MULTIMODAL LEARNING WITH MISSING MODALITIES VIA MULTI-LEVEL ALIGNMENT FOR DRUG–TARGET AFFINITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Drug–target affinity (DTA) prediction estimates the binding strength between a small-molecule drug and a target protein. Existing methods typically rely on either sequence or structure information, limiting the joint exploitation of complementary supervision from heterogeneous datasets. Label heterogeneity is often overlooked, leading to the conflation of distinct experimental measurements into a single prediction target. To address these issues, we propose M3A-DTA, a missing-modality, multimodal, and multi-assay framework that learns from any available subset of drug sequence, protein sequence, ligand structure, and protein structure. Modality-specific pretrained encoders provide chemically and biologically informed representations. Through multimodal alignment, adaptive fusion, and task-specific supervision, M3A-DTA learns robust representations that accommodate incomplete modalities. The design enables joint training on sequence-centric data together with structure-centric complexes. We specify evaluation under complete, naturally missing, and deliberately masked modalities, with cold-drug, cold-target, and cold-pair splits. M3A-DTA achieves strong performance across multiple evaluation settings while demonstrating broader applicability than most existing methods.

1 INTRODUCTION

Drug–target affinity prediction supports virtual screening by estimating continuous binding strength, but available evidence is heterogeneous. Sequence data are abundant, whereas three-dimensional complexes are sparse; nevertheless, most sequence- and structure-based methods assume a fixed input schema (Ozturk et al., 2018; Nguyen et al., 2021; Stepniewska-Dziubinska et al., 2018; Li et al., 2021). This prevents a single model from learning jointly from sequence-only and structure-rich records or operating reliably when a modality is absent.

Heterogeneity also occurs in the labels and representations. Distinct physical measurements depend on different assay assumptions and should not be pooled indiscriminately (Cheng & Prusoff, 1973; Tang et al., 2014). Meanwhile, global fusion discards the fact that a sequence token and a structural node may describe the same ligand atom or protein residue. Pretrained molecular and protein models provide strong priors (Chithrananda et al., 2020; Lin et al., 2023; Ji et al., 2024), but their element-level correspondence remains underused.

We therefore propose M3A-DTA (Figure 1), which treats modalities as optional typed evidence. Element-level alignment connects corresponding atoms and residues, availability-aware fusion supports arbitrary valid modality subsets, and measurement-specific heads preserve label semantics while sharing molecular interaction representations.

Our contributions are as follows:

- We formulate DTA as missing-modality prediction with one model for sequence-only, structure-only, and joint evidence.
- We align sequence and structure through known atom- and residue-level correspondence.
- We separate physical measurement heads, censored-label losses, and composite-score supervision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

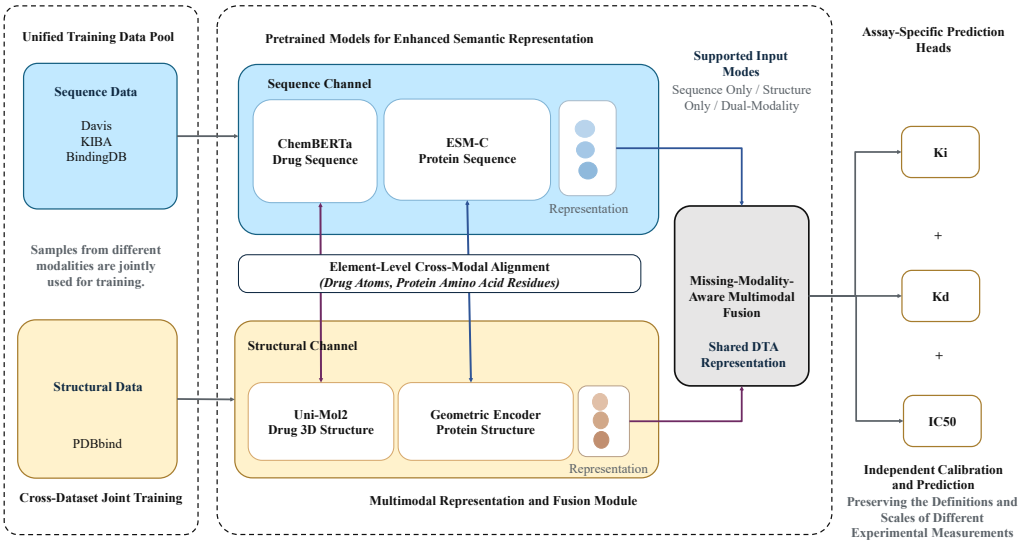

Figure 1: Overview of M3A-DTA. Four optional ligand and protein views are encoded separately, aligned at corresponding atoms or residues when paired evidence exists, fused with explicit availability information, and routed to measurement-specific prediction heads. Dashed branches denote missing modalities.

