Review

Not peer-reviewed version

Drug Discovery in the Era of Artificial Intelligence: From Target Identification to Clinical Trials

Yoshitaka Inoue [*,†] , Nan Hao [†] , Yingzhou Lu , Tianfan Fu , Augustin Luna [*]

Posted Date: 2 June 2026

doi: 10.20944/preprints202606.0091.v1

Keywords: artificial intelligence; machine learning; deep learning; drug discovery; drug development; drug-target interaction; molecule generation; clinical trials; biomedical data science; pharmacogenomics

 doi:10.20944/preprints202606.0091.v1

Review

Drug Discovery in the Era of Artificial Intelligence: From Target Identification to Clinical Trials

Yoshitaka Inoue [1,2,3,*,†], Nan Hao [4,†], Yingzhou Lu [5], Tianfan Fu [6] and Augustin Luna [2,3,*]

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA
[2] Computational Biology Branch, National Library of Medicine, Bethesda, USA
[3] Developmental Therapeutics Branch, National Cancer Institute, Bethesda, USA
[4] Renaissance School of Medicine, Stony Brook University, Stony Brook, USA
[5] School of Medicine, Stanford University, Stanford, USA
[6] State Key Laboratory for Novel Software Technology at Nanjing University, School of Computer Science, Nanjing University, Nanjing, China
* Correspondence: inoue019@umn.edu (Y.I.); augustin.luna@nih.gov (A.L.)
† These authors contributed equally to this work.

Abstract

Drug discovery remains costly and time-consuming, requiring substantial effort across target identification, lead optimization, and clinical validation. The rapid growth of artificial intelligence (AI) and machine learning (ML), particularly deep learning (DL), has created new opportunities to accelerate these processes by extracting complex patterns from large-scale biological and chemical data. As diverse datasets and modeling approaches continue to expand, there is a growing need for a structured understanding of how AI/ML methods contribute across the drug discovery pipeline. In this review, we provide a stage-wise synthesis of AI/ML applications in drug discovery, spanning target identification, target development, lead identification, lead optimization, and clinical trials. We highlight key methodological paradigms, including representation learning, graph-based modeling, and generative approaches, and discuss how different data modalities (e.g., omics, chemical structures, and pharmacogenomics) can shape computational model design and performance. We further examine critical challenges such as data bias, distributional shifts, quality, as well as the trade-off between predictive performance and mechanistic interpretability. Overall, while AI/ML methods have demonstrated substantial promise in accelerating drug discovery, their effectiveness remains constrained by data limitations and generalization challenges. Future progress will depend on improved data integration, robust evaluation across datasets, and the development of models that balance predictive accuracy with biological interpretability.

Keywords: artificial intelligence; machine learning; deep learning; drug discovery; molecular generation; virtual screening; lead optimization; clinical trials; pharmacogenomics

1. Introduction

Drug discovery is a complex, time-consuming, and expensive process [1]. Machine learning (ML) and artificial intelligence (AI) technologies have emerged as a powerful tool to support this workflow, leveraging biomedical data to accelerate development, reduce costs, and inform decision-making at multiple stages [2]. Although several reviews have addressed ML applications in specific aspects of drug discovery [3,4], a comprehensive, end-to-end perspective is still needed. In this review, we present an integrated overview of ML across the drug development pipeline, following the progression of a potential therapeutic from early target identification to clinical trials. This approach highlights how ML methods contribute at each stage and identifies practical strategies that have been implemented in current research. The foundation of these approaches relies on diverse datasets, summarized in Table 1, and on a range of ML models, illustrated in Figure 1. Figure 2 further shows how these models

 doi:10.20944/preprints202606.0091.v1

integrate across the drug discovery process [5], and Box 1 defines essential ML terminology. In addition to the stage-wise review, we conclude with a section on challenges and future directions, emphasizing benchmark bias, class imbalance issues, data distribution shifts that affect generalizability, mechanistic interpretability, and the constraints of deploying Large Language Models (LLMs) in clinical settings.

Table 1. Databases Related To Drug Discovery. Data are categorized by function and stage: Target Identification (●), Target Development (●), Lead Identification (●), Lead Optimization (●) and Clinical Trials (●), represented by distinct colored dots in the table. (C/R) denotes whether the dataset has been used for classification (C) and/or regression (R) tasks; "Other" indicates use in other task settings.

Name	Usage / ML Application Task	Limitations (see Section 3.4)
● UK Biobank [6]	Large-scale genotype-phenotype cohort; Application: C/R [7,8]	Sampling bias due to non-representative cohort (e.g., excludes people of age > 69 and < 40) [9]
● MGnify [10]	Microbiome genomic annotations; Application: Other	Sampling bias of particular taxa [11]
●● National Cancer Institute 60 (NCI60) [12] ●● Genomics of Drug Sensitivity in Cancer (GDSC) [15] ●● Cancer Cell Line Encyclopedia (CCLE) [16]	Large-scale drug response measurements across cancer cell lines; Application: C/R [13]	Cross-dataset inconsistency (e.g., AUC vs IC50) [14]
●● CellMiner Cross Database (CellMinerCDB) [17] ●● Therapeutics Data Commons (TDC) [19]	Cross-dataset multi-omics & drug response integration; Application: CellMinerCDB: C/R [13], TDC: C/R [18]	Cross-dataset inconsistency (e.g., AUC vs IC50) [14]
● Davis kinase inhibitors DB [20] ● Kinase Inhibitor Bioactivity Data (KIBA) [23] ● BindingDB [24]	Experimental binding affinity datasets for protein-ligand interactions; Application: C/R [21]	Sampling bias and class imbalance [22]
●● ChEMBL [25] ●● DrugTargetCommons (DTC) [31]	Curated bioactivity/literature databases; Application: ChEMBL: C/R [26,27], DTC: R [28]	Coverage bias [29] and similarity bias [30]
●● OpenTargets [32]	Integrated target-disease association and prioritization; Application: C [33]	Evidence weighting subjectivity, data source imbalance [32]
●● ZINC ligand discovery database [34]	Virtual chemical library for ligand screening and generation; Application: C/R [35]	Coverage bias due to incomplete representation of chemical or biological space [29]
● MoleculeNet [36]	Benchmark datasets for evaluating molecular property prediction; Application: C/R [36]	Coverage bias due to incomplete representation of chemical or biological space [29]
● PK-DB [37]	ADME/PK experimental data; Application: C/R [38]	Data incompleteness, heterogeneous reporting, and aggregation bias [37]
●● RCSB Protein Data Bank (PDB) [39] ● PDBBind [42]	High-quality protein 3D structures with protein-ligand binding benchmark; Application: R [40]	Bias toward crystallizable, stable, and well-folded proteins [41]
● Uniclust [43] ● Uniref [45]	Clustered protein sequences with different similarity levels available; Application: Other	Selection bias due to uneven homolog availability, over-representing well-characterized proteins [44]
● ClinicalTrials.gov [46]	Clinical trial registry and metadata for study design and outcomes; Application: Other	Selection bias, information / classification bias, confounding bias [47,48]

The following sections follow the stages of drug development:

- **Target Identification** examines ML applications for discovering therapeutic targets; we briefly discuss key data sources, including CRISPR screens, GWAS, and single-cell RNA sequencing (Section 2).
- **Target Development** explores ML-driven methods to predict how drugs interact with their targets and cellular systems, guiding molecule design (Section 3).
- **Lead Identification** introduces computational techniques for molecule generation and prioritization, including applications in targeted cancer therapy (Section 4).
- **Lead Optimization** addresses the refinement of candidate molecules, covering toxicity prediction, pharmacokinetic modeling, and optimization of safety and efficacy profiles (Section 5).
- **Clinical Trials** considers the use of ML to predict patient outcomes, optimize trial site selection, match patients to trials, and identify similarities across studies to enhance trial success (Section 6).

 doi:10.20944/preprints202606.0091.v1

We include recent preprints (e.g., arXiv) to capture the rapidly evolving landscape of AI in drug discovery; however, these works have not undergone peer review and should be interpreted with appropriate caution.

Box 1. Essential Terminology in Advanced AI and ML

- **Bidirectional Encoder Representations from Transformers (BERT):** A transformer-based language model that encodes bidirectional context from text sequences.
- **Contrastive Learning:** A representation learning paradigm that distinguishes samples by maximizing similarity between positive pairs and dissimilarity between negatives.
- **Convolutional Neural Network (CNN):** A neural network architecture that extracts hierarchical features using convolutional filters.
- **Diffusion:** A generative modeling framework that learns to generate data by iteratively denoising from a noise distribution.
- **Equivariant Network:** A model whose outputs transform consistently under geometric transformations of the input.
- **Equivariant Graph Neural Network (EGNN):** A graph neural network that preserves equivariance in 3D space (i.e., if the input is translated or rotated, the output changes consistently under the same transformation), enabling geometry-aware molecular modeling.
- **Geometric Deep Learning (GDL):** A class of methods that extend deep learning to non-Euclidean domains such as graphs and manifolds.
- **Genetic Algorithm (GA):** An optimization method based on evolutionary principles such as mutation, crossover, and selection.
- **Graph Neural Network (GNN):** A neural network designed to learn from graph-structured data via message passing between nodes.
- **Markov Chain Monte Carlo (MCMC):** A class of algorithms for sampling from complex probability distributions using Markov chains.
- **Neural Ordinary Differential Equations (Neural-ODE):** A class of continuous-depth neural network models in which a neural network learns the derivative, or rate of change, of a system's hidden state, and an ODE solver is used to evolve that state over time.
- **Recurrent Neural Network (RNN):** A neural network architecture for modeling sequential data with recurrent connections.
- **Representation Learning:** A paradigm in which models learn compact and informative feature representations (i.e., embeddings) from raw data.
- **Special Euclidean group (SE(n)):** The group of distance- and orientation-preserving transformations in n-dimensional space, including rotations and translations; for example, SE(3) denotes all such transformations in 3D space.

2. Target Identification

Target identification marks the initial stage of drug discovery, where the primary challenge lies in connecting genetic variation to disease mechanisms and experimentally validating causal genes. Among impactful approaches are genome-wide association studies (GWAS), CRISPR-based functional genomics, and single-cell RNA sequencing (scRNA-seq). These methods are complementary: GWAS offers breadth by surveying genetic variation across populations, CRISPR provides causal functional validation, and scRNA-seq delivers cellular resolution. Integrative data portals such as OpenTargets [32] support this process by presenting disease-gene associations based on aforementioned data sources, including GWAS, CRISPR, and expression data, amongst others, to facilitate systematic target prioritization. Separately, individual studies such as Morris et al integrate these methodologies to understand human genetic variants for blood cell traits using single-cell CRISPR screens, showcasing their importance in modern target discovery [49,50].

 doi:10.20944/preprints202606.0091.v1

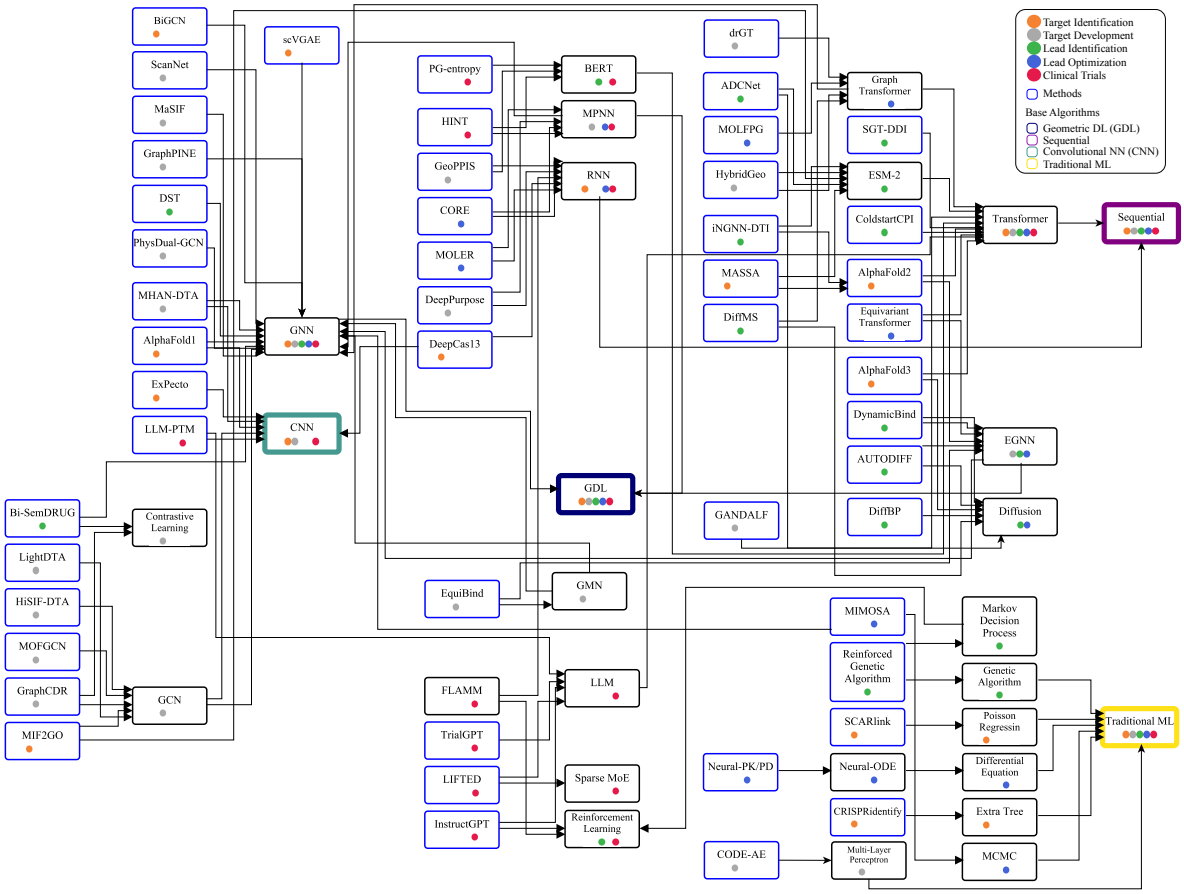

Figure 1. ML algorithms across drug discovery. Colored dots indicate stages—Target Identification (●), Target Development (●), Lead Identification (●), Lead Optimization (●), Clinical Trials (●). Blue-outlined labels denote methods; colored labels denote base algorithms.

