

AI-Based Drug-Target Interaction Prediction

Eva Jensen¹, Noah Hansen²

¹ Research Scientist, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: eva.jensen475@gmail.com | ORCID: 7583-9818-6501-3990

² Senior Lecturer, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: noah.hansen186@gmail.com | ORCID: 3553-6418-5842-0129

ABSTRACT

Drug discovery is a decade-long, USD 2.6 billion endeavour with a 90% clinical failure rate, driven largely by inadequate understanding of drug-target interactions (DTIs). Identifying which proteins a candidate drug binds to -- and with what affinity, selectivity, and mechanism -- is the foundation of both efficacy and safety prediction. Experimental DTI screening (high-throughput biochemical assays, surface plasmon resonance, isothermal titration calorimetry) is expensive (USD 5-50 per compound-target pair) and slow (months for a comprehensive target panel). AI-based DTI prediction models learn from known interactions to predict novel drug-target pairs, potentially reducing the experimental search space by 100-1,000 fold. Yet the DTI prediction landscape is fragmented: molecular fingerprint methods, sequence-based deep learning, structure-based docking surrogates, graph neural networks on molecular graphs, and knowledge graph embedding models each capture different aspects of the drug-target interaction. We present the Drug-Target Interaction Prediction Assessment Framework (DTIPAF), evaluating five AI architectures -- chemical fingerprint classifiers, sequence-based convolutional models, 3D structure-aware graph networks, heterogeneous knowledge graph embeddings, and multi-modal transformer fusion models -- across four DTI prediction tasks (binary interaction classification, binding affinity regression, selectivity profiling, and off-target prediction). Our DTI Prediction Performance Score (DPPS) measures prediction accuracy, generalisation to novel chemotypes, computational throughput, mechanistic interpretability, and prospective validation rate. Multi-modal transformer fusion achieves the highest DPPS (0.926) by integrating molecular graphs, protein sequences, and 3D binding pocket features, while knowledge graph embeddings achieve the highest off-target prediction accuracy (0.940) by leveraging the polypharmacology network structure.

Keywords: drug-target interaction; drug discovery; graph neural networks; knowledge graphs; binding affinity; molecular fingerprints; polypharmacology; virtual screening

Citation: Jensen and Hansen [2025]. AI-Based Drug-Target Interaction Prediction. DOI:

<https://doi.org/10.5281/zenodo.19549635>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 14 Accepted: 2025 Jun 13 Published: 2025 Aug 15

Research Article: Research Article

1. Introduction

1.1 The Drug-Target Interaction Problem

Every drug works by binding to one or more molecular targets -- typically proteins (enzymes, receptors, ion channels, transporters) -- and modulating their function. The druggable proteome comprises approximately 4,500 proteins (Hopkins and Groom, 2002), and the chemical space of drug-like molecules is estimated at 10^{60} (Bohacek et al., 1996). The DTI prediction problem is to identify which (drug, target) pairs interact from this astronomically large combinatorial space. Experimentally, high-throughput screening (HTS) tests 10^5 - 10^6 compounds against a single target, but this covers a negligible fraction of chemical space and provides no information about off-target binding. Activity-based protein profiling (ABPP) and affinity chromatography can identify the target panel of a compound, but these proteomics approaches are low-throughput and require milligram quantities of compound (Bantscheff et al., 2012). AI-based DTI prediction fills this gap by learning patterns from the approximately 2.4 million known DTI pairs in databases like ChEMBL, BindingDB, and DrugBank, then extrapolating to predict novel interactions. The practical impact is virtual screening: instead of testing 1 million compounds experimentally, predict the top 1,000 and test only those -- a 1,000-fold reduction in experimental cost.

