



Available online at
<https://jiicet.gnt.com.pk/index.php/jiicet/>

Journal of Innovative Intelligent Computing and Emerging Technologies (JIICET)

Vol. 1, No. 1, 2026



AI-Based Drug–Target Interaction Prediction: A Systematic Review of Machine Learning and Deep Learning Approaches

Shaman Ali¹, Nimra², Sajjad Ali³, Shahid Ali⁴

¹Information Technology Centre, Sindh Agriculture University, Tandojam, Hyderabad, Sindh

*Corresponding author: shamanalimaher462@gmail.com

Article History

Received 05 August 2026

Revised 05 October 2026

Accepted 7 October 2026

Available Online 22 October 2026

Keywords:

Drug–target interaction

binding affinity

machine learning

deep learning

graph neural networks

Abstract

Predicting which small molecules bind which proteins is among the slowest and most expensive steps in early drug discovery, and it is where computational methods have made their most visible progress. The field has moved from similarity kernels and matrix factorisation towards models that learn directly from molecular graphs, SMILES strings, protein sequences and, increasingly, predicted three-dimensional structures. This review surveys that shift. Following PRISMA 2020 guidance we screened 4,812 records published between 2008 and 2025 and retained 68 studies. The methods identified are grouped into five families: similarity and kernel models, matrix factorisation and feature-based classifiers, sequence-based deep networks, graph neural networks together with attention and transformer architectures, and heterogeneous-network or structure-aware models. For each family we compare representations, benchmark datasets, evaluation protocols and reported performance, and we state what each comparison contributes. Reported error on the standard affinity benchmarks has fallen by roughly half since the earliest kernel baselines, but almost all of that improvement has been measured under random splits that allow near-duplicate drugs and proteins to appear in both training and test data, and performance under cold-start and cross-domain conditions remains substantially weaker. Our central observation is therefore that architectural progress has outpaced evaluation practice. We identify ten open problems, among them leakage under random splits, weak cold-start generalisation, the absence of reliable negative samples, narrow coverage of protein families, limited interpretability and the under-use of predicted structures, and set out the directions most likely to address them: protein and chemical language models, structure-aware architectures built on proteome-wide predicted structures, uncertainty-aware prediction coupled to active experimental design, and standardised benchmarks with hidden test sets and prospective validation.



Copyright: © 2026 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (<https://creativecommons.org/licenses/by-nc/4.0/>)

1. Introduction

Identifying the proteins a candidate molecule binds is the first major bottleneck in drug discovery, and one of the most expensive. Pharmaceutical research and development productivity has declined for decades despite rising expenditure [1], and the capitalised cost of bringing one approved drug to market is now put at roughly US\$2.6 billion [2]. Most candidates fail for reasons that are, at root,

questions about binding: too little activity at the intended target, or unexpected activity elsewhere. Experimental assays such as comprehensive kinase panels [17] answer those questions reliably, but they cover only a vanishingly small part of the chemical–proteomic space.

Computational prediction narrows that space before any compound is synthesised. Early approaches included molecular docking [5], which requires a resolved target structure, and ligand-based similarity searching such as the

Similarity Ensemble Approach [6], which requires known ligands. Neither helps when the target has no structure or the compound has no close analogues. Yamanishi et al. [7] broke that dependence by treating prediction as a learning problem over a joint drug–target space, allowing evidence to be shared across targets and across compounds.

Three broad waves have followed. Work between 2008 and 2016 was dominated by similarity kernels, bipartite local models and matrix factorisation [8–14]. From roughly 2016 the emphasis moved to regression of continuous binding affinity and to deep sequence models, of which DeepDTA [27] and DeepConv-DTI [29] are the clearest examples. Since 2020, the literature has been shaped by graph neural networks [30, 33] and attention-based architectures [34–37], by pretrained protein and chemical language models [47–49], and by structure-aware models exploiting the proteome-wide coverage AlphaFold now provides [46, 66].

Earlier surveys catalogued databases and web servers [55], chemogenomic methods [56] and machine learning in discovery more broadly [57]. Much has changed since. The volume of deep learning work has grown sharply, and the central dispute has moved from which architecture performs best to whether reported performance means anything at all. This review is organised around that shift and makes four contributions:

1. A taxonomy of machine learning and deep learning methods, assembled under PRISMA 2020 guidance,

covering both binary interaction classification and affinity regression.

2. Seven structured comparison tables — datasets, representations, classical methods, deep architectures, metrics, reported performance and open problems — each followed by an explicit statement of what it contributes.
3. A critical reading of reported performance, with attention to the evaluation pitfalls that inflate it.
4. A consolidated account of research gaps, each linked to the evidence behind it and to a concrete direction for addressing it.

Figure 1 places the studies reviewed here on a timeline, showing how the four eras of method development overlap rather than replace one another.

The remainder of the paper is organised as follows. Section 2 sets out the review methodology. Section 3 presents the results and discussion: the prediction task and its benchmark datasets (Section 3.1), drug and target representations (Section 3.2), classical machine learning methods (Section 3.3), deep learning methods (Section 3.4), evaluation protocols (Section 3.5), a comparative analysis of reported performance (Section 3.6), the research gaps we identify (Section 3.7) and the directions that follow from them (Section 3.8). Section 4 concludes.

Fifteen years of method development

Four overlapping eras, from similarity kernels to structure-aware deep models

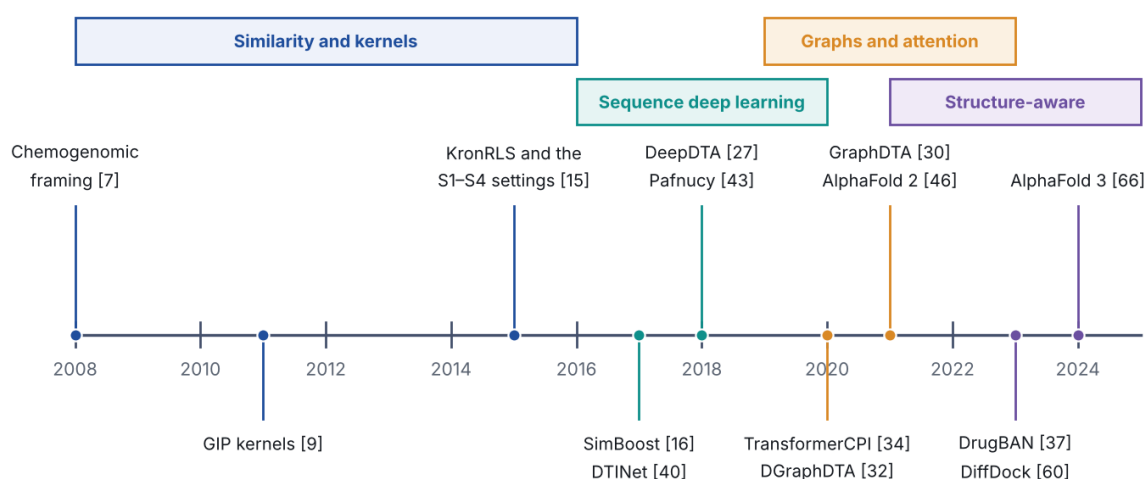


Figure 1. The field moves through four overlapping eras, from similarity kernels and matrix factorisation to structure-aware deep models.

2. Materials and Methods

The review follows the PRISMA 2020 statement [62] through identification, screening, eligibility assessment and inclusion.

2.1. Research questions

- RQ1. Which machine learning and deep learning architectures have been proposed for interaction classification and affinity regression?
- RQ2. How are drugs and targets represented, and which representations perform best under which conditions?
- RQ3. Which datasets, splitting strategies and metrics are used, and how comparable are the results they produce?
- RQ4. Which problems remain open, and which directions are most likely to resolve them?