2 Related Works

2.1 Sequence- and Structure-Based Drug–Target Affinity Prediction

Sequence-based DTA models exploit representations that are widely available and inexpensive to process. DeepDTA applies parallel convolutional encoders to ligand and protein strings (Ozturk et al., 2018), while GraphDTA replaces the ligand string encoder with a molecular graph network (Nguyen et al., 2021). Later methods emphasize local interaction modeling: MONN jointly predicts affinity and non-covalent contacts (Li et al., 2020); MolTrans uses substructure-aware Transformer interactions (Huang et al., 2021); and DrugBAN applies bilinear attention with domain adaptation (Bai et al., 2023). Recent sequence-centric systems such as PSICHIC and ProMoBind inject physicochemical constraints, predicted contacts, or binding-site pretraining without requiring an experimental complex at inference (Koh et al., 2023; Zhang et al., 2026). Their broad coverage is attractive, but sequence proximity is not spatial proximity, and performance can be inflated when related scaffolds or homologous targets cross random train–test boundaries (Pahikkala et al., 2015).

Structure-based methods instead operate on voxelized complexes, contact graphs, residue–atom neighborhoods, or equivariant geometric representations. Pafnucy uses three-dimensional convolutions over protein–ligand grids (Stepniewska-Dziubinska et al., 2018); IGN separates intra- and intermolecular graph convolutions (Jiang et al., 2021); and SIGN encodes distance and angle information with interactive graph pooling (Li et al., 2021). TankBind models candidate pockets with trigonometric constraints (Lu et al., 2022), whereas DynamicBind couples equivariant diffusion with ligand-conditioned protein flexibility (Lu et al., 2024). More recent co-folding systems, including Boltz-2, jointly predict complex structure and affinity (Passaro et al., 2025). These methods expose physical geometry directly, but experimental complexes are much scarcer than biochemical measurements, and predicted or docked poses introduce their own uncertainty. M3A-DTA therefore treats sequence and structure as complementary, optional evidence: either path remains independently predictive, while paired observations provide cross-modal alignment supervision.

2.2 Multimodal Learning with Incomplete Observation

Conventional multimodal fusion often assumes that every modality is present for every sample. When this assumption fails, discarding incomplete records wastes supervision, zero filling confounds absence with content, and maintaining one model per availability pattern fragments both statistical strength and deployment. ModDrop established stochastic modality removal as a way to learn cross-modal correlations while retaining robustness to missing signals (Neverova et al., 2016). SMIL later formalized learning with severely missing modalities in both training and testing and used Bayesian meta-learning to accommodate incomplete examples (Ma et al., 2021). Related representation-factorization methods such as MISA separate modality-invariant from modality-specific information (Hazarika et al., 2020).

DTA presents a particularly structured version of this problem: sequence evidence is common, structural evidence is sparse, and paired views are concentrated in particular sources. Consequently, a model can misuse modality availability as a proxy for source or label distribution. M3A-DTA combines stochastic modality dropout, learned missing tokens, explicit availability embeddings, balanced sampling, and prediction consistency. Unlike generic missing-modality methods, it additionally exploits deterministic atom and residue correspondence whenever sequence and structural views describe the same molecular element.

3 Methods

3.1 Problem formulation

Let a sample describe a ligand–protein pair with a set of available observations

$$\mathcal{X} = \{X_d^{seq}, X_p^{seq}, X_d^{str}, X_p^{str}\},$$

where the subscript denotes drug or protein and the superscript denotes sequence or structure. X_d^{seq} is a SMILES sequence or equivalent atom-ordered molecular representation; X_p^{seq} is an amino-acid sequence; X_d^{str} contains ligand atoms, bonds, and three-dimensional coordinates; and X_p^{str} contains protein residue or atom coordinates, optionally restricted to a binding pocket. Each input has an availability value $m_v \in \{0, 1\}$. We require at least one representation of the drug and one representation of the protein, but do not require both sequence and structure views for either entity.