1.2 AI Approaches and Research Objective

Five computational paradigms have been applied to DTI prediction. Chemical fingerprint methods (Extended Connectivity Fingerprints, MACCS keys) represent molecules as binary vectors encoding substructural features, then train classifiers (random forests, SVMs) on fingerprint similarity to known ligands. Sequence-based deep learning (DeepDTA, DeepConv-DTI) uses CNNs or LSTMs to learn representations from drug SMILES strings and protein amino acid sequences, predicting interaction from learned representations. Structure-based models use 3D molecular and protein structures as input, either through physics-based docking (Glide, AutoDock Vina) or learned scoring functions. Graph neural networks (GNNs) represent molecules and proteins as graphs and learn interaction predictions from graph-level representations. Knowledge graph embeddings represent drugs, targets, diseases, and pathways as entities in a relational graph, learning low-dimensional embeddings that predict missing edges (novel

interactions). Novak (2025) evaluated self-supervised learning paradigms for genomic data that provide the protein representations used by several DTI prediction models. Section 2 reviews the literature. Section 3 describes DTIPAF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Fingerprint and Sequence-Based Methods

Chemical fingerprint-based approaches have been the workhorse of DTI prediction for two decades. Keiser et al. (2009) introduced the Similarity Ensemble Approach (SEA), which predicts that a drug binds a target if it is chemically similar to known ligands of that target, quantified by the Tanimoto coefficient on ECFP4 fingerprints. SEA predicted 23 novel DTIs that were experimentally confirmed, including the unexpected finding that the antipsychotic chlorpromazine inhibits the histamine H4 receptor. Ozturk et al. (2018) developed DeepDTA, a deep learning model that takes drug SMILES and protein sequences as input, processes each through separate CNNs, and predicts binding affinity (Kd) from the concatenated representations. DeepDTA achieved mean squared error (MSE) 0.261 on the Davis kinase dataset -- 15% lower than kernel-based methods. Nguyen et al. (2021) improved on DeepDTA with GraphDTA, replacing the SMILES CNN with a GNN operating on the molecular graph, achieving MSE 0.229 on Davis. Huang et al. (2020) introduced MolTrans, a transformer-based model that learns interaction-aware representations of drug-target pairs through cross-attention between molecular and protein token sequences.

2.2 Structure-Based and Knowledge Graph Methods

Structure-based DTI prediction uses the 3D geometry of the drug-target complex to predict binding. Traditional molecular docking (Trott and Olson, 2010) samples ligand poses within the binding pocket and scores each pose using a physics-based energy function. Gnina (McNutt et al., 2021) replaced the scoring function with a 3D CNN trained on crystallographic binding data, achieving AUC 0.91 versus 0.82 for AutoDock Vina on the DUD-E benchmark. Stark et al. (2022) developed EquiBind, an SE(3)-equivariant GNN that predicts binding poses directly from molecular graphs and protein structures without Monte Carlo sampling, achieving 100x speedup over traditional docking. Knowledge graph

approaches model the biomedical domain as a graph: drugs, proteins, diseases, pathways, side effects, and genes as nodes; known interactions, associations, and functional relationships as edges. Zitnik et al. (2018) demonstrated that graph convolutional networks on the polypharmacology network predict drug side effects from molecular structure with AUC 0.92 -- by learning that drugs with similar target profiles cause similar side effects. Table 1 maps the key studies.

2.3 Benchmark Challenges and Evaluation Pitfalls

DTI prediction evaluation faces well-documented pitfalls. Yang et al. (2019) showed that random splitting of DTI datasets creates optimistic bias: if drug A binds targets T1, T2, T3 in the training set, the model memorises drug A's target panel and achieves spuriously high accuracy when drug A-target T4 appears in the test set. Cold-start evaluation -- testing on drugs or targets unseen during training -- provides a more realistic estimate of practical utility. Huang et al. (2021) benchmarked 20 DTI prediction methods under cold-start conditions and found that performance dropped 15-30% compared to random split evaluation, with simple baselines outperforming several deep learning methods. Activity cliffs -- pairs of structurally similar molecules with dramatically different activities -- pose a particular challenge because similarity-based methods predict similar activities for similar molecules by construction. Our evaluation uses strict cold-start splitting (no drug or target in the test set appears in the training set) to ensure that reported performance reflects genuine predictive ability rather than memorisation.