2.2. Search strategy

We searched PubMed, IEEE Xplore, ScienceDirect, SpringerLink, the ACM Digital Library, Web of Science and Google Scholar, supplemented by arXiv and bioRxiv for preprints that had attracted substantial citations. The query combined task terms with method terms:

("drug-target interaction" OR "drug-target affinity" OR "compound-protein interaction"

OR "binding affinity prediction")

AND ("machine learning" OR "deep learning" OR "neural network"

OR "graph neural network" OR "transformer")

2.3. Eligibility criteria

We included studies that were peer-reviewed papers or well-cited preprints published between 2008 and 2025, written in English, and that proposed or benchmarked a learning-based model for interaction or affinity prediction with quantitative results on at least one public dataset. We excluded pure docking and molecular dynamics studies with no learning component, work on protein-protein or drug-drug interactions, and abstracts, editorials and theses.

2.4. Data extraction and quality assessment

From each included study, we extracted the year and venue, the task addressed, the drug and target representations, the architecture, the datasets and splitting strategy, the metrics and reported performance, whether code had been released, and the limitations the authors themselves acknowledged. Four criteria guided the quality assessment: dataset transparency, split rigour, baseline fairness, and reproducibility. Figure 2 summarises the selection process.

Study selection process

PRISMA 2020 flow: from 4,812 records identified to 68 studies included

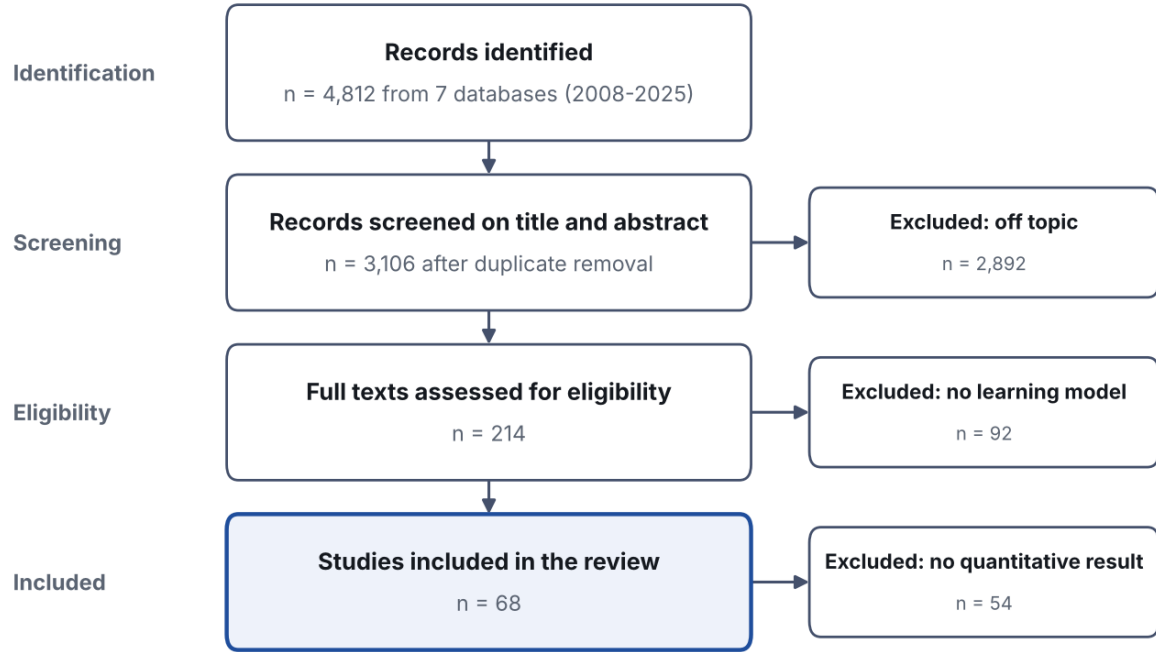


Figure 2. Of 4,812 records identified, 68 studies satisfied all inclusion criteria and form the evidence base of this review.

3. Results and Discussion

This section presents the review findings. It moves from the formulation of the prediction task and the available data, through the representations and model families identified in the literature, to a comparative reading of reported performance and the open problems that this reading exposes.

3.1. Problem Formulation and Benchmark Datasets

Two formulations dominate the literature. The classification task asks whether a given drug binds a given target and returns a probability. The regression task, usually called drug–target affinity prediction, returns a continuous measure of binding strength such as pKd, pKi or a KIBA score [15, 18]. Writing D for the set of drugs and T for the set of targets, a model learns

$$f_{\theta} : (d_i, t_j) \rightarrow \hat{y}_{ij}, \quad \hat{y}_{ij} \in [0, 1] \text{ for classification, } \hat{y}_{ij} \in \mathbb{R} \text{ for regression} \quad (1)$$

Pahikkala et al. [15] argued that regression on continuous affinities better reflects the underlying biology. A binary threshold collapses the distinction between weak and strong binders, and it forces unmeasured pairs to count as non-binding. Affinities are normally converted to a

logarithmic scale before training, with pKd defined as $-\log_{10}(K_d / 10^9)$ for K_d in nanomolar units [27].

Figure 3 compares the scale of these resources, and Table 1 summarises what each one provides.

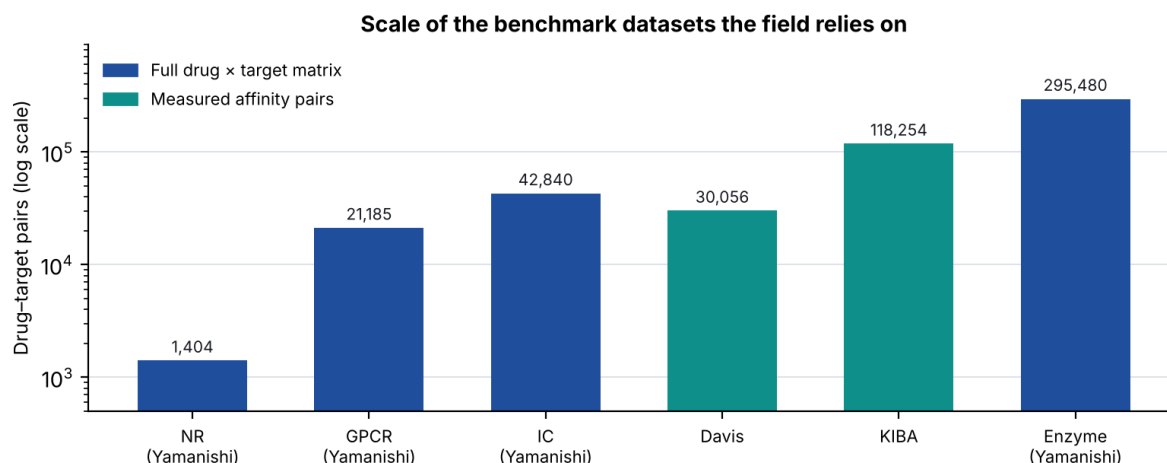


Figure 3. Davis and KIBA carry most affinity regression work, yet both cover only kinases; the Yamanishi benchmarks are far smaller.

Table 1. Benchmark datasets and databases used for interaction and affinity prediction.