2.1. Understanding Biological Functions of Possible Targets

Population-level discovery. GWAS systematically links common genetic variants to disease phenotypes in large cohorts such as the UK Biobank ($N = 500{,}000$) [6]. Recent computational advances extend the utility of GWAS by helping pinpoint functional non-coding variants: the method SCARlink maps GWAS signals to enhancer-gene links using regularized Poisson regression [51], while ExPecto leverages a deep convolutional network trained to predict chromatin features, then uses a linear model to infer tissue-specific expression changes from sequence variants, thereby prioritizing causal GWAS variants [52]. Together, these approaches refine association loci into mechanistic regulatory hypotheses.

Beyond variant-to-gene mapping, recent work shows that integrating multiple protein data modalities (e.g., structure and sequence) can help prioritize understudied targets and infer protein function when direct experimental evidence is limited. MASSA jointly pre-trains on protein sequence, structure, and functional annotations to learn transferable protein embeddings that improve downstream protein-property, protein-protein interaction, and protein-ligand prediction tasks [53]. Similarly, MIF2GO sequentially integrates protein data from six modalities, including homology, interaction, domain, pathway, and subcellular localization signals, and demonstrates strong robustness, especially in cases where data for the different modalities is available [54]. For target identification, these multimodal representations are valuable because they move beyond sequence-only function inference and provide a principled way to connect molecular structure and sequence information when nominating plausible disease-relevant targets.

Causal validation of gene function. CRISPR-based technologies enable direct perturbation of candidate genes to test their functional relevance. To support such applications, computational tools reduce noise and improve specificity. For example, CRISPRidentify uses a dataset of positive and

 doi:10.20944/preprints202606.0091.v1

negative CRISPR arrays (clusters of repeats plus spacers) to train a decision tree-based classifier on features such as repeat similarity, spacer-length uniformity, and repeat hairpin stability, yielding a lower false-positive rate than heuristic repeat-search methods [55]. DeepCas13 likewise leverages large-scale Cas13d proliferation screens and a convolutional-recurrent neural network to model guide sequence and RNA secondary-structure features, enabling accurate prediction of both on-target knockdown efficiency and guide-dependent off-target viability effects [56]. By integrating experimental screens with computational predictions, researchers can more efficiently confirm or refute GWAS-derived hypotheses.

Cell-type-specific resolution. Applications of scRNA-seq for drug discovery have been previously reviewed [57]; its applications complement population and perturbation studies by profiling expression at single-cell resolution, thereby uncovering disease-relevant heterogeneity. One central challenge being addressed through ML with scRNA-seq is in addressing dropout-induced sparsity [58], which motivates advanced imputation methods, such as MAGIC, ALRA, and DeepImpute [59]. Newer methods, such as scVGAE, utilize graph-based methods; scVGAE leverages a variational graph autoencoder with a Zero-Inflated Negative Binomial model [60], while BiGCN employs bi-directional graph convolutional networks [61]. The generalizability of these methods remains an open issue, as evident in evaluation variability across datasets, while evaluations based on clustering can suffer from overdependence on ground truth and the potential for discrepancies between the ground truth labels and the true intrinsic clustering structure [62].

2.2. Protein Structure Prediction

Protein structure prediction (PSP) has emerged as a powerful tool for determining protein 3D structures from amino acid sequences, which is essential for rational understanding of drug target interactions. PSP has been a major and fundamental challenge in bioinformatics. Before DL-based methods, PSP made vast strides using evolutionary coupling analysis, which inferred spatial relationships by studying co-evolution of amino acids [63]. Further breakthroughs came with AlphaFold [64], which achieved unprecedented accuracy in CASP13's free modeling category [65] by integrating evolutionary information with DL and attention mechanisms. Subsequent versions, AlphaFold2 [66] and AlphaFold3 [67], further improved predictions using Graph Neural Networks, transformers, and diffusion models. The success of these efforts has relied on large, community-driven datasets. For example, AlphaFold initially leveraged experimentally determined protein structures from the RCSB Protein Data Bank (PDB) [39], along with evolutionary information derived from multiple sequence alignments such as Uniclust [43]; its performance was evaluated in community-wide benchmarks such as the Critical Assessment of Structure Prediction (CASP) [65]. AlphaFold2 expanded to include multiple sequence alignments from UniRef90 [45], BFD [68], and MGnify [10], improving structure prediction accuracy. Despite advances, these methods continue to have limitations, such as: (1) stereochemical errors, including chirality violations and atomic clashes in large complexes, (2) mis-modeling of disordered regions due to the diffusion model, and (3) inability to capture protein dynamics, such as open conformations of E3 ubiquitin ligases [67]. These challenges highlight areas for future improvement in PSP for both structural biology and drug discovery applications.

 doi:10.20944/preprints202606.0091.v1

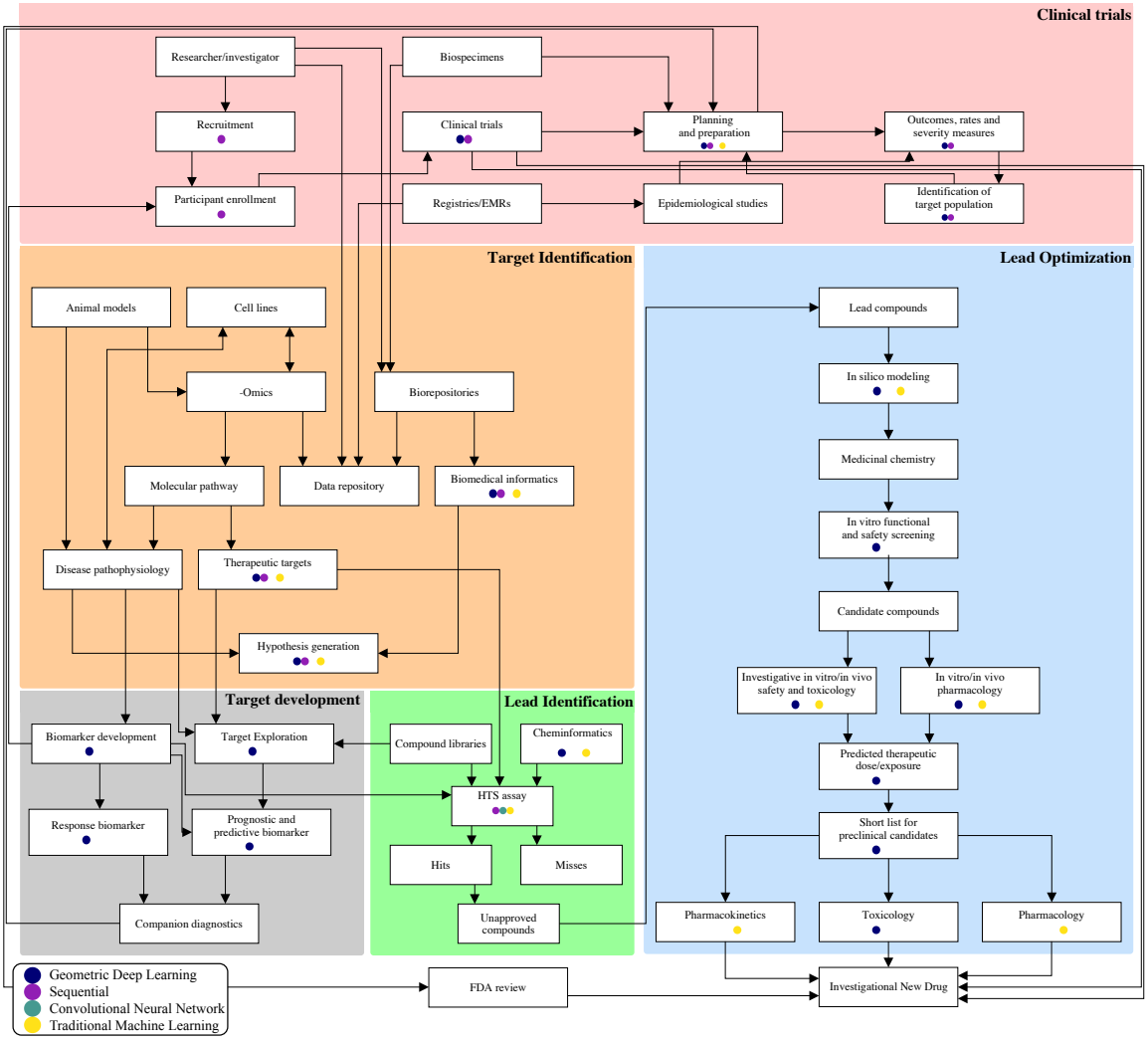

Figure 2. The Drug Discovery Process with Machine Learning (ML) Methods. This diagram illustrates the drug discovery process and highlights the integration of various ML models at different stages. The stages include Clinical Trials, Target Identification, Lead Identification, Lead Optimization, and Target Development. The diagram highlights the integration of ML techniques across these stages, represented by color-coded dots (● for geometric deep learning (GDL), ● for Sequential, ● for convolutional neural network (CNN), and ● for Traditional ML), demonstrating their pivotal role in enhancing the drug discovery process.

3. Target Development

3.1. Drug-Target Interaction Prediction

The primary tasks in target development encompass understanding drug-target interactions, predicting protein-drug binding, and forecasting drug responses. In this context, drug-target interaction (DTI) prediction determines whether and how strongly a drug molecule interacts with a specific protein target, typically formulated as a binary classification or regression problem. Databases such as KIBA [23], Davis [20], and BindingDB [24] provide essential data for training predictive models and serve as a benchmarking foundation. Additional widely used resources include ChEMBL [25] and DrugTargetCommons [31], which provide large-scale curated ligand binding affinity and enzyme inhibition data for drug-target interaction modeling. In addition, platforms such as the Therapeutics Data Commons [19] provide standardized benchmarks and data portals for AI-driven drug discovery tasks. Many ML approaches have been used to tackle this immense challenge, as recently reviewed [69].

DTI prediction methods can be broadly categorized into representation-based, network-based, structure-based, and multimodal interaction models. Amongst others, from our work, we highlight two approaches, DeepPurpose [21] and DrugAgent [18]. For representation learning, DeepPurpose [21]

 doi:10.20944/preprints202606.0091.v1

generates embeddings for both drugs and proteins using Deep Neural Networks, a convolutional neural network (CNN), a Recurrent Neural Network (RNN), and transformers within an encoder-decoder framework to predict binding affinities. In contrast, to enhance model explainability beyond accuracy alone, DrugAgent [18] integrates scientific literature, ML predictions, and prior knowledge graphs to clarify why a drug is effective against a specific target, following a trend to incorporate information about previous research into such efforts. Consequently, data-rich settings often benefit from end-to-end neural encoders for affinity prediction, whereas scenarios requiring mechanistic justification or decision transparency can leverage knowledge-integrated pipelines. In practice, these strategies are complementary: learned embeddings provide strong predictive signals, while external evidence improves interpretability and supports target-mechanism plausibility.

Structure-based methods have become increasingly important for modeling protein-small molecule interactions. These methods tend to leverage three-dimensional protein-ligand complexes to directly predict binding affinity, with datasets such as PDBbind [42] and benchmark suites such as the Comparative Assessment of Scoring Functions (CASF) [70] for evaluating protein-ligand interactions enabling the development of deep learning models that capture geometric and physicochemical interactions [71].

HybridGeo [71] is a representative structure-based geometric deep learning framework that explicitly models three-dimensional interactions in protein-ligand complexes. It employs dual-view graphs to distinguish between intra-molecular covalent interactions and inter-molecular non-covalent interactions, and it introduces hybrid message-passing strategies that use spatial distances both as message features and as aggregation weights. HybridGeo's main limitation is its static representation of protein-ligand complexes. It uses distance-threshold-based graphs rather than explicitly modeling interaction types, electronic effects, or conformational dynamics, which may affect generalization to novel proteins and ligands.

Beyond structure-based methods, network-based approaches incorporate relational biological information to enhance protein representations. HiSIF-DTA [72] is a hierarchical semantic information fusion framework that integrates low-order structural semantics from residue-level protein graphs with high-order functional semantics from protein-protein interaction (PPI) networks. Specifically, each protein is represented by a residue-level graph constructed from structural information such as contact maps, while its functional context is modeled as a node in a PPI graph. The framework couples these two levels through cross-level information propagation, thereby enriching protein embeddings with both local structural and global interaction-context information. The resulting protein representations are then combined with drug graph embeddings for binding affinity prediction. However, this hierarchical fusion may incur substantial computational cost, as it requires extracting residue-level graph representations for all proteins in each training batch before integrating them with the PPI network.

