Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Model Rankings in Drug–Target Affinity Prediction Vary Across Evaluation Regimes: A Controlled Multi-Seed Study of Model-Comparison Reliability

Haobo Kui

2993272783@qq.com

Harbin Normal University

Xinyuan Cui

Harbin Normal University

Research Article

Keywords: drug–target affinity prediction, cold-start generalization, model-comparison reliability, evaluation protocol, ranking stability, reproducibility

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11338813/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Drug–target affinity (DTA) prediction directly supports compound screening and target prioritization, but different deployment settings pose different generalization demands. We therefore asked whether conclusions from model comparisons remain reliable across evaluation regimes. Using cleaned KIBA (118 036 records) and Davis (30 056 records), we constructed warm, cold-target (defined by full protein sequence), and cold-drug settings under a unified protocol. Three approximately capacity-matched cross-modal fusion models and two controlled architectural baselines were compared with identical splits, training budgets, and evaluation procedures, using three paired model seeds. The study comprised 126 prespecified training runs. Held-out test sets were opened only after training and model selection were complete, with record-level, unique-input, and integrity audits providing end-to-end traceability. Relative performance and nominal rankings changed markedly across regimes: the three-seed mean C-index rankings for KIBA cold-target versus cold-drug gave Kendall's τ = -1.00; the most consistent cross-dataset model-family differences emerged under cold-drug; and multi-seed analysis showed that a single run did not fully characterize the complete method ordering. A post-hoc VEGFR2 case, locked by performance-independent criteria before target-specific predictions were inspected, extended this pattern to compound prioritization for an unseen target. Together, these results support scenario-matched, multi-seed, auditable DTA model comparison as a more reliable basis for model selection and interpretation.

1. Introduction

Drug–target affinity (DTA) prediction aims to estimate the continuous binding strength between small molecules and target proteins and is an important component of candidate-compound screening and target prioritization in computational drug discovery[1–2]. In practical applications, DTA models face different prediction tasks. Predictions for drugs and targets already observed during training primarily represent interpolation under seen-entity conditions, whereas predictions for new compounds or new targets require generalization to unseen entities along different dimensions. Consequently, a model that performs well under one benchmark split does not necessarily retain the same relative advantage under another target generalization scenario. For drug-discovery-oriented model selection, the key question is therefore not only which model obtains a higher score, but also whether that relative conclusion remains reliable when the evaluation regime changes.

Public benchmark datasets such as KIBA and Davis have driven the development of deep-learning methods for DTA prediction [3–4]. DeepDTA established an early DTA regression framework using one-dimensional convolutional representations of SMILES strings and protein sequences [5], and GraphDTA subsequently introduced molecular graph neural networks to encode drug topology [6]. Cross-domain adaptation, Transformers, self-attention, and cross-modal interaction mechanisms have since been further explored for DTA/DTI modeling [7–13]. At the same time, generalization to unseen entities has received sustained attention. ELECTRA-DTA, ColdDTA, DebiasedDTA, and the more recent CS-DTA have examined new-drug and/or new-target scenarios from the perspectives of cold-start evaluation,

debiasing, and strict generalization protocols [14–17]. These studies establish cold-start evaluation as an important component of DTA generalization assessment. Beyond improving predictive performance under a specific protocol, however, whether the relative ordering of models remains stable across generalization regimes such as warm, cold-drug, and cold-target warrants systematic investigation as a distinct model-comparison problem.

Recent work has further advanced DTA along two complementary directions: predictive architecture improvement and generalization-oriented evaluation design. On the architectural side, a recent multimodal DTA study in this journal integrated ChemBERTa, graph neural networks, ESM-2, and three-dimensional structural modeling to improve predictive performance under standard evaluation settings [31], while EDTA combined a parallel CNN–BiLSTM architecture with attention to balance predictive performance and computational efficiency [32]. This journal has also reported ResBiGAAT for protein–ligand affinity prediction with emphasis on external-data generalization [34], as well as an application study using deep-learning-based DTI prediction for SARS-CoV-2 antiviral drug repurposing [33]. On the evaluation side, similarity-aware evaluation (SAE) reconsidered generalization on low-similarity samples [28], PCIM-DTA examined multiple scenarios including warm, cold-drug, cold-target, cold-both, and scaffold-drug [29], and SIGMA-DTA further investigated adaptive interaction modeling under cold-start and low-similarity conditions [30]. These studies strengthen representation learning, computational efficiency, application validation, or generalization assessment, but their primary objective is not to test systematically whether relative conclusions among different models remain stable across evaluation regimes under a unified and controlled protocol. We therefore address a complementary question: when training budget, data splitting, evaluation procedure, and model seeds are jointly controlled, do changes among warm, cold-drug, and cold-target alter the relative ordering of models, and are those ordering conclusions reproducible across repeated training runs?

This question is jointly influenced by the evaluation protocol, training stochasticity, and the choice of metric. A warm split allows drug and target entities to recur across training and test sets and therefore primarily measures interpolation among seen entities; this is not equivalent to data leakage. More direct evaluation contamination occurs when identical model inputs cross training, validation, and test boundaries, or when held-out test information is used for label scaling, model selection, or hyperparameter tuning. Previous studies have shown that train–test similarity, dataset bias, evaluation protocols, and performance metrics can materially affect model-comparison conclusions [18–22]. In addition, when several methods perform similarly, the complete ordering produced by a single training run may be affected by stochastic training variability. Ranking-oriented metrics such as C-index and absolute-error metrics such as RMSE or MAE may also support different model-selection conclusions. Reporting only a single split, a single run, or a single metric may therefore be insufficient to characterize the stability of comparative conclusions.

To examine these issues under controlled conditions, we used three cross-modal fusion strategies—feature concatenation, element-wise product, and token-level cross-attention—as approximately

capacity-matched experimental models. The three models provide different routes for integrating drug and protein representations while keeping the drug encoder, protein encoder, training protocol, and regression head as consistent as possible. We also included two controlled reimplementations, DeepDTA-style and GraphDTA-GCN-style, as sequence-based and graph-based architectural references, respectively. Together, these models form controlled experimental objects for testing whether relative conclusions among methods change with the intended generalization scenario under a common training and evaluation protocol.

Specifically, we established a unified three-regime evaluation framework on KIBA and Davis comprising warm, sequence-defined cold-target, and cold-drug settings. All splits ensured that identical model inputs did not cross training, validation, and test boundaries. Cold-drug isolated drug entities using canonical SMILES, whereas cold-target defined target identity by the full protein sequence rather than by the original target ID. Label scaling was fitted on the training set only, and the held-out test remained sealed until all prespecified training and model selection had been completed. The full experimental matrix covered 126 prespecified training runs and combined three paired model seeds, record-level and unique-input evaluation levels, complementary raw-scale error metrics and C-index, and auditable held-out-test provenance.

We investigated three related questions. First, do the relative rankings of DTA methods remain stable across warm, cold-target, and cold-drug evaluation regimes? Second, can the complete ordering of methods within the same regime be stably reproduced across independent training seeds? Third, do conclusions about the relative performance of model families and fusion strategies depend on the intended generalization scenario and evaluation metric? The results show clear shifts in nominal model rankings across regimes. The three-seed mean rankings for KIBA cold-target and cold-drug were completely reversed, while seed-wise results showed that this mean-level change should not be interpreted as an identical reversal in every individual training run. The most consistent cross-dataset model-family differences occurred under cold-drug generalization, whereas relative performance under warm and cold-target conditions depended more strongly on the specific dataset, model, and metric.

The central contribution of this study is a unified, controlled, multi-seed, and auditable design for comparing DTA models while explicitly tying model selection to the intended generalization setting. By evaluating warm, cold-target, and cold-drug scenarios on both KIBA and Davis, we quantify cross-regime ranking shifts and combine them with paired model seeds, complementary metrics, record-level and unique-input analyses, and a sealed held-out-test audit. This design provides reproducible evidence for distinguishing model-comparison conclusions that are consistent across scenarios from those that depend on the evaluation conditions. The overall study workflow is summarized in Fig. 1.

2. Materials and methods

2.1. Data cleaning and three evaluation regimes

We used two DTA benchmark datasets, KIBA [3] and Davis [4]. The raw inputs consisted of drug SMILES strings, target protein sequences, and continuous affinity labels. All data splits were generated once and frozen before formal training; no subsequent repartitioning or resampling was performed during training.