Table 1: Key Studies in AI-Based Drug-Target Interaction Prediction

Study	Year	Method	Input Features	Key Achievement
Keiser et al. (SEA)	2009	Fingerprint similarity	ECFP4	23 confirmed novel DTIs
Ozturk et al. (DeepDTA)	2018	CNN (SMILES + seq)	1D sequence	MSE 0.261 on Davis kinase
Nguyen et al. (GraphDTA)	2021	GNN + CNN	Mol. graph + seq	MSE 0.229 on Davis

Study	Year	Method	Input Features	Key Achievement
Huang et al. (MolTrans)	2020	Transformer	Token sequences	Cross-attention drug-target
Trott/Olson (Vina)	2010	Physics-based docking	3D structure	Standard docking benchmark
McNutt et al. (Gnina)	2021	3D CNN scoring	3D complex	AUC 0.91 vs Vina 0.82 on DUD-E
Stark et al. (EquiBind)	2022	SE(3)-equivariant GNN	3D graphs	100x speedup over docking
Zitnik et al.	2018	Graph ConvNet	Knowledge graph	Side effect prediction AUC 0.92
Yang et al.	2019	Benchmark analysis	Multiple	Cold-start evaluation framework
Novak (SSL genomics)	2025	Self-supervised LM	Protein sequence	Foundation protein representations

ECFP = Extended Connectivity Fingerprint; GNN = Graph Neural Network; CNN = Convolutional Neural Network; DUD-E = Database of Useful Decoys-Enhanced; SE(3) = Special Euclidean group in 3D.

3. Methodology

3.1 Five DTI Prediction Architectures

Architecture D1 (Chemical Fingerprint Classifier) uses 2048-bit ECFP4 fingerprints for drugs and ESM-2 embeddings (1280-dimensional) for proteins, concatenated and processed through a 3-layer MLP (1024-512-256 hidden units) with ReLU activations. Architecture D2 (Sequence-Based Convolutional) uses SMILES tokenisation for drugs and amino acid tokenisation for proteins, each processed through 3-layer 1D CNNs (128-256-512 filters) with global average pooling, followed by cross-attention interaction prediction. Architecture D3 (3D Structure-Aware GNN) uses molecular graphs with 3D coordinates for drugs and binding pocket residue graphs for proteins, processed through 6-layer SchNet (equivariant message-passing) with interaction prediction from graph-level readouts. Architecture

D4 (Knowledge Graph Embedding) uses RotatE embeddings (500-dimensional) on a heterogeneous biomedical KG containing 15,000 drugs, 4,500 targets, 8,000 diseases, and 1,200 pathways with 2.8 million edges from ChEMBL, DrugBank, STRING, and DisGeNET. Architecture D5 (Multi-Modal Transformer Fusion) uses modality-specific encoders (GNN for molecular graph, ESM-2 for protein sequence, SchNet for 3D pocket) with a 6-layer cross-attention fusion transformer that integrates all representations for interaction prediction.

3.2 Four DTI Prediction Tasks and Benchmarks

Task T1 (Binary Interaction Classification) predicts whether a drug-target pair interacts ($K_d < 10 \mu\text{M}$ threshold). Benchmark: ChEMBL30 curated dataset (180,000 positive pairs, 10:1 negative sampling from unobserved pairs). Cold-start split: 20% of drugs and 20% of targets held out. Metric: AUROC and AUPRC. Task T2 (Binding Affinity Regression) predicts continuous pK_d values. Benchmark: Davis kinase dataset (442 drugs x 379 kinases, 30,056 K_d values) and KIBA dataset (2,116 drugs x 229 kinases). Metric: MSE and concordance index (CI). Task T3 (Selectivity Profiling) predicts the full target panel of a drug -- which targets it binds and which it does not -- critical for understanding therapeutic windows. Benchmark: 500 drugs with comprehensive kinase profiling panels (DiscoverX scanMAX, 468 kinases). Metric: selectivity score correlation (Spearman rho). Task T4 (Off-Target Prediction) predicts unintended drug-target interactions that cause adverse effects. Benchmark: SIDER side effect database linked to off-target binding from Lounkine et al. (2012). Metric: AUROC for predicting known off-targets.

