representations to continuous latent spaces, transforming the generation problem into latent space sampling and optimization, thus providing a natural computational framework for scaffold hopping and multi-objective optimization. With the growth of model scale and data volume, Transformer architectures based on self-attention mechanisms have become the mainstream framework owing to their capacity for explicitly modeling long-range dependencies, demonstrating powerful versatility across small molecule generation, protein sequence design, and nucleic acid modeling. More recently, the emergence of diffusion models and flow matching models has further advanced generative design toward structural fidelity and physical consistency; these models learn the continuous evolution from noise distributions to real molecular distributions, naturally incorporating geometric constraints and energetic preferences during generation, thereby exhibiting unique advantages in three-dimensional structure generation and interaction interface engineering.

4.2. Generative design of small molecule drugs

By learning the distributions of large-scale molecular data, generative models can navigate directly within the latent chemical space to produce candidates with desired biological activity, drug-likeness,

and synthetic accessibility, thereby transcending the bottlenecks of traditional virtual screening constrained by the size and diversity of existing compound libraries [149,150] (Figure 5A).

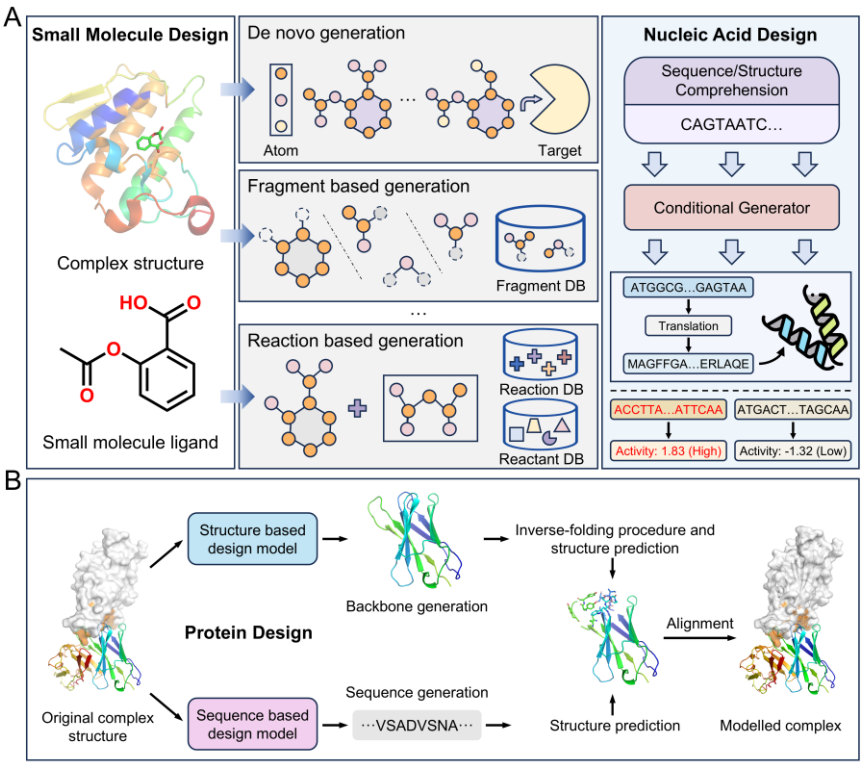

Figure 5. Generative AI applications across diverse therapeutic modalities. **(A)** Small molecule & nucleic acid design. The left panel illustrates small molecule generation strategies, including de novo design, fragment-based growing, and reaction-based generation. The right panel depicts nucleic acid design optimization, covering sequence/structure comprehension and conditional generation for specific translational activities; **(B)** Protein design. A generative workflow for protein therapeutics, highlighting the interplay between structure-based design, backbone generation, and sequence optimization to create functional protein binders.

De novo molecular generation constitutes the core application in this field. Early methods relied on RNNs and LSTMs for autoregressive generation; for instance, the curriculum learning strategy implemented by Guo and colleagues within the REINVENT platform accelerated learning [151], while the iterative optimization workflow proposed by Mokaya and colleagues enabled generation based on both known and unseen molecular features [152]. With the increasing incorporation of structural information, generative models are now designed to interact with protein binding pockets. Specifically, ResGen employs a multiscale hierarchical autoregressive approach to generate three-dimensional molecules directly within binding pockets [153], SurfGen utilizes graph neural networks to encode spatial interactions between pocket topology and ligand atoms [154], KGDiff integrates chemical knowledge guidance into the denoising process of diffusion models [155], and Tree-Invent combines reinforcement learning with topological tree constraints for de novo generation and scaffold hopping [156]. In terms of prospective applications, PI3Kγ inhibitors generated by Moret and colleagues using chemical language models that integrate structural and bioactivity information demonstrated effective inhibition of the PI3K/Akt pathway in brain cancer cell models [157]. Furthermore, the GENTRL platform developed by

in silico Medicine, combining VAEs with reinforcement learning, designed novel triazolopyridine candidates targeting the DDR1 kinase within only 21 days, with lead compounds exhibiting IC_{50} values at the 10 nM level and favorable pharmacokinetic profiles in pulmonary fibrosis and tumor models [158].

Scaffold hopping and fragment-based molecular generation represent key strategies for generating molecules with novel bioactivity in modern drug discovery. Lim and colleagues employed VAEs-based graph generation models for scaffold-centric molecular design [159], SyntaLinker Hybrid combines fragment hybridization with transfer learning [160], and DeepHop utilizes protein sequence information and 3D molecular conformations for cross-scaffold generation [161]. In applications requiring precise linker design such as PROTACs and ADCs, methods including DeLinker [162], DiffLinker [163], and Linker-GPT [164] leverage GNNs and Transformer architectures to precisely control linker length, hydrophobicity, and spatial conformation.

Pharmacophore-guided molecular generation enhances the fidelity of generated molecules within active sites by enforcing interaction patterns such as hydrogen-bond donors/acceptors, hydrophobic centers, and charge distributions [165,166]. TransPharmer employs ligand-based pharmacophore fingerprints as conditional inputs to GPT architectures [167], DEVELOP embeds three-dimensional pharmacophore information into graph neural networks to ensure sustained alignment with pharmacophore constraints [168], and PGMG leverages pharmacophore features to overcome data scarcity [169]. Omics-guided molecular generation transcends the dependence on predefined targets: PaccMannRL uses target cell transcriptomic profiles as conditional inputs to generate candidate molecules capable of modulating specific gene expression patterns [170], BiCEV [171] utilizes differential gene expression data to design small molecules with multi-target modulation potential, and G2D-Diff [172] designs small molecule therapeutics with targeted anticancer activity based on specific cancer genotypes. Reaction-aware molecular generation addresses the critical bottleneck of synthetic accessibility: the Synthesia [173] framework guides structural modifications through retrosynthetic analysis, substantially reducing the generation of unsynthesizable compounds; Uni-RXN integrates reaction classification, condition modeling, and conditional molecular generation into a unified model [174].

In summary, the generative design of small molecules has evolved from early sequence models relying on SMILES or two-dimensional topology to systematic methodologies integrating three-dimensional structures, interaction patterns, and chemical rules, advancing molecular design from mere structural feasibility toward integrated functional controllability and synthetic operability.

4.3. Generative design of protein and peptide therapeutics

Compared to small molecules, protein and peptide therapeutics possess unique advantages in targeting complex biological interfaces, modulating protein-protein interactions, and achieving high-specificity therapeutic functions, having emerged as pivotal modalities for oncology, biologics, and gene editing applications. However, the vast sequence space, stringent conformational constraints, and multidimensional functional requirements of proteins and peptides render traditional empirical screening methods inefficient. The rapid development of generative AI models provides novel computational tools for de novo design and directed optimization of protein and peptide therapeutics, enabling a paradigm shift from stochastic screening to controlled generation [175] (Figure 5B).

In peptide therapeutic design, sequence-based and structure-based strategies form complementary approaches. Sequence-based methods leverage protein language models (PLMs) to learn the statistical

distributions of amino acid sequences and their implicit associations with function, proving particularly suitable for tumor-associated targets lacking stable binding pockets or exhibiting intrinsic disorder. PepPrCLIP combines PLMs with contrastive learning frameworks and has been successfully applied to conformationally diverse or intrinsically disordered targets, including UltraID enzyme-inhibitory peptides and peptide binders against the oncogenic fusion protein SS18-SSX1 [176]. PepMLM fine-tunes ESM-2 through cross-sequence masking strategies to reconstruct complete binding regions without structural input [177]. PepINVENT extends generative peptide design to broader chemical space including non-natural amino acids [178]. Structure-based peptide generation models explicitly model peptide-protein interactions in three-dimensional space: PepMimic employs latent space diffusion models to simultaneously generate peptide sequences and three-dimensional structures, successfully designing peptides with nanomolar affinity against targets including PD-L1, CD38, BCMA, HER2, and CD4 [179]. For cyclic peptide design, AfCycDesign modifies the relative position encoding of AlphaFold2 to incorporate cyclization constraints, identifying binders with nanomolar IC_{50} values against MDM2 and Keap1 among over 10,000 structurally diverse cyclic peptide backbones [180]; RFpeptides, built upon RFdiffusion and RoseTTAFold2, produced nanomolar-affinity cyclic peptides against MCL1, MDM2, GABARAP, and RbtA, with multiple crystal structures showing excellent agreement with designed models (C_α RMSD < 1.5 Å) [181].

In the domain of de novo protein binder design, the integration of deep learning models with traditional protein design workflows has achieved multiple breakthroughs. Starting from the known structures of cancer-associated targets, Cao and colleagues employed free amino acid matching and spatial alignment algorithms to identify compatible binder positions, producing stable binding proteins capable of recognizing EGFR and CD276 [182]. The Baker laboratory utilized RFdiffusion with precise geometric constraints to de novo design protein binders capable of specifically binding and neutralizing multiple three-finger snake venom toxin subtypes [183], and designed miniprotein binders targeting PD-L1 comprising only approximately 60 amino acids, exhibiting picomolar affinity, high thermal stability (T_m > 90 ℃), and superior tumor tissue penetration compared to conventional monoclonal antibodies [184]. BindCraft, as an automated deep learning-based workflow, leverages AlphaFold2 to directly optimize protein-protein interfaces during generation, validated across multiple complex interface targets [185]. Most recently, BoltzGen [186] has emerged as a fully open-source, all-atom generative model that unifies structure prediction and binder design, enabling the flexible creation of diverse modalities—such as nanobodies and cyclic peptides—against a wide range of biomolecular targets. The Chai-2 [187] model has pushed the limits of experimental reliability, achieving a transformative 16% hit rate in fully de novo antibody design and a 68% success rate in miniprotein design, effectively moving the field from large-scale screening toward high-precision, atomic-level molecular engineering.