LightDTA [73] instead prioritizes computational efficiency, using random-walk-based embeddings on PPI networks to reduce memory cost and inference time. To compensate for the loss of fine-grained molecular information, it distills structural and biochemical knowledge from a more expressive teacher model into a lightweight student model, although this comes at the cost of reduced fine-grained biochemical interpretability.

Moreover, multimodal interaction models aim to integrate heterogeneous representations and explicitly model cross-entity interactions. MHAN-DTA [74] is a multiscale hybrid attention framework that combines, for each entity, sequence-based and graph-based representations through cross-modal attention, specifically fusing protein residue sequences with pocket contact graphs and drug SMILES with molecular graphs. It then applies a cross-entity attention module to model fine-grained interactions between the fused drug representation and the fused protein-pocket representation. In addition, MHAN-DTA uses self-attention over the full protein sequence so that residues in the local binding pocket can aggregate information from remote residues outside the pocket, providing a more global protein context than pocket-only models. A limitation, however, is that this design relies on binding-

 doi:10.20944/preprints202606.0091.v1

complex or pocket information and may be less effective for proteins with highly complex domain organization or specialized active-site chemistry, such as metal-coordinated binding environments.

In addition, recent work has explored physics-informed graph models that incorporate explicit physical interaction terms into affinity prediction. PhysDual-GCN [75] combines graph-based ligand and protein representations with energy terms related to atomic attraction or repulsion to approximate docking-derived binding scores; the study focuses on DYRK2, a kinase with potential relevance to Alzheimer's disease. Part of its current limitations is that it was evaluated only against docking-derived references using a very small ligand set, and that its main architecture does not explicitly model the full 3D protein structure.

3.2. Binding Site Prediction

While DTI prediction determines whether and how strongly molecules interact, binding site prediction focuses on the spatial aspect by identifying specific regions on proteins where these interactions physically occur. Because affinity prediction alone does not provide explicit spatial context, binding-site prediction complements DTI modeling by localizing interactions to specific residues or pockets. This structural analysis reveals the locations, or "pockets," where ligands, proteins, peptides, etc., dock with the target protein, providing crucial information for drug design and optimization. Work in this area is often divided by representation choices (e.g., surface-based descriptors, atom-level spatio-chemical fields, and equivariant pose modeling) because they differ in data prerequisites (surface meshes vs. high-quality atomic coordinates), resolution (interface-level vs. residue/atom-level), and downstream goals (broad pocket discovery vs. residue prioritization vs. fast pose generation). We highlight methods using these different strategies.

From a surface-based perspective, MaSIF (molecular surface interaction fingerprinting) [76] employs geometric deep learning (GDL) (see Box 1) to generate unique surface-based fingerprints for biomolecular interactions, focusing on protein-protein and protein-ligand binding. MaSIF leverages neural networks to capture dynamic molecular surface properties, including electrostatic potential, hydrophobicity, and hydrogen bonding patterns, offering advantages over traditional molecular descriptors that rely on static representations. GeoPPIS [77], while similarly motivated by the need to better capture three-dimensional structural information, addresses a different problem and representation level. Rather than learning from surface patches, GeoPPIS formulates protein-protein interaction site prediction as a residue-level task on a geometry-aware k-nearest-neighbor graph, where both node and edge descriptors include scalar and vector geometric features. In addition, unlike MaSIF's emphasis on surface fingerprinting, GeoPPIS explicitly targets the limitations of current PPIS predictors by combining geometric graph learning with relative solvent accessibility-guided transfer learning to reduce overfitting and with clustering/ensemble post-prediction strategies to improve prediction stability. Together, these studies show that geometry-aware learning can be useful across different structural abstractions, from surface patches to residue graphs, although they are designed for different prediction settings and outputs.

In contrast, ScanNet (spatio-chemical arrangement of neighbors neural network) [78] uses GDL to analyze spatio-chemical (SC) properties at atomic and amino acid levels for predicting functional sites in proteins. The model constructs local coordinate frames centered on each heavy atom, oriented by covalent bonds, while processing neighboring atoms as point clouds with specific coordinates and chemical attributes. Through trainable SC filters, Gaussian kernels, and sparse bilinear products, ScanNet preserves spatial relationships across physical and attribute space, ensuring predictions remain invariant to Euclidean transformations. This atom-level framing sharpens residue-level site delineation when high-quality coordinates are available.

Meanwhile, EquiBind [79] further advances binding site prediction by employing SE(3)-equivariant GDL to model how drug-like molecules bind to protein targets. The model predicts receptor binding sites and ligand orientations while accounting for ligand flexibility, which is critical because drug-like molecules often change conformation during binding. Compared to traditional methods that require extensive sampling, EquiBind achieves up to 100× faster predictions by combining graph matching

 doi:10.20944/preprints202606.0091.v1

networks with E(3)-equivariant neural networks (see Box 1). Accordingly, this pose-centric view favors rapid structure-based design workflows that require plausible placements at scale.

Although significant advances have been made, these methods still face substantial challenges: surface-based approaches often lack sufficient chemical context; atom-level spatio-chemical methods require high-quality structural data and still struggle with induced fit and long-range interaction effects; equivariant pose-modeling, while geometrically expressive, remains computationally intensive, demands large structural datasets, and often generalizes poorly to novel protein pockets or ligands. Across all classes, there is a persistent dependency on static structures, limited accounting for receptor/ligand flexibility, and reduced performance on unseen protein families or chemically diverse ligands.

3.3. Drug-Response Prediction

A related challenge is drug-response prediction (DRP), which must reconcile three requirements: integrating heterogeneous cell/drug evidence, maintaining generalization under distributional shift, and providing mechanistic interpretability that can be biologically vetted. Large-scale pharmacogenomic public resources underpin this task of which there are several including the Genomics of Drug Sensitivity in Cancer (GDSC) [15] that provides drug response measurements for over 1,000 cancer cell lines across hundreds of compounds with accompanying multi-omics data and the Cancer Cell Line Encyclopedia (CCLE) [16] offers complementary genetic and pharmacologic characterization at similar scale. CellMinerCDB [80] aggregates multiple pharmacogenomics datasets with molecular profiling to enable cross-dataset analyses. Large-scale efforts to associate phenotypes with drug response were pioneered by the National Cancer Institute since the 1990s and have been an intense area of research utilizing computational algorithms ever since. Here, we describe several newer methods that use DL [81].

By constructing similarity matrices from gene expression, copy number variation, mutations, and drug fingerprints, and propagating information over the resulting cell-drug graph, MOFGCN [82] learns latent features for sensitivity prediction and attains superior performance on both GDSC and CCLE. This design capitalizes on redundancy across modalities to denoise noisy features, and it is particularly effective when omics breadth is wide and measurement noise is idiosyncratic rather than systematic. In practice, similarity fusion trades some fine-grained mechanism tracing for stability and coverage. By contrast, when distributional variability dominates (e.g., shifts in lineage composition, assay platforms, or response sparsity), contrastive objectives can sharpen decision boundaries. GraphCDR [83] represents cell lines and drugs as nodes with responses as edges, and applies multi-task contrastive learning to separate sensitive versus resistant patterns explicitly. This objective improves generalization by forcing representation spaces to respect response discriminants across cohorts. The benefit is resilience to shift; the trade-off is that the latent separation is optimized for discrimination first, with mechanistic explanations emerging indirectly.

In a complementary approach, when interpretability and biological validation are prioritized, attention and importance propagation on heterogeneous graphs can make gene-level drivers explicit. In an example from our work, drGT [13] applies a graph attention network over drug-cell-gene relations, producing attention scores that quantify gene importance in each drug-cell context. This allows external validation (e.g., PubMed co-occurrence) and achieves 78% accuracy and 76% F1 for DNA-damaging compounds in NCI60 [12].

3.4. Data Biases and Method Selection

Drug response prediction (DRP) models are constrained by biases in pharmacogenomic datasets, affecting both predictive performance and out-of-distribution generalization. These biases arise from the composition of the dataset, the measurement of responses, and the distributional heterogeneity across studies [14,84–86].

First, dataset bias reflects the uneven representation of cancer types, molecular subtypes, and drug classes in pharmacogenomic resources [14,87]. Well-studied cell lines and drug classes (e.g., kinase inhibitors) are often overrepresented, leading models to learn lineage- or drug-specific correlations that may not generalize to underrepresented cancers or novel mechanisms of action. This issue is closely

 doi:10.20944/preprints202606.0091.v1

related to similarity bias [30], in which models exploit shared chemical or biological patterns rather than learning true drug-response or drug-target mechanisms. Consequently, models may perform well on random splits dominated by chemically or biologically similar samples, but fail to generalize across underrepresented protein families, unseen molecular scaffolds, or sparsely sampled regions of chemical space [36,88].

Second, label bias and measurement variability introduce additional challenges. There is no de facto drug response measurement; some databases use IC_{50} and, others, AUC, which are not directly interchangeable and also can depend on assay design (e.g., dose range and curve fitting algorithms). Differences in experimental protocols across datasets (e.g., GDSC vs. CCLE) introduce systematic discrepancies in response labels [14,89]. For example, in classification tasks, poor community standardization in defining sensitive and resistant responses in samples and weak responses can lead to implicit label noise and class imbalance.

Third, distributional shift limits real-world applicability. Models trained on cell lines often fail to generalize to patient data due to the data heterogeneity differences [90]. Moreover, drug response in vitro reflects controlled dose-response assays, whereas clinical outcomes occur in the presence of many other factors such as circulatory and immune systems, which can lead to mismatched label semantics [91]. Together, these discrepancies create inconsistencies in both features and labels, even for overlapping drugs or molecular profiles. To mitigate this problem, CODE-AE [90] performs domain adaptation between cell-line and patient domains by learning representations that are transferable across the two settings while accounting for confounding effects. In addition, GANDALF [91] uses a diffusion-based generative approach to synthesize patient-like data from cell-line data, thereby augmenting the limited patient training data.

These biases directly inform method selection. Similarity-based fusion is effective when multi-omics coverage is broad, and redundancy can reduce noise [82]. In contrast, contrastive graph-based methods improve robustness under distributional shift by enforcing discriminative structure across cohorts [92]. For methods that incorporate prior knowledge, results can depend on the availability and quality of included prior knowledge [13].

Data leakage is a common and often underappreciated problem in machine learning for biomedical prediction. In knowledge-graph link prediction, for example, a chronological split is often preferable when the goal is to predict future discoveries rather than randomly hide existing edges. This is because missing edges do not necessarily indicate a negative label; they could simply be edges that have yet to be verified. This problem is exacerbated by the scarcity of reported negative results, which makes it difficult to distinguish true non-interactions from unobserved or untested associations; attempts to address the positive bias remain limited but exist, as with the journal Access Microbiology [93]. However, many benchmark datasets do not provide sample-level temporal metadata [94]. Similarly, in drug response prediction, integrating multiple datasets can introduce leakage because the same or near-duplicate drug-cell line response measurements may appear across different sources with inconsistent identifiers [95]. As a result, explicitly removing all test-related labels or equivalent measurements from the training data requires a concerted effort.

Data quality is also a critical consideration for predictive modeling. Well-known curated resources such as ChEMBL, specifically consider this issue by combining manual and automated curation to standardize activity values and units, annotate assays and targets, and flag potential data-quality issues such as outliers, duplicates, and transcription errors [96]. These curation efforts substantially improve the accessibility, comparability, and integrity of bioactivity data. At the same time, because the underlying data are drawn from heterogeneous publications and assay settings, some residual uncertainty in assay comparability, target assignment, and measurement consistency can remain [97]. A recent publication by the ChEMBL group describes aspects related to data quality. For example, they describe ongoing efforts to provide greater consistency to pharmacokinetic data units with 91% of AUC data units being $ng\ h\ ml^{-1}$ [98]. Gibson et al. report on the usage of datasets for clinical prediction models, without clear provenance and major deficiencies based on criteria from TRIPOD+AI

 doi:10.20944/preprints202606.0091.v1

(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence [99]) used in over 100 studies [100]. This work highlights the need for researchers to understand the datasets they use and to remain vigilant about data quality.

4. Lead Identification

4.1. Virtual Screening with AI

Lead identification focuses on discovering promising drug candidates with desirable properties through ML tasks such as protein binding site prediction and molecule generation. Unlike target identification, this phase emphasizes finding and optimizing potential drug molecules that could become therapeutic agents. In this context, virtual screening searches for candidate compounds within existing chemical libraries, and, since this process can be highly time-consuming, AI-based approaches have increasingly become a chosen method for simulating and prioritizing candidates before wet-lab validation.

We contrast two complementary strategies: (1) data-driven convolutional scoring and (2) structure-based flexible receptor modeling, because they trade off coverage, pose realism (i.e., plausibility of binding pose orientation and conformation), and computation. One notable example is the Atomwise AIMS Program, which developed AtomNet [101], a CNN designed to explore chemical space and identify promising drug candidates. AtomNet achieved an impressive 73% success rate across 296 academic targets, with an average hit rate of 7.6%, far surpassing traditional methods (0.001-0.15%). The model performs consistently across different structure types (crystal: 5.6%, cryo-EM: 5.5%, homology: 5.1%) and can identify hits for 70% of targets even without known activities. Leveraging the massive computational power of 40,000 CPUs and 3,500 GPUs to screen 16 billion compounds, AtomNet achieved 91% hit confirmation rate based on experimental validation in dose-response assays, successfully targeting challenging proteins, including protein-protein interactions (74% success) and allosteric sites (79% success).