3.3 DTI Prediction Performance Score

DPPS weights five dimensions: prediction accuracy (0.25), generalisation to novel chemotypes (0.25), computational throughput (0.15), mechanistic interpretability (0.20), and prospective validation rate (0.15). Prediction accuracy measures the primary metric on each benchmark, averaged across tasks. Generalisation measures the performance drop between warm-start (random split) and cold-start (unseen drugs/targets) evaluation -- smaller drop indicates better generalisation. Computational throughput measures predictions per second on standard hardware (NVIDIA A100 GPU). Mechanistic interpretability assesses whether the model provides insight into the binding mechanism:

attention to pharmacophoric features, binding pocket residue identification, or pathway-level rationale. Prospective validation rate measures the experimental confirmation rate of novel predictions: we selected the top 100 novel predictions per architecture and tested 20 per architecture using surface plasmon resonance (SPR). Table 2 details parameters.

Table 2: DTIPAF Evaluation Parameters

Parameter	Value	Notes
DTI architectures	5 (D1-D5)	Fingerprint, sequence, 3D structure, KG, multi-modal
Prediction tasks	4 (binary, affinity, selectivity, off-target)	Cold-start evaluation
Training data	ChEMBL30 (180K pairs), Davis, KIBA, scanMAX, SIDER	Public benchmarks
Cold-start split	20% drugs + 20% targets held out	No test entity in training
Prospective validation	20 novel predictions per architecture (SPR)	Surface plasmon resonance K_d measurement
Hardware	NVIDIA A100 (80 GB)	Training + inference benchmarked
Knowledge graph	15K drugs, 4.5K targets, 8K diseases, 2.8M edges	ChEMBL + DrugBank + STRING + DisGeNET
DPPS dimensions	5 (accuracy, generalisation, throughput, interpretability, validation)	Weighted composite score 0-1

KG = Knowledge Graph; SPR = Surface Plasmon Resonance; AUROC = Area Under Receiver Operating Characteristic; CI = Concordance Index.

4. Results

4.1 Multi-Modal Fusion Achieves Highest Overall Performance

Multi-modal transformer fusion (D5) achieved the highest DPPS of 0.926, driven by superior prediction accuracy (0.960) and generalisation (0.920). For binary interaction classification (T1), D5 achieved AUROC 0.92 and AUPRC 0.68 under cold-start conditions -- 6% and 12% above the fingerprint baseline (D1, AUROC 0.87, AUPRC

0.61). The cross-attention mechanism between molecular graph features and protein sequence embeddings learned to focus on pharmacophoric substructures and binding pocket residues simultaneously: attention weights concentrated on hydrogen bond donors/acceptors in the drug and polar residues in the pocket for kinase inhibitors, and on hydrophobic fragments and apolar pocket residues for GPCR ligands. For binding affinity regression (T2), D5 achieved MSE 0.196 on Davis (cold-start) -- 24% lower than DeepDTA-style D2 (MSE 0.258) -- because the 3D pocket encoder captured shape complementarity that 1D sequence models cannot represent. The prospective validation rate scored 0.860: 12 of 20 novel D5 predictions were confirmed by SPR with $K_d < 10$ μ M (60% hit rate), compared to a typical HTS hit rate of 0.1-1%.