Data cleaning removed records with missing required fields, labels that could not be parsed as numeric values, or SMILES strings that could not be parsed by RDKit [24]. SMILES strings were canonicalized while retaining stereochemical information; no additional desalting, charge neutralization, or stereochemistry removal was performed. During cleaning and splitting, we did not further alter the affinity-label representation already present in the input files. The direction of the Davis labels was independently checked in the Section 2.9 case study using directly matchable raw Kd records. Exact duplicates were jointly defined by target ID, protein sequence, canonical SMILES, and label. Of the 118 254 original KIBA records, 218 exact duplicates were removed, leaving 118 036 records; all 30 056 Davis records were retained. After cleaning, KIBA contained 2 068 unique canonical SMILES and 229 unique protein sequences. Davis contained 68 unique drugs and 442 target IDs corresponding to only 379 unique protein sequences.

Because the actual model input consists of canonical SMILES and a protein sequence, their combination was defined as one model-input unit. The primary analysis used a keep_grouped policy: repeated measurements were retained, but records with the same model input were forced into the same data partition. When one input was associated with multiple labels, the primary dataset did not silently aggregate those records; the original records were retained and documented in the audit files. KIBA and Davis contained 288 and 546 same-input groups with conflicting labels, respectively.

The warm regime used input-group-aware splitting. Indivisible model-input groups were first defined by the pair (canonical SMILES, full protein sequence). With NumPy RandomState initialized at seed 42, these groups were randomly permuted and then assigned sequentially to the training, validation, and test partitions according to cumulative record counts, targeting an approximately 0.8/0.1/0.1 record-level ratio. All records belonging to the same model-input group therefore remained in a single partition, so identical model inputs never crossed split boundaries, although drug and protein entities themselves could recur across partitions.

For the cold-drug regime, canonical SMILES defined indivisible drug groups. Drug groups were randomly permuted with NumPy RandomState (seed 42) and assigned according to cumulative record counts to approximate the same 0.8/0.1/0.1 record-level ratio. Each canonical drug therefore belonged to exactly one partition, yielding zero canonical-SMILES overlap across training, validation, and test sets while allowing protein entities to recur across partitions.

The cold-target regime used a different construction to preserve the unseen-protein intent of the source split. Full protein sequences present only in the source test set and absent from the source training set formed the final cold-target test protein set; every sequence observed in the source training set remained in the training pool. Consequently, the single Davis sequence shared by the source training and test files

was explicitly retained in the training pool and excluded from the final test set. Unique protein sequences in the training pool were sorted, shuffled with NumPy RandomState (seed 42), and approximately 10% were sampled for validation, with the remainder used for training. Target identity was defined by the full protein sequence rather than by target ID, yielding strict sequence disjointness across the final training, validation, and test partitions.

The final KIBA training/validation/test sizes were 94 429/11 804/11 803 for warm, 85 975/12 421/19 640 for cold-target, and 94 491/11 745/11 800 for cold-drug. The corresponding Davis sizes were 24 051/3 000/3 005, 22 372/2 856/4 828, and 24 310/3 094/2 652. Because the cold-target test set inherits the source unseen-protein design and validation samples approximately 10% of the unique protein sequences in the training pool, its final record proportions were not generated by a direct record-level 0.8/0.1/0.1 target. Complete partition files, split metadata, and generation/verification scripts will be released with the formal configuration in the public code repository.

Post-split integrity checks confirmed zero cross-partition overlap of identical model inputs in all six formal splits. Canonical SMILES were strictly disjoint across training, validation, and test sets under cold-drug, and full protein sequences were strictly disjoint across the three partitions under cold-target. Training and evaluation subsequently read these frozen partitions directly; no resplitting, resampling, or split refitting was performed.

Affinity labels were min–max normalized separately for each dataset–regime combination. The scaler was fitted only on the training labels and then applied unchanged to validation and test data; validation or test values outside the training range were not clipped. Row-wise recomputation produced a maximum numerical discrepancy of approximately 1.1×10^{-16}. The formal protocol did not include cold-both (simultaneously unseen drug and target), because that setting requires additional edge removal or reassignment and therefore constitutes a separate data-reduction protocol.

2.2. Drug molecular graph representation and encoder

Each canonical SMILES [23] was converted into a two-dimensional molecular graph using RDKit [24]. Atoms were represented as nodes with 41-dimensional feature vectors encoding atom type, degree, formal charge, number of attached hydrogens, hybridization state, aromaticity, and scaled atomic mass. Bonds were expanded in both directions, with each directed edge represented by a 10-dimensional feature vector encoding bond type, stereochemistry, conjugation state, and ring membership. States outside the prespecified categories were encoded as all-zero vectors within the corresponding categorical block.

The primary drug encoder was a three-layer edge-aware graph attention network (GAT) [25] with hidden dimension 128 and four attention heads per layer, followed by BatchNorm, ReLU, and dropout ($p = 0.2$). The final node representations were aggregated by global sum pooling to produce a 128-dimensional drug-level vector. Atom-level token representations and valid-node masks were retained for AttentionFusion. Global mean pooling was used only in the prespecified sensitivity variant.

2.3. Protein sequence representation and encoder

Protein sequences were encoded using a 22-token vocabulary: PAD = 0, the 20 standard amino acids = 1–20, and unknown or other residues = 21. Each token was first mapped to a 128-dimensional embedding.

The embedded sequence was passed in parallel through three one-dimensional convolutional branches with kernel sizes 3, 5, and 7. Each branch produced 64 channels followed by ReLU. The three outputs were concatenated into a 192-dimensional residue-level representation and then projected through Linear(192→128), followed by LayerNorm, ReLU, and dropout (p = 0.2). Projected representations at PAD positions were explicitly zeroed, and masked global max pooling was performed only over non-PAD residues. The resulting 128-dimensional vector was then passed through a global MLP, Linear(128→128)→LayerNorm→ReLU→Dropout. The 128-dimensional residue-token representations before pooling were also retained for cross-modal attention.

In the primary configuration, the maximum protein length was 1 500 amino acids, with right-side truncation. Approximately 20.4% and 26.0% of KIBA and Davis records, respectively, contained sequences longer than 1 000 amino acids, whereas approximately 2.2% and 3.8% exceeded 1 500 amino acids. Thus, a small proportion of long sequences were still truncated under the 1 500-aa primary configuration. We prespecified max_len = 1000 as a single-factor sensitivity experiment to assess the effect of sequence-length choice.

2.4. Approximately capacity-matched cross-modal fusion and regression head

Both the drug and protein encoders produced 128-dimensional global representations while also providing atom-level and residue-level tokens. We compared three fusion modules, each producing a 256-dimensional output and having trainable parameter counts within a similar range.

ConcatFusion concatenated the two 128-dimensional global vectors into a 256-dimensional vector, followed by Linear(256→256)→ReLU→Dropout(0.2)→LayerNorm; the fusion module contained 66 304 parameters. ProductFusion first applied separate Linear(128→128) projections to the two modalities, computed their element-wise product, and then used Linear(128→256)→ReLU→Dropout(0.2)→LayerNorm; this module contained 66 560 parameters.

AttentionFusion implemented token-level atom-to-protein cross-attention, with drug atom tokens as queries and protein residue tokens as keys/values. Both modalities were projected from 128 dimensions into an 80-dimensional internal space and normalized with LayerNorm. Four attention heads were used, and protein PAD positions were masked through key_padding_mask. The attention output was combined with the query through a residual connection and normalization, followed by masked mean pooling over valid drug atoms. This interaction representation was then combined with the projected 80-dimensional

residual context term 0.5 × (drug_global + protein_global), followed by Linear(80→256)→ReLU→Dropout(0.2)→LayerNorm. The fusion module contained 68 288 parameters.

The three complete primary models contained 592 321, 592 577, and 594 305 trainable parameters, respectively. Each fusion module was followed by the same regression head: a 256-dimensional input passed through hidden layers of 512, 256, and 128 units, each using LayerNorm, ReLU, and dropout (p = 0.3), before producing a continuous affinity prediction. We therefore regard the three fusion strategies as approximately capacity-matched rather than strictly parameter-matched.

2.5. Training, model selection, and random-seed protocol

All formal experiments used the same frozen training protocol. AdamW [27] was used with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-5}, batch size 64, and a maximum of 200 epochs. The training loss was sample-level MSE, and gradients were clipped at a global norm of 5.0. Learning-rate scheduling used ReduceLROnPlateau with factor = 0.5, patience = 10, and a minimum learning rate of 1×10^{-6}.

Model selection was based exclusively on validation MSE, with early-stopping patience = 30 and min_delta = 0. The held-out test was not provided as input to the trainer. Training read only the fixed training and validation splits, and final test evaluation was performed in a separate stage.

All methods used the same three model seeds, 42, 123, and 2026, paired across methods for comparison. Random seeds for Python, NumPy, PyTorch, and CUDA were fixed; cuDNN deterministic mode was enabled and benchmark mode disabled. Because some GPU scatter-reduction operations in PyTorch Geometric may remain nondeterministic under specific hardware/software combinations, our objective was protocol-level reproducibility and multi-seed consistency rather than bitwise deterministic reproduction.