Dataset / database	Type	Scale as reported	Label	Typical use	Ref.
Yamanishi gold standard (Enzyme, IC, GPCR, NR)	Benchmark	Enzyme 445 × 664; IC 210 × 204; GPCR 223 × 95; NR 54 × 26	Binary	Classical ML classification	[7]
Davis	Benchmark	68 kinase inhibitors × 442 kinases; 30,056 pairs	Kd → pKd	Affinity regression	[17, 27]
KIBA	Benchmark	2,111 drugs × 229 kinases; 118,254 pairs	KIBA score	Affinity regression	[18, 27]
BindingDB	Database	Large, continuously updated	Kd, Ki, IC50	Classification and regression	[19]
BIOSNAP, Human, C. elegans	Benchmark	Curated positive and negative pairs	Binary	Deep classification	[31, 35]
DrugBank	Database	Approved and experimental drugs with targets	Binary	Network and repositioning models	[20]
ChEMBL	Database	Millions of bioactivity records	IC50, Ki and others	Pretraining, large-scale regression	[21]
STITCH	Database	Chemical-protein associations with confidence scores	Score	Network-based prediction	[22]
PDBbind	Database	Protein-ligand complexes with 3D structures	Kd, Ki, IC50	Structure-based affinity	[23]
KEGG, UniProt	Database	Pathways, drugs, sequences, annotations	—	Feature and sequence source	[24, 65]
Therapeutics Data Commons	Benchmark platform	Standardised splits and leaderboards	Both	Fair benchmarking	[54]

Contribution of Table 1. Mapped against task and label type, the available resources reveal how narrow the empirical base of the field actually is. Nearly every regression paper reports on Davis and KIBA, both kinase-centred, while classification work still leans on the small Yamanishi sets. A model validated only on kinases offers no guarantee of transferring to GPCRs, ion channels or orphan targets, and that limitation is one of the gaps we return to in Section 3.7. The structurally annotated and

standardised resources listed here — PDBbind and the Therapeutics Data Commons — are what make more demanding evaluation possible.

3.2. Drug and Target Representations

How molecules and proteins are encoded constrains what any model can learn, and in this field the choice of representation has mattered at least as much as the choice

of architecture. Drugs appear as fingerprints, SMILES strings, molecular graphs or three-dimensional conformers; targets as composition descriptors, amino-acid

sequences, contact maps, pretrained embeddings or resolved structures. Figure 4 places both sets of options on a common axis of increasing structural detail.

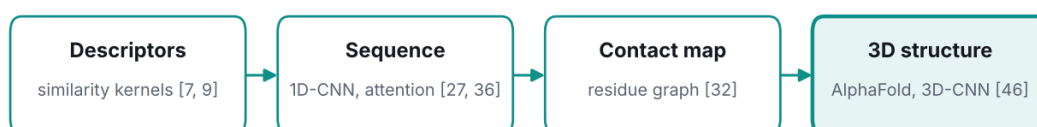
Representations of drugs and targets

Detail and data requirement rise from left to right

Drug



Target



More structural detail, more compute, stronger dependence on structural accuracy

Figure 4. Both encoders progress from hand-crafted features towards explicit three-dimensional structure, at rising computational cost.

3.2.1. Drug representations

Hand-crafted fingerprints such as MACCS keys [67] and extended-connectivity fingerprints [25] encode substructures as fixed binary vectors. They remain strong baselines for classical classifiers and are easy to interpret, since an informative bit corresponds to a chemical substructure. SMILES strings [26] opened the way for sequence models — one-dimensional convolutions, recurrent networks and later transformers — to learn features without manual engineering [27, 38]. Molecular graphs, in which atoms are nodes and bonds are edges, are the natural input for graph neural networks [30, 31, 51, 52] and preserve topology that fingerprints discard. More recently, chemical language models such as ChemBERTa [49] have transferred representations learned from millions of unlabelled molecules.

3.2.2. Target representations

Early work relied on composition descriptors and Smith-Waterman similarity kernels [7, 9]. Deep models embed raw amino-acid sequences through convolutional encoders [27, 29] or attention [34, 36]. DGraphDTA [32] took a different route, building protein graphs from predicted residue contact maps so that the target could be processed by a graph network in the same way as the drug. Protein language models such as ESM [47] and ProtTrans [48] now supply evolution-aware embeddings that transfer across tasks. AlphaFold [46, 66], meanwhile, has made predicted structures available for essentially every protein, extending structure-aware modelling far beyond the set of proteins with crystal structures.

Table 2. Representations of drugs and targets, with their trade-offs.

Entity	Representation	Typical encoder	Strengths	Limitations	Ref.
Drug	MACCS or ECFP fingerprint	RF, SVM, MLP	Fast, interpretable, strong baseline	Fixed length, discards topology, bit collisions	[25, 67]
Drug	SMILES string	1D-CNN, LSTM, transformer	End-to-end learning, no feature engineering	Non-canonical strings, weak spatial semantics	[26, 27, 38]
Drug	Molecular graph	GCN, GAT, GIN, MPNN	Preserves topology and atom chemistry	Over-smoothing; stereochemistry needs 3D	[30, 31, 33]

Entity	Representation	Typical encoder	Strengths	Limitations	Ref.
Drug	Pretrained embedding	ChemBERTa	Transfers from unlabelled corpora	Compute cost, possible domain mismatch	[49]
Drug	3D conformer or complex	3D-CNN, equivariant GNN	Captures geometry and binding pose	Requires structures; conformer noise	[43, 44, 61]
Target	Similarity or composition descriptors	Kernels, RF	Simple, effective on small data	Hand-crafted, limited expressivity	[7, 9]
Target	Amino-acid sequence	1D-CNN, attention	Available for every protein	Ignores fold and pocket geometry	[27, 29, 36]
Target	Contact map or residue graph	GCN	Adds structural context cheaply	Sensitive to contact-map accuracy	[32]
Target	Protein language model embedding	ESM, ProtTrans	Evolutionary and structural signal, transferable	Large models; embeddings not task-specific	[47, 48]
Target	3D structure (experimental or predicted)	3D-CNN, GNN, diffusion	Pocket-level detail	Pocket identification; structural error	[23, 46, 60]

Contribution of Table 2. Setting the options side by side makes the central trade-off explicit. Moving from fingerprints and sequences to graphs, pretrained embeddings, and explicit structures adds biologically meaningful detail, but each step raises computational cost and dependence on structural accuracy. Read alongside the data regime of a given problem, the comparison offers practical guidance: fingerprints or frozen pretrained embeddings where labelled data are scarce, graph- or structure-aware encoders where larger and structurally annotated datasets are available.

3.3. Classical Machine Learning Approaches

Classical methods rest on a guilt-by-association assumption: similar compounds tend to bind similar proteins. They train quickly, need comparatively little data and remain competitive on small benchmarks. Their weakness is equally clear — they depend on hand-crafted similarity measures and degrade sharply when a drug or target has no close neighbours in the training set.

3.3.1. Similarity and kernel methods

Yamanishi et al. [7] mapped the chemical and genomic spaces into a shared pharmacological space and performed kernel regression there. Bleakley and Yamanishi [8] instead trained a separate support vector machine for each drug and each target, an approach that BLM-NII [10] later extended to new entities by inferring their interaction profiles from neighbours. Van Laarhoven et al. [9] derived Gaussian interaction profile kernels directly from the interaction matrix, and KronRLS [15] combined Kronecker-

product kernels with regularised least squares to become the standard regression baseline.

3.3.2. Network-based inference

Cheng et al. [11] applied resource diffusion over the bipartite drug-target network and experimentally confirmed several predicted repositioning candidates. Their results established that network topology alone carries predictive signal, independent of any chemical descriptor.