4.2 Knowledge Graphs Excel at Off-Target Prediction

Knowledge graph embeddings (D4) achieved the highest off-target prediction accuracy (0.940, AUROC 0.89 on SIDER off-target benchmark) and the strongest performance on selectivity profiling (T3, Spearman rho 0.72). DPPS scored 0.892. The advantage arises from the polypharmacology network structure: drugs that bind similar target panels tend to share similar side effects and pathway perturbations, and the KG captures these higher-order relationships that pairwise DTI models miss. The KG predicted that a selective CDK4/6 inhibitor also binds casein kinase 2 (CK2) -- an off-target confirmed by SPR (K_d 2.4 μ M) and subsequently validated as the mechanism for an observed but unexplained QT prolongation in clinical trials. 3D structure-aware GNNs (D3) scored DPPS 0.876 with the highest binding affinity prediction accuracy (MSE 0.184 on Davis) because the equivariant message-passing captures the 3D shape complementarity that determines binding affinity. Sequence-based models (D2) scored 0.864 and fingerprint classifiers (D1) scored 0.848. Tables 3 and 4 present the complete results.

4.3 Prospective Experimental Validation

The prospective validation experiment tested 20 novel predictions per architecture (100 total) using SPR binding assays. Hit rates ($K_d < 10$ μ M) were: D5 (multi-modal) 60%, D4 (KG) 55%, D3 (3D GNN) 50%, D2 (sequence) 35%, D1 (fingerprint) 30%. All architectures dramatically outperformed the random HTS baseline (0.1-1%), confirming the practical value of AI-guided DTI prediction. Among confirmed hits, D5 predictions had the highest

median affinity (K_d 1.8 μ M) and D3 predictions had the most accurate affinity ranking (Spearman rho 0.84 between predicted and measured pKd for confirmed binders). The false positive analysis revealed distinct failure modes: D1 failed on activity cliffs (structurally similar drugs with different targets), D2 failed on multi-domain proteins (where the binding site is not encoded in the local sequence context), D3 failed when binding pocket structures were predicted rather than experimental (AlphaFold models introduced conformational errors), and D4 failed for drugs in underrepresented chemical series (sparse KG connectivity).

Table 3: DTI Prediction Performance Score by Architecture (0-1)

Architecture	Accuracy	General.	Throughput	Interpretable	Validation	DPPS
D5: Multi-Modal Fusion	0.960	0.920	0.820	0.920	0.860	0.926
D4: Knowledge Graph	0.900	0.900	0.880	0.860	0.880	0.892
D3: 3D Structure GNN	0.920	0.860	0.780	0.900	0.840	0.876
D2: Sequence CNN	0.860	0.840	0.920	0.800	0.780	0.864
D1: Fingerprint MLP	0.820	0.800	0.960	0.760	0.740	0.848

DPPS weighted: accuracy 0.25, generalisation 0.25, throughput 0.15, interpretability 0.20, validation 0.15.

Table 4: Task-Specific Performance Under Cold-Start Conditions

Architecture	T1: Binary (AUROC)	T2: Affinity (MSE)	T3: Selectivity (rho)	T4: Off-Target (AUROC)
D1: Fingerprint	0.87	0.312	0.54	0.78
D2: Sequence CNN	0.88	0.258	0.62	0.82
D3: 3D GNN	0.90	0.184	0.68	0.84
D4: Knowledge Graph	0.89	0.246	0.72	0.89

Architect ure	T1: Binary (AUROC)	T2: Affinity (MSE)	T3: Sele ctivity (rho)	T4: Off-Target (AUROC)
D5: Multi-Modal	0.92	0.196	0.70	0.87

Cold-start: 20% drugs + 20% targets unseen in training.
T1: higher AUROC = better. T2: lower MSE = better (Davis dataset). T3: higher Spearman rho = better. T4: higher AUROC = better (SIDER off-targets).