In contrast, RosettaVS [102] emphasizes structure-based virtual screening by explicitly modeling receptor flexibility and optimizing protein-ligand energetics, using full side-chain and partial backbone flexibility to capture induced conformational changes, and employing an improved physics-based force field (RosettaGenFF-VS) to evaluate binding poses and affinities. Using active learning, it can screen billions of compounds within a few days, successfully identifying novel binders for KLHDC2 (7 hits, 14% hit rate) and Nav1.7 (4 hits, 44% hit rate) with micromolar affinities. Predicted binding poses have been validated with X-ray structures, confirming the method's accuracy.

Despite the gains, significant limitations persist across these methods. Scaling to libraries with billions of entries imposes a heavy computational budget. Moreover, early enrichment (7.6-14%) [101,102] does not preclude substantial attrition during confirmatory assays, indicating a nontrivial false-positive burden. Coverage is further bounded by the finite portion of drug-like space represented in available libraries, while candidates still demand synthesis and experimental validation.

4.2. Molecule Generation

Molecule generation aims to create new and diverse molecular structures with specific desirable properties. These approaches are often applied in the context of large virtual chemical libraries such as ZINC [34], which support ligand screening and molecule generation. In practice, approaches in this field generally fall into two complementary categories: generative models (GMs) and combinatorial optimization (CO) methods, both designed to propose novel compounds that satisfy predefined criteria such as the Quantitative Estimate of Drug-likeness (QED), a metric that optimizes eight desirable molecular properties, including molecular weight [103].

A thorough review of this area has recently been presented by Zeng et al. [104]. Briefly, we cover two main areas based on our experiences. From a representation-learning standpoint, generative models learn to produce novel molecular structures from latent space representations, including autoregressive and diffusion-based approaches. An example is Autoregressive Diffusion Modeling

 doi:10.20944/preprints202606.0091.v1

for Structure-Based Drug Design (AUTODIFF) [105], which employs a conformal motif assembly strategy combined with an SE(3)-equivariant network to accurately model protein-ligand interactions, generating drug-like molecules that meet multiple critical criteria, including chemical and metabolic stability. By contrast, combinatorial optimization methods explore discrete chemical space using techniques such as genetic algorithms (GA) and reinforcement learning (RL). The Reinforced Genetic Algorithm (RGA) [106] integrates GA with RL, using neural networks to guide mutation and crossover operations rather than relying on random selection, while incorporating 3D structural information from both protein targets and ligands. A separate related innovation, the Differentiable Scaffolding Tree (DST) [107], transforms discrete chemical structures into differentiable representations, allowing gradient-based optimization directly on chemical graphs; DST leverages graph neural networks to backpropagate gradients, providing interpretable insights into how structural features relate to molecular properties.

Recent advances have increasingly leveraged diffusion models as a powerful generative framework for molecular design, due to their ability to model complex distributions over molecular structures. In particular, conditional diffusion models have emerged as a dominant paradigm, enabling molecule generation conditioned on protein pocket geometry, binding affinity objectives, or physicochemical properties. For example, structure-based diffusion models such as DiffBP [108] generate molecular structures conditioned on 3D protein binding sites, learning joint distributions over atom types and coordinates in an equivariant framework. Beyond structure-based settings, conditional diffusion has also been applied to other modalities; for instance, DiffMS [109] formulates molecular generation as a graph diffusion process conditioned on mass spectra and chemical formula constraints, enabling structure elucidation from experimental data.

While conditional generation improves control over desired properties, many approaches still assume static protein structures, limiting their ability to capture realistic binding processes. DynamicBind [110] formulates ligand binding as an SE(3)-equivariant diffusion-based docking process that starts from an apo-like (ligand-free) protein structure and an initially placed ligand, and iteratively refines both components toward a ligand-bound holo-like (ligand-bound) complex. At each diffusion step, the model predicts rigid-body transformations (translation and rotation) and torsional updates for ligands, while simultaneously adjusting protein backbone and side-chain conformations, effectively learning an energy landscape that can transition from apo to holo states without explicit molecular dynamics simulation [110].

ColdstartCPI [111] instead operates in a sequence-based setting but introduces induced-fit modeling through joint compound-protein representations. It combines Mol2Vec [112] and ProtTrans [113] embeddings with a Transformer architecture that performs self-attention over concatenated substructure features (e.g., local chemical fragments identified from SMILES) and residue features (e.g., amino-acid embeddings derived from protein sequences), allowing compound and protein representations to be updated jointly. This design explicitly models inter- and intra-molecular interactions, enabling feature adaptation across binding partners and improving generalization when predicting interactions involving previously unseen compounds, proteins, or compound-protein pairs.

Similarly, iNGNN-DTI [114] models drug-target interaction at the graph level by constructing both molecular and protein graphs and encoding them with a k-subgraph GNN extractor, in which each node representation is updated by pooling information from the k-hop subgraph surrounding that node. This allows the model to use k-subgraphs as a way of building richer node and graph representations, thereby capturing higher-order local substructures that standard message passing may miss. To model cross-molecular interactions, iNGNN-DTI then applies a cross-attention-free transformer (cross-AFT) module, where drug-node features and protein-node features are combined through learned query, key, and value transformations with pairwise bias terms, so that each atom or residue can aggregate weighted information from its potential binding partner without standard dot-product attention. In addition, pretrained sequence embeddings from Chemformer [115] and ESM [116] are fused with the graph-derived features to provide complementary global molecular and protein context, yielding more interpretable residue-atom interaction patterns and improving

 doi:10.20944/preprints202606.0091.v1

generalization to unseen drugs, proteins, or drug-target pairs. Collectively, these approaches differ from conventional generative or optimization-based methods by explicitly modeling the coupling between ligand structure and protein response, either through geometric diffusion in 3D space or joint representation learning over molecular and protein features.

Despite these advances, from an algorithmic perspective, many molecule-generation methods still struggle with chemical validity and realistic 3D geometry [117], often producing invalid torsions, distorted rings, or other local structural artifacts. Many approaches—especially GA-style or RL-based search — rely on costly evaluations while exploring chemical space in a largely random-walk fashion, highlighting the need for further integration of multimodal data [104]. Finally, models frequently overfit to training distributions, yielding low-novelty molecules and being constrained by representation limits.

5. Lead Optimization

5.1. Molecule Optimization

Lead optimization focuses on refining drug candidates by improving molecular structure (related to training for molecular generation), assessing toxicity, and modeling pharmacokinetics. This aims to enhance the drug-like properties of existing candidates, a challenging task given the vastness of chemical space. In practice, approaches differ in how they explore and constrain chemical space, as well as the trade-off between molecule diversification.

For example, multi-constraint molecule sampling for molecule optimization (MIMOSA) [118], formulates molecule optimization as a Bayesian sampling problem, employing GNN-based proposals as a Markov Chain Monte Carlo kernel. Using three fundamental substructure operations (i.e., add, replace, and delete), MIMOSA achieved a 49.1% improvement over the second-best method, GA, on the ZINC dataset (Table 1) when optimizing for activity against the tested target and molecular properties. Separately, Copy & Refine (CORE) [119], optimizes molecules by deciding at each step whether to copy an existing substructure from the input molecule or generate a new one. CORE combines this strategy with scaffolding tree generation and adversarial training, where a discriminator network learns to distinguish real from generated molecules. This structure-preserving approach accelerates convergence around viable scaffolds, though it may remain close to the starting series unless diversity pressure is added explicitly. One way this is done is by using Molecule-Level Reward (MOLER) functions to introduce similarity rewards and size deviation penalties using the RL technique, policy-gradient optimization to adjust model parameters by directly increasing the expected reward of its outputs [120].

A critical issue in lead optimization is the activity cliff, where small structural modifications produce disproportionately large changes in potency [121,122]. Such cliffs violate the smooth structure-activity assumptions implicitly made by many generative and QSAR-style models and therefore expose failure modes that are hidden by good average metrics. MoleculeACE [123] systematically benchmarked 24 classical and deep learning models across bioactivity datasets from 30 macromolecular targets and showed that model performance degrades substantially on activity cliff compounds, with errors on these compounds often underestimated when relying solely on global metrics such as root mean square error.

5.2. Toxicity Prediction

Accurate toxicity prediction is essential for reducing drug development failures and ensuring patient safety in clinical trials. Drug toxicity is often organ-specific, affecting systems such as the liver, kidney, and cardiovascular system, which are major causes of drug attrition during drug discovery and adverse clinical outcomes [124,125]. Recent methods have emphasized both 3D geometry-aware modeling and enriched molecular representations. We describe two graph-based methods. First, Cremer et al. [126] demonstrated the use of Equivariant Graph Neural Networks (EGNNs) to predict molecular toxicity. By incorporating 3D structure information while maintaining rotational and translational invariance, EGNN ensures consistent predictions regardless of molecular orientation or position. Their

 doi:10.20944/preprints202606.0091.v1

evaluation across multiple MoleculeNet endpoints showed significant improvements over SMILES-based transformers: 14% on Tox21 (0.789 vs. 0.691), 18% on ToxCast (0.685 vs. 0.578), and 13% on BACE (0.832 vs. 0.739). Second, the Multi-level Fingerprint-based Graph Transformer (MolFPG) [127] integrates diverse molecular fingerprints with graph transformer architectures. MolFPG combined substructure detection (Morgan fingerprints), functional group identification (MACCS keys), and physicochemical property characterization (RDKit features). This feature fusion module results in better performance than other GNN architectures benchmarked to enhance toxicity prediction. However, challenges remain in translating computational predictions to human outcomes. For example, predicting tissue distribution, especially accumulation in vital organs or blood cells, remains difficult [128]. Such limitations hamper extrapolating from experimental models to human physiology.

In addition to single-drug toxicity prediction, drug-drug interactions (DDIs) represent a critical source of adverse effects in clinical settings. DDIs arise when the pharmacokinetic or pharmacodynamic properties of one drug are altered by another, potentially leading to reduced efficacy or increased toxicity [129,130]. SGT-DDI [129] is a multimodal framework that integrates two-dimensional molecular graph representations with three-dimensional geometric features through a Transformer-based architecture. It learns each drug from two complementary views: a graph encoder captures topological patterns from the 2D molecular structure, while a spatial encoder models its 3D conformational geometry. These modality-specific representations are fused through multi-head attention to produce context-aware drug embeddings for predicting both whether a DDI occurs and what type of interaction it represents. A limitation, however, is that the approach depends on the availability and quality of 3D conformational information, which may be uncertain for flexible molecules, and it may still underrepresent biological context beyond molecular structure alone, such as dosage, metabolism, or patient-specific factors.

Bi-SemDRUG [130] is a knowledge graph-based framework that learns drug representations from drug-centric subgraphs extracted from a large-scale biomedical knowledge graph. For each drug, it constructs two complementary subgraph views: a connectivity-based view that preserves local neighborhood structure through random-walk sampling, and an importance-based view that captures globally important biomedical entities through PageRank-based sampling. Bi-SemDRUG has two components: a low-order encoder for neighborhood-based pairwise biomedical relations and a high-order encoder for hypergraph-based multi-entity association patterns. Here, the modeled "interactions" are not restricted to physical drug-target binding, but more broadly denote typed biomedical relations in the knowledge graph and, ultimately, typed drug-drug interaction relations. The resulting representations are aligned and integrated through contrastive learning to generate drug embeddings for multi-type DDI prediction.

While toxicity prediction focuses on assessing the adverse effects of individual compounds, drug combination discovery addresses the complementary challenge of optimizing therapeutic efficacy through synergistic interactions. In this context, machine learning approaches have been widely adopted to predict drug synergy using large-scale combination screening data and computational models [131]. For example, DeepSynergy [132] employs a feedforward neural network that takes concatenated multi-modal features, which include chemical descriptors of drug pairs and gene expression profiles of cell lines, as input to predict drug synergy scores. Another method, TargetScore, uses a prior knowledge-based network analysis using phosphoproteomics measurements [133]. Kong et al. have recently published a more complete review of over 100 computational methods examining drug synergy [134].

5.3. Pharmacokinetic Modeling

Pharmacokinetic (PK) and Pharmacodynamic (PD) modeling play a crucial role in predicting how drugs are absorbed, distributed, metabolized, and eliminated over time. Traditional PK models rely on mathematical equations to describe the absorption, distribution, metabolism, and excretion (ADME) of compounds (often differential equations) [135]. Recent advances have introduced neural network-based approaches as alternatives. Lu et al. [136] applied neural ordinary differential equations (Neural-ODEs) [137] to PK modeling, demonstrating their ability to predict dosing regimens for

 doi:10.20944/preprints202606.0091.v1

trastuzumab emtansine (T-DM1), a conjugated monoclonal antibody used to treat breast cancer. Their Neural-PK model accurately predicted complete drug concentration profiles using only early PK data from the first 21-day dosing cycle, achieving an R^2 of 0.98 and a correlation coefficient of 0.99 on 133 test patients. In follow-up work, Neural-PK/PD [138] extended this type of approach to model drug effects and patient responses over time. The Neural-PK/PD architecture uses three encoders, PK, PD, and initial-condition encoders whose outputs condition a Neural-ODE module with separate PK and PD vector fields. The ODE submodule (implemented as a recurrent network) integrates these fields over time to produce PK and PD trajectories, while a dedicated dose input drives the PK dynamics. The authors of these studies caution that DL models trained on small sample sizes may be prone to overfitting, and suggest that pre-training using PK data could improve model performance. One such resource is the Pharmacokinetics Database (PK-DB) [37], which compiles PK data from clinical trials.