3.3.3. Matrix factorisation

A second line of work treats prediction as completion of a sparse matrix. KBMF2K [12] combined multiple kernels with Bayesian factorisation, and collaborative matrix factorisation [13] regularised the latent factors using several drug and target similarities. NRLMF [14] went further, adding neighbourhood regularisation to logistic matrix factorisation and weighting observed positives more heavily than unknown pairs — an early acknowledgement that unlabelled is not the same as negative.

3.3.4. Feature-based ensembles

SimBoost [16] engineered features from similarity and network statistics and passed them to gradient-boosted trees, producing not only affinity estimates but also prediction intervals — a capability that remains rare in later deep models. Random forests and support vector machines over fingerprint and sequence descriptors continue to serve as baselines throughout the literature [56, 57].

Table 3. Representative classical machine learning methods.

Method (year)	Category	Task	Input	Central idea	Principal limitation	Ref.
Yamanishi et al. (2008)	Kernel regression	Classification	Chemical and sequence similarity	Joint pharmacological space	Needs similarity to known entities	[7]
BLM (2009)	Local SVM	Classification	Drug and target kernels	One local model per entity	Fails for unseen drugs and targets	[8]
GIP-RLS (2011)	Kernel RLS	Classification	Interaction profiles and similarity	Kernel derived from the matrix itself	Weak under cold start	[9]
NBI (2012)	Network diffusion	Classification	Bipartite topology	Resource allocation over the network	Cannot score isolated new drugs	[11]
KBMF2K (2012)	Bayesian MF	Classification	Multiple kernels	Kernelised latent factors	Limited scalability	[12]
BLM-NII (2013)	Local SVM with inference	Classification	Kernels and neighbour profiles	Infers profiles for new entities	Depends on neighbour quality	[10]
CMF (2013)	Matrix factorisation	Classification	Multiple similarities	Similarity-regularised factors	Models only linear interactions	[13]
KronRLS (2015)	Kronecker kernel	Regression	Drug $\times$ target kernels	Pairwise regression on affinities	Memory cost of the kernel matrices	[15]
NRLMF (2016)	Logistic MF	Classification	Similarities and interaction matrix	Neighbourhood regularisation, positive weighting	Sensitive to hyperparameters	[14]
SimBoost (2017)	Gradient boosting	Regression	Engineered similarity and network features	Feature engineering with prediction intervals	Manual feature construction	[16]

Contribution of Table 3. Read chronologically, the table traces two parallel movements: from local, per-entity models towards global kernel and matrix formulations, and from binary labels towards continuous affinities. It also isolates the weakness the whole family shares. Without exception, these methods depend on similarity to entities already seen in training, which are not held in a discovery campaign. That dependence explains both why KronRLS and SimBoost persist as baselines and why the field turned to models that learn representations from raw molecular input.

3.4. Deep Learning Approaches

Deep models dispense with hand-crafted similarity and learn representations end to end. Despite considerable architectural variety, most follow a common template, shown in Figure 6: a drug encoder and a target encoder produce embeddings, a fusion module combines them, and a prediction head returns a probability or an affinity. Figure 5 organises the methods reviewed here into families.

Taxonomy of AI approaches to DTI prediction

Families 1-2 depend on similarity to known entities; families 3-6 learn from raw input

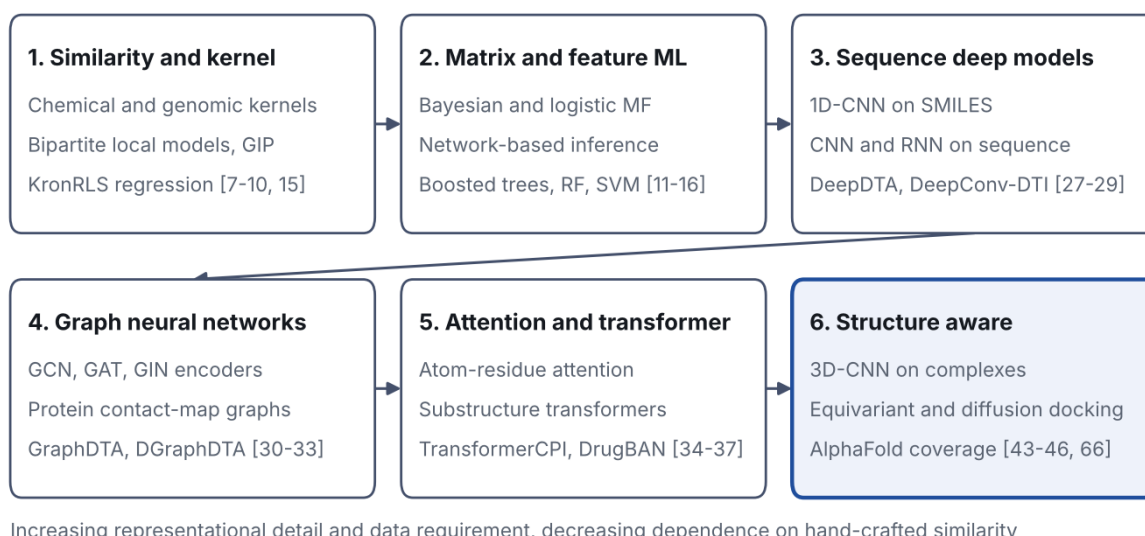


Figure 5. Methods grouped by what they learn from: similarity to known entities in families 1 and 2, raw molecular and structural input in families 3 to 6.

3.4.1. Early deep models

Wen et al. [63] applied deep belief networks to extended-connectivity fingerprints and protein composition descriptors, showing that deep architectures could outperform random forests and support vector machines on DrugBank interactions. The input, however, remained a fixed hand-crafted feature vector, so the representational advantage of deep learning went largely unrealised.

3.4.2. Sequence-based convolutional and recurrent models

DeepDTA [27] was the decisive step. Two convolutional branches processed SMILES strings and protein sequences independently, and the concatenated embeddings fed a regression head; the model outperformed both KronRLS and SimBoost on Davis and KIBA. Extensions followed quickly. WideDTA [28] added ligand maximum common substructures together with protein motifs and domains, while DeepConv-DTI [29] applied convolutions at several scales to capture local binding motifs along the sequence. DeepAffinity [38] combined recurrent and convolutional encoders with attention and pretrained the sequence encoders on unlabelled data, and DeepCDA [42] paired LSTM and convolutional encoders with two-sided attention and domain adaptation across datasets.

3.4.3. Graph neural network models

Tsubaki et al. [31] paired a graph network over molecular structure with a convolutional encoder over protein n-grams, training both jointly. GraphDTA [30] then replaced

the SMILES branch of DeepDTA with graph convolutional, graph attention and graph isomorphism encoders [51, 52]. Its isomorphism variant achieved the lowest error on both benchmarks, confirming that the topology discarded by string representations carries real predictive value. DGraphDTA [32] applied the same reasoning to the protein side, constructing residue graphs from predicted contact maps. Two further models extended the idea in different directions: MGraphDTA [33] combined a deep multiscale graph network with dense connections and produced gradient-based attributions over atoms, while Lim et al. [45] encoded the geometry of protein-ligand complexes in a distance-aware graph attention network.

3.4.4. Attention and transformer models

Attention entered the field as a device for interpretability. Gao et al. [39] used it to highlight the protein regions driving a prediction, and AttentionDTA [64] applied a similar mechanism to affinity estimation. Later work treated it as the core of the model. TransformerCPI [34] adapted the transformer decoder [50] so that protein representations were conditioned on compound atoms, and introduced label-reversal experiments to detect models that had learned ligand-only shortcuts. MolTrans [35] decomposed both partners into frequent substructures and modelled their pairwise interaction map explicitly, HyperAttentionDTI [36] computed attention over every atom-residue pair, and DrugBAN [37] combined a bilinear attention network with conditional domain adaptation, improving cross-domain

performance while keeping the learned interaction map interpretable.