The model predicts a vector

$$\hat{\mathbf{y}} = f_\theta(\mathcal{X}, \mathbf{m}) = (\hat{y}_{Kd}, \hat{y}_{Ki}, \hat{y}_{IC50}, \hat{y}_{KIBA}),$$

but a record contributes supervised loss only to the component indicated by a. At inference, users request the head matching the desired endpoint. If the endpoint is unspecified, the system returns all supported predictions with their modality-conditioned uncertainty rather than silently merging them.

3.2 Architecture overview

Figure 1 summarizes the four-stage architecture. First, modality-specific pretrained backbones encode each available input at element and global resolutions. Second, correspondence-aware projectors map ligand atoms and protein residues from different modalities into shared spaces. Third, an availability-aware hierarchical Transformer integrates evidence within each molecular entity and then models drug–protein interactions. Fourth, measurement-specific heads map the shared pair representation to the requested endpoint, while an auxiliary head accommodates composite activity supervision during training.

The same interface supports both a high-resolution token path and an efficient pooled-feature fallback. Observed representations are standardized with training-set statistics, projected to a common dimension, tagged by modality and availability, and processed jointly. The token path adds protein geometry and element-level correspondence; the fallback uses one global token per available

view. Architectural and optimization settings are consolidated in Table 2 rather than embedded in the method definition.

3.3 MODALITY-SPECIFIC PRETRAINED ENCODERS

Ligand sequence encoder. A pretrained molecular language model encodes canonical isomeric SMILES into contextual tokens and a pooled representation, transferring chemical syntax learned from unlabeled corpora (Chithrananda et al., 2020). We use ChemBERTa-100M-MLM. Tokenizer offsets and atom-mapped SMILES associate token states with atoms; non-atom symbols contribute only to global context.

Protein sequence encoder. A pretrained protein language model produces residue-level evolutionary and biochemical features (Rives et al., 2021; Lin et al., 2023). We use ESM Cambrian 600M (EvolutionaryScale, 2024); long sequences are processed in chunks while retaining absolute residue indices for structural alignment.

Ligand structure encoder. A three-dimensional molecular model integrates atom identity, bond topology, and conformer geometry while preserving one token per heavy atom. We instantiate it with Uni-Mol2-570M (Ji et al., 2024), use experimental coordinates when available, record generated-conformer provenance otherwise, and retain heavy-atom indices for alignment.

Protein structure encoder. A geometry-aware encoder represents each indexed residue using sequence identity and local geometry. We use invariant geometric attention. The same encoder accepts pocket-centered or full-chain inputs, and predicted structures carry confidence and source embeddings.

In the primary configuration, pretrained backbones are frozen and their row-aligned features are cached. Projectors, alignment and fusion modules, and prediction heads remain trainable.

3.4 ELEMENT-PRESERVING CROSS-MODAL ALIGNMENT

Global multimodal alignment only asks whether two pooled representations belong to the same molecule or protein. M3A-DTA instead exploits exact correspondence whenever it is known. Let $h_{d,i}^{seq}$ and $h_{d,i}^{str}$ represent the sequence-derived and structure-derived embeddings of ligand atom i. Let $h_{p,j}^{seq}$ and $h_{p,j}^{str}$ represent amino-acid residue j. These token pairs describe the same element in different contexts.

For ligands, correspondence is established through atom mapping. The canonical molecule is parsed once, and atom-map identifiers are propagated to the SMILES serialization and the coordinate-bearing molecule. Any conformer reordering is inverted before features are cached. Tokens without a unique atom referent participate only in the sequence branch. For proteins, PDB residues are mapped to the reference sequence by chain-aware alignment. Residue keys include chain identifier, residue number, and insertion code; unresolved residues and non-standard amino acids are represented with masks. A structural token is aligned only when its identity and sequence location are unambiguous.