Membrane protein targets, long considered challenging due to their high hydrophobicity and lipid bilayer dependence, have now achieved substantial progress through generative protein design. By combining motif-guided RFdiffusion with high-throughput screening platforms, small protein agonists and antagonists targeting therapeutically relevant GPCRs such as CXCR4 and GLP1R have been successfully designed [188]. The RSO method, integrating the predictive power of AlphaFold2 with differentiable sequence space, successfully transferred functional domains to soluble protein scaffolds [189]. MeMDLM, a masked diffusion language model trained on native membrane protein sequences, can directly generate novel sequences with accurate transmembrane topology [190]. In

therapeutic enzyme design, researchers employed the COMBS algorithm combined with backbone cyclization and D-amino acid incorporation strategies to produce proteins selectively binding anticoagulants and anticancer drugs; D-amino acid configuration variants exhibited extremely high drug clearance efficiency in mouse models while maintaining low immunogenicity [191]. Language models have also emerged as platforms for developing novel gene editing tools: by training on CRISPR-Cas atlases constructed from metagenomic datasets, researchers developed models capable of generating functional novel CRISPR-Cas proteins, with the resulting OpenCRISPR-1 editor maintaining precise genome editing activity in human cells [192].

In therapeutic antibody design, sequence-level and structure-level strategies have developed in parallel. At the sequence level, the PALM-H3 model can de novo generate CDRH3 sequences with target antigen specificity [193], while AbBFN2 supports multiple tasks including antibody sequence generation, annotation, optimization, and humanization under the Bayesian flow network paradigm [194]. Furthermore, MAGE enables the generation of antibodies with binding specificity against multiple antigens including SARS-CoV-2, H5N1 influenza, and RSV-A. At the structure-level, methods such as IgGM can handle antibody structure prediction, inverse sequence design, affinity maturation, and de novo generation within a unified framework [195]. The tFold system integrates antibody structure prediction, complex modeling, and epitope-specific antibody de novo design, successfully designing monoclonal antibodies with nanomolar affinity targeting influenza hemagglutinin, PD-1, PD-L1, and SARS-CoV-2 RBD [196]. Notably, the Baker laboratory systematically demonstrated the applicability of RFdiffusion in de novo antibody design, producing nanobodies and full-length IgGs against influenza virus, Clostridioides difficile toxin TcdB, SARS-CoV-2 RBD, and PHOX2B-HLA-C*07:02 complex, with cryo-EM structural analysis confirming near-atomic-level design consistency [197].

4.4. Generative design of nucleic acid therapeutics

Nucleic acid therapeutics, by virtue of their high programmability, precise targeting, and well-defined mechanisms of action, demonstrate broad potential in cancer treatment and gene regulation. However, nucleic acid function is not solely determined by linear sequence; spatial folding, local structural accessibility, and dynamic interactions with proteins or other nucleic acids collectively constitute a highly coupled functional basis. Traditional design approaches relying primarily on Watson-Crick base pairing rules and empirical screening prove insufficient for systematically capturing the complex mappings among sequence, structure, and function [165]. Generative AI provides a transformative computational paradigm for nucleic acid drug design (Figure 5A).

In the mRNA therapeutics domain, Transformer-based generative models such as GEMORNA [198] and GenerRNA [199] undergo pretraining on massive RNA datasets, learning sequence grammar and structural preferences spanning coding sequences and untranslated regions, transcending traditional codon bias optimization to enable de novo generation of novel coding sequences and regulatory elements. To address the specific challenges of coding sequence (CDS) optimization, tools like LinearDesign [200] leverage computational linguistics algorithms (e.g., lattice parsing) to globalize the search for sequences with optimal secondary structures and minimum free energy, significantly enhancing mRNA stability and half-life. Complementing this, CodonBERT [201] employs a LLMs framework to learn the complex "codon grammar" from massive protein-coding datasets, transcending simple codon bias to capture functional and evolutionary constraints. Models such as Smart5UTR and UTR-LM [202] demonstrate the

advantages of generative approaches in UTR optimization [203,204], marking the transition of mRNA design from empirical fragment assembly toward systematic generative engineering.

For RNA interference (RNAi) and antisense oligonucleotide (ASO) therapeutics, the key breakthrough in generative design is the explicit incorporation of the three-dimensional structural information of target RNA. Generative models incorporate RNA secondary and tertiary structures as design constraints, giving rise to the concept of structure-aware ASOs (3D-ASOs), transcending traditional Watson-Crick pairing to introduce non-canonical interactions such as Hoogsteen base pairs and base triplets, enabling ASOs to recognize and stably bind highly structured RNA targets [205]. By combining MD simulations with generative modeling, these methods reveal the pivotal roles of ASO length, conformational flexibility, and structural accessibility in binding efficiency [206], thereby rendering previously undruggable highly structured RNA targets amenable to rational design.

Having surveyed generative AI across small molecules (Section 4.2), proteins and peptides (Section 4.3), and nucleic acids (Section 4.4), a cross-modality perspective reveals both shared computational foundations and critical divergences (Table 3). Small-molecule generation currently benefits from the most mature data infrastructure and validation pipelines, whereas protein/peptide design is advancing rapidly through diffusion-based architectures, and nucleic acid design remains constrained by limited experimentally validated structure–function data. These modality-specific bottlenecks suggest that a one-size-fits-all generative framework is unlikely at present. Instead, continued progress will require architectures and training strategies tailored to the unique representational and evaluative demands of each therapeutic class.

Table 3. Comparative analysis of generative AI across therapeutic modalities in oncology.

Dimension	Small molecules	Peptides/Proteins	Nucleic acids
Molecular representation	SMILES strings, fingerprints, descriptors, molecular graphs, 3D conformers	Amino acid sequences, MSA, contact/distance maps, 3D atomic coordinates	Nucleotide sequences, secondary structures (dot-bracket notation), 3D tertiary folds
Data availability	Abundant: ChEMBL (~2.4M compounds), PubChem (~119M), ZINC (~37B); extensive bioactivity annotations	Moderate: PDB (~250K structures), UniProt (~250M sequences); limited binding affinity data for designed binders	Abundant: GenBank/ENA/DDBJ (~2.5 × 10^{13} bases; > 3.7 billion sequences), SRA (> 50 PB sequencing reads), RNAcentral (> 45M ncRNAs); limited experimentally validated functional annotations and structural data
Key evaluation metrics	Drug-likeness (QED), synthetic accessibility (SA), binding affinity (IC_{50}/K_i), ADMET, selectivity	Binding affinity (K_d), thermal stability (T_m), folding accuracy (RMSD, pLDDT), expression yield, immunogenicity	Target knockdown efficiency (%KD), hybridization specificity, metabolic stability, immunogenicity, delivery efficiency
Dominant generative architectures	VAEs, GANs, RL-guided Transformers, diffusion models, normalizing flows, *etc*.	Protein language models (e.g. ProGen, ESM), GNNs, structure-conditioned diffusion (e.g. RFdiffusion, Chroma), flow matching, *etc*.	RNA-specific Transformers, sequence-based LLMs, GNNs for secondary structure-aware generation, coarse-grained diffusion, *etc*.
Primary challenges	Synthesizability; off-target toxicity; multi-objective optimization (potency *vs*. ADMET); scaffold novelty	Immunogenicity; conformational flexibility; aggregation propensity; manufacturing scalability; limited experimental throughput	*In vivo* delivery and stability; off-target hybridization; innate immune activation; limited annotated training data
State of experimental validation	Well-established: binding assays, cell-based activity, *in vivo* efficacy; multiple AI-designed compounds in clinical trials	Emerging: display technologies, cryo-EM validation, SPR for affinity; first AI-designed binders entering preclinical	Early stage: cell-based knockdown, *in vivo* PK/PD, delivery system optimization; few AI-designed sequences validated in clinical trials

In aptamer design, generative models complete the computational loop from structure to sequence. RhoDesign takes target three-dimensional structures as input, employing geometric deep learning to generate nucleic acid sequences capable of stable folding and target binding [207]. AiDTA uses reinforcement learning to formalize aptamer design as a fragment assembly problem [165], and DAPTEV achieves continuous exploration of aptamer sequence space through latent space search and

evolutionary algorithms [208]. In the CRISPR-Cas domain, generative AI applications are evolving from local component optimization toward system-level design: the genome-scale foundation model Evo, trained on complete prokaryotic genomes, can de novo generate kilobase-length DNA sequences encoding Cas effector proteins and their cognate non-coding RNA components [209], representing the pinnacle of generative nucleic acid design—models no longer optimize individual molecules but directly generate complete, functionally coherent molecular systems.

5. AI-driven preclinical evaluation of oncology drugs

While virtual screening and generative AI successfully expand the available chemical space and produce novel candidates, the translational success of these designs hinges on rigorous preclinical assessment. The transition from *in silico* hits to clinical candidates requires satisfying complex multiparametric constraints that extend well beyond simple binding affinity. This section examines how AI bridges the gap between molecular design and clinical trials by systematically predicting pharmacokinetics, toxicology, and cross-species translatability.

5.1. AI-assisted prediction of ADMET properties

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties fundamentally govern the clinical viability of oncology drugs. AI methodologies now exploit intricate correlations between molecular structures and physiological endpoints to facilitate high-throughput screening of safety profiles. The technical landscape has evolved from classical machine learning to deep learning architectures. While traditional algorithms (e.g., Random Forest, XGBoost) remain effective for physicochemical parameters, recent GNNs and Transformer architectures have significantly augmented predictive prowess [210–212]. For instance, models like MSformer-ADMET utilize fragment-aware pretraining to surpass unimodal benchmarks, pinpointing toxicity-linked structural motifs to guide lead optimization [213]. Furthermore, multi-modal integration—combining SMILES sequences with 3D conformational graphs—has improved the generalization of predictions for complex endpoints such as hERG inhibition and drug-induced liver injury (DILI). Crucially, uncertainty-aware frameworks, such as posterior networks, are being adopted to distinguish on-target efficacy from off-target liabilities, providing confidence intervals that are essential for decision-making in safety assessment.