2.6. Evaluation metrics and robustness analysis for repeated inputs

Final evaluation followed a unified regression-metric protocol. Error metrics included MSE, RMSE, and MAE, together with the concordance index (C-index). C-index compared only sample pairs with different observed labels; tied predictions contributed 0.5 to concordance. An O(n log n) implementation was used and cross-checked against a reference implementation.

Models were trained on labels min–max scaled using the training set. Evaluation therefore produced both normalized-scale metrics and inverse-transformed raw-scale MSE, RMSE, and MAE. C-index is invariant to strictly monotonic linear scaling. The main results report raw-scale RMSE, MAE, MSE, and C-index as the primary interpretable metrics, while normalized metrics are retained for complete reproducibility and supplementary reporting.

In addition to the primary record-level analysis, we performed a unique-input robustness analysis. Records were grouped by (canonical SMILES, protein sequence); within each group, the median observed label and mean prediction were used to recompute the same evaluation metrics. This analysis assesses the influence of repeated-measurement weighting, is distinct from the 36 prespecified sensitivity runs, and was not used for model selection.

2.7. Held-out test sealing and integrity audit

To keep model development separate from final performance reporting, all split files and training-label scalers were frozen before formal training. Training code was allowed to read only the training and validation data, and every checkpoint recorded test_evaluated_during_training = False. The project entered the unified held-out-test phase only after all 126 prespecified training runs had finished and an acceptance snapshot had been generated. For every run, the checkpoint, record-level predictions, unique-input predictions, and evaluation metadata were incorporated into a SHA256 audit chain, providing a traceable path from frozen inputs to final results.

The integrity audit ultimately accepted all 126 runs. The generic checker covered 108 runs directly. The remaining 18 were GraphDTA-GCN-style runs whose parameter organization differed from the path assumptions of the generic checker; these were examined with an architecture-aware review that directly verified drug-branch parameters, training history, and checkpoint provenance. All 18 passed. This two-level audit preserved a common integrity standard while avoiding false flags caused by implementation differences, and it required neither retraining nor rerunning the held-out test.

2.8. Controlled baselines and prespecified sensitivity experiments

We included two controlled baselines matched for data splits and training budget and reimplemented them under the same data, seeds, and training protocol, providing directly comparable sequence-based and graph-based architectural references. DeepDTA-style used a character-level canonical-SMILES CNN together with a protein CNN, whereas GraphDTA-GCN-style used a three-layer 128-dimensional GCN [26], global max pooling, and a corresponding protein CNN. Throughout this paper, DeepDTA-style and GraphDTA-GCN-style [5–6] refer to these controlled reimplementations. Their purpose was to reduce confounding from cross-study differences in data, training budget, and evaluation protocol, rather than to reproduce the original repositories bit for bit or to construct a cross-study contemporary SOTA leaderboard. Complete layer-wise configurations, run settings, and software dependencies will be released with the code and frozen configurations.

The three primary fusion models were tightly clustered at 592 321–594 305 trainable parameters, providing an approximate capacity match that makes comparisons among fusion strategies more direct. DeepDTA-style and GraphDTA-GCN-style contained 422 849 and 415 809 trainable parameters, respectively, and served as architectural references under the common protocol. Because total parameter counts were not strictly matched across model families, differences between fusion models

and baselines are interpreted empirically for the prespecified architecture–capacity combinations rather than attributed to any fusion mechanism alone.

The prespecified sensitivity experiments contained two single-factor variants, both using ConcatFusion as the reference. protein_len_1000 changed only the maximum protein length from 1 500 to 1 000, whereas drug_pool_mean changed only graph-level pooling from sum to mean. The complete experimental matrix comprised 54 primary runs, 36 controlled-baseline runs, and 36 prespecified sensitivity runs, for a total of 126 runs.

2.9. Exploratory post-hoc application case study design

To connect benchmark-level evaluation with practical compound prioritization, we designed an exploratory post-hoc case study using the existing frozen held-out predictions. Candidate targets were considered only from the final Davis cold-target held-out test set. Before examining any target-specific Prediction_raw values, we applied performance-independent eligibility criteria: the full protein sequence had to be completely unseen during training; sequence-to-target-name mapping had to be one-to-one; held-out compound coverage had to be complete or sufficient; the target had to be pharmacologically relevant; and additional preprocessing confounders were avoided where possible. Among eligible targets, VEGFR2 was chosen as an illustrative pharmacologically relevant case and locked before any target-specific predictions were inspected. Here, “locked” means only that the target was fixed before the post-hoc case analysis began; the case study was not part of the original 126 prespecified training runs. VEGFR2 is a key receptor in VEGF signaling that regulates endothelial-cell function and angiogenesis and has long been a major pharmacological target of anti-angiogenic therapy [35]. Its sequence is 1 356 amino acids long, below the 1 500-amino-acid truncation threshold of the primary configuration, and the final frozen test set contained all 68 Davis compounds. The target-lock record and its SHA256 checksum were retained with the case-study artifacts.

For VEGFR2, all five methods used only previously generated, frozen sample-level held-out predictions; no model was retrained, no checkpoint was reselected, and no additional hyperparameter tuning was performed. For each drug–target pair, predictions from the three model seeds (42, 123, and 2026) were averaged before comparison with the reference ranking derived from observed affinity. The primary analysis reported Kendall’s τ-b across all 68 compounds and the overlap between observed and predicted top-10 compounds; raw-scale RMSE was used only as an auxiliary descriptive metric, and seed-wise results were retained for supplementary analysis. The direction of the frozen Davis labels was checked by exact string matching of raw SMILES: 45 of the 68 VEGFR2 compounds could be matched unambiguously to the original Kd records, with a maximum absolute difference of 4.98×10^{-10} between Label and $9 - \log10(Kd[nM])$. The remaining 23 compounds were not included in this string-level direct-match check. This audit was used only to verify label direction and played no role in target selection or model-performance calculation; its result was consistent with larger Label values representing stronger affinity.

This case study complements the main experiments at the application level and is explicitly exploratory and post hoc. It is not part of the original 126 prespecified training runs and was not used for model selection, hyperparameter tuning, or revision of the primary conclusions. Its purpose is to use fully frozen held-out predictions to map benchmark-level model differences onto a pharmacologically relevant target that was unseen during training and to examine how alternative model choices translate into differences in candidate-compound prioritization.

3. Results and discussion

3.1. Cross-regime performance and ranking shifts

Table 1 summarizes record-level raw-scale RMSE and C-index on the final held-out test for the five methods, reported as mean ± SD over three model seeds. Because the raw label scales differ between KIBA and Davis, RMSE is compared only within the same dataset.

Table 1
A. KIBA record-level held-out test performance (RMSE / C-index, mean ± SD, n = 3)

Method	Warm	Cold-target	Cold-drug
AttentionFusion	0.4601 ± 0.0008	0.6496 ± 0.0200	0.5809 ± 0.0057
	0.8447 ± 0.0003	0.6792 ± 0.0204	0.7676 ± 0.0045
ConcatFusion	0.4433 ± 0.0044	0.6645 ± 0.0127	0.5713 ± 0.0044
	0.8536 ± 0.0012	0.6664 ± 0.0167	0.7759 ± 0.0034
ProductFusion	0.4359 ± 0.0017	0.6731 ± 0.0257	0.5705 ± 0.0049
	0.8546 ± 0.0012	0.6690 ± 0.0021	0.7754 ± 0.0009
DeepDTA-style	0.4384 ± 0.0056	0.6136 ± 0.0212	0.5920 ± 0.0073
	0.8593 ± 0.0031	0.6838 ± 0.0268	0.7627 ± 0.0049
GraphDTA-GCN-style	0.4157 ± 0.0014	0.6184 ± 0.0069	0.5914 ± 0.0028
	0.8664 ± 0.0013	0.6934 ± 0.0109	0.7617 ± 0.0023

Note: In each cell, the first line is raw-scale RMSE (lower is better) and the second line is C-index (higher is better).

Table 1
B. Davis record-level held-out test performance (RMSE / C-index, mean ± SD, n = 3)

Method	Warm	Cold-target	Cold-drug
AttentionFusion	0.5140 ± 0.0082	0.6744 ± 0.0054	0.7991 ± 0.0676
	0.8480 ± 0.0131	0.7885 ± 0.0056	0.6803 ± 0.0356
ConcatFusion	0.4886 ± 0.0087	0.6756 ± 0.0025	0.7260 ± 0.0183
	0.8590 ± 0.0029	0.7919 ± 0.0065	0.7005 ± 0.0134
ProductFusion	0.4541 ± 0.0070	0.6671 ± 0.0053	0.7824 ± 0.0419
	0.8663 ± 0.0013	0.7984 ± 0.0012	0.7211 ± 0.0107
DeepDTA-style	0.4497 ± 0.0149	0.6770 ± 0.0090	0.8103 ± 0.0746
	0.8696 ± 0.0071	0.7862 ± 0.0017	0.6093 ± 0.1044
GraphDTA-GCN-style	0.4648 ± 0.0169	0.6762 ± 0.0097	0.8279 ± 0.0149
	0.8674 ± 0.0018	0.7897 ± 0.0054	0.6763 ± 0.0105

Note: In each cell, the first line is raw-scale RMSE (lower is better) and the second line is C-index (higher is better).