Each modality has its own projector P_v, followed by a shared normalization layer. For a matched element k, the alignment loss is

$$\mathcal{L}_{elem} = \frac{1}{|\mathcal{C}|} \sum_{(u,v,k)\in\mathcal{C}} \left(1 - \cos(P_u h_k^u, P_v h_k^v)\right),$$

where $\mathcal{C}$ is the set of valid atom and residue correspondences. This positive-pair loss is complemented by an in-batch contrastive term. Other atoms or residues from the same entity are hard negatives only when their chemical identity and local context differ; otherwise they are omitted to avoid false negatives among repeated elements. A symmetric temperature-scaled objective retrieves the structural token from its sequence counterpart and vice versa.

Alignment does not force the views to become identical. A private subspace is retained for modality-specific information, and only a designated shared subspace participates in $\mathcal{L}_{elem}$. The fused token is produced by gated residual exchange:

$$\tilde{h}_k^u = h_k^u + g_k^u \odot W_{v\to u} h_k^v, \qquad g_k^u = \sigma(W_g[h_k^u; h_k^v; m_u; m_v]).$$

When one view is absent or correspondence is uncertain, the residual exchange is skipped. This mechanism lets structure refine a sequence token without overwriting the evolutionary or syntactic information carried by the sequence encoder.

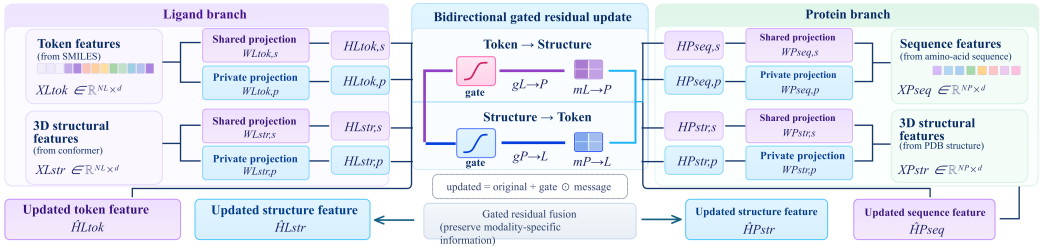

Figure 2: Element-level correspondence and gated exchange. The ligand panel maps atom-mapped sequence tokens to the same atoms in a 3D conformer, while the protein panel maps amino-acid positions to resolved structural residues. Valid correspondences support shared-space alignment and bidirectional gated residual updates; ambiguous or missing correspondences are masked.

3.5 Assay-aware prediction heads

The shared pair representation enters one independent head for each of the three physical measurement types defined in the problem formulation. Each head is a small pre-normalized Transformer operating on the complete fused token sequence, followed by a linear regressor on the head-specific CLS token. Independent attention layers allow each endpoint to emphasize different chemical and protein evidence. The heads share encoders and fusion parameters, enabling transfer from abundant tasks to sparse tasks without forcing a common final mapping.

For an exact label, the supervised loss for measurement type a is masked mean squared error:

$$\mathcal{L}_{exact} = \ell_{\mathrm{MSE}}(\hat{y}_a, y).$$

Only one physical head is normally observed per database row, so a one-hot target mask routes the gradient and ensures that every record activates exactly one physical head. Relation operators are handled directly in the loss. Because labels are represented on a negative logarithmic scale, inequality direction is reversed during transformation; a one-sided soft-margin loss penalizes only predictions that violate the feasible half-line. Approximate labels use a tolerance interval around y, with no penalty inside the interval and a smooth penalty outside it.

The composite activity score is trained through a separate source-specific head with masked mean squared error. Its gradients update the shared encoders and interaction representation, providing pairwise supervision, but the head is never used as a substitute for a physical measurement.

3.6 Training objective and optimization

The complete objective is

$$\mathcal{L} = \mathcal{L}_{assay} + \lambda_{elem}\mathcal{L}_{elem} + \lambda_{con}\mathcal{L}_{contrast} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{reg}\|\theta\|_2^2.$$

$\mathcal{L}_{assay}$ includes exactly observed physical targets and the separate composite-score auxiliary target. Alignment losses are computed only for paired views with valid correspondence, while consistency is computed only for legal modality subsets. Coefficients are tuned on validation data with the complete-modality and missing-modality objectives considered jointly. A model is not selected solely for its best complete-input RMSE if it collapses under absent structure.