5.2. AI-based pharmacodynamic evaluation and biomarker identification

Beyond safety, verifying target engagement and predicting efficacy in heterogeneous tumor microenvironments is paramount. AI frameworks are increasingly applied to integrate multifaceted pharmacodynamic readouts with multi-omics layers. Structure-enabled machine learning models now leverage high-confidence protein structures (e.g., from AlphaFold2) to estimate binding affinities with accuracy approaching experimental assays [214]. In parallel, deep learning models interrogate tumor transcriptomic profiles to infer response patterns associated with oncogenic signaling cascades (e.g., MAPK, PI3K–AKT), providing quantitative benchmarks for prioritizing targeted agents [215]. To address tumor heterogeneity, AI models incorporate patient-derived organoid (PDO) datasets for individualized efficacy projection [216]. Functional precision medicine platforms integrating multiparametric PDO readouts with AI-driven prioritization have successfully

identified regimens with superior *in vivo* tumor suppression [217]. Moreover, AI facilitates the discovery of composite biomarkers from proteogenomic data, linking phosphorylation states to therapeutic outcomes, thus enhancing the stratification of patient cohorts in preclinical studies [218,219].

5.3. AI-enhanced pathology imaging and phenotypic screening

Pathology imaging and high-content screening (HCS) provide the ground truth for validating computational predictions. AI has substantially advanced these modalities by automating feature extraction and quantifying morphological perturbations at scale [220,221]. Deep learning-based segmentation enables accurate delineation of tumor compartments, quantifying metrics such as immune cell infiltration and stromal remodeling [222]. In PDO models, AI-guided morphological classification has effectively stratified organoid phenotypes that correlate with chemotherapy responsiveness [223]. Furthermore, eco-evolutionary AI models integrating histological features with biological data have shown promise in predicting radiotherapy responses. Complementing tissue pathology, AI-enhanced HCS integrates morphological, proliferative, and apoptotic features to generate holistic efficacy profiles [224]. Advanced platforms like HCS-3DX combine automated microscopy with single-cell analytics to dissect complex phenotypes in 3D tumor models, bridging the gap between cellular assays and tissue-level responses [225].

5.4. AI modeling for cross-species data integration

A major bottleneck in translation is the biological divergence between animal models and humans. Recent advances in species-agnostic representation learning provide a methodological basis for aligning cross-species data. Orthology-independent transfer learning models construct unified latent spaces to integrate single-cell data from different species, facilitating the identification of conserved functional gene programs without explicit gene matching [226]. Deep learning architectures, such as SATURN, align cellular states across species by combining protein sequence embeddings with transcriptomic profiles [227]. In parallel, physiologically based pharmacokinetic (PBPK) modeling offers a quantitative framework for extrapolating toxicokinetic parameters from small mammals to humans [228]. These approaches emphasize representation harmonization, helping researchers distinguish model-specific artifacts from translatable therapeutic mechanisms.

6. Technical challenges and optimization strategies

While AI paradigms in preclinical evaluation have improved the screening of safety and efficacy, as discussed in the previous section, the translation from theoretical models to practical applications remains constrained by multiple systemic obstacles. These include heterogeneity in data quality, the "black box" nature of model interpretability, limited algorithmic generalization, and the inherent complexity of interdisciplinary collaboration.

6.1. Data quality, standardization, and scarcity

Data constitutes the cornerstone of AI models, with its quality and quantity determining the theoretical upper bound of predictive performance. The current drug discovery field confronts three

principal data-related challenges (Figure 6A). First, data quality and standardization issues are pervasive: although public databases such as ChEMBL, PubChem, and DrugBank provide vast repositories of bioactivity data, these datasets typically aggregate results from heterogeneous experimental conditions, assay methods, and species, introducing substantial noise and inconsistency [229,230], potentially causing models to capture spurious correlations. Second, data bias and imbalance severely distort predictions: in toxicity prediction and activity screening, negative data frequently go unreported or unpublished, causing datasets to skew heavily toward positive samples, leading models to produce high false positive rates in practical applications [5]. Third, data scarcity limits progress in niche domains: for rare cancer subtypes or novel targets (such as orphan GPCRs), high-quality labeled data available for training proves extremely scarce, severely constraining supervised learning model performance [231,232].

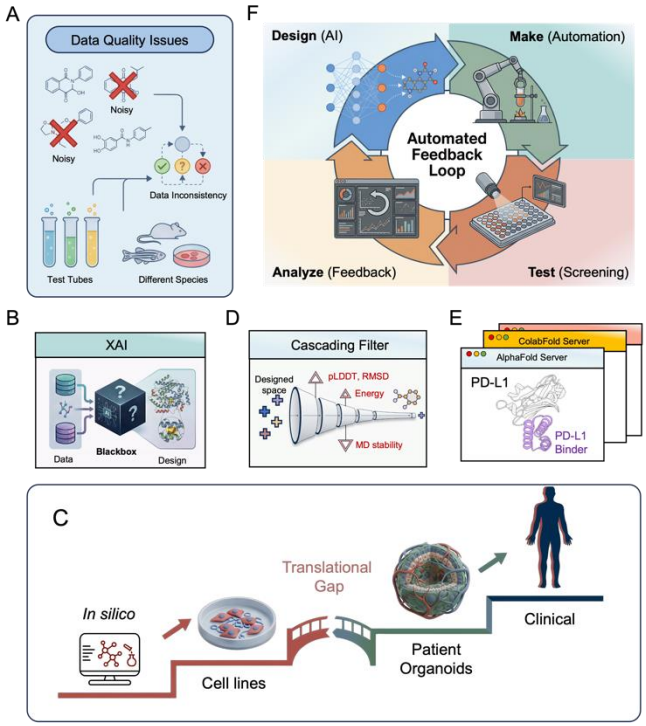

Figure 6. Technical challenges and strategic solutions in AI-driven drug discovery. **(A)** Data quality bottlenecks arising from noisy datasets, experimental inconsistencies, and heterogeneity across different species and assay conditions; **(B)** The transition from "Black Box" models to explainable AI (XAI), aiming to provide interpretability for molecular design decisions; **(C)** Illustration of the "Translational Gap" between in silico predictions and clinical reality; bridging this divide requires moving beyond simple cell lines to more physiologically relevant validation models like patient organoids; **(D)** A cascading filtration strategy that integrates physics-based metrics—including pLDDT confidence, RMSD, energy potentials, and MD stability—to refine the designed chemical space; **(E)** Democratization of advanced algorithms through accessible, user-friendly interfaces (e.g., AlphaFold servers) that facilitate protein binder design; **(F)** The automated "Design–Make–Test–Analyze" (DMTA) closed-loop system, where AI designs are coupled with robotic automation to accelerate feedback and optimization.

A range of complementary strategies are being developed to address each of these challenges. For data quality and standardization, the field is increasingly adopting rigorous data-curation

protocols and constructing standardized ontologies to ensure interoperability and reusability [230]. For negative data scarcity, which constitutes a particularly critical bottleneck, several mitigation approaches have emerged: synthetic negative data generation through decoy set construction (e.g., the DUD-E methodology employing property-matched decoys) provides a practical means to populate the inactive class; positive-unlabeled (PU) learning methods are specifically designed for datasets with confirmed positives but ambiguous negatives; and data augmentation strategies that generate hard negatives—structurally similar to actives but functionally inactive—further enhance model discrimination capability. For data scarcity in niche domains, transfer learning and few-shot learning have proven effective: by pretraining models on large-scale general datasets to learn universal chemical grammar and features, then fine-tuning on task-specific small datasets, performance in low-data regimes can be substantially improved [150,233]. To mitigate data silos caused by patient privacy regulations, federated learning (FL) [234] offers a complementary solution to data silos caused by patient privacy regulations, enabling collaborative model training across multiple institutions without sharing raw clinical data—particularly valuable for rare cancer subtypes where single-institution data is insufficient. Finally, active learning strategies allow algorithms to autonomously select the most informative samples for experimental validation, with this predict-validate-retrain closed-loop system maximizing model performance improvement while minimizing experimental costs [235], and can be applied across all three challenge areas, for instance by prioritizing informative negative sampling from unexplored chemical space to iteratively improve model discrimination.

6.2. Interpretability and reliability

Deep learning models, particularly deep neural networks, are commonly characterized as “black-box” systems: while they may achieve high predictive accuracy, they often fail to elucidate the underlying biological or chemical logic behind their decisions [236]. In healthcare and drug discovery, this lack of interpretability represents a critical deficiency: chemists and clinicians require not only the prediction (the “what”) but also the rationale (the “why”), for instance, which specific functional groups or structural motifs prompted a model to flag a molecule as toxic. Furthermore, when encountering data outside the training distribution, model performance often degrades precipitously, and the lack of capability to quantify prediction uncertainty renders this risk unacceptable in clinical safety contexts [237].

Explainable artificial intelligence represents a pivotal research direction (Figure 6B). Post-hoc interpretability methods (such as SHAP [238] and LIME [239]) are widely used to quantify the contribution of input features to prediction outputs, helping researchers discern the model’s decision-making logic. Deeper strategies involve developing intrinsically interpretable models, for example, Transformer architectures incorporating attention mechanisms can intuitively visualize key regions of focus when processing molecular sequences or protein structures, thereby fostering trust and informing molecular optimization. Integrating uncertainty quantification techniques allows models to estimate their own predictive confidence, which is critical for screening high-risk candidates. Sparse auto-encoder (SAE) methods enable interpretable analysis of high-dimensional embedding spaces within foundation models, providing deeper insights into their latent representations [240].