Such PK/PD models and related data can be of use to the growing research area of digital twins. Digital twins are virtual representations of biological systems that enable in silico simulations across scales, from individual cells to entire humans [139]. One recent use of PK-DB data sought improved diabetes management through such a digital twin study that developed a whole-body physiologically-based pharmacokinetic model [140]. There have been other reports on how clinical trial outcomes improve through the use of digital twins [141].

6. Clinical Trials

This section reviews applications of artificial intelligence (AI) in clinical trials, with a focus on patient outcome prediction, trial site selection, patient-trial matching, and trial similarity search, as well as the current state of LLM integration.

Before analyzing specific AI applications in clinical trials, we first characterized the overall landscape of trials employing ML/AI-related techniques, specifically, deep learning (DL), machine learning (ML), artificial intelligence (AI), and large language models (LLMs).

We retrieved a total of 520,050 clinical trials from ClinicalTrials.gov (as of December 2024) and performed separate keyword-based searches using the following exact keywords: “artificial intelligence”, “machine learning”, “deep learning”, and “large language model”. Each query targeted both the title and condition fields, such that any trial containing the corresponding keyword in these fields was identified. The retrieved dataset included the NCT Number, Study Title, Study Status, Conditions, and Phases.

To refine the dataset, we excluded studies with the following statuses: UNKNOWN, WITHDRAWN, WITHHELD, NOT YET RECRUITING, and NO LONGER AVAILABLE. These exclusion criteria were applied to remove trials without sufficient status information, as such entries may reflect incomplete reporting, which could introduce noise or bias in downstream analysis. After this preprocessing step, 408,530 trials remained. Among them, 3,746 trials (0.92%) contained at least one ML/AI-related term. Specifically, we identified 1,338 AI-related trials, 1,430 ML-related trials, 1,491 DL-related trials, and 96 LLM-related trials.

Figure 3 (a) presents a Venn diagram illustrating the overlap among these categories. The majority of studies involve ML and DL, together accounting for over 2,500 trials, followed by AI (over 1,000 trials) and LLMs (fewer than 100 trials). We further examined the clinical phases of the 3,746 AI-related trials identified through our keyword-based search described above. Among these, 275 trials had reported clinical trial phases. Because the extraction was based on exact keyword matches in the title and condition fields, the resulting phase counts should be interpreted as a broad indicator of AI-related trial records rather than a definitive estimate of AI methods as the primary study intervention. Clinical trial phases generally progress from initial safety assessment (Phase 1) to efficacy evaluation (Phase 2), large-scale confirmatory studies (Phase 3), and post-marketing real-world surveillance (Phase 4). As shown in Figure 3 (b), the largest proportion of studies is in Phase 2, while some DL-based studies have progressed to Phases 3 and 4. These Phase 4 records should be interpreted with caution, as keyword-based extraction may capture studies where deep learning is mentioned in a supportive or analytical context rather than as the primary intervention. For example, only a small subset of Phase 4 records explicitly describe deep

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2026 doi:10.20944/preprints202606.0091.v1

learning-based prediction models in the official ClinicalTrials.gov record, such as NCT05357326 (myopia intervention) and NCT04685642 (mood disorder treatment). Interestingly, most LLM-related trials are concentrated in Phase 2, followed by Phase 4. Consistent with the temporal trend in Figure 3(c), the number of AI-related trials has increased markedly in recent years.

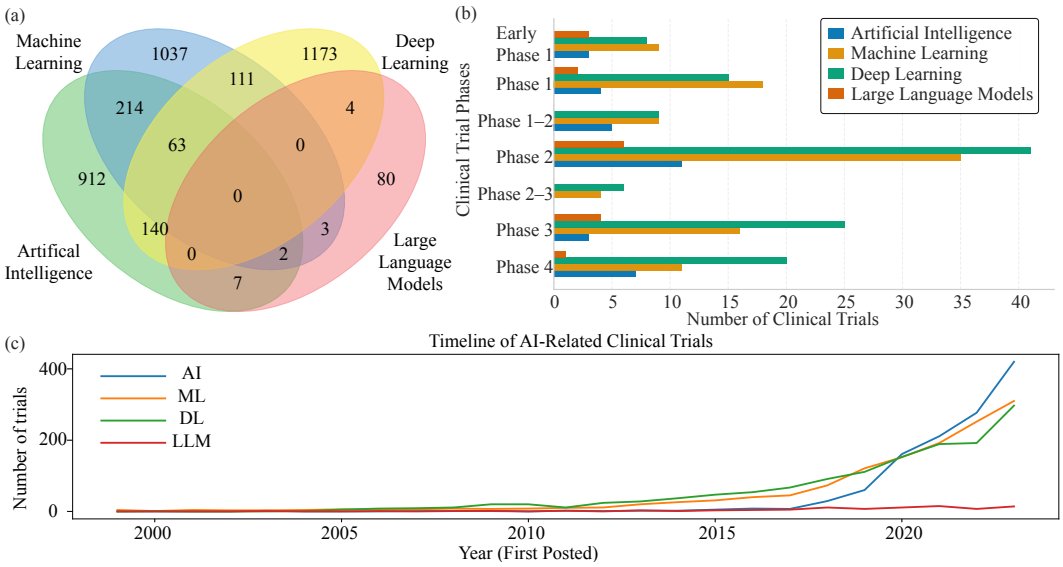

Figure 3. (a) Distribution and overlap of 3,746 AI-related clinical trials. Deep Learning (DL) and Machine Learning (ML) account for the largest numbers of unique trials, 1,173 and 1,037, respectively, followed by Artificial Intelligence (AI) with 912 trials. Large Language Models (LLMs) are referenced in a total of 96 trials, including 80 uniquely associated with LLMs. (b) Phase-wise distribution of AI technologies across clinical trial stages. The peak activity is observed in Phase 2, particularly for Deep Learning (~40 trials) and Machine Learning (~35 trials), with limited LLM-related trials observed across all phases. (c) Temporal trends of AI-related clinical trials based on the ClinicalTrials.gov "First Posted" date. Categories are not mutually exclusive.

6.1. Trial Site Selection

Trial site selection evaluates factors such as geographic location and costs, and selected sites must enroll participants while ensuring compliance with regulatory and study requirements. Since patients with varying demographics may respond differently, trials should reflect the overall population.

In this context, PG-entropy [142] is a 2-layer neural network that uses a learning-to-rank approach (i.e., training models to optimize the ordering of items in a ranking task) with a Plackett-Luce probabilistic ranking policy, and incorporates an entropy-based fairness reward inside the policy-learning objective to select a Top-K set of sites whose combined patient demographics are more evenly distributed across groups. Separately, the FRAMM (Fair Ranking with Missing Modalities) [143], algorithm uses a Q-value-style scoring network while not explicitly being a deep Q-value network (DQN). FRAMM scores each site based on its contribution to a reward that combines enrollment utility and fairness/diversity, and it trains this using the REINFORCE policy gradient algorithm, which utilizes Monte Carlo sampling to maximize the expected reward over Top-K rankings. The use of policy-gradient optimization instead of DQN aids in this one-shot ranking task. In both methods, the REINFORCE policy-gradient method is used to provide a way to compute the gradient without differentiating through the discrete choice of site selection.

While traditional site selection relies on historical data, modern approaches should consider temporal changes in population demographics and healthcare accessibility. Therefore, future developments should focus on adaptive modeling that responds to changing population socioeconomic characteristics and healthcare access patterns [144–146].

 doi:10.20944/preprints202606.0091.v1

6.2. Trial Outcome Prediction

Trial-outcome prediction estimates the likelihood of success based on disease characteristics, the intervention, and elements of study design. Choosing an effective model in practice requires addressing three key challenges: complex relationships among trial components, heterogeneous and partially missing modalities, and limited data availability. Here, we describe two methods, HINT and LIFTED, that tackle outcome prediction for trials.

HINT is a hierarchical interaction graph over drugs, diseases, and protocols, then utilizes a dynamic attentive GNN for prediction. HINT is tested against a benchmark dataset of 17,538 clinical trials, comprising 13,880 small-molecule drugs and 5,335 diseases [147], where each key data modality is embedded by a separate embedding model. For example, the Graph-based Attention Model (GRAM) model was used for disease codes mapped to a hierarchy, and the Clinical-BERT (Bidirectional Encoder Representations from Transformers) was used to embed sentences of eligibility criteria. Separately, the reliability of HINT was further improved by combining HINT with a calibrated reject/abstain layer for uncertainty-aware selective classification, which abstains on low-confidence cases [148].

However, this strategy can be vulnerable when relational fields are sparse or noisy, in which case a language-mediated approach can be more robust, as argued by the authors of Multi-Modal Mix-of-Experts for Outcome Prediction (LIFTED) [149]. To achieve this, LIFTED normalizes inputs using a unified transformer-based encoder to extract representations from these modal-specific language descriptions, thereby avoiding the need for a modal-specific encoder (such as with HINT), which may limit method use with new data modalities. LIFTED learns cross-modal patterns using noise-resilient encoders (augmentation plus consistency loss) and sparse gating (sparse Mixture-of-Experts with noisy Top-K routing). These choices improve robustness to data noise and, by unifying heterogeneous fields into natural-language descriptions with a shared encoder, may also mitigate the practical effects of schema changes and data missingness.

6.3. Patient-Trial Matching

Patient-trial matching connects individuals to appropriate clinical trials by aligning their medical records with eligibility criteria [150].

Owing to LLMs' ability to understand text, LLMs can be employed to aid in patient selection. For example, den Hamer et al. studied the use of LLM-assisted pre-screening that used prompts combining one-shot, selection-inference, and chain-of-thought techniques to decide which eligibility criteria can be checked from a patient summary and evaluate those criteria [151]. Using 10 synthetic patient profiles, each representing a separate cancer, the authors identified 146 relevant trials for the disease and 4,135 criteria. Criterion is screenable if the profiles contained sufficient information to evaluate the criterion (e.g., gene alteration presence would be screenable if in the summary). The LLM correctly identified screenability for 72% (2,994/4,135) of the criteria and answered 72% (341/471) of the screenable criteria correctly. Trial-level recall was 0.5 in the purely automated setting. Before a trial was excluded entirely, a physician examined each of the flagged trial dropout criteria. The precision reached 1.0, with a precision of 0.71. This workflow reduced the number of criteria a physician needed to check by ~90%, to <10% (328/4,135).

Another related study showed similar improvements. TrialGPT implemented an LLM-based pipeline with retrieval, matching, and ranking modules, evaluated on the SIGIR 2016 and TREC 2021/2022 oncology cohorts [152]. In retrieval, GPT-generated keywords were issued to a hybrid BM25 plus MedCPT (lexical and semantic, respectively) retriever and fused via reciprocal rank fusion; using GPT-4 keywords, this reached 86.2% recall at top-500. In criterion-level matching, TrialGPT generated rationales, located supporting sentences in the patient note, and assigned eligibility labels; three domain experts rated explanations 87.8% correct across 105 patient-trial pairs (1,015 criteria). For trial-level ranking, aggregating criterion predictions outperformed baselines (e.g., BioLinkBERT). In a clinician pilot study with 36 trial pairs in a crossover design (18 with vs 18 without per annotator),

 doi:10.20944/preprints202606.0091.v1

there was a non-significant trend to higher annotation accuracy with TrialGPT (97.2% vs 91.7%) and a 42.6% mean reduction in screening time.

Despite reported gains, current LLM-based trial-matching systems still struggle with unclear or ambiguous eligibility criteria, so a physician-in-the-loop remains necessary to curb misclassification [151]. Furthermore, real-world matching often depends on longitudinal notes, lab values, and multimodal data that the above evaluations did not cover [152].

6.4. Challenges of LLMs in Clinical Tasks

Despite their strong performance in tasks such as patient-trial matching and eligibility screening, LLMs face several challenges that limit their safe and effective deployment in clinical settings [153,154]. Hallucinations remain a critical issue [155–157]. In some yet unclear instances, LLMs may generate plausible but incorrect interpretations of patient records or eligibility criteria, potentially leading to incorrect inclusion or exclusion decisions. In settings such as clinical trial recruitment, these errors can compromise patient safety; such issues necessitate continued human oversight and verification. Second, ethical and privacy concerns arise when applying LLMs to sensitive clinical data [153]. Patient records regularly contain legally protected health information (PHI). The use of LLMs raises questions about data security, consent, and compliance with healthcare data protection regulations such as the Health Insurance Portability and Accountability Act Liability (HIPAA) law in the United States [158,159].

Integration with electronic health record (EHR) systems is another and more practical challenge [160,161]. Clinical data systems are heterogeneous, incomplete, and often stored in unstructured formats (e.g., physician notes), making the reliable extraction and standardization of data difficult. Additionally, extensive real-world deployment would require interoperability with various hospital information systems, which work that typically lies outside the scope of research. Lastly, these systems may inherit biases from training data [162], potentially leading to systematic exclusion or misclassification of underrepresented patient groups (e.g., 94.6% of the UKBioBank participants are classified as "white") [9]. Addressing this issue requires careful dataset curation, bias-aware model design, and human-in-the-loop validation.

Separately, the emerging paradigm of using LLMs as evaluators (i.e., “LLM-as-a-judge”) has gained attention for tasks such as evidence assessment and decision support [163]. In this scenario, LLMs are being used to evaluate the output of other AI systems. In part, this is driven by the issues in the availability of expert human reviewers to evaluate the large number of methodologies being created, as well as to address variability that can exist between human reviewers [164]. However, the reliability of such approaches is imperfect [165]. Additionally, LLM-based judgments can be sensitive to prompt design and may lack consistency and domain-specific grounding [166], particularly when evaluating complex content with clinical terms and jargon. As a result, LLM use in life-and-death decision-making requires careful validation and should be complemented with expert oversight.