KIBA and Davis both showed clear, complementary patterns of scenario-specific strengths. On KIBA warm, GraphDTA-GCN-style achieved both the lowest RMSE and the highest C-index. Under KIBA cold-target, DeepDTA-style had the lowest RMSE, whereas GraphDTA-GCN-style had the highest C-index; under KIBA cold-drug, ProductFusion achieved the lowest RMSE and ConcatFusion the highest C-index. On Davis, DeepDTA-style nominally led both mean RMSE and C-index under warm, ProductFusion nominally led both metrics under cold-target, and under cold-drug ConcatFusion achieved the lowest RMSE while ProductFusion achieved the highest C-index. These complementary strengths across deployment scenarios and evaluation objectives provide direct empirical support for multi-regime model comparison.

Figure 2 presents the RMSE and C-index of all five methods side by side under the three evaluation regimes. Because the two metrics do not always favor the same model, the subsequent ranking analysis is restricted to nominal rankings based on mean C-index.

Using the three-seed mean C-index to obtain nominal rankings of the five methods for each dataset–regime combination, Kendall's τ for KIBA warm–cold-target, warm–cold-drug, and cold-target–cold-drug was + 0.60, -0.60, and − 1.00, respectively; the corresponding Davis values were − 0.20, -0.40, and + 0.80. The τ = -1.00 result for KIBA cold-target versus cold-drug is the clearest example of a ranking shift across generalization scenarios and highlights why a single evaluation setting cannot stand in for relative performance under different deployment goals. This statistic describes ordering based on three-

seed mean performance; the seed-wise analysis below further characterizes its training variability. These cross-regime ranking shifts are visualized in Fig. 3.

3.2. Seed-wise ranking variability across training runs

Seed-wise Kendall's τ values were not fully consistent with several ranking results based on mean C-index. For KIBA, the three seed-wise values were + 0.40, + 0.20, and + 0.40 for warm–cold-target; -0.40, 0.00, and − 0.40 for warm–cold-drug; and − 0.60, -0.40, and + 0.20 for cold-target–cold-drug. The corresponding values for the three Davis regime pairs were (+ 0.20, -0.20, + 0.60), (0.00, 0.00, 0.00), and (+ 0.40, + 0.80, + 0.40). Thus, the mean τ = -1.00 between KIBA cold-target and cold-drug does not imply that all three independent training runs exhibited a complete reversal.

Multi-seed analysis further exposed the training variability hidden by mean rankings. KIBA produced three distinct complete method orderings under each of warm, cold-target, and cold-drug; Davis produced three, two, and three, respectively. None of the six dataset–regime combinations yielded an identical complete ordering across all three seeds. This result illustrates the value of paired multi-seed evaluation: it makes ranking uncertainty visible when it would otherwise be missed by a single run. We therefore treat rankings based on mean performance as descriptive nominal rankings and interpret differences between neighboring methods together with mean ± SD and seed-wise results.

Cross-regime ranking shifts are among the most direct findings of this study. The mean-ranking Kendall's τ = -1.00 between KIBA cold-target and cold-drug provides the clearest example, whereas the seed-wise τ values show that such shifts cannot be stably summarized by any single training run. Considering the mean ranking together with paired-seed results allows us to distinguish overall cross-regime ordering changes from stochastic orderings arising in individual runs, thereby supporting a more robust interpretation of relative differences among methods.

Warm, cold-target, and cold-drug correspond to different generalization objectives rather than a simple hierarchy of difficulty. Warm primarily measures interpolation among seen entities, whereas cold-drug and cold-target assess extrapolation to unseen drugs and unseen protein sequences, respectively. Previous CPI/DTA studies have shown that train–test similarity, cold-start protocols, spurious correlations, and benchmark bias can influence generalization assessment [14–22]. Under a unified frozen protocol, our study further shows that the relative ordering of the same controlled methods also changes with the intended generalization scenario, indicating that model comparison should be aligned with the expected deployment setting.

Using three paired model seeds made training variability directly observable under a common experimental protocol and provided a more robust basis for interpreting cross-regime ranking shifts than any single run. We therefore report mean ± SD, paired-seed differences, and nominal rankings together, so that model comparisons retain information about both average trends and training variation. Because only three seeds were used, these comparisons remain descriptive and are not interpreted as formal tests of statistical significance or equivalence.

3.3. The most consistent cross-dataset model-family differences occurred under cold-drug generalization

The most consistent cross-dataset pattern occurred under cold-drug. On Davis, ConcatFusion achieved an RMSE of 0.7260 ± 0.0183, lower than 0.8103 ± 0.0746 for DeepDTA-style and 0.8279 ± 0.0149 for GraphDTA-GCN-style. ProductFusion achieved a C-index of 0.7211 ± 0.0107, higher than 0.6763 ± 0.0105 for GraphDTA-GCN-style and 0.6093 ± 0.1044 for DeepDTA-style. On KIBA, ProductFusion achieved an RMSE of 0.5705 ± 0.0049, lower than the approximately 0.592 values of both baselines; ConcatFusion achieved a C-index of 0.7759 ± 0.0034, higher than 0.7627 ± 0.0049 for DeepDTA-style and 0.7617 ± 0.0023 for GraphDTA-GCN-style.

Descriptive within-seed paired C-index differences strengthened the cross-dataset signal under cold-drug: all three ProductFusion-minus-GraphDTA-GCN-style differences were positive on Davis, and all three ConcatFusion-minus-GraphDTA-GCN-style differences were positive on KIBA. Cold-drug therefore provided the most reproducible cross-dataset model-family differences in this study. Because the two controlled baselines contained approximately 29%–30% fewer parameters, these findings are interpreted for the prespecified architecture–capacity combinations rather than as advantages attributable solely to the fusion mechanism. These cold-drug comparisons are summarized in Fig. 4.

The most consistent model-family differences across the two datasets were concentrated under cold-drug, providing the strongest cross-dataset evidence in this study for generalization to unseen compounds. The pattern may reflect interactions among drug representation, model inductive bias, capacity, and distribution shift. Although the present experiments do not disentangle these mechanisms, the consistent direction across datasets supports treating cold-drug as a core evaluation scenario in DTA model comparison rather than substituting warm results for it.

This observation is also consistent with previous generalization studies showing that stricter splits alter the information structure available to models and that different architectures need not respond in the same way [16–22]. Accordingly, we emphasize the empirical pattern repeatedly observed under cold-drug conditions without making causal claims about its mechanistic origin beyond what the data support.

The three primary fusion models contained approximately 592k–594k parameters and were therefore closely capacity-matched, providing strong control for direct comparisons among the fusion strategies. DeepDTA-style and GraphDTA-GCN-style contained approximately 423k and 416k parameters, respectively, and served as cross-model-family architectural references under the same protocol.

Parameter count alone did not show a monotonic relationship with performance. The smaller GraphDTA-GCN-style achieved the best mean performance on KIBA warm, whereas the most consistent cross-dataset advantage of the fusion models over the baselines emerged under cold-drug. The observed pattern is therefore better interpreted as the joint empirical effect of architectural characteristics,

capacity, and evaluation regime. Future work could separate capacity more finely by introducing widened, parameter-matched baselines.

3.4. Complementary strengths across metrics and fusion strategies

Davis cold-drug further illustrates that the evaluation objective can affect model selection: ProductFusion achieved the highest C-index, whereas ConcatFusion achieved the lowest RMSE. C-index places greater emphasis on relative ranking of candidate compounds, while RMSE/MAE more directly measure quantitative affinity error. Thus, even within the same dataset and generalization regime, different evaluation metrics may support different model choices. We therefore report both ranking metrics and raw-scale error metrics rather than defining a globally best model independent of the application objective.

Considering only the three primary fusion models, ProductFusion achieved the lowest raw-scale RMSE in four of the six dataset–regime combinations and the highest nominal C-index in four; ConcatFusion achieved the lowest RMSE on Davis cold-drug and the highest C-index on KIBA cold-drug; AttentionFusion led C-index on KIBA cold-target. Overall, ProductFusion and ConcatFusion showed broad competitiveness across settings, while AttentionFusion contributed a complementary advantage in a specific cold-target scenario. Under the present approximate capacity control, relatively simple concatenation and element-wise interaction were therefore already competitive across regimes, whereas the more complex token-level cross-attention did not automatically translate into a consistent generalization gain.