6.3. Bridging the computation-to-application gap: enhancing prediction reliability

A substantial "valley of death" exists between AI predictions and wet-lab validation (Figure 6C). On one hand, computational models typically operate as black boxes, optimizing synthetic numerical scores rather than genuine biological activity. This discrepancy arises because training datasets are frequently aggregated from heterogeneous sources with varying experimental biases, causing models to learn statistical artifacts rather than fundamental chemical rules. Consequently, a model may achieve high accuracy on benchmark datasets yet fail to generate thermodynamically stable or synthetically feasible ligands in real-world applications [241]. To ameliorate this disconnect and ground AI predictions in physical reality, a paradigm shift toward physics-informed AI is imperative. A major strategy involves hybridizing deep learning architectures with physics-based energy functions. For instance, TorsionNet employs a deep neural network to rapidly predict small-molecule torsional energy profiles with quantum mechanical accuracy, enabling its integration as a conformational constraint during molecular generation to ensure that candidate molecules adopt physically plausible geometries [242]. In the domain of binding affinity prediction, physics-informed scoring functions such as PLANET incorporate molecular mechanics force field terms—including bond, angle, dihedral, Lennard-Jones, and Coulomb interactions—into a multi-objective graph neural network training framework, effectively regularizing learned representations with established physical laws [243]. Beyond scoring, target-aware diffusion-based generative models are emerging as a complementary strategy; for example, dual diffusion frameworks that jointly model ligand geometry and atomic interactions within binding pockets incorporate structural and energetic constraints during the denoising process, biasing generation toward geometrically and energetically favorable poses [244]. Collectively, these approaches enable AI agents to explore chemical space that is not only diverse but also grounded in physical reality (Figure 6D).

Furthermore, static structure prediction often falls short of capturing the dynamic nature of protein-ligand interactions. As such, integrating MD simulations into AI workflows is essential. By training models on MD trajectories rather than static crystal structures, AI can learn to identify cryptic binding pockets and account for protein flexibility. While the high computational cost of generating long-timescale MD data remains a bottleneck, future optimizations may rely on AI-accelerated neural network potentials to achieve *ab initio* accuracy at a fraction of the cost, enabling the rigorous assessment of ligand residence time and binding stability prior to synthesis. On the experimental side, the validation paradigm is shifting from simple 2D cell lines to patient-derived organoids, which better recapitulate tumor heterogeneity. Integrating AI-driven high-content image analysis with automated organoid screening creates a high-fidelity feedback loop, ensuring that computationally designed molecules are validated in physiologically relevant systems before animal testing.

Looking further ahead, quantum computing (QC) presents a potentially transformative frontier in drug discovery [241]. Classical approximation methods possess inherent limitations when handling the precise electronic structures of transition metal cofactors or complex reaction mechanisms, while quantum algorithms theoretically offer the promise of simulating these systems with unprecedented accuracy.

6.4. Accelerating translation: automation, accessibility, and novel modalities

Even with physically rigorous models, bridging the translational gap requires deep interdisciplinary integration and the reengineering of drug discovery workflows. A major obstacle remains the

technical barrier separating computational experts from experimentalists. While platforms such as AlphaFold server [245] and ColabFold server [246] have democratized access to protein structure prediction (Figure 6E), a notable lack of similarly user-friendly platforms exists for small molecule design and nucleic acid engineering. Developing intuitive, GUI-based AI tools, analogous to those successfully deployed in image analysis and genomics, proves essential for empowering biologists and chemists to integrate advanced algorithms into their daily research without requiring extensive coding expertise.

Beyond accessibility, establishing automated closed-loop "Design-Make-Test-Analyze" systems is crucial for validating AI designs at scale (Figure 6F). By directly coupling generative AI models with automated synthesis platforms and high-throughput screening robotics, researchers can compress iterative cycles from months to days. Such automation serves a dual purpose: it verifies the synthetic feasibility of AI designs while generating standardized, high-quality wet-lab data to refine models through bias correction. However, general-purpose automation platforms that allow users to autonomously design and manage tasks remain scarce; while initiatives such as SaprotHub [247] demonstrate the potential of cloud-based laboratory automation, their current applications are primarily confined to enzyme engineering in synthetic biology. This limitation highlights the urgent need for broader platforms encompassing small molecule synthesis and pharmacological screening.

Ultimately, the true test of AI utility lies in its capacity to conquer historically recalcitrant targets. For undruggable proteins with intrinsically disordered regions, such as MYC or KRAS, traditional occupancy-driven inhibition often proves ineffective. However, AI is now demonstrating its potential in designing novel therapeutic modalities, such as PROTACs and molecular glues, which leverage induced proximity to degrade pathogenic proteins. Concurrently, by analyzing large-scale genetic screening data to apply synthetic lethality principles, AI is identifying tumor-specific dependencies, providing new entry points for precision oncology even in the absence of conventional targets.

Taken together, the challenges and strategies discussed across Sections 6.1–6.4 reveal a field in which different AI technologies coexist at vastly different stages of maturity. Table 4 provides a consolidated mapping of representative technologies to their estimated technology readiness levels (TRL), from early-stage foundation models that remain confined to academic benchmarks, to AI-designed compounds that have entered phase I/II clinical trials. This heterogeneous landscape underscores that while select applications have achieved translational impact, the majority of AI-driven methodologies still face considerable distance from clinical deployment—a gap that the automation, interpretability, and data strategies outlined above are specifically designed to close.

Table 4. Technology readiness level mapping of AI technologies in oncology drug discovery.

TRL Level	Description	Representative AI Technologies	Current Status
Early Research (TRL 1–3)	Concept demonstration and proof-of-concept	Single-cell foundation models for target discovery; multi-scale biological simulation frameworks; quantum ML for electronic structure; AI-driven neoantigen vaccine design with TCR modeling	Demonstrated in academic benchmarks; no prospective pharmaceutical validation
Academic Validation (TRL 4–5)	Validated in controlled academic settings	Diffusion models for de novo protein design; GNN-based perturbation prediction; physics-informed scoring functions; generative models for nucleic acid therapeutics	Emerging prospective wet-lab validation; growing reproducibility
Industry Adoption (TRL 6–7)	Integrated into pharmaceutical pipeline workflows	GNNs/Transformer-based virtual screening and property prediction; AlphaFold for SBDD; ADMET prediction models; RL-guided lead optimization; molecular generative models	Actively used in pharma R&D; AI-nominated compounds in preclinical stages
Clinical-Trial Linked (TRL 8–9)	Directly supporting clinical compounds	AI-designed kinase inhibitors;AI-guided drug repurposing; AI-optimized antibody therapeutics; contrastive learning for genome-wide screening	Multiple AI-originated compounds in phase I/II clinical trials; first approvals anticipated

7. Conclusions

AI has transcended its role as a mere accelerator to become the foundational architect of a new paradigm in oncology drug discovery. As detailed in this review, the integration of AI is driving a decisive shift from a historical reliance on serendipity and empirical screening toward a future of data-driven, rational engineering.

The transformative power of AI is most evident in its dual capacity to decode biological complexity and expand chemical possibilities. By untangling multidimensional omics data, AI is revealing high-value targets—such as synthetic lethal partners and cryptic immune checkpoints—that were previously invisible to reductionist approaches. Simultaneously, the rise of generative AI has fundamentally altered therapeutic design: we are no longer limited to screening finite libraries but are now empowered to design de novo molecules, proteins, and nucleic acids with precise functional profiles, thereby challenging the very concept of undruggable targets.

However, the path to clinical translation remains obstructed by the "reality gap" between computational predictions and biological outcomes. The future success of AI in oncology will depend not solely on algorithmic sophistication, but on the robustness of the underlying data infrastructure and the seamless integration of verification loops. Addressing the "black box" nature of deep learning through XAI, while grounding predictions in physical reality through physics-informed machine learning, represents a critical next step. Moreover, current AI models typically operate at a single biological scale; developing foundation models capable of bridging molecular perturbations, cellular signaling, and tissue-level drug responses remains a fundamental open problem whose resolution would enable genuinely systems-level drug response modeling.

Looking forward, the ultimate trajectory of this field lies in the convergence of the digital and physical worlds. The establishment of automated DMTA loops—where AI agents autonomously steer wet-lab experimentation—promises to overcome the data scarcity bottleneck and iteratively refine models for greater predictive accuracy. The next frontier extends this vision toward fully closed-loop autonomous discovery systems: self-driving laboratories in which generative models design candidates, robotic platforms execute synthesis and high-throughput assays, and AI interprets results to iteratively refine hypotheses—all without human intervention. While individual components of this pipeline exist, their seamless end-to-end integration remains an unsolved engineering and algorithmic challenge. As these technologies mature, AI will not only shorten development timelines but will fundamentally redefine the boundaries of precision medicine, delivering novel, effective therapies to cancer patients with unprecedented speed and precision.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT for language polishing and readability improvement across the manuscript, and used Nano Banana Pro (Gemini 3 Pro Image) to assist with minor non-data illustrative elements in parts of Figures 2 and 6. After using these tools, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of the manuscript and the figures.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (2022YFC3400501); the National Natural Science Foundation of China (82425104).

Authors' contribution

Conceptualization, Honglin Li; investigation, Jianxin Tang, Jinhang Xu, Wenqing Zhang, and Daohong Gong; writing—original draft preparation, Jianxin Tang, Jinhang Xu, Wenqing Zhang, and Daohong Gong; writing—review and editing, Honglin Li and Xiaolong Cheng; visualization, Jianxin Tang, Jinhang Xu, Wenqing Zhang, Daohong Gong, and Qixing Huang; supervision, Honglin Li; funding acquisition, Honglin Li. Jianxin Tang, Jinhang Xu, Wenqing Zhang, and Daohong Gong contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Prof. Honglin Li holds the position of Editorial Board Members for *Advanced Cancer Research* and has not peer reviewed or made any editorial decisions for this paper.