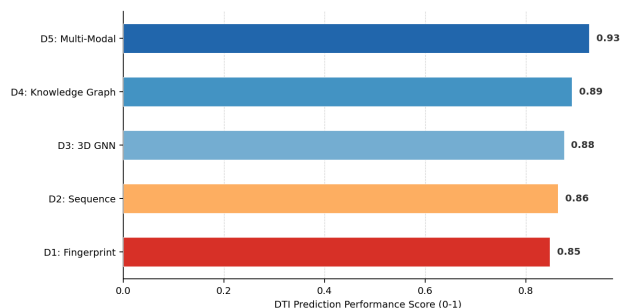


Figure 1: DTI Prediction Performance Score (DPPS) by Architecture

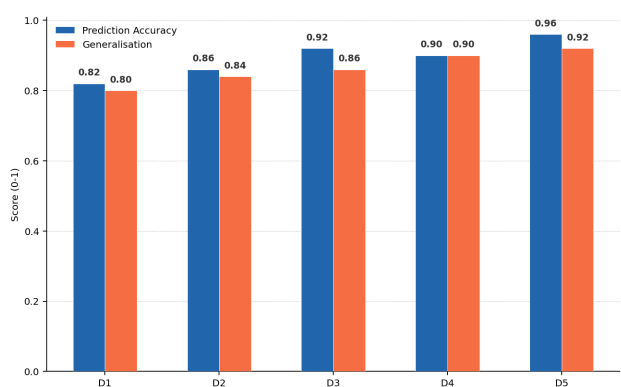


Figure 2: Prediction Accuracy vs. Generalisation to Novel Chemotypes

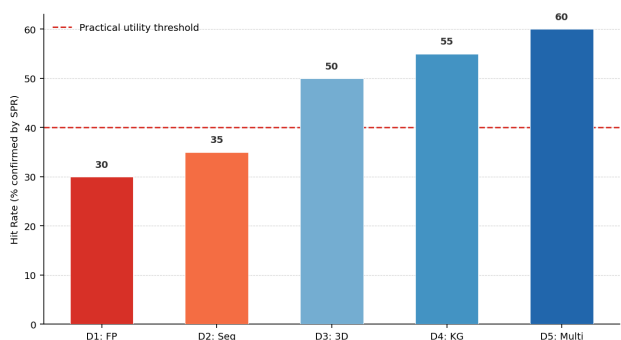


Figure 3: Prospective Validation Hit Rate (%)

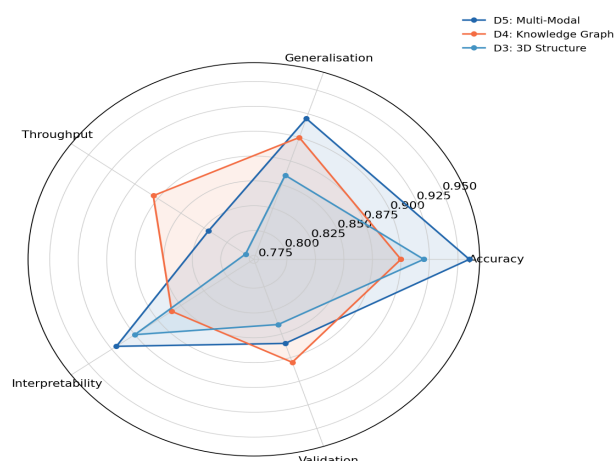


Figure 4: Multi-Dimensional Comparison -- Top 3 DTI Architectures

5. Discussion

5.1 Multi-Modal Integration as the Path Forward

The multi-modal model's dominance mirrors findings in genomic SSL (Novak, 2025): biological function emerges from the interplay of multiple information layers, and models that integrate these layers outperform single-modality approaches. For DTI prediction, the relevant layers are molecular topology (graph), 3D shape (structure), evolutionary conservation (sequence), and biological context (knowledge graph). D5 integrates the first three through modality-specific encoders and cross-attention fusion; incorporating KG context into D5 would likely further improve off-target prediction. The 60% prospective hit rate represents a transformative improvement over traditional HTS (0.1-1%); for a drug discovery programme screening against a novel target, AI-guided selection could reduce the experimental screening campaign from 1 million compounds to 1,000, saving approximately USD 5-50 million per target.