This comparison is specific to the current 80-dimensional, one-way atom-to-protein token-level cross-attention implementation. Its purpose is to delineate the settings in which the present fusion strategies are most useful, not to generalize the conclusion to other attention architectures.

3.5. Exploratory post-hoc case study: compound prioritization for a pharmacologically relevant unseen target

Following the performance-independent eligibility criteria described in Section 2.9, VEGFR2 was locked as the unseen-target case before the post-hoc analysis began and before any target-specific predictions were inspected. Its full 1 356-amino-acid sequence was absent from the cold-target training set and mapped only to VEGFR2 in the Davis metadata; the final frozen held-out test contained 68 unique canonical compounds. All 15 formal prediction files (5 methods × 3 seeds) were based on exactly the same 4 828 Davis cold-target frozen test records, with the same 68 VEGFR2 records present in every file.

Using frozen predictions averaged across the three seeds, ConcatFusion, ProductFusion, AttentionFusion, DeepDTA-style, and GraphDTA-GCN-style achieved Kendall's τ-b values of 0.645, 0.718, 0.597, 0.742, and 0.779, respectively, across the 68 VEGFR2 compounds. Their observed Top-10 overlaps

were 4/10, 5/10, 4/10, 5/10, and 7/10, and the auxiliary raw-scale RMSE values were 0.857, 0.664, 0.888, 0.581, and 0.428. All five methods showed positive rank correlations while producing distinguishable compound-prioritization profiles, indicating that differences seen in aggregate model comparisons can carry through to candidate ranking for a specific unseen target. The corresponding compound-prioritization profiles are shown in Fig. 5.

The target-specific ordering for VEGFR2 did not fully match the overall mean C-index ordering under Davis cold-target. By overall mean C-index, the methods were ordered ProductFusion, ConcatFusion, GraphDTA-GCN-style, AttentionFusion, and DeepDTA-style; the descriptive VEGFR2 ordering by τ-b was GraphDTA-GCN-style, DeepDTA-style, ProductFusion, ConcatFusion, and AttentionFusion. Because VEGFR2 had been locked before any target-specific Prediction_raw values were examined, this difference did not arise from outcome-based target selection. Among the observed top-10 compounds, V01 (observed rank 1) was ranked 1st by both DeepDTA-style and GraphDTA-GCN-style but 12th by ConcatFusion and 15th by AttentionFusion. V10 (observed rank 10) was ranked 6th by GraphDTA-GCN-style but 25th and 23rd by ConcatFusion and AttentionFusion, respectively. The case offers a concrete example of how aggregate benchmark conclusions can translate into different prioritization profiles for a specific unseen target, reinforcing the practical value of scenario-matched model selection. As a single-target exploratory post-hoc analysis, it is intended as an application-oriented illustration rather than as evidence of a universal pattern across unseen targets.

3.6. Robustness to repeated inputs and prespecified design variants

All 126 held-out evaluations generated both record-level and unique-input predictions. At the unique-input level, records with the same (canonical SMILES, protein sequence) were aggregated using the median observed label and mean prediction to assess the influence of repeated-measurement weighting. The record-level analysis is the primary analysis in the main text; complete unique-input mean ± SD results are reported in the supplementary material and were not used for model selection. The prespecified single-factor sensitivity results are summarized in Table 2.

Table 2
C-index values for prespecified single-factor sensitivity experiments (three-seed mean)

Dataset	Setting	Concat (primary)	Mean pooling	Δ	max_len = 1000	Δ
Davis	warm	0.8590	0.8679	+ 0.0089	0.8619	+ 0.0029
Davis	cold-target	0.7919	0.7919	+ 0.0000	0.7894	-0.0025
Davis	cold-drug	0.7005	0.7007	+ 0.0002	0.6787	-0.0218
KIBA	warm	0.8536	0.8521	-0.0015	0.8554	+ 0.0018
KIBA	cold-target	0.6664	0.6912	+ 0.0248	0.6723	+ 0.0059
KIBA	cold-drug	0.7759	0.7755	-0.0004	0.7780	+ 0.0021

Across the six dataset–regime combinations, the absolute C-index change caused by either prespecified single-factor modification was no greater than 0.025. The largest increase was + 0.0248 for mean pooling under KIBA cold-target, and the largest decrease was − 0.0218 for max_len = 1000 under Davis cold-drug. Because both variants used ConcatFusion as the reference, these results indicate that its main cross-regime pattern was broadly robust to the two individual design choices—graph-level pooling and maximum protein length—and that neither factor alone explains the observed regime dependence. This sensitivity conclusion is restricted to ConcatFusion.

3.7. Data, split, and held-out-test integrity audit

The data and split audits provided an additional structural safeguard for the main results. After cleaning, KIBA decreased from 118 254 source records to 118 036, while Davis retained all 30 056 records. Across all six frozen splits, identical model inputs had zero cross-partition overlap; canonical SMILES were disjoint under cold-drug, and full protein sequences were disjoint under cold-target. Row-wise recomputation of the training-set scaler remained within machine floating-point precision.

In Davis, 442 target IDs corresponded to 379 unique protein sequences. A cold-target split based on target ID could therefore incorrectly treat entities with identical sequences as different targets. The formal cold-target split consequently defined target identity by the full protein sequence.

The formal experiment covered 42 dataset–regime–method/variant combinations, each with three paired model seeds. The 54 primary runs, 36 controlled-baseline runs, and 36 prespecified sensitivity runs together yielded 126 final held-out evaluations. All 126 runs were evaluated and retained in the final aggregation, producing 42 record-level groups and 42 unique-input groups, each containing all three seeds.

The two-level integrity audit successfully accepted all 126/126 runs: the generic checker passed 108 directly, and the remaining 18 GraphDTA-GCN-style runs underwent architecture-aware provenance and checkpoint verification to the same standard.

Formal training and final testing were strictly separated. Record-level predictions, unique-input predictions, and evaluation metadata from all 126 runs were incorporated into an independent hash chain, allowing every final result to be traced back to its frozen artifact. The architecture-aware review also made the audit system compatible with different model implementations. Together with zero cross-partition input overlap, train-only scaling, and a sealed held-out test, this provenance design provides direct evidence for the integrity and reproducibility of the comparison and for the absence of test-based tuning.

3.8. Relation to existing work

As outlined in the Introduction, DTA research has advanced along several fronts, including cold-start evaluation, similarity-aware assessment, multi-scenario generalization, multimodal architectures, computational efficiency, and application validation [7],[14–18],[28–34]. These directions complement the present study. Architecture-led work primarily seeks better predictive performance or stronger unseen-entity generalization within a given protocol, whereas our evaluation-led perspective asks how relative model-family rankings, seed-level ordering stability, and metric dependence change across generalization scenarios under fixed and auditable training and testing procedures. In this way, the study moves beyond model performance alone to the reliability of the conclusions drawn from model comparisons, providing a reproducible empirical basis for scenario-matched DTA model selection.

4. Limitations

This study focuses on two established, kinase-dominated DTA benchmarks, KIBA and Davis, and uses three paired model seeds per condition. This controlled scope makes cross-regime changes in relative performance easier to isolate under a common protocol. The three primary fusion models were tightly approximately capacity-matched, while the two controlled baselines served as smaller architectural references; cross-model-family differences are therefore interpreted as the joint effect of architecture and capacity. Conclusions about AttentionFusion are limited to the present 80-dimensional, one-way atom-to-protein cross-attention implementation. Only a small fraction of sequences longer than 1 500 amino acids were truncated in the primary configuration, and the prespecified length and pooling sensitivity experiments showed that the main ConcatFusion cross-regime pattern was broadly robust to these two single-factor choices. Natural extensions include additional external datasets, more model seeds, cold-both, scaffold- or structure-similarity-based splits, and wet-lab validation. The VEGFR2 case, locked before target-specific predictions were inspected, is a single-target post-hoc example intended to show how aggregate benchmark conclusions can map onto concrete candidate rankings; it provides application-level interpretation rather than replacing larger prospective validation.

5. Conclusions

We established a unified and auditable DTA evaluation design on the cleaned KIBA and Davis datasets, covering warm, cold-target defined by full protein sequence, and cold-drug generalization regimes. Under

a fixed training protocol, three approximately capacity-matched cross-modal fusion strategies and two controlled architectural baselines were compared using identical data splits, training budgets, and evaluation procedures. The complete study comprised 126 prespecified training runs with three paired model seeds per condition; the held-out test was kept strictly separate from training and model selection, while record-level, unique-input, sensitivity, and SHA256 integrity-audit outputs were generated in parallel.