We optimize trainable projectors, alignment layers, fusion blocks, and prediction heads with AdamW (Loshchilov & Hutter, 2019). Validation error controls learning-rate reduction and early stopping, gradient clipping stabilizes updates, and mixed precision is used on compatible accelerators. The exact training configuration and search ranges are reported in Table 2.

3.7 Inference and uncertainty under changing evidence

At inference, the model accepts any supported availability mask without retraining. For a sequence-only pair, it returns predictions using the ligand and protein sequence paths. When ligand-conformer or protein-structure evidence is supplied, the same pair is reevaluated with the additional input. We estimate uncertainty using an ensemble of independently seeded fusion models and report the mean and standard deviation for the selected assay head. Ensemble members are loaded sequentially, so GPU memory does not scale with ensemble size.

Evaluation considers both calibration within each observation regime and stability across regimes. We measure prediction shift when structure is added, error reduction from additional evidence, and confidence–error correlation. Large shifts without an accuracy gain indicate that the structural branch may encode source bias rather than useful geometry.

4 Experiments

4.1 Research questions

The experiments are organized around six questions. **RQ1:** Can one model match modality-specific baselines when only sequences or only structures are available? **RQ2:** Does joint training on heterogeneous sources improve generalization to cold drugs, cold targets, and unseen pairs? **RQ3:** Does element-level atom/residue alignment improve prediction beyond global multimodal fusion? **RQ4:** Does separating physical measurement types reduce error and cross-assay interference relative to one pooled affinity head? **RQ5:** Do pretrained encoders provide complementary gains, and are those gains retained under missing modalities? **RQ6:** Does performance improve gracefully as additional modalities become available?

4.2 Data splits and leakage control

We report four split regimes following recommendations for realistic drug–target generalization (Pahikkala et al., 2015). A warm random split measures interpolation among familiar chemical and protein families. A cold-drug split assigns Bemis–Murcko scaffolds (Bemis & Murcko, 1996) to disjoint partitions. A cold-target split clusters proteins by sequence identity and holds out entire clusters. A cold-pair split prevents pair overlap while seeking to minimize close ligand and target neighbors across partitions. For structure data, we additionally evaluate a split based jointly on ligand scaffold and protein-family similarity.

All modalities belonging to the same biological pair remain in the same partition. This includes alternative structures, conformers, assay replicates, and entries duplicated across datasets. Clustering and scaffold assignment occur before the train/validation/test split. Pretrained encoders are fixed models trained independently of our affinity labels, but no supervised feature normalization or calibration statistic is computed outside the training partition.

The main model-selection criterion is the macro-average of standardized validation RMSE across physical measurement types and required modality regimes. We also report overall RMSE for continuity with earlier experiments. Hyperparameters are selected once on validation data and fixed before test evaluation. The final comparison uses five random seeds and summarizes the mean, standard deviation, and paired bootstrap confidence intervals over held-out pairs.

4.3 Experimental setup

All large pretrained backbones are frozen in the primary configuration, and cached features are standardized using training-partition statistics only. Table 2 consolidates architecture and optimization

Table 1: Main results on the PDBbind v2020 full set under the random split. Lower is better for RMSE and MAE; higher is better for Pearson, Spearman, R^2, and concordance index (CI). Best values are shown in bold and second-best values are underlined. N/A indicates unavailable results.

Method	**RMSE** ↓	**MAE** ↓	**Pearson** ↑	**Spearman** ↑	R^2 ↑	**CI** ↑
Pafnucy (Stepniewska-Dziubinska et al., 2018)	1.435	1.144	0.635	0.587	0.348	0.707
IGN (Jiang et al., 2021)	1.404	1.116	0.662	0.638	0.385	0.730
SIGN (Li et al., 2021)	1.373	1.086	0.685	0.656	0.398	0.736
TankBind (Lu et al., 2022)	1.345	1.060	0.718	0.689	0.404	0.750
DrugBAN (Bai et al., 2023)	1.480	1.159	0.657	0.612	0.319	0.720
PSICHIC (Koh et al., 2023)	1.314	1.015	0.710	0.686	<u>0.428</u>	<u>0.751</u>
DynamicBind (Lu et al., 2024)	<u>1.253</u>	<u>1.013</u>	0.689	0.656	N/A	N/A
ProteinReDiff (Nguyen et al., 2024)	1.443	1.168	0.721	0.639	N/A	N/A
Boltz-2 (Passaro et al., 2025)	N/A	N/A	0.609	0.600	N/A	N/A
ProMoBind (Zhang et al., 2026)	1.268	1.020	<u>0.738</u>	<u>0.707</u>	N/A	N/A
Ours	**1.225**	**0.898**	**0.758**	**0.740**	**0.590**	**0.783**

Table 2: Primary hyperparameter configuration.