References

[1] Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics* 2019, 20:273–286.
[2] Thomas DW. Clinical development success rates 2006–2015. *BIO Ind. Anal.* 2016, 1:16.
[3] 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.
[4] Singh S, Kaur N, Gehlot A. Application of artificial intelligence in drug design: a review. *Comput. Biol. Med.* 2024, 179:108810.
[5] Tran TTV, Wibowo SA, Tayara H, Chong KT. Artificial intelligence in drug toxicity prediction: recent advances, challenges, and future perspectives. *J. Chem. Inf. Model.* 2023, 63(9):2628–2643.
[6] Sun Q, Wang H, Xie J, Wang L, Mu J, *et al.* Computer-aided drug discovery for undruggable targets. *Chem. Rev.* 2025, 125(13):6309–6365.
[7] Tan P, Chen X, Zhang H, Wei Q, Luo K. Artificial intelligence aids in development of nanomedicines for cancer management. *Semin. Cancer Biol.* 2023, 89:61–75.
[8] Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* 2019, 119(18):10520–10594.
[9] Lorente JS, Sokolov AV, Ferguson G, Schiöth HB, Hauser AS, *et al.* GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery* 2025, 24(6):458–479.
[10] Whitfield JR, Soucek L. MYC in cancer: from undruggable target to clinical trials. *Nat. Rev. Drug Discovery* 2025, 24(6):445–457.
[11] Zhang O, Lin H, Zhang X, Wang X, Wu Z, *et al.* Graph neural networks in modern AI-Aided drug discovery. *Chem. Rev.* 2025, 125(20):10001–10103.
[12] Perez-Lopez R, Laleh GN, Mahmood F, Kather JN. A guide to artificial intelligence for cancer researchers. *Nat. Rev. Cancer* 2024, 24(6):427–441.

[13] Gonçalves E, Ryan CJ, Adams DJ. Synthetic lethality in cancer drug discovery: challenges and opportunities. *Nat. Rev. Drug Discovery* 2026, 25(1):22–38.

[14] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 2019, 25(1):44–56.

[15] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* 2019, 18(6):463–477.

[16] Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, *et al.* Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* 2021, 124(4):686–696.

[17] Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. *Briefings Bioinf.* 2021, 22(2):1515–1530.

[18] Perdomo-Quinteiro P, Belmonte-Hernández A. Knowledge Graphs for drug repurposing: a review of databases and methods. *Briefings Bioinf.* 2024, 25(6):bbae461.

[19] Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discovery* 2022, 12(1):31–46.

[20] Nijman SM. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS Lett.* 2011, 585(1):1–6.

[21] Huang A, Garraway LA, Ashworth A, Weber B. Synthetic lethality as an engine for cancer drug target discovery. *Nat. Rev. Drug Discovery* 2020, 19(1):23–38.

[22] Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 2005, 434(7035):917–921.

[23] Lord CJ, Ashworth A. PARP inhibitors: synthetic lethality in the clinic. *Science* 2017, 355(6330):1152–1158.

[24] Murai J, Huang SY, Das BB, Renaud A, Zhang Y, *et al.* Trapping of PARP1 and PARP2 by clinical PARP inhibitors. *Cancer Res.* 2012, 72(21):5588–5599.

[25] Kim G, Ison G, McKee AE, Zhang H, Tang S, *et al.* FDA approval summary: olaparib monotherapy in patients with deleterious germline BRCA-mutated advanced ovarian cancer treated with three or more lines of chemotherapy. *Clin. Cancer Res.* 2015, 21(19):4257–4261.

[26] Kryukov GV, Wilson FH, Ruth JR, Paulk J, Tsherniak A, *et al.* MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells. *Science* 2016, 351(6278):1214–1218.

[27] Mavrakis KJ, McDonald III ER, Schlabach MR, Billy E, Hoffman GR, *et al.* Disordered methionine metabolism in MTAP/CDKN2A-deleted cancers leads to dependence on PRMT5. *Science* 2016, 351(6278):1208–1213.

[28] Kalev P, Hyer ML, Gross S, Konteatis Z, Chen CC, *et al.* MAT2A inhibition blocks the growth of MTAP-deleted cancer cells by reducing PRMT5-dependent mRNA splicing and inducing DNA damage. *Cancer Cell* 2021, 39(2):209–224.

[29] Liany H, Jeyasekharan A, Rajan V. Predicting synthetic lethal interactions using heterogeneous data sources. *Bioinformatics* 2020, 36(7):2209–2216.

[30] Cai R, Chen X, Fang Y, Wu M, Hao Y. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics* 2020, 36(16):4458–4465.

[31] Long Y, Wu M, Liu Y, Zheng J, Kwoh CK, *et al.* Graph contextualized attention network for predicting synthetic lethality in human cancers. *Bioinformatics* 2021, 37(16):2432–2440.

[32] Wan F, Li S, Tian T, Lei Y, Zhao D, *et al.* Exp2sl: a machine learning framework for cell-line-specific synthetic lethality prediction. *Front. Pharmacol.* 2020, 11:112.

[33] Li F, Hu Q, Zhang X, Sun R, Liu Z, *et al.* DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nat. Commun.* 2022, 13(1):7133.

[34] Li B, Ran T, Chen H. 3D based generative PROTAC linker design with reinforcement learning. *Briefings Bioinf.* 2023, 24(5):bbad323.

[35] Guo J, Knuth F, Margreitter C, Janet JP, Papadopoulos K, *et al.* Link-INVENT: generative linker design with reinforcement learning. *Digital Discovery* 2023, 2(2):392–408.

[36] Ge J, Li S, Weng G, Wang H, Fang M, *et al.* PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Res.* 2025, 53(D1):D1510–D1515.

[37] Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. *Science* 2018, 359(6382):1350–1355.

[38] Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, *et al.* Five-year survival and correlates among patients with advanced melanoma, renal cell carcinoma, or non–small cell lung cancer treated with nivolumab. *JAMA Oncol.* 2019, 5(10):1411–1420.

[39] Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell* 2017, 168(4):707–723.

[40] Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* 2018, 24(5):541–550.

[41] Noessner E. DGK-α: a checkpoint in cancer-mediated immuno-inhibition and target for immunotherapy. *Front. Cell Dev. Biol.* 2017, 5:16.

[42] Prinz PU, Mendler AN, Masouris I, Durner L, Oberneder R, *et al.* High DGK-α and disabled MAPK pathways cause dysfunction of human tumor-infiltrating $CD8^+$ T cells that is reversible by pharmacologic intervention. *J. Immunol.* 2012, 188(2):5990–6000.

[43] Wichroski M, Benci J, Liu S, Chupak L, Fang J, *et al.* DGKα/ζ inhibitors combine with PD-1 checkpoint therapy to promote T cell–mediated antitumor immunity. *Sci. Transl. Med.* 2023, 15(719):eadh1892.

[44] Grünenfelder DC, Velaparthi U, Warrier JS, Chupak L, Darne CP, *et al.* Design, Synthesis, and T Cell Checkpoint Combination Potential of First-In-Class DGKα/ζ Inhibitor BMS-986408. *J. Med. Chem.* 2025, 68(20):21840–21859.

[45] Majeti R, Chao MP, Alizadeh AA, Pang WW, Jaiswal S, *et al.* CD47 is an adverse prognostic factor and therapeutic antibody target on human acute myeloid leukemia stem cells. *Cell* 2009, 138(2):286–299.

[46] Logtenberg ME, Jansen JM, Raaben M, Toebes M, Franke K, *et al.* Glutaminyl cyclase is an enzymatic modifier of the CD47-SIRPα axis and a target for cancer immunotherapy. *Nat. Med.* 2019, 25(4):612–619.

[47] Schumacher TN, Kesmir C, Buuren MM. Biomarkers in cancer immunotherapy. *Cancer Cell* 2015, 27(1):12–14.

[48] Zheng L, Qin S, Si W, Wang A, Xing B, *et al.* Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 2021, 374(6574):abe6474.

[49] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021, 596(7873):583–589.

[50] Wu KE, Yost K, Daniel B, Belk J, Xia Y, *et al.* TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. In *Proceeding of Machine Learning in Computational Biology*, Seattle, USA, September 5–6, 2024, pp. 194–229.

[51] Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* 2021, 22(2):71–88.

[52] Saxena M, Burg SH, Melief CJ, Bhardwaj N. Therapeutic cancer vaccines. *Nat. Rev. Cancer* 2021, 21:360–378.

[53] Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* 2015, 348(6230):69–74.

[54] Yarchoan M, Johnson III BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* 2017, 17(4):209–222.

[55] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* 2015, 348(6230):124–128.

[56] Wells DK, Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KC, *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 2020, 183(3):818–834.

[57] Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 2018, 46(D1):D419–D427.

[58] Vita R, Blazeska N, Marrama D, Duesing S, Bennett J, *et al.* The immune epitope database (IEDB): 2024 update. *Nucleic Acids Res.* 2025, 53(D1):D436–D443.

[59] Zhang J, Ma W, Yao H. Accurate TCR-pMHC interaction prediction using a BERT-based transfer learning method. *Briefings Bioinf.* 2024, 25(1):bbad436.

[60] Jensen MF, Nielsen M. NetTCR 2.2-Improved TCR specificity predictions by combining pan-and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *bioRxiv* 2023.

[61] Sterner R, Sterner R. CAR-T cell therapy: current limitations and potential strategies. *Blood Cancer J.* 2021, 11(4):69.

[62] MacKay M, Afshinnekoo E, Rub J, Hassan C, Khunte M, *et al.* The therapeutic landscape for cells engineered with chimeric antigen receptors. *Nat. Biotechnol.* 2020, 38(2):233–244.

[63] Dannenfelser R, Allen GM, VanderSluis B, Koegel AK, Levinson S, *et al.* Discriminatory power of combinatorial antigen recognition in cancer T cell therapies. *Cell Syst.* 2020, 11(3):215–228.

[64] Albert BA, Yang Y, Shao XM, Singh D, Smith KN, *et al.* Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nat. Mach. Intell.* 2023, 5(8):861–872.

[65] Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 2020, 48(W1):W449–W454.

[66] Weber A, Born J, Rodriguez Martínez M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* 2021, 37(Supplement_1):i237–i244.

[67] Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* 2021, 3(10):864–875.

[68] Chung Y, Ha JH, Im KC, Lee JS. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 17–21, 2024, pp. 11591–11600.

[69] Zhang D, Schroeder A, Yan H, Yang H, Hu J, *et al.* Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat. Biotechnol.* 2024, 42(9):1372–1377.

[70] Zou J, Xiao K, Chen Z, Pei J, Xu J, *et al.* Predicting Spatial Transcriptomics from H&E Image by Pretrained Contrastive Alignment Learning. *bioRxiv* 2025.

[71] Wang C, Chan AS, Fu X, Ghazanfar S, Kim J, *et al.* Benchmarking the translational potential of spatial gene expression prediction from histology. *Nat. Commun.* 2025, 16(1):1544.