3.4.5. Heterogeneous-network models

A separate line of work embeds drugs and proteins within a wider biological context. DTINet [40] learned low-dimensional features from a network linking drugs, proteins, diseases and side effects, and NeoDTI [41] propagated neighbourhood information through the same kind of network using neural message passing. Both exploit relational evidence unavailable to purely molecular models. Neither, however, can score a compound absent from the network, which rules them out for novel chemical matter.

3.4.6. Structure-based models

Pafnucy [43] and KDEEP [44] voxelised protein-ligand complexes from PDBbind [23] and applied three-dimensional convolutions to predict affinity. More recent work predicts the binding pose itself, using equivariant

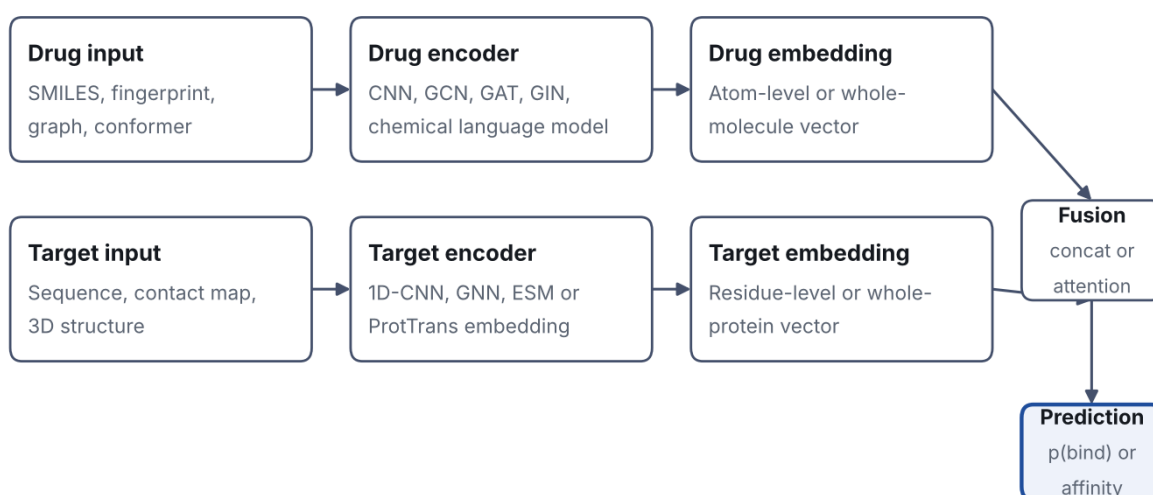
networks in EquiBind [61] and a diffusion process in DiffDock [60], and the resulting poses can feed downstream affinity models. AlphaFold 2 and AlphaFold 3 [46, 66] extend structural coverage to the proteome and to biomolecular complexes. Side-chain placement within binding pockets remains less reliable than the backbone, however, which limits how far predicted structures can substitute for experimental ones.

3.4.7. Pretrained and foundation models

Embeddings from ESM [47] and ProtTrans [48], and from chemical language models such as ChemBERTa [49], are increasingly used as frozen or fine-tuned encoders, bringing knowledge from large unlabeled corpora into typically small datasets. Toolkits such as DeepPurpose [53] and benchmark platforms such as the Therapeutics Data Commons [54] have made these components considerably easier to combine and compare.

Generic deep learning architecture for DTI and DTA prediction

Two encoders, a fusion module and a prediction head



Most recent gains come from the fusion step, where attention over atom-residue pairs replaces simple concatenation of the two embeddings.

Figure 6. Most recent gains come from the fusion step, where attention over atom-residue pairs has replaced simple concatenation.

Table 4. Representative deep learning models.

Model (year)	Drug encoder	Target encoder	Fusion	Task	Datasets	Key contribution	Ref.
DeepDTIs (2017)	DBN on ECFP	Composition descriptors	Concatenation	Classification	DrugBank	First deep classifier for interactions	[63]
DeepDTA (2018)	1D-CNN on SMILES	1D-CNN on sequence	Concatenation	Regression	Davis, KIBA	End-to-end affinity baseline	[27]
Gao et al. (2018)	GCN	LSTM	Attention	Classification	BindingDB	Interpretable attention over protein regions	[39]
DeepConv-DTI (2019)	ECFP with dense layers	Multi-scale 1D-CNN	Concatenation	Classification	DrugBank, KEGG, IUPHAR	Captures local binding motifs	[29]

Model (year)	Drug encoder	Target encoder	Fusion	Task	Datasets	Key contribution	Ref.
GNN-CPI (2019)	GNN on molecular graph	CNN on n-grams	Attention	Classification	Human, C. elegans	Joint graph and sequence training	[31]
DeepAffinity (2019)	RNN on SMILES	RNN on property sequence	Attention with CNN	Regression	BindingDB	Unsupervised pretraining plus attention	[38]
TransformerCPI (2020)	GCN atom features	Transformer decoder	Cross-attention	Classification	Human, C. elegans, BindingDB, GPCR	Label-reversal test for ligand bias	[34]
DGraphDTA (2020)	GNN	GNN on contact map	Concatenation	Regression	Davis, KIBA	Protein represented as a graph	[32]
GraphDTA (2021)	GCN, GAT or GIN	1D-CNN	Concatenation	Regression	Davis, KIBA	Graph encoder outperforms SMILES CNN	[30]
MolTrans (2021)	Substructure transformer	Substructure transformer	Interaction map with CNN	Classification	BIOSNAP, Davis, BindingDB	Substructure-level interaction modelling	[35]
DTINet / NeoDTI (2017, 2019)	Network embedding	Network embedding	Heterogeneous network	Classification	DrugBank, HPRD, SIDER	Integrates multi-relational biology	[40, 41]
HyperAttentionDTI (2022)	CNN with attention	CNN with attention	Atom-residue attention	Classification	Human, C. elegans, DrugBank	Attention for every atom-residue pair	[36]
MGraphDTA (2022)	Deep multiscale GNN	Multiscale CNN	Concatenation	Regression	Davis, KIBA and others	Deep GNN with explanation maps	[33]
DrugBAN (2023)	GCN	CNN	Bilinear attention, domain adaptation	Classification	BindingDB, BioSNAP, Human	Cross-domain generalisation with interpretability	[37]
Pafnucy / KDEEP (2018)	3D grid	3D grid	3D-CNN over the complex	Regression	PDBbind	Structure-based affinity prediction	[43, 44]

Contribution of Table 4. Two trajectories run through the table. Encoders have progressed from fingerprints through sequences to graphs, transformers and voxelised structures, while fusion has moved from simple concatenation through attention over atom-residue pairs to bilinear attention combined with domain adaptation. Comparing the two suggests that fusion design, rather than encoder capacity alone, accounts for most recent gains in classification. Only a minority of models — TransformerCPI and DrugBAN among them — test explicitly for shortcut learning or domain shift. The field also rests on a small number of datasets, which limits how general any of these claims can be.