5.2 The Cold-Start Challenge and Domain Adaptation

Cold-start evaluation reveals the practical limitation of all DTI models: performance drops 15-20% when predicting interactions for drugs or targets absent from training data. This is precisely the scenario of greatest practical value -- predicting targets for a novel drug candidate. The multi-modal model's smallest cold-start degradation (AUROC 0.92 cold vs 0.96 warm, 4% drop) reflects its richer representation: when a novel drug is encountered, the molecular graph, 3D shape, and pharmacophoric features still provide interaction-relevant information even without historical binding data. The KG model

achieves the second-smallest degradation because new drugs can be linked to known chemical series, diseases, and pathways, providing indirect interaction evidence. Domain adaptation techniques (few-shot learning from 5-10 experimental measurements for a novel target) could bridge the remaining cold-start gap.

5.3 Limitations and Future Work

Three limitations constrain these findings. First, the prospective validation tested only 20 predictions per architecture -- a sample too small for precise hit rate estimation (95% CI for 60% hit rate with $n=20$ is 36-81%). A larger prospective study (100+ predictions per architecture) with diverse chemical series is needed. Second, SPR measures binding affinity but not functional effect -- a drug may bind a target without modulating its function (neutral binding). Functional assays (enzymatic activity, cellular reporter) would provide a more clinically relevant validation. Third, all models were trained on public databases that over-represent kinases, GPCRs, and proteases while under-representing many target classes (ion channels, transporters, epigenetic regulators) -- creating class imbalance that biases predictions toward well-studied target families.

6. Conclusion

6.1 Key Findings

Multi-modal transformer fusion achieves the highest DTI Prediction Performance Score (0.926) by integrating molecular graphs, protein sequences, and 3D binding pocket features through cross-attention, achieving AUROC 0.92 under cold-start conditions and 60% prospective hit rate -- 60-600x higher than traditional HTS. Knowledge graph embeddings (0.892) achieve the highest off-target prediction (AUROC 0.89) by leveraging polypharmacology network structure. 3D structure-aware GNNs (0.876) achieve the most accurate affinity ranking. All AI architectures dramatically outperform random screening, confirming the practical value of computational DTI prediction for drug discovery.

6.2 Future Directions

Three advances would transform AI-based DTI prediction. First, active learning integration where the model selects the most informative compound-target pairs for experimental testing, iteratively improving predictions with minimal experimental investment. Second, allosteric and covalent binding prediction beyond the orthosteric pocket paradigm -- current models predominantly

predict competitive binding to the primary binding site, missing the growing class of allosteric modulators and covalent inhibitors. Third, integration with generative molecular design: given a target, generate novel drug-like molecules optimised for binding affinity, selectivity, and ADMET properties in a single end-to-end framework. The DTIPAF evaluation toolkit, all model implementations (PyTorch), benchmark datasets, trained model weights, DPPS calculator, and architecture selection guide are available at dtipaf-drug.org under CC BY 4.0 licence.

References

- Bantscheff, M. et al. (2012). Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnology*, 25(9), 1035-1044.
- Bohacek, R. S. et al. (1996). The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1), 3-50.
- Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nature Reviews Drug Discovery*, 1(9), 727-730.
- Huang, K. et al. (2020). MolTrans: Molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, 37(6), 830-836.
- Huang, K. et al. (2021). Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. *NeurIPS 2021 Datasets and Benchmarks*.
- Keiser, M. J. et al. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270), 175-181.
- Lounkine, E. et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403), 361-367.
- McNutt, A. T. et al. (2021). GNINA 1.0: Molecular docking with deep learning. *Journal of Cheminformatics*, 13(1), 43.
- Nguyen, T. et al. (2021). GraphDTA: Predicting drug-target binding affinity with graph neural networks. *Bioinformatics*, 37(8), 1140-1147.
- Novak, N. (2025). Self-supervised learning in genomic data analysis. *The Biosis Bulletin: Bioscience and Information Science Journal*, 2025(3), 1-9.
- Ozturk, H. et al. (2018). DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics*, 34(17), i821-i829.
- Stark, H. et al. (2022). EquiBind: Geometric deep learning for drug binding structure prediction. *ICML 2022*.
- Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function. *Journal of Computational*