[72] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* 2018, 24(10):1559–1567.

[73] Kather JN, Pearson AT, Halama N, Jäger D, Krause J, *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 2019, 25(7):1054–1056.

[74] Xu H, Usuyama N, Bagga J, Zhang S, Rao R, *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 2024, 630(8015):181–188.

[75] Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* 2024, 30(3):850–862.

[76] Lu MY, Chen B, Williamson DF, Chen RJ, Liang I, *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* 2024, 30(3):863–874.

[77] Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, *et al.* How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024, 187(25):7045–7063.

[78] Adduri AK, Gautam D, Bevilacqua B, Imran A, Shah R, *et al.* Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv* 2025.

[79] Dong M, Adduri A, Gautam D, Carpenter C, Shah R, *et al.* Stack: in-context learning of single-cell biology. *bioRxiv* 2026.

[80] Cui H, Wang C, Maan H, Pang K, Luo F, *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* 2024, 21(8):1470–1480.

[81] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, *et al.* Transfer learning enables predictions in network biology. *Nature* 2023, 618(7965):616–624.

[82] Hao M, Gong J, Zeng X, Liu C, Guo Y, *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* 2024, 21(8):1481–1491.

[83] Rosen Y, Roohani Y, Agarwal A, Samotorčan L, Consortium TS, *et al.* Universal cell embeddings: a foundation model for cell biology. *bioRxiv* 2023.

[84] Zeng Y, Xie J, Shangguan N, Wei Z, Li W, *et al.* CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat. Commun.* 2025, 16(1):4679.

[85] Kedzierska KZ, Crawford L, Amini AP, Lu AX. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* 2025, 26(1):101.

[86] Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat. Methods* 2019, 16(8):715–721.

[87] Li L, You Y, Fu Y, Liao W, Fan X, *et al.* A systematic comparison of single-cell perturbation response prediction models. *bioRxiv* 2024.

[88] Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 2023, 614(7949):742–751.

[89] Osorio D, Zhong Y, Li G, Xu Q, Yang Y, *et al.* scTenifoldKnk: an efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. *Patterns* 2022, 3(3):100434.

[90] Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 2017, 6:e26726.

[91] Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 2020, 36(2):603–610.

[92] Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, *et al.* A foundation model for clinician-centered drug repurposing. *Nat. Med.* 2024, 30(12):3601–3613.

[93] Alsentzer E, Li MM, Kobren SN, Noori A, Network UD, *et al.* Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Med.* 2025, 8(1):380.

[94] Jia Y, Gao B, Tan J, Zheng J, Hong X, *et al.* Deep contrastive learning enables genome-wide virtual screening. *Science* 2026, 391(6781):eads9530.

[95] Li MM, Huang Y, Sumathipala M, Liang MQ, Valdeolivas A, *et al.* Contextual AI models for single-cell protein biology. *Nat. Methods* 2024, 21(8):1546–1557.

[96] Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, *et al.* Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv* 2023, arXiv:2305.17100v2.

[97] Thaingtamtanha T, Ravichandran R, Gentile F. On the application of artificial intelligence in virtual screening. *Expert Opin. Drug Discovery.* 2025, 20(7):845–857.

[98] Caba K, Tran-Nguyen VK, Rahman T, Ballester PJ. Comprehensive machine learning boosts structure-based virtual screening for PARP1 inhibitors. *J. Cheminf.* 2024, 16(1):40.

[99] Chen W, Zhuang X, Chen Y, Shen L, Yang H, *et al.* Discovery of potent and selective CDK2 inhibitors with high safety and favorable bioavailability for the treatment of cancer. *Eur. J. Med. Chem.* 2025, 290:117503.

[100] Lyu J, Irwin JJ, Shoichet BK. Modeling the expansion of virtual screening libraries. *Nat. Chem. Biol.* 2023, 19(6):712–718.

[101] Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, *et al.* A Critical assessment of docking programs and scoring functions. *J. Med. Chem.* 2006, 49(20):5912–5931.

[102] Meng X, Zhang H, Mezei M, Cui M. Molecular Docking: a powerful approach for structure-based drug discovery. *Curr. Comput.-Aided Drug Des.* 2011, 7(2):146–157.

[103] Agu PC, Afiukwa CA, Orji OU, Ezeh EM, Ofoke IH, *et al.* Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Sci. Rep.* 2023, 13(1):13398.

[104] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, *et al.* QSAR Modeling: where have you been? Where are you going to? *J. Med. Chem.* 2014, 57(12):4977–5010.

[105] Wang E, Sun H, Wang J, Wang Z, Liu H, *et al.* End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem Rev.* 2019, 119(16):9478–9508.

[106] Moon S, Hwang SY, Lim J, Kim YW. PIGNet2: a versatile deep learning-based protein–ligand interaction prediction model for binding affinity scoring and virtual screening. *Digital Discovery* 2024, 3(2):287–299.

[107] Dey R, Brocidiacono M, Koirala K, Tropsha A, Popov KI. Extending machine learning model for implicit solvation to free energy calculations. *arXiv* 2025, arXiv:2510.20103.

[108] Siebenmorgen T, Menezes F, Benassou S, Merdivan E, Didi K, *et al.* MISATO: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nat. Comput. Sci.* 2024, 4(5):367–378.

[109] Korlepara DB, Srivastava R, Pal PK, Raza SH, Kumar V, *et al.* PLAS-20k: Extended dataset of protein-ligand affinities from MD simulations for machine learning applications. *Sci. Data* 2024, 11(1):180.

[110] Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature* 2023, 616(7958):673–685.

[111] Li F, Ackloo S, Arrowsmith CH, Ban F, Barden CJ, *et al.* CACHE Challenge #1: Targeting the WDR Domain of LRRK2, A Parkinson's Disease Associated Protein. *J. Chem. Inf. Model.* 2024, 64(22):8521–8536.

[112] Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* 2018, 58(2):287–296.

[113] McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, *et al.* GNINA 1.0: molecular docking with deep learning. *J. Cheminf.* 2021, 13(1):43.

[114] Cao D, Chen G, Jiang J, Yu J, Zhang R, *et al.* Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat. Mach. Intell.* 2024, 6(6):688–700.

[115] Stärk H, Ganea O, Pattanaik L, Barzilay DR, Jaakkola T. EquiBind: geometric deep learning for drug binding structure prediction. In *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, July 17–23, 2022, pp. 20503–20521.

[116] Townshend RJ, Eismann S, Watkins AM, Rangan R, Karelina M, *et al.* Geometric deep learning of RNA structure. *Science* 2021, 373(6558):1047–1051.

[117] Yu J, Li Z, Chen G, Kong X, Hu J, *et al.* Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nat. Comput. Sci.* 2023, 3(10):860–872.

[118] Cao D, Chen M, Zhang R, Wang Z, Huang M, *et al.* SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nat. Methods* 2025, 22(2):310–322.

[119] Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: diffusion steps, twists, and turns for molecular docking. *arXiv* 2023, arXiv:2210.01776.

[120] Vinogradov V, Nguyen KT, Steshin S, Izmailov I, Doronichev A. BIOPTIC B1 ultra-high-throughput virtual screening system discovers LRRK2 ligands in vast chemical space. *J. Chem. Inf. Model.* 2025, 65(19):9927–9936.

[121] Rocha MN, Sousa DS, Mendes FR, Santos HS, Marinho GS, *et al.* Ligand and structure-based virtual screening approaches in drug discovery: minireview. *Mol. Diversity* 2025, 29(3):2799–2809.
[122] Heid E, Greenman KP, Chung Y, Li S, Graff DE, *et al.* Chemprop: a machine learning package for chemical property prediction. *J. Chem. Inf. Model.* 2024, 64(1):9–17.
[123] Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, *et al.* Discovery of a structural class of antibiotics with explainable deep learning. *Nature* 2024, 626(7997):177–185.
[124] Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, *et al.* Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* 2023, 19(11):1342–1350.
[125] Mueller R, Dawson ES, Meiler J, Rodriguez AL, Chauder BA, *et al.* Discovery of 2-(2-benzoxazoyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu5): from an artificial neural network virtual screen to an *in vivo* tool compound. *ChemMedChem.* 2012, 7(3):406–414.
[126] Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discovery.* 2021, 16(9):949–959.
[127] Mao Y, Ghosh S, Pal R. AdapTor: adaptive topological regression for quantitative structure-activity relationship modeling. *J. Cheminf.* 2025, 17(1):128.
[128] Badkul A, Xie L, Zhang S, Xie L. Multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology. *Nat. Mach. Intell.* 2025, pp. 1–11.
[129] Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* 2016, 59(9):4035–4061.
[130] Crivelli-Decker JE, Beckwith Z, Tom G, Le L, Khuttan S, *et al.* Machine learning guided AQFEP: a fast and efficient absolute free energy perturbation solution for virtual screening. *J. Chem. Theory Comput.* 2024, 20(16):7188–7198.
[131] Gusev F, Gutkin E, Kurnikova MG, Isayev O. Active learning guided drug design lead optimization based on relative binding free energy modeling. *J. Chem. Inf. Model.* 2023, 63(2):583–594.
[132] Min Y, Wei Y, Wang P, Wang X, Li H, *et al.* From static to dynamic structures: improving binding affinity prediction with graph-based deep learning. *Adv. Sci.* 2024, 11(40):2405404.
[133] Wu F, Jin S, Jiang Y, Jin X, Tang B, *et al.* Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. *Adv. Sci.* 2022, 9(33):2203796.
[134] Doerr S, Majewski M, Pérez A, Krämer A, Clementi C, *et al.* TorchMD: a deep learning framework for molecular simulations. *J. Chem. Theory Comput.* 2021, 17(4):2355–2363.
[135] Zariquiey SF, Galvelis R, Gallicchio E, Chodera JD, Markland TE, *et al.* Enhancing protein-ligand binding affinity predictions using neural network potentials. *J. Chem. Inf. Model.* 2024, 64(5):1481–1485.
[136] Ormeño F, General IJ. Convergence and equilibrium in molecular dynamics simulations. *Commun. Chem.* 2024, 7(1):26.
[137] Wei H, McCammon JA. Structure and dynamics in drug discovery. *npj Drug Discovery* 2024, 1(1):1.
[138] Zhao X, Ahn D, Nam G, Kwon J, Song S, *et al.* Identification of crocetin as a dual agonist of GPR40 and GPR120 responsible for the antidiabetic effect of saffron. *Nutrients* 2023, 15(22):4774.