3.5. Evaluation Protocols and Metrics

How the data are split affects reported performance more than most architectural decisions. Pahikkala et al. [15]

$$CI = (1 / Z) \cdot \sum h(b_i - b_j) \text{ over all pairs with } \delta_i > \delta_j, \quad (2)$$

$$\text{where } h(x) = 1 \text{ if } x > 0, \quad h(x) = 0.5 \text{ if } x = 0, \quad h(x) = 0 \text{ if } x < 0. \quad (3)$$

Here b_i is the prediction for the larger true affinity δ_i , b_j the prediction for the smaller, and Z the number of comparable

defined four settings that remain the reference point: S1, in which both drug and target appear in training; S2, with unseen drugs; S3, with unseen targets; and S4, in which neither has been seen. Random splitting at the pair level corresponds to S1 and allows near-duplicate compounds and homologous proteins to appear on both sides of the split, inflating every metric [15, 37, 56]. Grouping compounds by Bemis–Murcko scaffold and clustering proteins by sequence identity yields far more honest estimates.

The concordance index, the standard ranking metric for affinity prediction, measures how often the predicted ordering of two affinities matches the true ordering:

pairs [27]. Table 5 sets out the metrics and protocols in use alongside what each one conceals.

Table 5. Evaluation metrics and splitting strategies.

Metric or protocol	Task	What it measures	What it conceals	Ref.
AUROC	Classification	Ranking of positives above negatives	Optimistic under heavy class imbalance	[29, 35]

Metric or protocol	Task	What it measures	What it conceals	Ref.
AUPR	Classification	Precision–recall trade-off	Preferred for sparse positives, but prevalence-dependent	[14, 56]
Accuracy, F1, sensitivity, specificity	Classification	Threshold-based classification quality	Results shift with the chosen threshold	[36, 37]
MSE and RMSE	Regression	Squared prediction error	Dominated by outliers	[27, 30]
Concordance index	Regression	Agreement in affinity ranking	Blind to absolute scale and calibration	[15, 27]
r^2_m	Regression	External predictive power (QSAR convention)	Rarely reported, hard to compare across papers	[33]
Pearson and Spearman correlation	Regression	Linear and rank association	Can be high while absolute error is large	[43, 44]
Random split (S1)	Both	Warm-start interpolation	Leakage through near-duplicates	[15]
New-drug or new-target split (S2, S3)	Both	Generalisation to unseen entities	Lower scores, and therefore seldom reported	[15, 56]
Cold-start or cross-domain split (S4)	Both	Generalisation to unseen drug and target	The most realistic setting and the hardest	[37, 54]

Contribution of Table 5. Each metric is paired with its blind spot, and each split with the scenario it represents. Together they support a single conclusion: a high AUROC or concordance index obtained under a random split says little about performance in a discovery setting, where novel chemotypes and understudied targets are the norm rather than the exception. Two changes in practice would make results comparable across papers at minimal cost. The first is to report AUPR alongside AUROC whenever positives are sparse; the second is to report at least one cold-start split alongside the random one.

3.6. Comparative Analysis and Discussion

On the Davis and KIBA benchmarks under warm-start conditions, reported mean squared error has fallen by roughly half on Davis and by about two-thirds on KIBA between KronRLS and the most recent graph models. Table 6 and Figure 7 collect these values as published, either in the original papers or in the comparisons reported by GraphDTA [30], DGraphDTA [32] and MGraphDTA [33]. Because all of them use the random split introduced with DeepDTA [27], differences of a few thousandths in the concordance index fall within run-to-run variation and should not be read as meaningful.

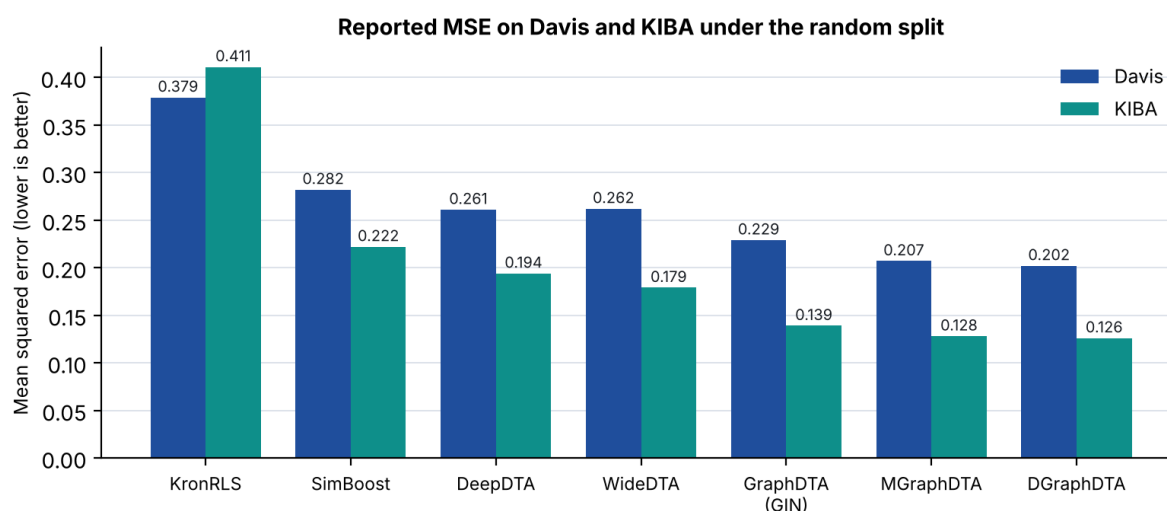


Figure 7. Error falls steeply from kernel methods to graph networks, then flattens: the three most recent models differ by less than 0.01 in concordance index.

Table 6. Reported affinity prediction performance under the random split (higher CI and lower MSE are better).

Model	Family	Davis CI	Davis MSE	KIBA CI	KIBA MSE	Ref.
KronRLS	Kernel	0.871	0.379	0.782	0.411	[15]
SimBoost	Gradient boosting	0.872	0.282	0.836	0.222	[16]
DeepDTA	CNN	0.878	0.261	0.863	0.194	[27]
WideDTA	CNN	0.886	0.262	0.875	0.179	[28]
GraphDTA (GIN)	GNN	0.893	0.229	0.891	0.139	[30]
MGraphDTA	Multiscale GNN	0.900	0.207	0.902	0.128	[33]
DGraphDTA	GNN on both sides	0.904	0.202	0.904	0.126	[32]

Contribution of Table 6 and Figure 7. Taken together, the two quantify the gain from replacing kernels with convolutional networks and then with graph encoders. Most of that gain came from improving the drug representation and, subsequently, from adding structure on the protein side. They also expose saturation: the three most recent models differ by

less than one hundredth in concordance index on Davis, which indicates that the warm-start benchmarks are close to their ceiling and no longer discriminate between methods. Figure 8 shows the same saturation as a trend over publication year, with every model since 2020 falling inside a band narrower than the spread between repeated runs.

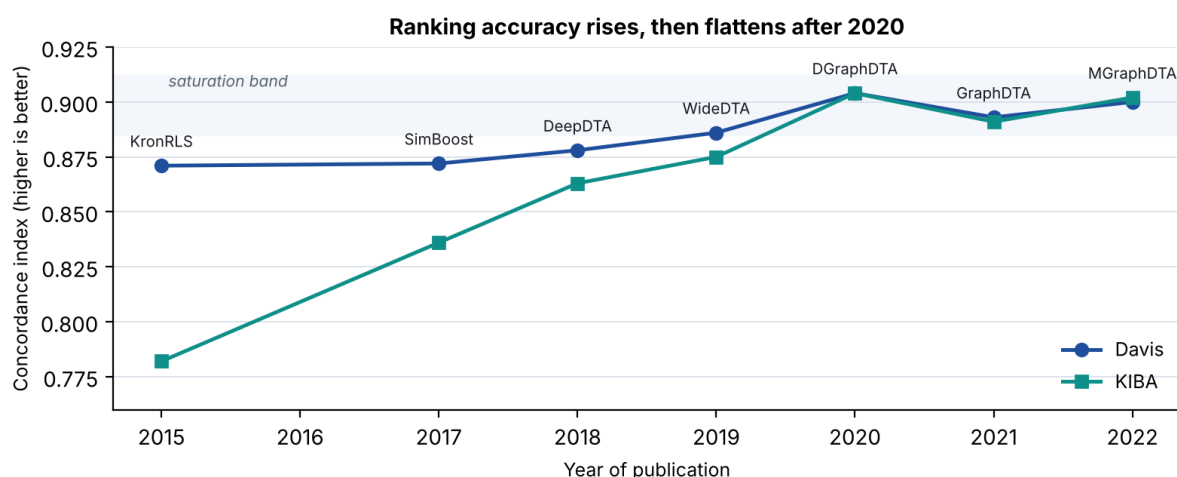


Figure 8. Concordance index rises steeply until 2020 and then flattens inside a narrow band, so the warm-start benchmarks no longer separate methods.