7. Discussion

We provide an overview of AI applications in drug discovery, highlighting state-of-the-art methods and key datasets, all contextualized within the stages of drug development.

The drug discovery landscape is dominated by two complementary AI technologies: graph neural networks (GNNs) and large language models (LLMs). GNNs are particularly well-suited for chemical structures, naturally representing atoms and bonds as graphs and excelling at capturing relational biomedical data. LLMs, on the other hand, have shown strong performance in processing complex textual data, such as clinical trial reports, and generating sophisticated embeddings from biological data, as demonstrated in recent advances in gene expression analysis [167]. Together, these technologies are driving a paradigm shift in biological data interpretation.

Despite the availability of numerous high-quality datasets (Table 1), significant challenges remain. These include ensuring data quality, meeting the high computational demands of state-of-the-art models, and improving model interpretability. For example, models like AlphaFold require over 100

 doi:10.20944/preprints202606.0091.v1

GPUs to process approximately 170,000 protein structures, highlighting accessibility and scalability concerns in research settings.

7.1. Challenges and Future Directions

Across the drug discovery pipeline, the primary limitation of current AI approaches lies not in model architecture, but in the mismatch between how models are trained and how they are ultimately used. Most models are developed on biased and heterogeneous datasets, while real-world applications require generalization across diverse and clinically constrained settings.

A central issue is data bias, whereby existing datasets disproportionately represent well-studied targets and diseases, such as kinases and breast cancer, leading models to learn patterns driven by data availability rather than underlying biology [14,128]. This bias limits applicability to less-studied targets and rare diseases, where data remain scarce. As a result, model capabilities and performance often reflect areas of historical research focus.

Closely related is the problem of distribution mismatch. Drug discovery data combine heterogeneous measurements (e.g., Resazurin/Syto60 vs. CellTiter-Glo [15]) and modalities (e.g., RNA-seq vs. microarray). Similarly, models trained on cell-line data are sometimes applied to patient populations, introducing substantial distribution shift due to the data heterogeneity differences [14,85,90]. As a result, models trained and evaluated under standard random train-test splits primarily capture interpolation within known data distributions, while failing to generalize to unseen targets, chemical scaffolds, or patient cohorts [36,168,169]. Consequently, robust performance across heterogeneous settings remains difficult to achieve without improved data harmonization and more rigorous evaluation on out-of-distribution data (e.g., training on large-scale pan-cancer datasets with testing on equivalent rare cancer data) [170–172].

Interpretability remains another key bottleneck for deployment. While recent models have improved in providing post hoc explanations [173,174], they rarely offer mechanistic insights or calibrated uncertainty at a level required for clinical or biological decision-making [173,175]. Clinicians require explanations that are both biologically grounded and reliable, yet such capabilities remain limited [175]. Separately, many clinically relevant scenarios, including rare diseases and low-data settings, remain largely unaddressed [176,177]. Prior knowledge in the form of curated databases and text content from scientific publications is one mitigation strategy [13,133,178,179].

Large language models (LLMs) built on abundant text collections introduce a complementary opportunity by enabling access to information hidden in scientific literature and improved accessibility for clinicians through natural language interfaces [156]. However, their deployment in clinical workflows is constrained by fundamental reliability issues, including hallucination, sensitivity to prompt variation, and lack of verifiable reasoning [180–182]. These limitations can lead to incorrect or overconfident decisions. As a result, LLMs are currently best positioned as clinician-in-the-loop tools rather than autonomous systems [154].

Combined, these challenges indicate that many of the bottlenecks in AI-driven drug discovery are structural rather than algorithmic. Progress will depend both on developing new architectures as well as addressing evident challenges in data bias, cross-domain generalization, and model architecture that conforms to biological and clinical constraints.

Funding: This research was supported in part by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health (NIH) (ZIALM240126). Tianfan Fu is supported by Young Scientists Fund (C Class) of the National Natural Science Foundation of China (Grant No. 62506154), the Fundamental Research Funds for the Central Universities and Nanjing University International Collaboration Initiative (Grant No. 020214380129) and the "111 Center"(No. B26023).

Author Contributions: Y.I., N.H., Y.L., T.F., and A.L. contributed to conceptualization, methodology, and literature synthesis. Y.I. performed the clinical trial analysis. A.L. and T.F. supervised the project. Y.I., Y.L., T.F., and A.L. drafted the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

 doi:10.20944/preprints202606.0091.v1

Conflicts of Interest: The authors declare no competing interests.

Use of Artificial Intelligence: During the preparation of this work, the authors used ChatGPT (GPT-5.3, OpenAI) and Gemini (Google) for manuscript structure verification and stylistic refinement. These tools were used solely to assist with language and organization. All outputs were carefully reviewed and edited by the authors, who take full responsibility for the content of this publication.

References

1. Mohs, R.C.; Greig, N.H. Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **2017**, *3*, 651–657.
2. Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C.W.; Xiao, C.; Sun, J.; Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nature chemical biology* **2022**, *18*, 1033–1036.
3. Serrano, D.R.; Luciano, F.C.; Anaya, B.J.; Ongoren, B.; Kara, A.; Molina, G.; Ramirez, B.I.; Sánchez-Guirales, S.A.; Simon, J.A.; Tomietto, G.; et al. Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine. *Pharmaceutics* **2024**, *16*, 1328.
4. Kanakia, A.; Sale, M.; Zhao, L.; Zhou, Z. AI In Action: Redefining Drug Discovery and Development. *Clinical and Translational Science* **2025**, *18*, e70149.
5. Wagner, J.; Dahlem, A.M.; Hudson, L.D.; Terry, S.F.; Altman, R.B.; Gilliland, C.T.; DeFeo, C.; Austin, C.P. A dynamic map for learning, communicating, navigating and improving therapeutic development. *Nature reviews Drug discovery* **2018**, *17*, 150–150.
6. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **2015**, *12*, e1001779.
7. Dolezalova, N.; Cairo, M.; Despotovic, A.; Booth, A.T.; Reed, A.B.; Morelli, D.; Plans, D. Development of a dynamic type 2 diabetes risk prediction tool: a uk biobank study. *arXiv preprint arXiv:2104.10108* **2021**.
8. Dabbah, M.A.; Reed, A.B.; Booth, A.T.; Yassaee, A.; Despotovic, A.; Klasmer, B.; Binning, E.; Aral, M.; Plans, D.; Morelli, D.; et al. Machine learning approach to dynamic risk modeling of mortality in COVID-19: A UK Biobank study. *Scientific reports* **2021**, *11*, 16936.
9. Fry, A.; Littlejohns, T.J.; Sudlow, C.; Doherty, N.; Adamska, L.; Sprosen, T.; Collins, R.; Allen, N.E. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **2017**, *186*, 1026–1034. https://doi.org/10.1093/aje/kwx246.
10. Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M.L.; Burdett, T.; Burgin, J.; Caballero-Pérez, J.; Cochrane, G.; Colwell, L.J.; et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **2023**, *51*, D753–D759.
11. McLaren, M.R.; Willis, A.D.; Callahan, B.J. Consistent and correctable bias in metagenomic sequencing experiments. *elife* **2019**, *8*, e46923.
12. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* **2006**, *6*, 813–823.
13. Inoue, Y.; Lee, H.; Fu, T.; Kuang, R.; Luna, A. drGT: Attention-Guided Gene Assessment of Drug Response Utilizing a Drug-Cell-Gene Heterogeneous Network, 2026, [arXiv:cs.LG/2405.08979].
14. Safikhani, Z.; El-Hachem, N.; Quevedo, R.; Smirnov, P.; Goldenberg, A.; Birkbak, N.J.; Mason, C.; Hatzis, C.; Shi, L.; Aerts, H.J.; et al. Assessment of pharmacogenomic agreement. *F1000Research* **2016**, *5*, 825.
15. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **2012**, *41*, D955–D961.
16. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G.V.; Lo, C.C.; McDonald III, E.R.; Barretina, J.; Gelfand, E.T.; Bielski, C.M.; Li, H.; et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **2019**, *569*, 503–508.
17. Reinhold, W.C.; Wilson, K.; Elloumi, F.; Bradwell, K.R.; Ceribelli, M.; Varma, S.; Wang, Y.; Duveau, D.; Menon, N.; Trepel, J.; et al. CellMinerCDB: NCATS Is a Web-Based Portal Integrating Public Cancer Cell Line Databases for Pharmacogenomic Explorations. *Cancer research* **2023**, *83*, 1941–1952.
18. Inoue, Y.; Song, T.; Fu, T. DrugAgent: Explainable Drug Repurposing Agent with Large Language Model-based Reasoning. *arXiv preprint arXiv:2408.13378* **2024**.

 doi:10.20944/preprints202606.0091.v1

19. Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C.W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* **2021**.
20. Davis, M.I.; Hunt, J.P.; Herrgard, S.; Ciceri, P.; Wodicka, L.M.; Pallares, G.; Hocker, M.; Treiber, D.K.; Zarrinkar, P.P. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology* **2011**, *29*, 1046–1051.
21. Huang, K.; Fu, T.; Glass, L.M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **2020**, *36*, 5545–5547.
22. Ahmad, B.; Ouahada, K.; Hamam, H. Machine learning for drug-target interaction prediction: A comprehensive review of models, challenges, and computational strategies. *Computational and Structural Biotechnology Journal* **2026**.
23. Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling* **2014**, *54*, 735–743.
24. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **2007**, *35*, D198–D201.
25. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* **2019**, *47*, D930–D940.
26. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J.K.; Ceulemans, H.; Clevert, D.A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science* **2018**, *9*, 5441–5451.
27. Dao, Q.T.; Do, T.M.D.; Thai, Q.M.; Tran, P.T.; Ngo, S.T.; Nguyen, T.H. Identifying Potential BACE1 Inhibitors from the ChEMBL Database Using Machine Learning and Atomistic Simulation Approaches. *ACS Omega* **2026**.
28. Mukherjee, S.; Ghosh, M.; Basuchowdhuri, P. Deep Graph Convolutional Network and LSTM based approach for predicting drug-target binding affinity. *arXiv preprint arXiv:2201.06872* **2022**.
29. Kretschmer, F.; Seipp, J.; Ludwig, M.; Klau, G.W.; Böcker, S. Coverage bias in small molecule machine learning. *Nature communications* **2025**, *16*, 554.
30. Son, H.; Lee, S.; Kim, J.; Park, H.; Hwang, M.H.; Yi, G.S. BASE: A web service for providing compound-protein binding affinity prediction datasets with reduced similarity bias. *BMC bioinformatics* **2024**, *25*, 340.
31. Tanoli, Z.; Alam, Z.; Vähä-Koskela, M.; Ravikumar, B.; Malyutina, A.; Jaiswal, A.; Tang, J.; Wennerberg, K.; Aittokallio, T. Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. *Database* **2018**, *2018*, bay083.
32. Buniello, A.; Suveges, D.; Cruz-Castillo, C.; Llinares, M.B.; Cornu, H.; Lopez, I.; Tsukanov, K.; Roldán-Romero, J.M.; Mehta, C.; Fumis, L.; et al. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic acids research* **2025**, *53*, D1467–D1475.
33. Han, Y.; Klinger, K.; Rajpal, D.K.; Zhu, C.; Teeple, E. Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform. *BMC bioinformatics* **2022**, *23*, 232.
34. Sterling, T.; Irwin, J.J. ZINC 15–ligand discovery for everyone. *Journal of chemical information and modeling* **2015**, *55*, 2324–2337.
35. Liu, Y.; Kashima, H. Chemical property prediction under experimental biases. *Scientific Reports* **2022**, *12*, 8206.
36. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; Gomes, J.; Geniesse, C.; Pappu, A.S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.
37. Grzegorzewski, J.; Brandhorst, J.; Green, K.; Eleftheriadou, D.; Duport, Y.; Barthorscht, F.; Köller, A.; Ke, D.Y.J.; De Angelis, S.; König, M. PK-DB: pharmacokinetics database for individualized and stratified computational modeling. *Nucleic acids research* **2021**, *49*, D1358–D1364.
38. Chou, W.C.; Lin, Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences* **2023**, *191*, 1–14.
39. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.
40. Ballester, P.J.; Mitchell, J.B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169–1175.