Three complementary findings emerged. First, relative performance and nominal rankings changed substantially with the warm, cold-target, and cold-drug generalization objectives; for KIBA, the three-seed mean rankings for cold-target versus cold-drug reached Kendall's $\tau = -1.00$. Second, cold-drug yielded the most consistent cross-dataset model-family differences, with ConcatFusion and ProductFusion showing complementary strengths in RMSE and C-index. Third, the multi-seed results revealed training variability in the complete method ordering, supporting the joint use of repeated runs, mean ± SD, and paired-seed results for model comparison. Among the three approximately capacity-matched fusion models, ProductFusion and ConcatFusion remained broadly competitive across most settings, while AttentionFusion showed a scenario-specific advantage on KIBA cold-target.

Overall, this study extends DTA model comparison beyond a single benchmark score to four dimensions: generalization scenario, training variability, evaluation objective, and result traceability. The exploratory VEGFR2 case further shows that aggregate benchmark rankings can translate into different compound-prioritization profiles for a specific unseen target. By combining unified splitting, controlled training, multi-seed replication, complementary metrics, record-level and unique-input analyses, and sealed held-out-test auditing, the evaluation design provides a more complete empirical basis for scenario-matched model selection, fair comparison, and reproducible interpretation.

Declarations

CRediT authorship contribution statement

Haobo Kui: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Supervision, Project administration, Funding acquisition, Resources, Writing – original draft, Writing – review & editing.

Xinyuan Cui: Methodology, Formal analysis, Investigation, Data curation, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Undergraduate Innovation and Entrepreneurship Training Program of Heilongjiang Province (Provincial Guidance Project) under Grant No. S202610231053.

Author Contribution

H.K. conceptualized the study, designed the methodology, developed the software, conducted formal analysis and investigation, curated the data, prepared visualizations, supervised the project, acquired funding and resources, and wrote the original manuscript. X.C.contributed to methodology, formal analysis, investigation, data curation, validation, and visualization, and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Data Availability

The original KIBA and Davis data sources are cited in the corresponding references and are not redistributed in the repository. To support the auditability and reproducibility of this study, frozen split indices and cryptographic hashes, formal experimental configurations, run-level and aggregated results for all 126 prespecified evaluations, split-generation and verification scripts, complete controlled-baseline implementations, the VEGFR2 case-study code, integrity-audit materials, and the formal software-environment specification are publicly available at GitHub (https://github.com/qingqingyuanyuan/dta-model-comparison-reliability), release v1.0.0. The corresponding release is permanently archived on Zenodo (DOI: 10.5281/zenodo.22860429). Record-level and unique-input prediction artifacts generated in the formal evaluation have been retained as part of the frozen project archive and can be provided upon reasonable request.

References

1. Santos R et al (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16(1):19–34
2. Vamathevan J et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18(6):463–477
3. Tang J, Szwajda A, Shakyawar S et al (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 54(3):735–743
4. Davis MI et al (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051
5. Öztürk H, Özgür A, Özkirimli E (2018) DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 34(17):i821–i829

6. Nguyen T et al (2021) GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 37(8):1140–1147
7. Abbasi K et al (2020) DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics 36(17):4633–4642
8. Chen L 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
9. Huang K et al (2021) MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics 37(6):830–836
10. Zhao Q et al (2023) AttentionDTA: predicting drug–target binding affinity with attention mechanism. IEEE/ACM Trans Comput Biol Bioinform 20(2):852–863
11. Li S et al (2020) MONN: a multi-objective neural network for predicting compound–protein interactions and affinities. Cell Syst 10(4):308–322e11
12. Bian J, Zhang X, Zhang X, Xu D, Wang G (2023) MCANet: shared-weight-based MultiheadCrossAttention network for drug-target interaction prediction. Brief Bioinform 24(2):bbad082
13. Liu Y, Xia X, Gong Y, Song B, Zeng X (2024) SSR-DTA: Substructure-aware multi-layer graph neural networks for drug-target binding affinity prediction. Artif Intell Med 157:102983
14. Wang J, Wen NF, Wang C, Zhao L, Cheng L (2022) ELECTRA-DTA: a new compound–protein binding affinity prediction model based on contextualized sequence encoding. J Cheminform 14:14
15. Fang K et al (2023) ColdDTA: utilizing data augmentation and attention-based feature fusion for drug–target binding affinity prediction. Comput Biol Med 164:107372
16. Özçelik R, Bag A, Özkirimli E et al (2023) A framework for improving the generalizability of drug–target affinity prediction models (DebiasedDTA). J Comput Biol 30(11):1226–1239
17. Jiang Z, Dai H, Chen Y, Qiao B, Sun S (2026) CS-DTA: a language model-driven framework for robust drug–target affinity prediction under strict cold-start scenarios. Front Chem 14:1834317
18. Yaseen A, Amin I, Akhter N, Ben-Hur A, Minhas F (2022) Insights into performance evaluation of compound–protein interaction prediction methods. Bioinformatics 38(Suppl 2):ii75–ii81
19. Mayr A et al (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 9(24):5441–5451
20. Ezzat A, Wu M, Li X-L, Kwoh C-K (2019) Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 20(4):1337–1357
21. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26(9):1169–1175
22. Rohrer SG, Baumann K (2008) Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. J Chem Inf Model 48(4):704–718

23. Weininger D (1988) SMILES, a chemical language and information system. J Chem Inf Comput Sci 28(1):31–36
24. Landrum G (2013) RDKit: open-source cheminformatics software. https://www.rdkit.org
25. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. Proc ICLR
26. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. Proc ICLR
27. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. Proc ICLR
28. Zhang C, Hu Z, Jiang C, Chen W, Xu J, Zhang S (2025) Rethinking the generalization of drug target affinity prediction algorithms via similarity-aware evaluation. ICLR
29. Zhou Z, Chen M, Zhou W, Xiao Q (2026) PCIM-DTA: pairwise conditional interaction modeling for drug–target affinity prediction under cold-start scenarios. Bioinformatics 42(9):btag646
30. Zhou Z, Chen M, Zhou W, Xiao Q (2026) Similarity-adaptive interaction routing for drug–target affinity prediction. J Biomed Inf 182:105100
31. Xu Y, Xiao X, Lin W (2026) A drug–target affinity prediction model integrating multimodal feature fusion and structural modeling. Comput Biol Chem 124:109053. https://doi.org/10.1016/j.compbiolchem.2026.109053
32. Abdul Cader JM, Newton MAH, Sattar A (2026) Efficient drug–target affinity prediction via interaction features and parallel CNN–BiLSTM with attention. Comput Biol Chem 121:108862. https://doi.org/10.1016/j.compbiolchem.2025.108862
33. de Souza JG, Fernandes MAC, Barbosa RM (2026) Deep learning-based prediction of drug–target interactions between antiviral drugs and SARS-CoV-2 proteins using an image-based representation approach. Comput Biol Chem 125:109264. https://doi.org/10.1016/j.compbiolchem.2026.109264
34. Abdelkader GA, Njimbouom SN, Oh TJ, Kim JD, ResBiGAAT (2023) Residual Bi-GRU with attention for protein-ligand binding affinity prediction. Comput Biol Chem 107:107969. https://doi.org/10.1016/j.compbiolchem.2023.107969
35. Olsson AK, Dimberg A, Kreuger J, Claesson-Welsh L (2006) VEGF receptor signalling - in control of vascular function. Nat Rev Mol Cell Biol 7(5):359–371. https://doi.org/10.1038/nrm1911