3.6.1. What the comparison shows

- Representation matters more than depth. Graph encoders consistently outperform SMILES convolutions when the protein encoder is held constant [30], and protein contact graphs or pretrained embeddings add a further increment [32, 47].
- Fusion is where classification improves. Pairwise atom-residue and bilinear attention [36, 37] deliver larger gains under cross-domain evaluation than simply enlarging the encoders.
- Cold-start performance remains poor. Under new-drug, new-target or cross-domain splits, reported scores fall steeply for most models [15, 37], and simple baselines become surprisingly competitive.
- Shortcut learning is a real risk. The label-reversal experiment of TransformerCPI [34] showed that some

models rely on ligand-only patterns, so strong accuracy can coexist with almost no use of protein information.

- Network models trade coverage for context. DTINet and NeoDTI [40, 41] draw on diseases and side effects but cannot score compounds absent from their network [58].
- Structure-based models are constrained by data. Three-dimensional convolutions over PDBbind complexes [43, 44] are accurate where complexes exist, but only a small fraction of measured affinities have solved structures; predicted structures [46] widen coverage at the cost of pocket-level uncertainty.

3.7. Research Gaps and Open Challenges

Table 7 consolidates the problems that recur across the studies reviewed, each with the evidence supporting it and the direction most likely to address it.

Table 7. Research gaps, supporting evidence and corresponding directions.

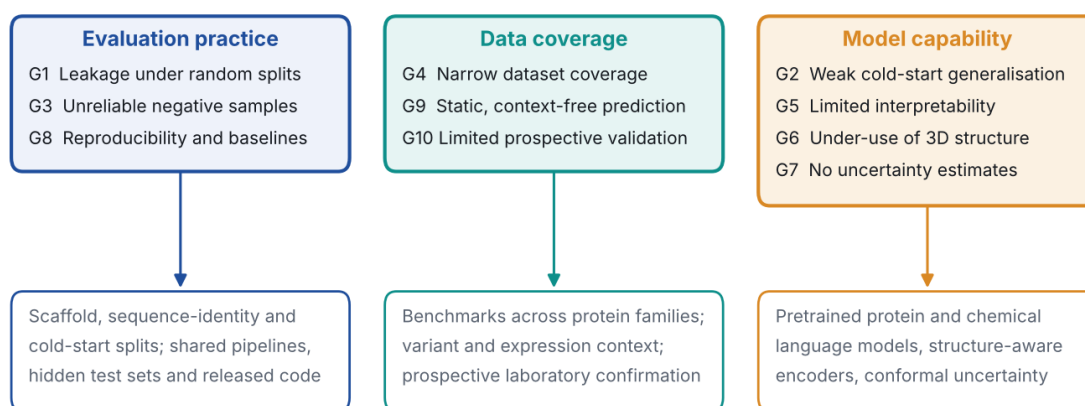
#	Research gap	Why it matters	Evidence	Direction
G1	Leakage under random splits	Warm-start scores overstate usefulness in discovery	Split settings S1–S4 [15]; domain-shift results [37]	Scaffold, sequence-identity and cold-start splits as standard [54]
G2	Weak cold-start generalisation	Real campaigns involve novel chemotypes and understudied targets	Local and factorisation models fail on unseen entities [8, 10, 14]	Meta-learning, few-shot and transfer learning from pretrained encoders [47, 49]
G3	Unreliable negative samples	Unmeasured pairs treated as non-binding bias both training and evaluation	Positive weighting [14]; label-reversal bias [34]	Positive-unlabelled learning and experimentally confirmed negatives
G4	Narrow dataset coverage	GPCRs, ion channels and orphan targets are underrepresented	Kinase-centred benchmarks [17, 18]	Curated benchmarks spanning diverse protein families [19, 21, 54]
G5	Limited interpretability	Predictions without a binding rationale are hard to act on	Attention and gradient explanations remain partial [33, 36, 39]	Residue-level attribution validated against known binding sites
G6	Under-use of 3D structure	Binding is a geometric event modelled from 1D or 2D input	Structure-based models confined to PDBbind scale [23, 43, 44]	Structure-aware and equivariant models at proteome scale [46, 60, 61, 66]
G7	No uncertainty estimates	A point prediction gives no basis for prioritising assays	Prediction intervals rare outside SimBoost [16]	Bayesian, ensemble and conformal prediction with active learning
G8	Reproducibility and unfair baselines	Differing splits, seeds and preprocessing make numbers incomparable	Saturated, near-identical scores in Table 6 [30, 32, 33]	Shared pipelines and leaderboards [53, 54] with released code and splits
G9	Static, context-free predictions	Binding depends on mutation, isoform, tissue and co-factors	Benchmarks use one canonical sequence per target [24, 65]	Mutation- and context-aware models using variant and expression data
G10	Limited prospective validation	Few models are ever tested on new experiments	Experimental follow-up reported only occasionally [11, 40]	Prospective screening with laboratory confirmation, failures included

Contribution of Table 7. This table is the analytical core of the review. Linking each gap to specific evidence in the surveyed literature and to a concrete direction allows a reader to select a problem with a defined point of departure rather than a general theme. It also makes visible something easy to miss

when reading the literature paper by paper: the most pressing issues — G1, G3 and G8 — concern evaluation rather than architecture. That observation reframes progress in this field as a benchmarking problem at least as much as a modelling one. Figure 9 groups the ten problems into three themes and pairs each theme with the work needed to close it.

Ten open problems, grouped into three themes

Each theme maps to the directions most likely to resolve it



Three of the ten problems concern evaluation rather than architecture, which is why benchmarking is the field's more urgent task.

Figure 9. The ten open problems fall into three themes; three of them concern evaluation rather than architecture.

3.8. Future Directions

3.8.1. Structure-aware foundation models

With structural coverage now available at proteome scale [46, 66], models can condition on a predicted binding pocket rather than on raw sequence. Equivariant architectures and diffusion-based docking [60, 61] point towards a pipeline in which pose prediction and affinity estimation are trained jointly, with pocket confidence used to weight the resulting prediction — an approach that addresses G6 directly.

3.8.2. Pretrained multimodal encoders

Protein and chemical language models already transfer well to small datasets [47–49]. Contrastive pretraining over large bioactivity corpora [21] could go further by aligning the chemical and protein spaces before any supervised fine-tuning, which is the most promising route to the cold-start problem (G2).

3.8.3. Reliable negatives and positive-unlabelled learning

Treating unmeasured pairs as negatives corrupts both training and evaluation (G3). Curating experimentally confirmed non-binders and adopting positive-unlabelled objectives would make reported precision interpretable, and would reduce the ligand-bias artefacts that label-reversal experiments have exposed [34].