 doi:10.20944/preprints202606.0091.v1

41. Derry, A.; Carpenter, K.A.; Altman, R.B. Training data composition affects performance of protein structure analysis algorithms. In Proceedings of the PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022. World Scientific, 2021, pp. 10–21.
42. Wang, R.; Fang, X.; Lu, Y.; Yang, C.Y.; Wang, S. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry* **2005**, *48*, 4111–4119.
43. Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M.J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **2017**, *45*, D170–D176.
44. Orlando, G.; Raimondi, D.; Vranken, W. Observation selection bias in contact prediction and its implications for structural bioinformatics. *Scientific Reports* **2016**, *6*, 36679.
45. Suzek, B.E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C.H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **2007**, *23*, 1282–1288.
46. Zarin, D.A.; Tse, T.; Williams, R.J.; Califf, R.M.; Ide, N.C. The ClinicalTrials. gov results database—update and key issues. *New England Journal of Medicine* **2011**, *364*, 852–860.
47. Lambert, J. Statistics in Brief: How to Assess Bias in Clinical Studies? *Clinical Orthopaedics and Related Research* **2011**, *469*, 1794–1796. https://doi.org/10.1007/s11999-010-1538-7.
48. Criscuolo, C.; Dolci, T.; Salnitri, M. Towards assessing data bias in clinical trials. In Proceedings of the VLDB Workshop on Data Management and Analytics for Medicine and Healthcare. Springer, 2022, pp. 57–74.
49. Morris, J.A.; Caragine, C.; Daniloski, Z.; Domingo, J.; Barry, T.; Lu, L.; Davis, K.; Ziosi, M.; Glinos, D.A.; Hao, S.; et al. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **2023**, *380*, eadh7699.
50. You, Y.; Lai, X.; Pan, Y.; Zheng, H.; Vera, J.; Liu, S.; Deng, S.; Zhang, L. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduction and Targeted Therapy* **2022**, *7*, 156.
51. Mitra, S.; Malik, R.; Wong, W.; Rahman, A.; Hartemink, A.J.; Pritykin, Y.; Dey, K.K.; Leslie, C.S. Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature genetics* **2024**, *56*, 627–636.
52. Zhou, J.; Theesfeld, C.L.; Yao, K.; Chen, K.M.; Wong, A.K.; Troyanskaya, O.G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics* **2018**, *50*, 1171–1179.
53. Hu, F.; Hu, Y.; Zhang, W.; Huang, H.; Pan, Y.; Yin, P. A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. *Advanced Science* **2023**, *10*, 2301223.
54. Ma, W.; Bi, X.; Jiang, H.; Wei, Z.; Zhang, S. Annotating protein functions via fusing multiple biological modalities. *Communications Biology* **2024**, *7*, 1705.
55. Mitrofanov, A.; Alkhnbashi, O.S.; Shmakov, S.A.; Makarova, K.S.; Koonin, E.V.; Backofen, R. CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic acids research* **2021**, *49*, e20–e20.
56. Cheng, X.; Li, Z.; Shan, R.; Li, Z.; Wang, S.; Zhao, W.; Zhang, H.; Chao, L.; Peng, J.; Fei, T.; et al. Modeling CRISPR-Cas13d on-target and off-target effects using machine learning approaches. *Nature communications* **2023**, *14*, 752.
57. Van de Sande, B.; Lee, J.S.; Mutasa-Gottgens, E.; Naughton, B.; Bacon, W.; Manning, J.; Wang, Y.; Pollard, J.; Mendez, M.; Hill, J.; et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery* **2023**, *22*, 496–520.
58. Kharchenko, P.V.; Silberstein, L.; Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **2014**, *11*, 740–742.
59. Hou, W.; Ji, Z.; Ji, H.; Hicks, S.C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome biology* **2020**, *21*, 218.
60. Inoue, Y. scVGAE: A Novel Approach using ZINB-Based Variational Graph Autoencoder for Single-Cell RNA-Seq Imputation. *arXiv preprint arXiv:2403.08959* **2024**.
61. Inoue, Y.; Kulman, E.; Kuang, R. BiGCN: Leveraging Cell and Gene Similarities for Single-cell Transcriptome Imputation with Bi-Graph Convolutional Networks. *bioRxiv* **2024**, pp. 2024–04.
62. Cheng, Y.; Ma, X.; Yuan, L.; Sun, Z.; Wang, P. Evaluating imputation methods for single-cell RNA-seq data. *BMC bioinformatics* **2023**, *24*, 302.
63. Marks, D.S.; Colwell, L.J.; Sheridan, R.; Hopf, T.A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PloS one* **2011**, *6*, e28766.
64. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.

 doi:10.20944/preprints202606.0091.v1

65. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* **2019**, *87*, 1011–1020.
66. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
67. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.
68. Steinegger, M.; Mirdita, M.; Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods* **2019**, *16*, 603–606.
69. Xu, L.; Ru, X.; Song, R. Application of machine learning for drug–target interaction prediction. *Frontiers in genetics* **2021**, *12*, 680117.
70. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *Journal of chemical information and modeling* **2018**, *59*, 895–913.
71. Li, J.; Jiang, H.; Ma, W.; Bi, X.; Chen, R.; Lu, W.; Cai, Q.; Yang, F.; Wei, Z.; Zhang, S. Geometric deep learning for protein-ligand affinity prediction with hybrid message passing strategies. *IEEE Journal of Biomedical and Health Informatics* **2025**.
72. Bi, X.; Zhang, S.; Ma, W.; Jiang, H.; Wei, Z. HiSIF-DTA: a hierarchical semantic information fusion framework for drug-target affinity prediction. *IEEE journal of biomedical and health informatics* **2023**, *29*, 1579–1590.
73. Huang, X.; Bi, X.; Xing, N.; Ma, W.; Jiang, H.; Cai, Q.; Lu, W.; Yang, F.; Wei, Z.; Zhang, S. LightDTA: lightweight drug-target affinity prediction via random-walk network embedding and knowledge distillation. *Molecular Diversity* **2026**, pp. 1–24.
74. Li, J.; Bi, X.; Ma, W.; Jiang, H.; Liu, S.; Lu, Y.; Wei, Z.; Zhang, S. MHAN-DTA: a multiscale hybrid attention network for drug-target affinity prediction. *IEEE Journal of Biomedical and Health Informatics* **2024**, *29*, 7910–7921.
75. Gider, V.; Budak, C. A physics-informed graph neural network to approximate docking-based binding affinity for DYRK2 in Alzheimer's drug repurposing. *Scientific Reports* **2026**.
76. Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **2020**, *17*, 184–192.
77. Xu, W.; Xie, Y.; Jiang, H.; Fu, Y.; Liu, H.; Wei, Z.; Zhang, S. Geometry-Aware Protein–Protein Binding Site Prediction Using Geometric Learning and Pretraining Strategies. *Journal of Chemical Information and Modeling* **2025**, *65*, 12088–12098.
78. Tubiana, J.; Schneidman-Duhovny, D.; Wolfson, H.J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods* **2022**, *19*, 730–739.
79. Stärk, H.; Ganea, O.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In Proceedings of the International conference on machine learning. PMLR, 2022, pp. 20503–20521.
80. Luna, A.; Elloumi, F.; Varma, S.; Wang, Y.; Rajapakse, V.N.; Aladjem, M.I.; Robert, J.; Sander, C.; Pommier, Y.; Reinhold, W.C. CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics. *Nucleic acids research* **2021**, *49*, D1083–D1093.
81. Adam, G.; Rampášek, L.; Safikhani, Z.; Smirnov, P.; Haibe-Kains, B.; Goldenberg, A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology* **2020**, *4*, 19.
82. Peng, W.; Chen, T.; Dai, W. Predicting drug response based on multi-omics fusion and graph convolution. *IEEE Journal of Biomedical and Health Informatics* **2021**, *26*, 1384–1393.
83. Liu, X.; Song, C.; Huang, F.; Fu, H.; Xiao, W.; Zhang, W. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Briefings in Bioinformatics* **2022**, *23*, bbab457.
84. Li, X.; Das, T.; Bhattarai, K.; Rajaganapathy, S.; Buchner, V.C.; Wang, Y.; Su, C.; Sun, L.; Wang, L.; Cerhan, J.R.; et al. Leveraging multi-source data to resolve inconsistency across pharmacogenomic datasets in drug sensitivity prediction. In Proceedings of the AMIA Annual Symposium Proceedings, 2025, Vol. 2024, p. 744.
85. Rahman, R.; Dhruba, S.R.; Matlock, K.; De-Niz, C.; Ghosh, S.; Pal, R. Evaluating the consistency of large-scale pharmacogenomic studies. *Briefings in Bioinformatics* **2019**, *20*, 1734–1753.
86. Elloumi, F.; Reinhold, W.C.; Varma, S.; Wang, Y.; Kinali, M.; Arakawa, Y.; Inoue, Y.; Aladjem, M.I.; Pommier, Y.; Luna, A. CellMiner cross-database (CellMinerCDB) version 2.2 for explorations of patient-derived cancer cell line pharmacogenomics. *Nucleic Acids Research* **2026**, *54*, D1345–D1354.
87. Dost, K.; Pullar-Strecker, Z.; Brydon, L.; Zhang, K.; Hafner, J.; Riddle, P.J.; Wicker, J.S. Combatting over-specialization bias in growing chemical databases. *Journal of Cheminformatics* **2023**, *15*, 53.

 doi:10.20944/preprints202606.0091.v1

88. Nguyen, T.M.; Nguyen, T.; Tran, T. Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring. *Briefings in Bioinformatics* **2022**, *23*, bbac269.
89. Rajapakse, V.N.; Luna, A.; Yamade, M.; Loman, L.; Varma, S.; Sunshine, M.; Iorio, F.; Sousa, F.G.; Elloumi, F.; Aladjem, M.I.; et al. CellMinerCDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. *Iscience* **2018**, *10*, 247–264.
90. He, D.; Liu, Q.; Wu, Y.; Xie, L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence* **2022**, *4*, 879–892.
91. Jayagopal, A.; Zhang, Y.; Walsh, R.J.; Tan, T.Z.; Jeyasekharan, A.D.; Rajan, V. GANDALF: Generative AttentioN based Data Augmentation and predictive modeLing Framework for personalized cancer treatment. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
92. Peng, W.; Liu, H.; Dai, W.; Yu, N.; Wang, J. Predicting cancer drug response using parallel heterogeneous graph convolutional networks with neighborhood interactions. *Bioinformatics* **2022**, *38*, 4546–4553.
93. Bik, E.M. Publishing negative results is good for science, 2024.
94. Briere, G.; Stosskopf, T.; Loire, B.; Baudot, A. Benchmarking the Impact of Data Leakage on the Performance of Knowledge Graph Embedding Models for Biomedical Link Prediction. *bioRxiv* **2025**, pp. 2025–01.
95. Asiaee, A.; Strauch, J.; Azinfar, L.; Pal, S.; Pua, H.H.; Long, J.P.; Coombes, K.R. Widespread data leakage inflates performance estimates in cancer drug response prediction. *bioRxiv* **2026**, pp. 2026–02.
96. Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J.P. Activity, assay and target data curation and quality in the ChEMBL database. *Journal of computer-aided molecular design* **2015**, *29*, 885–896.
97. Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public K i data. *Journal of medicinal chemistry* **2012**, *55*, 5165–5173.
98. Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E.J.; Blackshaw, J.; Corbett, S.; De Veij, M.; Ioannidis, H.; Lopez, D.M.; Mosquera, J.F.; et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research* **2024**, *52*, D1180–D1192.
99. Collins, G.S.; Moons, K.G.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; Van Smeden, M.; et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj* **2024**, *385*.
100. Gibson, A.D.; White, N.M.; Collins, G.S.; Barnett, A.G. Evidence of unreliable data and poor data provenance in clinical prediction model research and clinical practice. *medRxiv* **2026**, pp. 2026–02.
101. Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* **2015**.
102. Zhou, G.; Rusnac, D.V.; Park, H.; Canzani, D.; Nguyen, H.M.; Stewart, L.; Bush, M.F.; Nguyen, P.T.; Wulff, H.; Yarov-Yarovoy, V.; et al. An artificial intelligence accelerated virtual screening platform for drug discovery. *Nature Communications* **2024**, *15*, 7761.
103. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nature chemistry* **2012**, *4*, 90–98.
104. Zeng, X.; Wang, F.; Luo, Y.; Kang, S.g.; Tang, J.; Lightstone, F.C.; Fang, E.F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine* **2022**, *3*.
105. Li, X.; Wang, P.; Fu, T.; Gao, W.; Li, C.; Shi, L.; Liu, J. AUTODIFF: Autoregressive Diffusion Modeling for Structure-based Drug Design. *arXiv preprint arXiv:2404.02003* **2024**.
106. Fu, T.; Gao, W.; Coley, C.; Sun, J. Reinforced genetic algorithm for structure-based drug design. *Advances in Neural Information Processing Systems* **2022**, *35*, 12325–12338.
107. Fu, T.; Gao, W.; Xiao, C.; Yasonik, J.; Coley, C.W.; Sun, J. Differentiable scaffolding tree for molecular optimization. *arXiv preprint arXiv:2109.10469* **2021**.
108. Lin, H.; Huang, Y.; Zhang, O.; Ma, S.; Liu, M.; Li, X.; Wu, L.; Wang, J.; Hou, T.; Li, S.Z. Diffbp: Generative diffusion of 3d molecules for target protein binding. *Chemical Science* **2025**, *16*, 1417–1431.
109. Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; Coley, C.W. Diffms: Diffusion generation of molecules conditioned on mass spectra. *arXiv preprint arXiv:2502.09571* **2025**.
110. Lu, W.; Zhang, J.; Huang, W.; Zhang, Z.; Jia, X.; Wang, Z.; Shi, L.; Li, C.; Wolynes, P.G.; Zheng, S. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications* **2024**, *15*, 1071.
111. Zhao, Q.; Zhao, H.; Guo, L.; Zheng, K.; Li, Y.; Ling, Q.; Tang, J.; Li, Y.; Wang, J. ColdstartCPI: Induced-fit theory-guided DTI predictive model with improved generalization performance. *Nature Communications* **2025**, *16*, 6436.