Figures

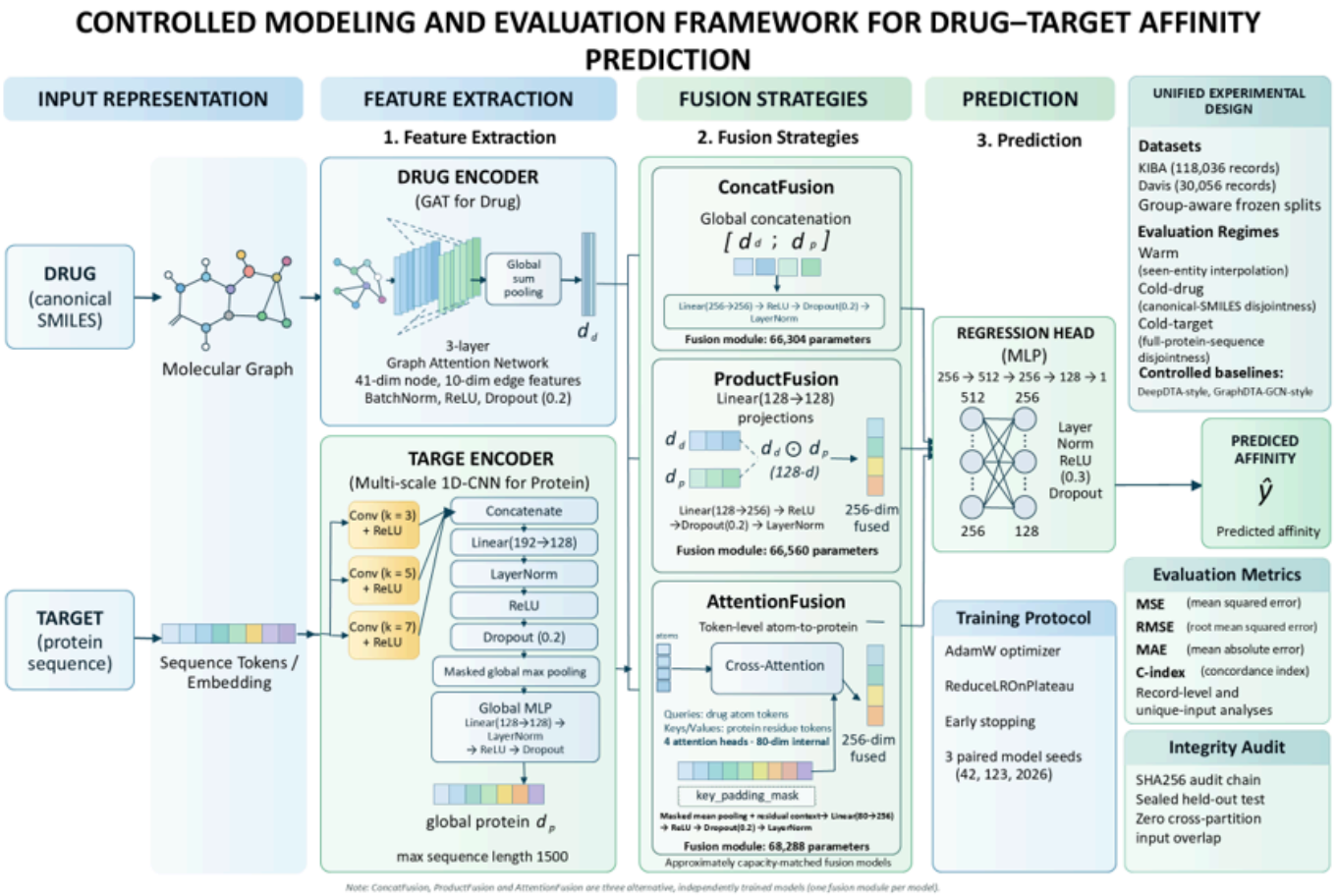


Figure 1

Study design and formal evaluation workflow. Data splitting, model comparison, and final held-out test evaluation all followed a frozen protocol; target identity under cold-target was defined by the full protein sequence.

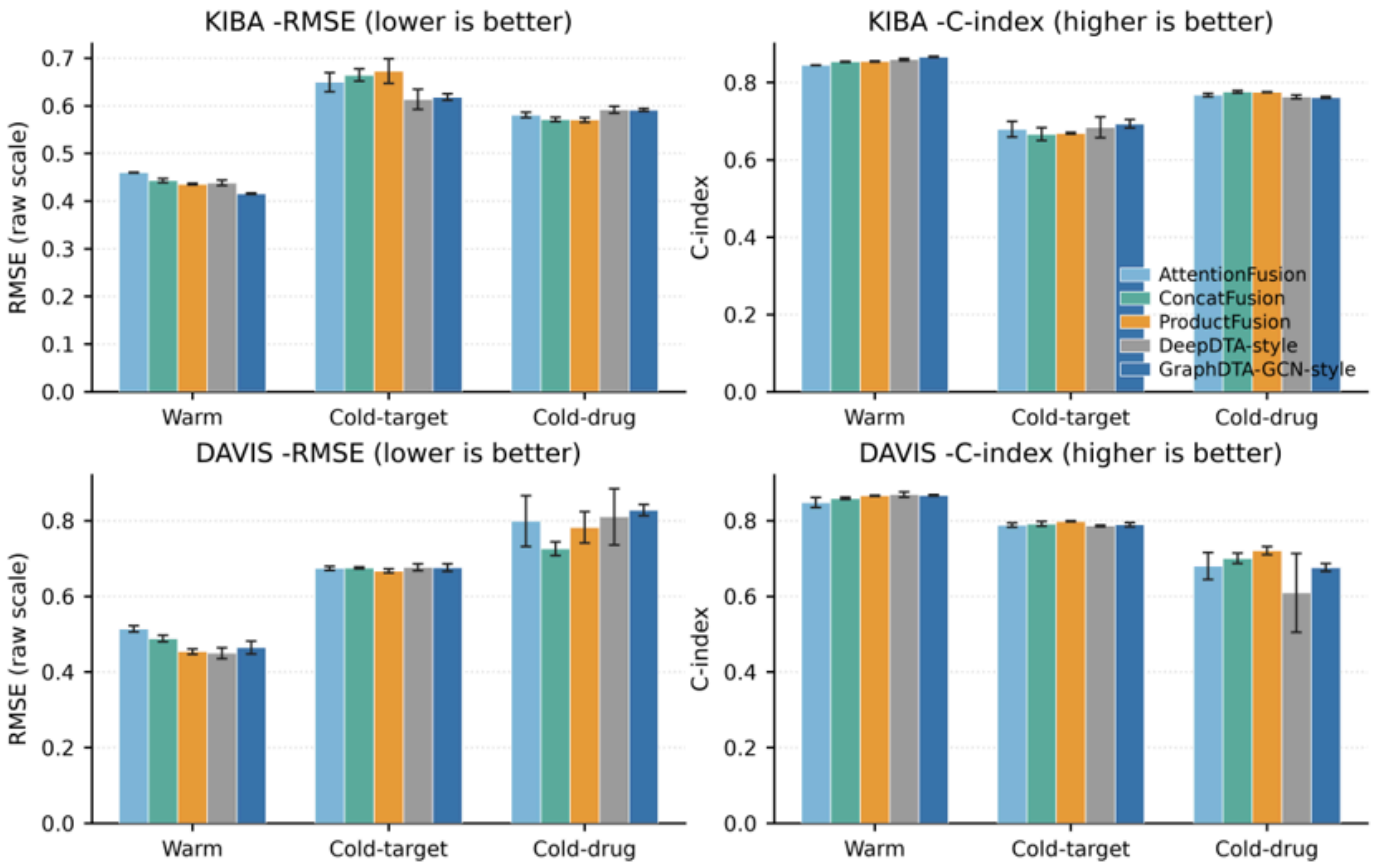


Figure 2

Overview of held-out test performance for the five methods under warm, cold-target, and cold-drug conditions. Error bars indicate the standard deviation across three model seeds.

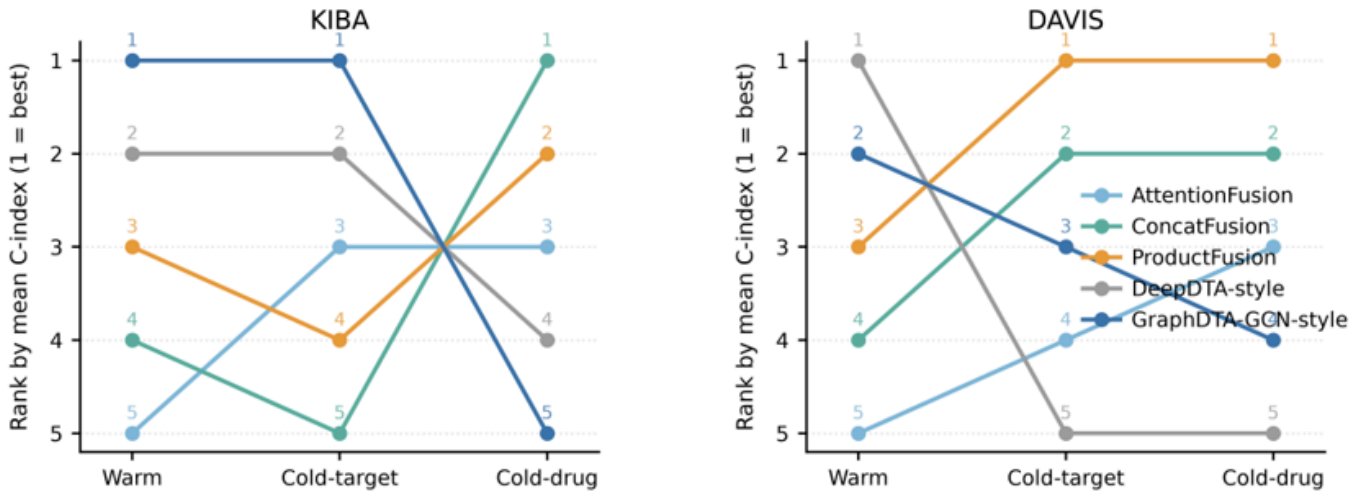


Figure 3

Nominal ranking shifts of the five methods based on mean C-index across three model seeds. Rank 1 denotes the highest mean C-index.