Component	**Setting**
Backbone update	Frozen; cached features
Shared embedding dimension	256
Fusion / prediction blocks	2 Transformer layers
Attention heads / FFN width	8 / 512
Dropout	0.1
Batch size	64
Optimizer	AdamW
Learning rate / weight decay	10^{-4} / 10^{-4}
Maximum epochs	400
Gradient-norm clipping	1.0

settings that are deliberately omitted from Section 3. Loss weights and modality-dropout probabilities are selected on validation data and fixed before test evaluation.

4.4 Baselines

Baselines cover sequence, structure, and multimodal families. Sequence-based comparisons include DeepDTA (Ozturk et al., 2018), GraphDTA (Nguyen et al., 2021), DrugBAN (Bai et al., 2023), PSICHIC (Koh et al., 2023), and ProMoBind (Zhang et al., 2026). Structure-based comparisons include Pafnucy (Stepniewska-Dziubinska et al., 2018), IGN (Jiang et al., 2021), SIGN (Li et al., 2021), TankBind (Lu et al., 2022), DynamicBind (Lu et al., 2024), and Boltz-2 (Passaro et al., 2025). Multimodal controls include global feature concatenation, gated late fusion, and a complete-modality Transformer without modality dropout. Missing-input controls use zero filling, mean imputation, and separate models trained for each observation pattern. Table 1 reports the warm-split comparison.

For measurement handling, we compare four alternatives: one pooled regressor, separate single-measurement models, a shared encoder with linear heads, and our Transformer heads with relation-aware loss. Composite scores are excluded from physical-head evaluation in every method. For fair comparison, all pretrained-backbone baselines use the same frozen checkpoints where architecturally possible, and parameter counts, training examples, and inference inputs are reported.

4.5 Missing-modality evaluation

We distinguish natural missingness from controlled masking. Natural missingness evaluates each dataset with the modalities it genuinely provides. Controlled masking starts from examples with all four views and evaluates every legal subset containing at least one drug and one protein view. The

Table 3: Violation rate on one-sided censored labels. Lower is better.

Method	**Cens. viol. rate ↓**
Pooled head	–
Dataset-specific heads	–
Three linear measurement heads	–
Three nonlinear measurement heads	75.0%
Three Transformer measurement heads	–
Full relation-aware M3A-DTA	–

central regimes are sequence only (X_d^{seq}, X_p^{seq}), structure only (X_d^{str}, X_p^{str}), ligand structure plus protein sequence, ligand sequence plus protein structure, and all modalities.

Three quantities complement ordinary prediction error. First, the **retention ratio** compares a masked regime's performance with complete-input performance. Second, **enrichment gain** measures the change when a structure is added to a sequence-only example. Third, **prediction instability** measures the absolute change between nested modality sets after controlling for whether the change moves toward the ground truth. We also stratify by protein-structure confidence, ligand-conformer source, sequence length, and pocket coverage.

During evaluation, the same trained checkpoint is used for every modality regime. Training or selecting a different model for each regime would not test the central claim of a unified predictor. Separate-regime models are included only as upper-bound or capacity-matched baselines.

Relation-aware handling is summarized by the censored-label violation rate in Table 3; point-estimation results for the physical measurement heads are reported with the main results rather than repeated in a separate assay-specific subsection.

5 Conclusion and Discussion

We presented M3A-DTA, a unified framework for drug–target affinity prediction with heterogeneous and incomplete evidence. Pretrained molecular, protein, and geometric encoders supply complementary representations, while explicit atom and residue correspondence aligns sequence and structure at the element level. Availability-aware fusion and missing-modality training allow the same model to operate with sequence-only, structure-only, mixed, or complete inputs. Measurement-specific heads preserve the meanings of distinct physical readouts, and composite activity scores contribute through a separate auxiliary objective.