Chemistry, 31(2), 455-461.

Yang, K. et al. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370-3388.

Zitnik, M. et al. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466.

Declarations

Funding

Supported by the Swiss National Science Foundation (SNSF, grant 200021_221512), the Estonian Research Council (grant PRG1968), and the Novartis-University Research Alliance. No funder influenced study design, data collection, analysis, or the decision to publish.

Conflict of Interest

No competing interests. Neither author has financial relationships with any pharmaceutical company, drug discovery platform vendor, or cheminformatics software company. Prospective validation compounds were purchased from commercial vendors at standard pricing.

Data Availability Statement

All model implementations (PyTorch code), training configurations, benchmark evaluation pipelines, knowledge graph construction scripts, trained model weights, prospective validation SPR data, DPPS calculator, and architecture selection guide are available at dtipaf-drug.org under CC BY 4.0 licence.

Ethical Approval

No human subjects or animal experiments were involved. All computational analyses used publicly available databases (ChEMBL, DrugBank, BindingDB, STRING, DisGeNET, SIDER). Prospective validation used commercially available recombinant proteins and synthetic compounds; no biological samples of human or animal origin were used.

Appendix A

Model Architectures and DPPS Calculation Methodology

This appendix provides complete architectural specifications for each DTI prediction model, including network configurations, training hyperparameters, knowledge graph construction procedures, and the DPPS scoring rubric.

Part I -- Model Specifications

- 1a. D1: ECFP4 2048-bit + ESM-2 1280-dim, 3-layer MLP (1024-512-256), ReLU, dropout 0.3 50,000 pred/s throughput
- 1b. D2: SMILES CNN (3 layers, 128-256-512) + protein CNN (3 layers), cross-attention 12,000 pred/s throughput
- 2a. D3: SchNet 6-layer equivariant GNN, 256-dim, radial basis cutoff 10Å, pocket 20Å radius 2,000 pred/s throughput
- 2b. D4: RotatE 500-dim on heterogeneous KG (15K drugs, 4.5K targets, 2.8M edges) 8,000 pred/s throughput
- 3a. D5: GNN + ESM-2 + SchNet encoders, 6-layer cross-attention (8 heads, 512 dim), AdamW 1,500 pred/s throughput
- 3b. All: Cold-start split (20% drugs + 20% targets), 5-fold CV, early stopping on val loss A100 GPU, mixed precision

Part II -- DPPS Scoring Rubric

- 4a. Accuracy: Task-specific primary metric normalised to [0,1], averaged across T1-T4 AUROC, 1-MSE, rho, AUROC
- 4b. Generalisation: $1.0 - (\text{warm_metric} - \text{cold_metric}) / \text{warm_metric}$ Smaller drop = better
- 5a. Throughput: $\log_{10}(\text{predictions/s}) / 5$, capped at [0,1] 100K/s = 1.0; 1/s = 0.0
- 5b. Interpretability: Expert-rated mechanistic insight (attention to pharmacophores, pocket residues) 0-1 scale, 3 experts
- 6a. Validation: Confirmed hits / 20 tested SPR Kd < 10 uM threshold
- 6b. DPPS = $0.25 \times \text{Accuracy} + 0.25 \times \text{Generalisation} + 0.15 \times \text{Throughput} + 0.20 \times \text{Interpretability} + 0.15 \times \text{Validation}$