[139] Gori DN, Barrionuevo EM, Alberca LN, Sbaraglini ML, Llanos MA, *et al.* Discovery of trypanosoma cruzi carbonic anhydrase inhibitors by a combination of ligand- and structure-based virtual screening. *J. Chem. Inf. Model.* 2025, 65(10):4980–4993.
[140] Zhou S, Yin S, Yang S, Wang Y, Feng P. Identification of novel HIF2α inhibitors: a structure-based virtual screening approach. *J. Enzyme Inhib. Med. Chem.* 2026, 41(1):2606435.
[141] An Y, Lim J, Glavatskikh M, Wang X, Norris-Drouin J, *et al.* In silico fragment-based discovery of CIB1-directed anti-tumor agents by FRASE-bot. *Nat. Commun.* 2024, 15(1):5564.
[142] Wang J, Yuan F, Kendre M, He Z, Dong S, *et al.* Rational design of allosteric inhibitors targeting C797S mutant EGFR in NSCLC: an integrative in silico and *in vitro* study. *Front. Oncol.* 2025, 15:1590779.
[143] Chang H, Zhang Z, Tian J, Bai T, Xiao Z, *et al.* Machine learning-based virtual screening and identification of the fourth-generation EGFR inhibitors. *ACS Omega* 2024, 9(2):2314–2324.
[144] Feng C, Ge Y, Wang S, Li M, Chen Q, *et al.* Discovery of small-molecule PD-L1 inhibitors via virtual screening and their immune-mediated anti-tumor effects. *Pharmaceuticals* 2025, 18(8):1209.
[145] Feng W, Liu L, Li L, Du P, Yuan Z, *et al.* Design and discovery of POLQ helicase domain inhibitors by virtual screening and machine learning. *Med. Chem. Res.* 2025, 34(6):1377–1391.
[146] Wang Y, Wang C, Liu J, Sun D, Meng F, *et al.* Discovery of 3-hydroxymethyl-azetidine derivatives as potent polymerase theta inhibitors. *Bioorg. Med. Chem.* 2024, 103:117662.
[147] Shneyderman A, Hammer SS, Remmel HL, Veviorskiy A, Alawi KM, *et al.* Evaluation of (Z)-endoxifen as a potential therapy for glioblastoma multiforme through computational and experimental analyses. *Sci. Rep.* 2025, 15(1):38225.
[148] Sebastian AM, Peter D. Artificial intelligence in cancer research: trends, challenges and future directions. *Life* 2022, 12(12):1991.
[149] Gangwal A, Ansari A, Ahmad I, Azad AK, Kumarasamy V, *et al.* Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Front. Pharmacol.* 2024, 15:1331062.
[150] Zhang K, Yang X, Wang Y, Yu Y, Huang N, *et al.* Artificial intelligence in drug development. *Nat. Med.* 2025, 31(1):45–59.
[151] Guo J, Fialková V, Arango JD, Margreitter C, Janet JP, *et al.* Improving de novo molecular design with curriculum learning. *Nat. Mach. Intell.* 2022, 4(6):555–563.
[152] Mokaya M, Imrie F, Hoorn WP, Kalisz A, Bradley AR, *et al.* Testing the limits of SMILES-based de novo molecular generation with curriculum and deep reinforcement learning. *Nat. Mach. Intell.* 2023, 5(4):386–394.
[153] Zhang O, Zhang J, Jin J, Zhang X, Hu R, *et al.* ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nat. Mach. Intell.* 2023, 5(9):1020–1030.
[154] Zhang O, Wang T, Weng G, Jiang D, Wang N, *et al.* Learning on topological surface and geometric structure for 3D molecular generation. *Nat. Comput. Sci.* 2023, 3(10):849–859.
[155] Qian H, Huang W, Tu S, Xu L. KGDiff: towards explainable target-aware molecule generation with knowledge guidance. *Briefings Bioinf.* 2024, 25(1):bbad435.
[156] Xu M, Chen H. Tree-Invent: a novel multipurpose molecular generative model constrained with a topological tree. *J. Chem. Inf. Model.* 2023, 63(22):7067–7082.

[157] Moret M, Angona PI, Cotos L, Yan S, Atz K, *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* 2023, 14(1):114.
[158] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 2019, 37(9):1038–1040.
[159] Lim J, Hwang SY, Moon S, Kim S, Kim WY. Scaffold-based molecular design with a graph generative model. *Chem. Sci.* 2020, 11(4):1153–1164.
[160] Hu L, Yang Y, Zheng S, Xu J, Ran T, *et al.* Kinase inhibitor scaffold hopping with deep learning approaches. *J. Chem. Inf. Model.* 2021, 61(10):4900–4912.
[161] Zheng S, Lei Z, Ai H, Chen H, Deng D, *et al.* Deep scaffold hopping with multimodal transformer neural networks. *J. Cheminf.* 2021, 13(1):87.
[162] Imrie F, Bradley AR, Schaar M, Deane CM. Deep generative models for 3D linker design. *J. Chem. Inf. Model.* 2020, 60(4):1983–1995.
[163] Igashov I, Stärk H, Vignac C, Schneuing A, Satorras VG, *et al.* Equivariant 3D-conditional diffusion model for molecular linker design. *Nat. Mach. Intell.* 2024, 6(4):417–427.
[164] Su A, Luo Y, Zhang C, Duan H. Linker-GPT: design of Antibody-drug conjugates linkers with molecular generators and reinforcement learning. *Sci. Rep.* 2025, 15(1):20525.
[165] Guo G, Guo L, Qian J, He X, Qian X, *et al.* De novo design of protein-binding aptamers through deep reinforcement learning assembly of nucleic acid fragments. *bioRxiv* 2025.
[166] Bennett NR, Watson JL, Ragotte RJ, Borst AJ, See DL, *et al.* Atomically accurate de novo design of antibodies with RFdiffusion. *Nature* 2026, 649(8095):183–193.
[167] Xie W, Zhang J, Xie Q, Gong C, Ren Y, *et al.* Accelerating discovery of bioactive ligands with pharmacophore-informed generative models. *Nat. Commun.* 2025, 16(1):2391.
[168] Imrie F, Hadfield TE, Bradley AR, Deane CM. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* 2021, 12(43):14577–14589.
[169] Zhu H, Zhou R, Cao D, Tang J, Li M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nat. Commun.* 2023, 14(1):6234.
[170] Born J, Manica M, Oskooei A, Cadow J, Markert G, *et al.* PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* 2021, 24(4):102269.
[171] Pravalphruekul N, Piriyajitakonkij M, Phunchongharn P, Piyayotai S. De novo design of molecules with multiaction potential from differential gene expression using variational autoencoder. *J. Chem. Inf. Model.* 2023, 63(13):3999–4011.
[172] Kim H, Bae B, Park M, Shin Y, Ideker T, *et al.* A genotype-to-drug diffusion model for generation of tailored anti-cancer small molecules. *Nat. Commun.* 2025, 16(1):5628.
[173] Dolfus U, Briem H, Rarey M. Synthesis-aware generation of structural analogues. *J. Chem. Inf. Model.* 2022, 62(15):3565–3576.
[174] Qiang B, Zhou Y, Ding Y, Liu N, Song S, *et al.* Bridging the gap between chemical reaction pretraining and conditional molecule generation with a unified model. *Nat. Mach. Intell.* 2023, 5(12):1476–1485.
[175] Notin P, Rollins N, Gal Y, Sander C, Marks D. Machine learning for functional protein design. *Nat. Biotechnol.* 2024, 42(2):216–228.

[176] Bhat S, Palepu K, Hong L, Mao J, Ye T, *et al.* De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Sci. Adv.* 2025, 11(4):eadr8638.

[177] Chen LT, Quinn Z, Dumas M, Peng C, Hong L, *et al.* Target sequence-conditioned design of peptide binders using masked language modeling. *Nat. Biotechnol.* 2025, pp. 1–9.

[178] Geylan G, Paul Janet J, Tibo A, He J, Patronov A, *et al.* PepINVENT: generative peptide design beyond natural amino acids. *Chem. Sci.* 2025, 16(20):8682–8696.

[179] Kong X, Jiao R, Lin H, Guo R, Huang W, *et al.* Peptide design through binding interface mimicry with PepMimic. *Nat. Biomed. Eng.* 2025, pp. 1–16.

[180] Rettie SA, Campbell KV, Bera AK, Kang A, Kozlov S, *et al.* Cyclic peptide structure prediction and design using AlphaFold2. *Nat. Commun.* 2025, 16(1):4730.

[181] Rettie SA, Juergens D, Adebomi V, Bueso YF, Zhao Q, *et al.* Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. *Nat. Chem. Biol.* 2025, pp. 1–9.

[182] Xia Z, Jin Q, Long Z, He Y, Liu F, *et al.* Targeting overexpressed antigens in glioblastoma via CAR T cells with computationally designed high-affinity protein binders. *Nat. Biomed. Eng.* 2024, 8(12):1634–1650.

[183] Torres VS, Valle BM, Mackessy SP, Menzies SK, Casewell NR, *et al.* De novo designed proteins neutralize lethal snake venom toxins. *Nature* 2025, 639(80):225–231.

[184] Yang W, Hicks DR, Ghosh A, Schwartze TA, Conventry B, *et al.* Design of high-affinity binders to immune modulating receptors for cancer immunotherapy. *Nat. Commun.* 2025, 16(1):2001.

[185] Pacesa M, Nickel L, Schellhaas C, Schmidt J, Pyatova E, *et al.* One-shot design of functional protein binders with BindCraft. *Nature* 2025, 646(8084):483–492.

[186] Stark H, Faltings F, Choi M, Xie Y, Hur E, *et al.* BoltzGen: toward universal binder design. *bioRxiv* 2025.

[187] Team CD, Boitreaud J, Dent J, Geisz D, McPartlon M, *et al.* Zero-shot antibody design in a 24-well plate. *bioRxiv* 2025.