 doi:10.20944/preprints202606.0091.v1

112. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* **2018**, *58*, 27–35.
113. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 7112–7127.
114. Sun, Y.; Li, Y.Y.; Leung, C.K.; Hu, P. iNGNN-DTI: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models. *Bioinformatics* **2024**, *40*, btae135.
115. Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E.J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.
116. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.
117. Swanson, K.; Liu, G.; Catacutan, D.B.; Arnold, A.; Zou, J.; Stokes, J.M. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence* **2024**, *6*, 338–353.
118. Fu, T.; Xiao, C.; Li, X.; Glass, L.M.; Sun, J. Mimosa: Multi-constraint molecule sampling for molecule optimization. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 125–133.
119. Fu, T.; Xiao, C.; Sun, J. Core: Automatic molecule optimization using copy & refine strategy. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 638–645.
120. Fu, T.; Xiao, C.; Glass, L.M.; Sun, J. MOLER: Incorporate molecule-level reward to enhance deep generative model for molecule optimization. *IEEE transactions on knowledge and data engineering* **2021**, *34*, 5459–5471.
121. Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2012**, *55*, 2932–2942.
122. Maggiora, G.M. On outliers and activity cliffs why QSAR often disappoints, 2006.
123. Van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling* **2022**, *62*, 5938–5951.
124. Patel, C.N.; Kumar, S.P.; Rawal, R.M.; Patel, D.P.; Gonzalez, F.J.; Pandya, H.A. A multiparametric organ toxicity predictor for drug discovery. *Toxicology mechanisms and methods* **2020**, *30*, 159–166.
125. Lin, Z.; Will, Y. Evaluation of drugs with specific organ toxicities in organ-specific cell lines. *Toxicological sciences* **2012**, *126*, 114–127.
126. Cremer, J.; Medrano Sandonas, L.; Tkatchenko, A.; Clevert, D.A.; De Fabritiis, G. Equivariant graph neural networks for toxicity prediction. *Chemical Research in Toxicology* **2023**, *36*, 1561–1573.
127. Teng, S.; Yin, C.; Wang, Y.; Chen, X.; Yan, Z.; Cui, L.; Wei, L. MolFPG: multi-level fingerprint-based graph transformer for accurate and robust drug toxicity prediction. *Computers in Biology and Medicine* **2023**, *164*, 106904.
128. Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B* **2022**, *12*, 3049–3062.
129. Zhu, H.W.; Yi, H.C.; You, Z.H.; He, S.H.; Shang, X. A Multimodal Information Fusion Framework for Accurate and Robust Prediction of Drug-Drug Interactions. *Information Fusion* **2025**, p. 103981.
130. Bi, X.; Ma, W.; Jiang, H.; Cai, Q.; Nie, J.; Wei, Z.; Zhang, S. Subgraph-Focused Biomedical Knowledge Embedding with Bi-Semantic Encoder for Multi-Type Drug-Drug Interaction Prediction. *Information Fusion* **2025**, p. 104109.
131. Rationalizing combination therapies. *Nature Medicine* **2017**, *23*, 1113. https://doi.org/10.1038/nm.4426.
132. Preuer, K.; Lewis, R.P.; Hochreiter, S.; Bender, A.; Bulusu, K.C.; Klambauer, G. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **2018**, *34*, 1538–1546.
133. Wang, H.; Luna, A.; Yan, G.; Li, X.; Babur, O.; Mills, G.B.; Sander, C.; Korkut, A. Targeting adaptation to cancer treatment by drug combinations. *bioRxiv* **2021**, pp. 2021–04.
134. Kong, W.; Midena, G.; Chen, Y.; Athanasiadis, P.; Wang, T.; Rousu, J.; He, L.; Aittokallio, T. Systematic review of computational methods for drug combination prediction. *Computational and structural biotechnology journal* **2022**, *20*, 2807–2814.
135. Li, Y.; Meng, Q.; Yang, M.; Liu, D.; Hou, X.; Tang, L.; Wang, X.; Lyu, Y.; Chen, X.; Liu, K.; et al. Current trends in drug metabolism and pharmacokinetics. *Acta Pharmaceutica Sinica B* **2019**, *9*, 1113–1144.
136. Lu, J.; Deng, K.; Zhang, X.; Liu, G.; Guan, Y. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. *iScience* **2021**, *24*.

 doi:10.20944/preprints202606.0091.v1

137. Chen, R.T.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D.K. Neural ordinary differential equations. In Proceedings of the Advances in Neural Information Processing Systems, 2018, Vol. 31.
138. Lu, J.; Bender, B.; Jin, J.Y.; Guan, Y. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature Machine Intelligence* **2021**, *3*, 696–704.
139. Laubenbacher, R.; Sluka, J.P.; Glazier, J.A. Using digital twins in viral infection. *Science* **2021**, *371*, 1105–1106.
140. Elias, M.; König, M. A digital twin of glimepiride for personalized and stratified diabetes treatment. *Frontiers in Pharmacology* **2025**, *16*, 1686415.
141. Shen, M.d.; Chen, S.b.; Ding, X.d. The effectiveness of digital twins in promoting precision health across the entire population: a systematic review. *NPJ Digital Medicine* **2024**, *7*, 145.
142. Srinivasa, R.S.; Qian, C.; Theodorou, B.; Spaeder, J.; Xiao, C.; Glass, L.; Sun, J. Clinical trial site matching with improved diversity using fair policy learning. *arXiv preprint arXiv:2204.06501* **2022**.
143. Theodorou, B.; Glass, L.; Xiao, C.; Sun, J. FRAMM: Fair ranking with missing modalities for clinical trial site selection. *Patterns* **2024**, *5*.
144. Atkinson, J.A.; Wells, R.; Page, A.; Dominello, A.; Haines, M.; Wilson, A. Applications of system dynamics modelling to support health policy. *Public Health Research and Practice* **2015**, *25*, e2531531.
145. Chen, X.; Orom, H.; Hay, J.L.; Waters, E.A.; Schofield, E.; Li, Y.; Kiviniemi, M.T. Differences in rural and urban health information access and use. *The Journal of Rural Health* **2019**, *35*, 405–417.
146. Weiss, D.; Nelson, A.; Vargas-Ruiz, C.; Gligorić, K.; Bavadekar, S.; Gabrilovich, E.; Bertozzi-Villa, A.; Rozier, J.; Gibson, H.; Shekel, T.; et al. Global maps of travel time to healthcare facilities. *Nature medicine* **2020**, *26*, 1835–1838.
147. Fu, T.; Huang, K.; Xiao, C.; Glass, L.M.; Sun, J. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns* **2022**, *3*, 100445.
148. Chen, T.; Hao, N.; Lu, Y.; Rechem, C.V. Uncertainty Quantification on Clinical Trial Outcome Prediction, 2024, [arXiv:cs.LG/2401.03482].
149. Zheng, W.; Peng, D.; Xu, H.; Zhu, H.; Fu, T.; Yao, H. Multimodal clinical trial outcome prediction with large language models. *arXiv preprint arXiv:2402.06512* **2024**.
150. Klein, H.; Mazor, T.; Siegel, E.; Trukhanov, P.; Ovalle, A.; Vecchio Fitz, C.D.; Zwiesler, Z.; Kumari, P.; Van Der Veen, B.; Marriott, E.; et al. MatchMiner: an open-source platform for cancer precision medicine. *npj Precision Oncology* **2022**, *6*, 69.
151. Hamer, D.M.d.; Schoor, P.; Polak, T.B.; Kapitan, D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. *arXiv preprint arXiv:2304.07396* **2023**.
152. Jin, Q.; Wang, Z.; Floudas, C.S.; Chen, F.; Gong, C.; Bracken-Clarke, D.; Xue, E.; Yang, Y.; Sun, J.; Lu, Z. Matching patients to clinical trials with large language models. *Nature communications* **2024**, *15*, 9074.
153. Shool, S.; Adimi, S.; Saboori Amleshi, R.; Bitaraf, E.; Golpira, R.; Tara, M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making* **2025**, *25*, 117.
154. Rao, A.S.; Esmail, K.P.; Lee, R.S.; Jiang, S.; Arraiza Carlo, B.; Gill, J.; Khanna, P.; Kalmowitz, E.; Montagnese, B.; Heydari, K.; et al. Large language model performance and clinical reasoning tasks. *JAMA Network Open* **2026**, *9*, e264003.
155. Zhou, J.; Li, H.; Chen, S.; Chen, Z.; Han, Z.; Gao, X. Large language models in biomedicine and healthcare. *npj Artificial Intelligence* **2025**, *1*, 44.
156. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J.A.; Pimenta, D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ digital medicine* **2025**, *8*, 274.
157. Bean, A.M.; Payne, R.E.; Parsons, G.; Kirk, H.R.; Ciro, J.; Mosquera-Gómez, R.; Hincapié M, S.; Ekanayaka, A.S.; Tarassenko, L.; Rocher, L.; et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine* **2026**, pp. 1–7.
158. Zhong, X.; Li, S.; Chen, Z.; Ge, L.; Yu, D.; Wang, S.; You, L.; Shang, H. Considerations for patient privacy of Large Language Models in health care: scoping review. *Journal of Medical Internet Research* **2025**, *27*, e76571.
159. Schoolcraft, D.; Meltzer, A.C.; Sangal, R.; Terry, A.T.; Robertson, K.; Buckland, D.; Motalib, S.; Genes, N.; Vukmir, R.; Waseem, T.; et al. Health Insurance Portability and Accountability Act Liability in the Age of Generative Artificial Intelligence. *JACEP Open* **2026**, *7*, 100317.
160. Shah, S.V. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open* **2024**, *7*, e2425953.

 doi:10.20944/preprints202606.0091.v1

161. Croxford, E.; Gao, Y.; First, E.; Pellegrino, N.; Schnier, M.; Caskey, J.; Oguss, M.; Wills, G.; Chen, G.; Dligach, D.; et al. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine* **2025**, *8*, 640.
162. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915* **2024**.
163. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on llm-as-a-judge. *The Innovation* **2024**.
164. Genovese, A.; Hegstrom, L.; Prabha, S.; Gomez-Cabello, C.A.; Haider, S.A.; Collaco, B.; Wood, N.G.; Forte, A.J. Artificial Authority: The Promise and Perils of LLM Judges in Healthcare. *Bioengineering* **2026**, *13*, 108.
165. Laskar, M.T.R.; Jahan, I.; Dolatabadi, E.; Peng, C.; Hoque, E.; Huang, J. Improving automatic evaluation of large language models (LLMs) in biomedical relation extraction via LLMs-as-the-judge. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 25483–25497.
166. Wang, L.; Chen, X.; Deng, X.; Wen, H.; You, M.; Liu, W.; Li, Q.; Li, J. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ digital medicine* **2024**, *7*, 41.
167. Chen, Y.; Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv* **2023**.
168. Steshin, S. Lo-hi: Practical ml drug discovery benchmark. *Advances in Neural Information Processing Systems* **2023**, *36*, 64526–64554.
169. Sheridan, R.P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of chemical information and modeling* **2013**, *53*, 783–790.
170. Talevi, A.; Morales, J.F.; Hather, G.; Podichetty, J.T.; Kim, S.; Bloomingdale, P.C.; Kim, S.; Burton, J.; Brown, J.D.; Winterstein, A.G.; et al. Machine learning in drug discovery and development part 1: a primer. *CPT: pharmacometrics & systems pharmacology* **2020**, *9*, 129–142.
171. Fooladi, H.; Vu, T.N.L.; Mathea, M.; Kirchmair, J. Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. *Journal of Chemical Information and Modeling* **2025**, *65*, 9871–9891.
172. Ji, Y.; Zhang, L.; Wu, J.; Wu, B.; Li, L.; Huang, L.K.; Xu, T.; Rong, Y.; Ren, J.; Xue, D.; et al. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 8023–8031.
173. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion* **2021**, *76*, 243–297.
174. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **2019**, *1*, 206–215.
175. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. In Proceedings of the Machine learning for healthcare conference. PMLR, 2019, pp. 359–380.
176. Chen, Y.; Zhang, L.; Zhong, C.; Li, S.; Li, R.; Jia, Z.; Huang, P.; Wang, S.; He, Z.; Lin, H.; et al. Rare Cancer Explorer 1.0 (RaCE 1.0): a dedicated database and analytical platform focused on rare cancers. *Nucleic Acids Research* **2026**, *54*, D1590–D1607.
177. Pillai, R.K.; Jayasree, K. Rare cancers: Challenges & issues. *Indian Journal of Medical Research* **2017**, *145*, 17–27.
178. Thapa, K.; Kinali, M.; Pei, S.; Luna, A.; Babur, Ö. Strategies to include prior knowledge in omics analysis with deep neural networks. *Patterns* **2025**, *6*.
179. Babur, Ö.; Luna, A.; Korkut, A.; Durupinar, F.; Siper, M.C.; Dogrusoz, U.; Jacome, A.S.V.; Peckner, R.; Christianson, K.E.; Jaffe, J.D.; et al. Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns* **2021**, *2*.
180. Omar, M.; Sorin, V.; Collins, J.D.; Reich, D.; Freeman, R.; Gavin, N.; Charney, A.; Stump, L.; Bragazzi, N.L.; Nadkarni, G.N.; et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Medicine* **2025**, *5*, 330.
181. Ayoub, M.; Zhao, H.; Li, L.; Yang, D.; Hussain, S.; Wahid, J.A. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. *Communications Medicine* **2026**.

 doi:10.20944/preprints202606.0091.v1

182. Ding, C.; Bian, M.; Chen, P.; Zhang, H.; Li, T.; Liu, L.; Chen, J.; Li, Z.; Zhong, Y.; Liu, Y.; et al. Building a human-verified clinical reasoning dataset via a human LLM hybrid pipeline for trustworthy medical AI. *arXiv preprint arXiv:2505.06912* **2025**.