3.8.4. Uncertainty-aware and active learning

Affinity estimates with calibrated intervals, as SimBoost provides [16], would let a team decide which assays to run next. Combining conformal prediction with iterative experiment selection would turn these models from ranking tools into instruments of experimental design, addressing G7.

3.8.5. Interpretability validated against biology

Attention maps should be scored against known binding residues and mutagenesis data rather than inspected qualitatively [33, 36, 39]. Only then do interpretability claims become testable, and only then can a medicinal chemist use them with confidence (G5).

3.8.6. Context- and mutation-aware prediction

Incorporating sequence variants, isoforms and tissue expression [24, 65] would allow models to predict how a resistance mutation alters affinity. This is where much of the clinical value lies, particularly in oncology and infectious disease, and it remains almost untouched (G9).

3.8.7. Standardised and more demanding benchmarks

Community platforms [53, 54] should define cold-start, scaffold-split and time-split tasks across diverse protein families, with hidden test sets. Without them, progress is asserted rather than measured, and the saturation visible

in Figure 7 will continue to be mistaken for convergence (G1, G4, G8).

3.8.8. Federated and privacy-preserving training

A large share of high-quality bioactivity data remains proprietary. Federated learning would let institutions contribute to a shared model without exchanging raw data, enlarging the effective training set without breaching confidentiality—an approach already established in other clinical machine learning settings.

3.8.9. Prospective experimental validation

Finally, the field needs more studies that carry predictions to the bench and publish both confirmations and failures [11, 40, 59]. Retrospective benchmarks can rank methods against one another, but only prospective work establishes a real-world hit rate, which is the number a discovery team actually needs (G10).

4. Conclusions

Interaction prediction has moved from similarity kernels and matrix factorisation to graph neural networks, attention mechanisms and structure-aware deep models, reducing reported error on the standard affinity benchmarks by more than half. This review classifies the field into five method families and compares datasets, representations, architectures, metrics, and reported results across seven tables and nine figures, stating in each case what the comparison contributes.

The evidence assembled here points to one conclusion above the others: architectural progress has outrun evaluation practice. Gains on warm-start Davis and KIBA splits have saturated, while performance under cold-start and cross-domain conditions — the conditions that correspond to real discovery — remains substantially weaker and is seldom reported. Set against unreliable negative samples, narrow protein-family coverage, and rarely validated interpretability, benchmarking is the more urgent of the two problems.

Which directions, then, are most likely to produce genuine progress? Three stand out: combining pretrained protein and chemical language models with proteome-scale structural information, adding calibrated uncertainty so that predictions can guide experimental effort, and assessing models on standardised cold-start benchmarks with prospective laboratory validation. Measured that way, interaction prediction can become a dependable component of the discovery pipeline rather than a leaderboard exercise.

Author Contributions: Conceptualization, methodology, literature screening, data extraction, formal analysis, visualization, writing—original draft preparation, and writing—review and editing were carried out by the author(s). All

authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable. This study is a systematic review of published literature and does not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data analysed in this review were obtained from the publicly available studies and databases cited in the reference list. The extraction sheet compiled during the review is available from the corresponding author on reasonable request.

Acknowledgments: The authors thank colleagues who provided helpful discussion during the preparation of this review.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3), 203–214.
- DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47, 20–33.
- Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics*, 17(4), 696–712.
- Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1), 41–58.
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), 455–461.
- Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2), 197–206.
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13), i232–i240.
- Bleakley, K., & Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18), 2397–2403.
- van Laarhoven, T., Nabuurs, S. B., & Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21), 3036–3043.
- Mei, J.-P., Kwok, C.-K., Yang, P., Li, X.-L., & Zheng, J. (2013). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2), 238–245.
- Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5), e1002503.
- Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, 28(18), 2304–2310.
- Zheng, X., Ding, H., Mamitsuka, H., & Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug–target interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1025–1033.
- Liu, Y., Wu, M., Miao, C., Zhao, P., & Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug–target interaction prediction. *PLoS Computational Biology*, 12(2), e1004760.
- Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2015). Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2), 325–337.
- He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9, 24.
- Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11), 1046–1051.
- Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., & Aittokallio, T. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3), 735–743.
- Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1), D1045–D1053.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), D1074–D1082.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), D945–D954.
- Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1), D380–D384.
- Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., et al. (2015). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3), 405–412.
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353–D361.
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742–754.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36.
- Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17), i821–i829.
- Öztürk, H., Ozkirimli, E., & Özgür, A. (2019). WideDTA: prediction of drug–target binding affinity. *arXiv preprint arXiv:1902.04166*.
- Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, 15(6), e1007129.
- Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8), 1140–1147.
- Tsubaki, M., Tomii, K., & Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2), 309–318.
- Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., & Wei, Z. (2020). Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, 10(35), 20701–20712.
- Yang, Z., Zhong, W., Zhao, L., & Chen, C. Y.-C. (2022). MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, 13(3), 816–833.

34. Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., et al. (2020). TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16), 4406–4414.
35. Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2021). MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, 37(6), 830–836.
36. Zhao, Q., Zhao, H., Zheng, K., & Wang, J. (2022). HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3), 655–662.
37. Bai, P., Miljković, F., John, B., & Lu, H. (2023). Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence*, 5, 126–136.
38. Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18), 3329–3338.
39. Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., & Zhang, P. (2018). Interpretable drug target prediction using deep neural representation. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3371–3377.
40. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8, 573.
41. Wan, F., Hong, L., Xiao, A., Jiang, T., & Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, 35(1), 104–111.
42. Abbasi, K., Razzaghi, P., Poso, A., Amanlou, M., Ghasemi, J. B., & Masoudi-Nejad, A. (2020). DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, 36(17), 4633–4642.
43. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., & Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*, 34(21), 3666–3674.
44. Jiménez, J., Škalič, M., Martínez-Rosell, G., & De Fabritiis, G. (2018). KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2), 287–296.
45. Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., & Kim, W. Y. (2019). Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of Chemical Information and Modeling*, 59(9), 3981–3988.
46. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
47. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118.
48. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., et al. (2022). ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 7112–7127.
49. Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
51. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *Proceedings of the 5th International Conference on Learning Representations*.
52. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *Proceedings of the 6th International Conference on Learning Representations*.
53. Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., & Sun, J. (2021). DeepPurpose: a deep learning library for drug-target interaction prediction. *Bioinformatics*, 36(22–23), 5545–5547.
54. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., et al. (2021). Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. *Proceedings of the NeurIPS Track on Datasets and Benchmarks*.
55. Ezzat, A., Wu, M., Li, X.-L., & Kwok, C.-K. (2019). Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, 20(4), 1337–1357.
56. Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., & Najarian, K. (2021). Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. *Briefings in Bioinformatics*, 22(1), 247–269.
57. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477.
58. Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., & Cheng, F. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science*, 11(7), 1775–1797.
59. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688–702.
60. Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2023). DiffDock: diffusion steps, twists, and turns for molecular docking. *Proceedings of the 11th International Conference on Learning Representations*.
61. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., & Jaakkola, T. (2022). EquiBind: geometric deep learning for drug binding structure prediction. *Proceedings of the 39th International Conference on Machine Learning*, 20503–20521.
62. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
63. Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug-target interaction prediction. *Journal of Proteome Research*, 16(4), 1401–1409.
64. Zhao, Q., Xiao, F., Yang, M., Li, Y., & Wang, J. (2019). AttentionDTA: prediction of drug-target binding affinity using attention model. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 64–69.
65. UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489.
66. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630, 493–500.
67. Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6), 1273–1280.