[188] Muratspahić E, Feldman D, Kim DE, Qu X, Bratovianu AM, *et al.* De novo design of miniprotein agonists and antagonists targeting G protein-coupled receptors. *bioRxiv* 2025.

[189] Goverde CA, Pacesa M, Goldbach N, Dornfeld LJ, Balbi PEM, *et al.* Computational design of soluble and functional membrane protein analogues. *Nature* 2024, 631(8020):449–458.

[190] Goel S, Thoutam V, Marroquin EM, Gokaslan A, Firouzbakht A, *et al.* MeMDLM: de novo membrane protein design with property-guided discrete diffusion. In *proceedings of the Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, Singapore, April 24–28, 2025.

[191] Lu L, Gou X, Tan SK, Mann SI, Yang H, *et al.* De novo design of drug-binding proteins with predictable binding energy and specificity. *Science* 2024, 384(6691):106–112.

[192] Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, *et al.* Design of highly functional genome editors by modelling CRISPR–Cas sequences. *Nature* 2025, 645(8080):518–525.

[193] He H, He B, Guan L, Zhao Y, Jiang F, *et al.* De novo generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. *Nat. Commun.* 2024, 15(1):6867.

[194] Guloglu B, Bragança M, Graves A, Cameron S, Atkinson T, *et al.* AbBFN2: a flexible antibody foundation model based on Bayesian flow networks. *bioRxiv* 2025.

[195] Wang R, Wu F, Shi J, Song Y, Kong Y, *et al.* A generative foundation model for antibody design. *bioRxiv* 2025.

[196] Wu F, Zhao Y, Wu J, Jiang B, He B, *et al.* De novo design of epitope-specific antibodies via a structure-driven computational workflow. *Nat. Commun.* 2025.

[197] Bennett NR, Watson JL, Ragotte RJ, Borst AJ, See DL, *et al.* Atomically accurate de novo design of antibodies with RFdiffusion. *Nature* 2026, 649(8095):183–193.

[198] Zhang H, Liu H, Xu Y, Huang H, Liu Y, *et al.* Deep generative models design mRNA sequences with enhanced translational capacity and stability. *Science* 2025, 390(6773):eadr8470.

[199] Zhao Y, Oono K, Takizawa H, Kotera M. GenerRNA: a generative pre-trained language model for de novo RNA design. *PLoS One* 2024, 19(10):e0310814.

[200] Zhang H, Zhang L, Lin A, Xu C, Li Z, *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* 2023, 621(7978):396–403.

[201] Ren Z, Jiang L, Di Y, Zhang D, Gong J, *et al.* CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics* 2024, 40(7):btae330.

[202] Chu Y, Yu D, Li Y, Huang K, Shen Y, *et al.* A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nat. Mach. Intell.* 2024, 6(4):449–460.

[203] Tang X, Huo M, Chen Y, Huang H, Qin S, *et al.* A novel deep generative model for mRNA vaccine development: Designing 5′ UTRs with N1-methyl-pseudouridine modification. *Acta Pharm. Sin. B* 2024, 14(4):1814–1826.

[204] Morrow AK, Thornal A, Flynn ED, Hoelzli E, Shan M, *et al.* ML-driven design of 3′ UTRs for mRNA stability. *bioRxiv* 2024.

[205] Li Y, Garcia G, Arumugaswami V, Guo F. Structure-based design of antisense oligonucleotides that inhibit SARS-CoV-2 replication. *bioRxiv* 2021.

[206] Hörberg J, Carlesso A, Reymer A. Mechanistic insights into ASO-RNA complexation: advancing antisense oligonucleotide design strategies. *Mol. Ther.-Nucl. Acids.* 2024, 35(4).

[207] Wong F, He D, Krishnan A, Hong L, Wang AZ, *et al.* Deep generative design of RNA aptamers using structural predictions. *Nat. Comput. Sci.* 2024, 4(11):829–839.

[208] Andress C, Kappel K, Villena ME, Cuperlovic-Culf M, Yan H, *et al.* DAPTEV: deep aptamer evolutionary modelling for COVID-19 drug design. *PLoS Comput. Biol.* 2023, 19(7):e1010774.

[209] Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* 2024, 386(6723):eado9336.

[210] Fang X, Liu L, Lei J, He D, Zhang S, *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* 2022, 4(2):127–134.

[211] Göller AH, Kuhnke L, Montanari F, Bonin A, Schneckener S, *et al.* Bayer's in silico ADMET platform: A journey of machine learning over the past two decades. *Drug Discovery Today* 2020, 25(9):1702–1709.

[212] Yoshikai Y, Mizuno T, Nemoto S, Kusuhara H. Difficulty in chirality recognition for transformer architectures learning chemical structures from string representations. *Nat. Commun.* 2024, 15(1):1197.

[213] Liu H, Zhu B, Nie S, Li H, Lin Y, *et al.* Advancing ADMET prediction through multiscale fragment-aware pretraining with MSformer-ADMET. *Briefings Bioinf.* 2025, 26(5):bbaf506.

[214] Gomes J, Ramsundar B, Feinberg EN, Pande VS. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv* 2017, arXiv:1703.10603.

[215] Yu G, Fan Q. Deep learning-driven drug response prediction and mechanistic insights in cancer genomics. *Sci. Rep.* 2025, 15(1):20824.

[216] Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020, 38(5):672–684.

[217] Chow KE, Rashid MM, Lim JJ, Chan S. Abstract B087: feasibility study of an *ex vivo* functional precision medicine platform, optim.AI^{TM}, in guiding treatment for pancreatic cancer. *Cancer Res.* 2025, 85(18_Supplement_3):B087.

[218] Li Z, Li Y, Xiang J, Wang X, Yang S, *et al.* AI-enabled virtual spatial proteomics from histopathology for interpretable biomarker discovery in lung cancer. *Nat. Med.* 2026, pp. 1–14.

[219] You Y, Lai X, Pan Y, Zheng H, Vera J, *et al.* Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct. Target. Ther.* 2022, 7(1):156.

[220] McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, *et al.* Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Med.* 2024, 7(1):114.

[221] Moshkov N, Bornholdt M, Benoit S, Smith M, McQuin C, *et al.* Learning representations for image-based profiling of perturbations. *Nat. Commun.* 2024, 15(1):1594.

[222] Rigamonti A, Viatore M, Polidori R, Rahal D, Erreni M, *et al.* Integrating AI-powered digital pathology and imaging mass cytometry identifies key classifiers of tumor cells, stroma, and immune cells in non-small cell lung cancer. *Cancer Res.* 2024, 84(7):1165–1177.

[223] Lee MR, Kang S, Lee J, Kong SY, Kim Y, *et al.* Organoid morphology-guided classification for oral cancer reveals prognosis. *Cell Rep. Med.* 2025, 6(5).

[224] Yu M, Li W, Yu Y, Zhao Y, Xiao L, *et al.* Deep learning large-scale drug discovery and repurposing. *Nat. Comput. Sci.* 2024, 4(8):600–614.

[225] Diosdi A, Toth T, Harmati M, Istvan G, Schrettner B, *et al.* HCS-3DX, a next-generation AI-driven automated 3D-oid high-content screening system. *Nat. Commun.* 2025, 16(1):8897.

[226] Park Y, Muttray NP, Hauschild AC. Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Briefinfs Bioinf.* 2024, 25(2):bbae004.

[227] Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, *et al.* Toward universal cell embeddings: Integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* 2024, 21(8):1492–1500.

[228] Villain L, Schaller S, Lefaudeux D, Lautz LS, Siccardi M, *et al.* Physiologically based kinetic modelling for species extrapolation of toxicokinetic data between small mammals: a systematic evaluation. *Environ. Int.* 2025, 207:110003.

[229] Kim S, Chen J, Cheng T, Gindulyte A, He J, *et al.* PubChem 2023 update. *Nucleic Acids Res.* 2023, 51(D1):D1373–D1380.

[230] Tanoli Z, Fernández-Torras A, Özcan UO, Kushnir A, Nader KM, *et al.* Computational drug repurposing: approaches, evaluation of in silico resources and case studies. *Nat. Rev. Drug Discovery* 2025, 24(7):521–542.

[231] Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery* 2017, 16(12):829–842.

[232] Lorente JS, Sokolov AV, Ferguson G, Schiöth HB, Hauser AS, *et al.* GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery* 2025, 24(6):458–479.
[233] Qi X, Zhao Y, Qi Z, Hou S, Chen J. Machine learning empowering drug discovery: applications, opportunities and challenges. *Molecules* 2024, 29(4):903.
[234] Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. *Comput. Ind. Eng.* 2020, 149:106854.
[235] Jiang K, Yan Z, Bernardo MD, Sgrizzi SR, Villiger L, *et al.* Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* 2024, 387(6732):eadr6006.
[236] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* 2019, 116(44):22071–22080.
[237] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* 2021, 110(3):457–506.
[238] Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* 2022, 96:101845.
[239] Garreau D, Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Palermo, Italy, August 26–28, 2020, pp. 1287–1296.
[240] Gujral O, Bafna M, Alm E, Berger B. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proc. Natl. Acad. Sci.* 2025, 122(34):e2506316122.
[241] Kovyrshin A, Tornberg L, Crain J, Mensa S, Tavernelli I, *et al.* Prioritizing quantum computing use cases in the drug discovery and development pipeline. *Drug Discovery Today* 2025, 30(3):104323.
[242] Rai BK, Sresht V, Yang Q, Unwalla R, Tu M, *et al.* TorsionNet: a deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *J. Chem. Inf. Model.* 2022, 62(4):785–800.
[243] Zhang X, Gao H, Wang H, Chen Z, Zhang Z, *et al.* PLANET: a Multi-objective graph neural network model for protein–ligand binding affinity prediction. *J. Chem. Inf. Model.* 2024, 64(7):2205–2220.
[244] Huang L, Xu T, Yu Y, Zhao P, Chen X, *et al.* A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* 2024, 15(1):2657.
[245] Abramson J, Adler J, Dunger J, Evans R, Green T, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024, 630(8016):493–500.
[246] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* 2022, 19(6):679–682.
[247] Su J, Li Z, Tao T, Han C, He Y, *et al.* Democratizing protein language model training, sharing and collaboration. *Nat. Biotechnol.* 2025, pp. 1–7.