The framework lets abundant sequence data provide statistical coverage and structure-rich complexes provide geometric grounding without requiring every training example or inference query to contain both. Its main limitations arise from uncertain predicted or docked structures, incomplete assay context, source-dependent missingness, and imperfect atom or residue correspondence. Frozen pretrained encoders also improve efficiency at the possible cost of task-specific adaptation. Consequently, cold-split evaluation, modality-only controls, correspondence ablations, and calibration under each availability regime remain essential.

Promising extensions include conformational ensembles, assay metadata as an additional modality, parameter-efficient backbone adaptation, and disagreement-based prioritization of informative experiments. More broadly, M3A-DTA treats heterogeneous biomedical evidence as complementary supervision: a unified predictor can benefit from it only by preserving what is observed, explicitly marking what is absent, and respecting what each measurement represents.

AI use statement

In this work, we used generative AI tools to assist with polishing the grammar and wording of the manuscript and with translation, to retrieve relevant literature, and to support the ideation and execution of research code. We did not use generative AI tools to draft sections of the paper, generate synthetic datasets, or prove mathematical claims. All AI-assisted outputs were reviewed and verified

by the authors. We take responsibility for the final content of this work, including all text, analyses, and artifacts produced with the aid of generative AI.

REFERENCES

Peizhen Bai, Filip Miljkovic, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023. doi: 10.1038/s42256-022-00605-1.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

Yung-Chi Cheng and William H. Prusoff. Relationship between the inhibition constant (k_i) and the concentration of inhibitor which causes 50 per cent inhibition (i_{50}) of an enzymatic reaction. *Biochemical Pharmacology*, 22(23):3099–3108, 1973. doi: 10.1016/0006-2952(73)90196-2.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020. doi: 10.48550/arXiv.2010.09885.

EvolutionaryScale. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. `https://www.evolutionaryscale.ai/blog/esm-cambrian`, 2024. Accessed 2026-09-26.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678.

Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021. doi: 10.1093/bioinformatics/btaa880.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dejun Jiang et al. InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of Medicinal Chemistry*, 64(24):18209–18232, 2021. doi: 10.1021/acs.jmedchem.1c01830.

Huan Yee Koh, Anh T. N. Nguyen, Shirui Pan, Lauren T. May, and Geoffrey I. Webb. PSICHIC: Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *bioRxiv*, 2023. doi: 10.1101/2023.09.17.558145.

Shuangli Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. A unified framework for modeling non-covalent interactions based on graph neural networks. *Cell Systems*, 10(4):308–322.e11, 2020. doi: 10.1016/j.cels.2020.03.002.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein–ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 975–985, 2021. doi: 10.1145/3447548.3467311.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi: 10.1126/science.ade2574.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. TANKBind: Trigonometry-aware neural networks for drug–protein binding structure prediction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7236–7249, 2022.

Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G. Wolynes, and Shuangjia Zheng. DynamicBind: Predicting ligand-specific protein–ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15:1071, 2024. doi: 10.1038/s41467-024-45461-2.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Thin Nguyen, Hang Le, Thomas P. Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021. doi: 10.1093/bioinformatics/btaa921.

Viet Thanh Duy Nguyen, Nhan D. Nguyen, and Truong Son Hy. ProteinReDiff: Complex-based ligand-binding proteins redesign by equivariant diffusion-based generative models. *Structural Dynamics*, 11(6):064102, 2024. doi: 10.1063/4.0000271.

Hakime Ozturk, Arzucan Ozgur, and Elif Ozkirimli. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. doi: 10.1093/bioinformatics/bty593.

Tapio Pahikkala, Antti Airola, Sami Pietila, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2015. doi: 10.1093/bib/bbu010.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Marta M. Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018. doi: 10.1093/bioinformatics/bty374.

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014. doi: 10.1021/ci400709d.

Shuo Zhang, Li Xie, Daniel Tiourine, and Lei Xie. Sequence-based drug–target binding site pretraining enables cryptic pocket detection and improves binding affinity and kinetics prediction. *Journal of Cheminformatics*, 18(1):93, 2026. doi: 10.1186/s13321-026-01227-0.