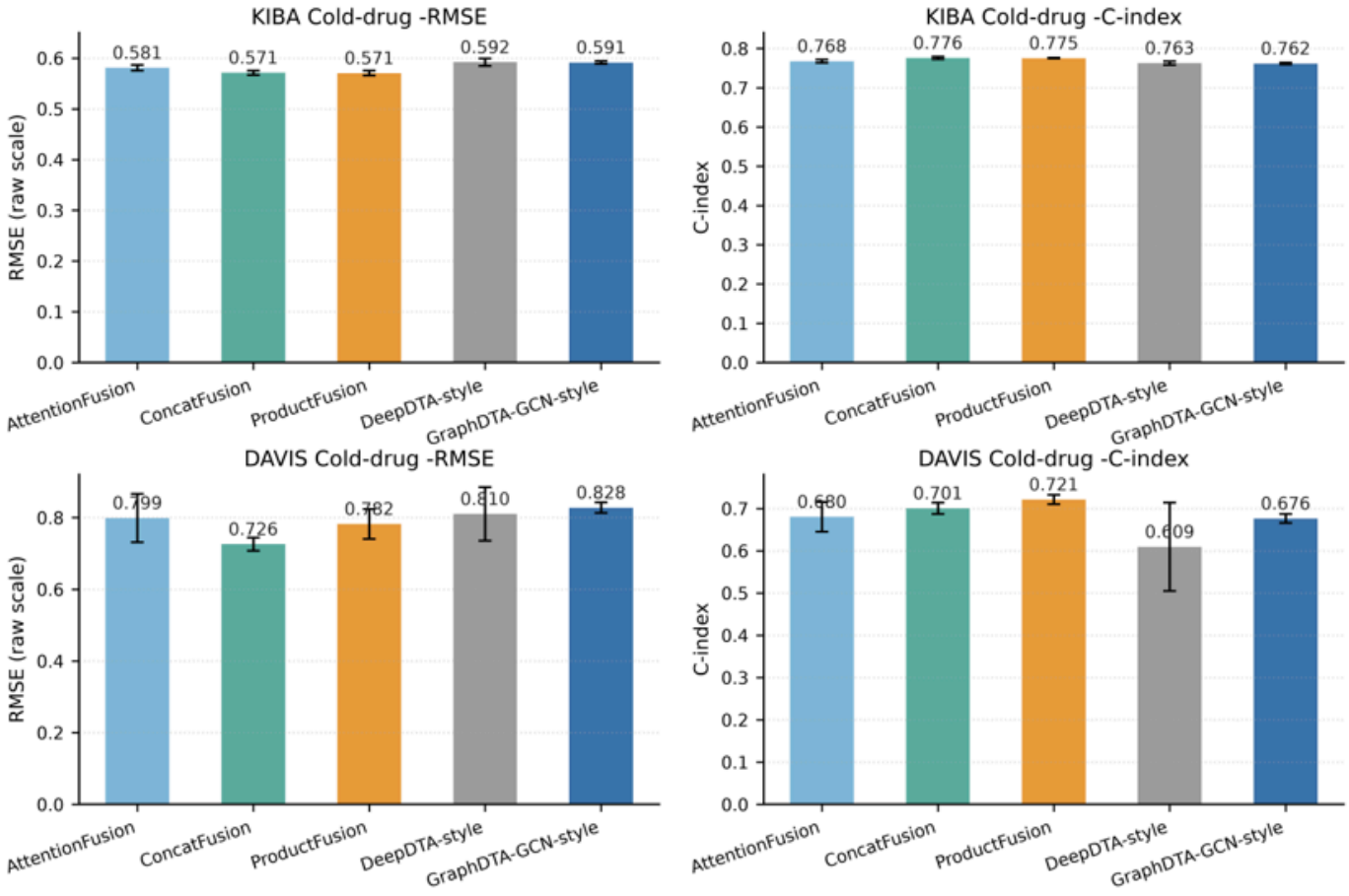


Figure 4

Record-level RMSE and C-index of the three fusion models and two controlled baselines under cold-drug conditions. Error bars indicate the standard deviation across three model seeds.

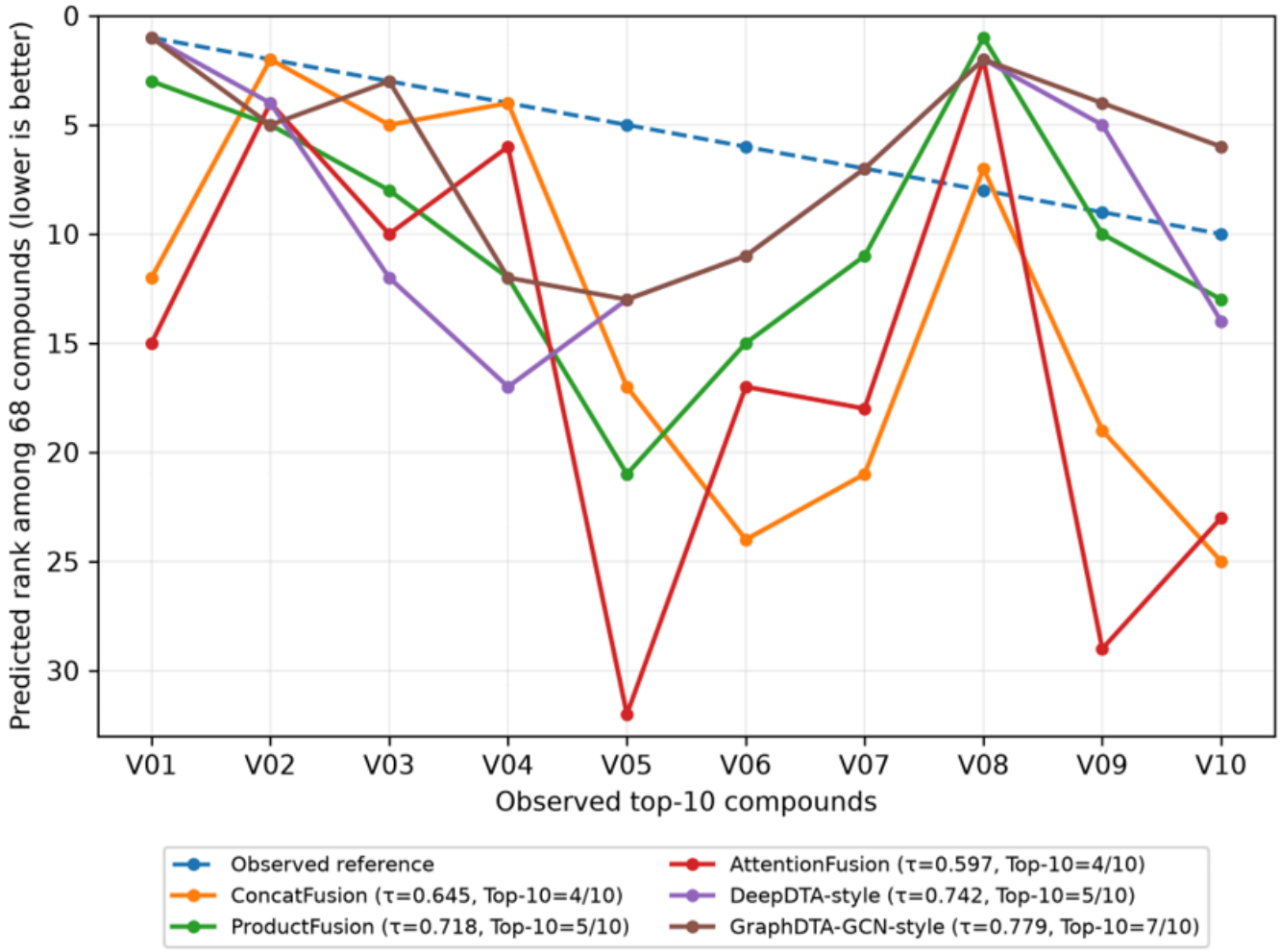


Figure 5

VEGFR2 unseen-target compound-prioritization profiles. V01–V10 denote the observed top-10 compounds ordered by measured affinity. The y-axis shows each compound's predicted rank among all 68 VEGFR2 compounds using predictions averaged across the three model seeds (rank 1 = highest predicted affinity). The dashed line denotes the observed reference rank, and the legend reports Kendall's τ-b and Top-10 overlap for each method across all 68 compounds. The mapping from V01–V10 to canonical SMILES and observed Label is provided in Supplementary Table S3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx