RESEARCH | Article in Press

Controlled evaluation of architectural, classifier, and training refinements in moltrans-based drug-target interaction prediction

Hao Pang · Fiseha Berhanu Tesema · Tianxiang Cui · Yuan Cheng · Yanwen Mao

Received: 25 May 2026 / Accepted: 31 August 2026

Abstract

Drug–target interaction (DTI) prediction is a central task in computational drug discovery, but performance improvements can be difficult to attribute when architectural modifications and training changes are introduced simultaneously. This study evaluates a Bidirectional Cross-Attention and Global Aggregation DTI model (BCAG-DTI) through eight controlled configurations that separate bidirectional cross-attention, global average–max pooling, classifier design, and optimisation strategy. All principal experiments use fixed training, validation, and test partitions and five matched random seeds on BindingDB, BIOSNAP, and DAVIS. Relative to the MolTrans baseline, the complete BCAG-DTI configuration improves mean area under the receiver operating characteristic curve (AUROC) from 0.8815 to 0.9063 on BindingDB, from 0.8631 to 0.8891 on BIOSNAP, and from 0.8808 to 0.8955 on DAVIS. The corresponding gains in area under the precision–recall curve (AUPRC) are 0.0831, 0.0346, and 0.0619, respectively. Controlled ablation shows that the enhanced classifier achieves the highest mean AUROC on BindingDB and BIOSNAP and the highest mean AUPRC and F1-score on all three datasets, whereas the cross-attention-plus-pooling configuration with enhanced training achieves the highest mean AUROC on DAVIS. The enhanced training strategy also provides substantial improvements, while cross-attention alone reduces performance under the baseline training configuration. In BIOSNAP robustness experiments, BCAG-DTI improves MolTrans for unseen drugs, unseen proteins, and 70–90% missing-interaction settings, whereas its AUROC and F1-score are slightly lower at 95% missing data. As an external same-split reference, CPI-GGS evaluated on the fixed BIOSNAP partitions achieves 0.8619 ± 0.0023 AUROC, 0.8645 ± 0.0042 AUPRC, and 0.7938 ± 0.0028 F1-score, compared with 0.8891 ± 0.0078, 0.8992 ± 0.0067, and 0.8178 ± 0.0094 for BCAG-DTI. This comparison is interpreted in the context of different input preprocessing pipelines and substantial differences in model capacity. Attention case analysis further indicates that cross-attention weights should be treated as model-internal allocation patterns rather than validated binding contacts. Overall, the results show that classifier design and optimisation account for a substantial portion of the improvement over MolTrans, while the contribution of cross-modal architectural components is optimisation-sensitive and dataset-dependent.

Keywords Drug–target interaction prediction · MolTrans · Cross-attention · Ablation study · Transformer · Optimisation · Bioinformatics · Representation learning

https://doi.org/10.1186/s12859-026-06634-6

Published online: 21 September 2026

Background

Drug discovery is a costly and time-intensive process that requires the exploration of a vast chemical space and extensive experimental validation [1, 2]. Computational prediction of drug–target interactions (DTIs) provides an efficient means of prioritising candidate compounds and protein targets for subsequent experimental investigation [3]. With the increasing availability of molecular, protein, and bioactivity data, machine-learning and deep-learning methods have become important tools for DTI prediction [4, 5].

Early deep-learning approaches typically represented drugs using the Simplified Molecular Input Line Entry System (SMILES) and proteins using amino-acid sequences. Convolutional and recurrent neural networks were then used to learn representations for interaction prediction [6–8]. Subsequent methods have incorporated molecular graphs, attention mechanisms, Transformer architectures, and pretrained or structure-aware representations to capture richer relationships between drugs and protein targets.

More recent DTI and compound–protein interaction (CPI) models increasingly employ explicit cross-modal interaction mechanisms, including cross-attention and interaction-aware pooling, together with graph- and structure-based representations [9–14]. Explicit interaction modelling is therefore already established in the field. The present study addresses a narrower question: how do cross-attention, global aggregation, classifier design, and optimisation affect a MolTrans-based sequence model when they are evaluated under a controlled experimental protocol?

MolTrans [15] provides a complementary sequence-based framework. It decomposes drug and protein sequences into frequent consecutive subsequences (FCSs), processes them using independent Transformer encoders, and constructs a dense pairwise interaction map followed by convolutional decoding. This design provides a suitable foundation for investigating how different representation and optimisation choices affect DTI prediction within a common sequence-processing pipeline.

Despite the increasing use of cross-attention and global feature aggregation, their contribution within a MolTrans-based framework is not necessarily independent of classifier design and optimisation strategy. Architectural modifications are often accompanied by changes in decoder capacity, regularisation, learning-rate scheduling, or other training choices. Consequently, improvements observed for a complete model may arise from several interacting factors rather than from a single architectural component. A controlled evaluation is therefore needed to determine how cross-attention, global sequence aggregation, classifier design, and optimisation individually and jointly affect predictive performance.

To investigate this question, we consider a Bidirectional Cross-Attention and Global Aggregation DTI model (BCAG-DTI) built on the MolTrans backbone. Bidirectional cross-attention conditions drug representations on protein features and protein representations on drug features before construction of the pairwise interaction map. A local interaction branch captures pairwise interaction patterns using a two-dimensional convolutional neural network (2D-CNN), while a global branch aggregates sequence-level information through masked average and max pooling. The resulting local and global representations are combined for DTI classification. Because the framework relies on sequence-derived representations rather than experimentally determined binding structures or atom–residue contact annotations, attention weights are treated as model-internal interaction patterns rather than direct evidence of physical binding contacts.

The main contributions of this study are as follows:

- We establish a controlled eight-configuration experimental framework that separately evaluates bidirectional cross-attention, global aggregation, classifier design, and optimisation within a common MolTrans backbone.
- We perform five matched-seed experiments on BindingDB, the Biomedical Network Dataset Collection (BIOSNAP), and the DAVIS kinase–inhibitor dataset, reporting mean performance and standard deviation for the principal comparisons.
- We evaluate generalisation on BIOSNAP under unseen-drug and unseen-protein conditions, together with 70–95% missing-interaction settings, using identical evaluation protocols for MolTrans and BCAG-DTI.

- We provide an external comparison with CPI-GGS, a compound–protein interaction model based on graphs and sequences [16], reproduced using the same fixed BIOSNAP data partitions in five deterministic runs with seeds 1–5.
- We analyse representative cross-attention patterns together with available pharmacological and structural evidence to examine their correspondence with known structural context and to clarify the limitations of attention-based interpretation.

Related work

Sequence-, graph-, and transformer-based models

Early deep-learning approaches to drug–target interaction prediction primarily relied on sequence representations. DeepDTA [7], DeepConv-DTI [8], and WideDTA [17] use convolutional architectures to learn representations from drug SMILES strings and protein sequences. These methods demonstrated the effectiveness of end-to-end representation learning but largely model drug and protein features independently before their integration.

Graph-based approaches subsequently incorporated molecular topology into drug representations. GNN-CPI and GraphDTA employ graph neural networks (GNNs) to represent molecular structures while learning complementary protein features [18, 19]. Such approaches provide a richer representation of molecular connectivity than sequence-only drug encodings.

Transformer-based models further improved the ability to capture long-range dependencies. MolTrans [15] decomposes drug and protein sequences into frequent consecutive subsequences (FCSs), processes them using independent Transformer encoders, and constructs a dense drug–protein interaction map for subsequent convolutional feature extraction. This architecture provides the backbone adopted in the present study.

Interaction-aware DTI modelling

Beyond independent drug and protein encoding, several recent approaches explicitly model interactions between the two modalities. MCL-DTI integrates multimodal drug information with protein representations using bidirectional multi-head cross-attention [9]. CAT-DTI employs cross-attention to capture drug–target interaction features and incorporates domain adaptation to improve performance under distribution shifts [10]. FMCA-DTI extracts multiple drug and protein fragments and uses shared-weight multi-head cross-attention to learn interactions between them [11].

GraphormerDTI combines Graph Transformer-based molecular representations with protein sequence features and applies an attention mechanism to model drug–protein interactions [12]. CPI-GGS combines graph-based compound representations with sequence-based protein encoding and an attention-based interaction module [16]. GGAP-CPI employs ligand and protein representations together with multi-head cross-attention pooling to construct compound–protein interaction representations [14].

These studies demonstrate that explicit cross-modal interaction modelling is already established in DTI and CPI prediction. The bidirectional cross-attention considered in this study is therefore used as a refinement of the MolTrans representation pipeline rather than as a new interaction-learning mechanism.

Structure-aware and generalisation-oriented models

Recent studies have also explored structural representations and evaluation settings designed to assess generalisation beyond randomly partitioned interaction pairs. GraphormerDTI uses a Graph Transformer to encode molecular topology and evaluates its ability to predict interactions involving previously unseen molecules [12].

iNGNN-DTI represents drugs and targets using nested graph neural networks, incorporates pretrained molecular and protein representations, and constructs protein graphs using predicted three-dimensional structures. Cross-attention is then used to model interactions between drug and target substructures [13].

GGAP-CPI addresses compound–protein interaction prediction using large-scale heterogeneous bioactivity data and evaluates the model under several challenging scenarios, including rare-protein prediction, transfer learning, and virtual screening [14]. These methods illustrate the increasing emphasis on structural information and out-of-distribution evaluation in DTI and CPI modelling.

The present study addresses a more restricted setting. It focuses on sequence-based MolTrans representations and investigates how cross-attention, global aggregation, classifier design, and optimisation affect performance when the underlying data partitions and evaluation protocol are controlled. Consequently, comparisons with graph- or structure-aware approaches are interpreted in the context of their different input representations and experimental protocols.

Methods

Figure 1 provides an overview of the BCAG-DTI pipeline. The framework builds on the MolTrans backbone and incorporates bidirectional cross-attention, local interaction modelling, global sequence aggregation, and an enhanced classifier. The contributions of these components are evaluated individually and in combination through the controlled configurations described in Section 3.8.

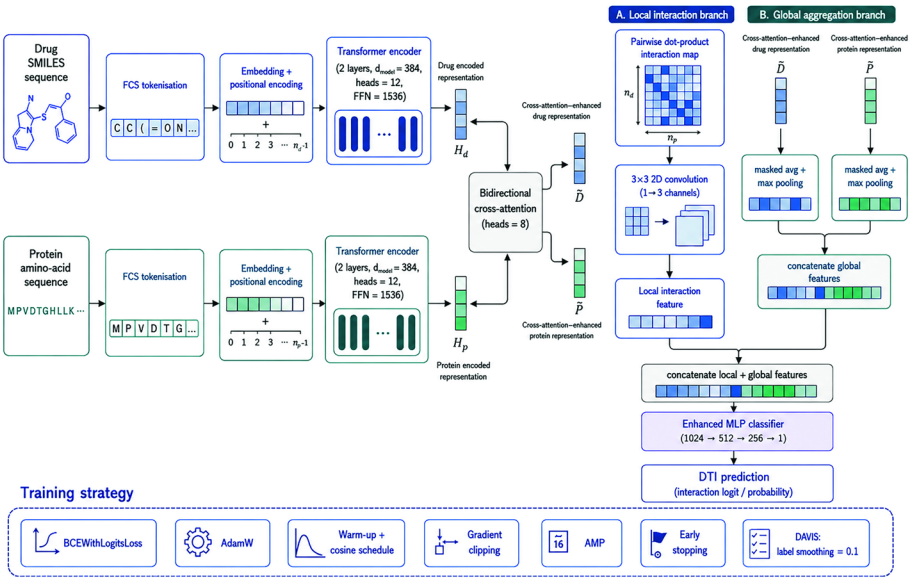

Fig. 1 Overview of the BCAG-DTI architecture. Drug and protein sequences undergo FCS-derived substructure tokenisation, shown as "FCS tokenisation" in the figure, using the substructure vocabularies implemented in MolTrans. The resulting tokens are embedded with positional information and processed independently by Transformer encoders. Bidirectional cross-attention produces cross-attention-enhanced drug and protein representations. The local interaction branch constructs a pairwise dot-product interaction map and processes it using a 3 × 3 two-dimensional convolution with three output channels. In parallel, the global aggregation branch applies masked average and max pooling separately to the drug and protein representations and concatenates the resulting pooled features. The local interaction feature and global representation are then concatenated and passed to the enhanced multilayer perceptron (MLP) classifier to produce the DTI prediction. The training-strategy panel summarises the enhanced training configuration used by A2, A7, and A8; the baseline-training configurations use the settings reported in Table 1

Problem formulation

Given a drug SMILES sequence d and a protein amino-acid sequence p, the model learns a scoring function

$$z_\theta : \mathcal{D} \times \mathcal{P} \to \mathbb{R}, \tag{1}$$

where $z_\theta(d, p)$ denotes the predicted interaction logit. The corresponding interaction probability is obtained as

$$\hat{y} = \sigma\left(z_\theta(d, p)\right), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $y \in \{0, 1\}$ is the ground-truth interaction label. Performance is evaluated using the area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), and F1-score.

MolTrans backbone

The backbone follows the MolTrans sequence-processing design. MolTrans uses frequent consecutive subsequence (FCS) mining to derive recurring drug and protein substructures [15]. In the released MolTrans implementation, the corresponding substructure vocabularies are provided as ESPF assets and applied using the subword tokenisation procedure implemented in the original code. We follow the same implementation and therefore refer to this representation as FCS-derived substructure tokenisation. The resulting drug and protein tokens are embedded with positional information and processed using independent Transformer encoders.

The embedding dimension is $E = 384$. Each encoder contains two Transformer layers with 12 self-attention heads and a feed-forward intermediate dimension of 1536. The maximum drug and protein token lengths are 50 and 545, respectively. The drug and protein vocabularies contain 23,532 and 16,693 tokens, respectively.

Let the encoded drug and protein representations be

$$\begin{aligned} D &\in \mathbb{R}^{B \times L_d \times E}, \\ P &\in \mathbb{R}^{B \times L_p \times E}, \end{aligned} \tag{3}$$

where B denotes the batch size, L_d and L_p denote the drug and protein token lengths, respectively, and E is the embedding dimension.

Bidirectional cross-attention

For configurations containing cross-attention, drug tokens attend to protein representations and protein tokens attend to drug representations. Using $\mathrm{MHA}(Q, K, V)$ to denote multi-head attention with query Q, key K, and value V, the cross-modal representations are computed as

$$\begin{aligned} \widetilde{D} &= \mathrm{LN}\left(D + \mathrm{MHA}(D, P, P)\right), \\ \widetilde{P} &= \mathrm{LN}\left(P + \mathrm{MHA}(P, D, D)\right), \end{aligned} \tag{4}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalisation. The cross-attention module uses eight attention heads, with padded positions excluded through attention masking.

This operation allows each modality to incorporate information from the other before pairwise interaction modelling. Because the model does not use experimentally annotated atom–residue contacts or binding-site supervision, the resulting attention weights are treated as internal model representations rather than physical binding contacts.

Pairwise interaction map and local features

The local interaction branch constructs a pairwise interaction map from the drug and protein representations. For configurations containing cross-attention, the map is computed as

$$M = \widetilde{D}\widetilde{P}^{\top} \in \mathbb{R}^{B\times L_d\times L_p}, \tag{5}$$

where the matrix multiplication is performed along the embedding dimension. For configurations without cross-attention, D and P are used directly in place of $\widetilde{D}$ and $\widetilde{P}$.

The resulting interaction map is expanded to a single input channel, regularised using dropout, and processed by a 3×3 two-dimensional convolution with three output channels. The convolutional output is then flattened to obtain the local interaction representation.

Global average–max aggregation

The global branch captures sequence-level information using masked average pooling and masked max pooling. For a sequence representation H and valid-token mask m, the pooled representations are defined as

$$\begin{aligned} h_{\text{avg}} &= \frac{\sum_i m_i h_i}{\sum_i m_i}, \\ h_{\text{max}} &= \max_{i:m_i=1} h_i. \end{aligned} \tag{6}$$

For configurations containing cross-attention, pooling is applied to $\widetilde{D}$ and $\widetilde{P}$; otherwise, it is applied to D and P. The global representation is then constructed as

$$G = \left[d_{\text{avg}}, d_{\text{max}}, p_{\text{avg}}, p_{\text{max}}\right]. \tag{7}$$

No learned attention-pooling operation is used. When global aggregation is enabled, G is concatenated with the flattened local interaction features before classification.

Classifier design

The MolTrans baseline classifier consists of linear layers with 512, 64, and 32 hidden units. ReLU activation is used together with Batch Normalisation in the first two hidden stages.

The enhanced classifier uses hidden dimensions of $1024 \rightarrow 512 \rightarrow 256$, followed by a scalar output layer. Each hidden layer is followed by Layer Normalisation, Gaussian Error Linear Unit (GELU) activation, and dropout. No residual connections are used in the enhanced classifier. Linear-layer weights are initialised using Xavier uniform initialisation.

Training strategies

Two training strategies are considered in the controlled experiments. The baseline strategy is used by A1, A3, A4, A5, and A6, whereas the enhanced strategy is used by A2, A7, and A8. The main optimisation and training settings for these strategies are summarised in Table 1.

As shown in Table 1, the baseline strategy uses Adam with a learning rate of 1×10^{-4}, a global batch size of 16, a dropout rate of 0.10, and a fixed training duration of 50 epochs. No weight decay, learning-rate warm-up or scheduling, automatic mixed precision (AMP), gradient clipping, early stopping, or label smoothing is used.

Although early stopping is disabled for the baseline configurations, model selection is still based on validation performance. Validation metrics are computed after every epoch, and the checkpoint is replaced whenever

Table 1 Training settings used for the baseline and enhanced training strategies

Setting	Baseline	Enhanced: BindingDB/BIOSNAP	Enhanced: DAVIS
Optimiser	Adam	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}	1.2×10^{-4}
Weight decay	0	1×10^{-5}	5×10^{-6}
Batch size	16	16	16
Maximum epochs	50	50	60
Warm-up epochs	None	3	5
LR schedule	None	Cosine	Cosine
Early stopping	None	Patience 15	Patience 25
Gradient clipping	None	1.0	1.0
AMP	No	Yes	Yes
Label smoothing	0	0	0.1
Dropout	0.10	0.15	0.10
Checkpoint criterion	Val. AUROC	Val. AUROC	Val. AUROC

Table 2 Controlled model configurations used to evaluate the effects of architectural components, classifier design, and training strategy

ID	Configuration	Cross-attn.	Global pool.	Enhanced classifier	Enhanced training
A1	MolTrans baseline	No	No	No	No
A2	Training strategy	No	No	No	Yes
A3	Cross-attention	Yes	No	No	No
A4	Global pooling	No	Yes	No	No
A5	Enhanced classifier	No	No	Yes	No
A6	Cross-attention + pooling	Yes	Yes	No	No
A7	Cross-attention + pooling + training	Yes	Yes	No	Yes
A8	Complete BCAG-DTI	Yes	Yes	Yes	Yes

validation AUROC strictly exceeds the previous best value. After all 50 epochs have been completed, the checkpoint with the highest validation AUROC is reloaded for test evaluation rather than using the final-epoch model.

For the enhanced strategy, the dataset-specific settings are also summarised in Table 1. On BindingDB and BIOSNAP, AdamW is used with a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, a dropout rate of 0.15, three warm-up epochs followed by cosine decay, and early stopping with patience 15. On DAVIS, the learning rate is 1.2×10^{-4}, weight decay is 5×10^{-6}, the dropout rate is 0.10, five warm-up epochs are used, and early-stopping patience is increased to 25. DAVIS also uses label smoothing of 0.1. Gradient clipping with a maximum norm of 1.0 and AMP are enabled for all enhanced-training configurations. The global batch size remains 16.

For all configurations, checkpoint selection is based on the highest validation AUROC. For A2, A7, and A8, training may terminate before the maximum number of epochs when the corresponding early-stopping patience is reached; the best validation-AUROC checkpoint is then reloaded for test evaluation.

Controlled experimental configurations

To isolate the effects of cross-attention, global aggregation, classifier design, and optimisation, eight controlled configurations are evaluated. Table 2 summarises the components included in each configuration. A1 represents the MolTrans baseline, A2–A7 introduce individual components or selected combinations, and A8 represents the complete BCAG-DTI configuration.

Table 3 Summary of the benchmark datasets used in the experiments

Metric	BindingDB	BIOSNAP	DAVIS
Number of drugs	7,165	4,510	68
Number of proteins	1,254	2,181	379
Number of interactions	32,601	27,483	11,103
Positive interactions	9,166	13,836	1,506
Negative interactions	23,435	13,647	9,597
Positive ratio (%)	28.12	50.34	13.56
Negative ratio (%)	71.88	49.66	86.44

Table 4 Fixed training, validation, and test partitions used for the three benchmark datasets

Dataset	Train	Validation	Test
BindingDB	12,668	6,644	13,289
BIOSNAP	19,238	2,748	5,497
DAVIS	2,086	3,006	6,011

Experiments

Datasets and fixed splits

We evaluate the framework on three widely used drug–target interaction benchmark datasets: BindingDB, BIOSNAP, and DAVIS [20–22]. The datasets differ substantially in scale and class balance, providing complementary settings for evaluating predictive performance.

BindingDB contains experimentally measured interactions between small molecules and protein targets. BIOSNAP provides a comparatively balanced collection of biomedical drug–target interactions, whereas DAVIS consists of kinase–inhibitor interactions and exhibits substantially greater class imbalance.

Table 3 summarises the principal characteristics of the three datasets.

Fixed training, validation, and test partitions are used throughout the controlled experiments so that all model configurations are evaluated under identical data conditions. The partition sizes are reported in Table 4.

All A1–A8 configurations use the same fixed partitions and five matched random seeds (1–5). The exact train, validation, and test split files used in all controlled experiments are provided in the accompanying code repository to enable reproduction of the reported results.

For the BIOSNAP robustness analysis, additional predefined partitions are used to evaluate generalisation to unseen drugs and unseen proteins, together with 70%, 80%, 90%, and 95% missing-interaction conditions. MolTrans (A1) and BCAG-DTI (A8) are evaluated using identical partitions and matched random seeds in each robustness setting.

Evaluation protocol

Each A1–A8 configuration is evaluated using five matched random seeds (1–5). Performance is reported as the mean and standard deviation of the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), and the F1-score.

AUROC measures ranking performance across both classes, whereas AUPRC is particularly informative for imbalanced datasets such as DAVIS. AUROC and AUPRC are computed directly from the predicted probabilities. The F1-score is computed using a classification threshold selected exclusively from the validation set.

For all A1–A8 configurations, validation metrics are evaluated after each epoch and model selection is based on validation AUROC. Whenever an epoch achieves a validation AUROC that strictly exceeds the previous best value, the corresponding model checkpoint is saved. At the same validation evaluation, the probability

threshold that maximises validation F1 is determined and stored together with that checkpoint. After training, the checkpoint with the highest validation AUROC is reloaded, and its associated validation-derived F1 threshold is applied unchanged to the test predictions. Thus, checkpoint selection is based on validation AUROC, whereas the F1 threshold associated with the selected checkpoint is determined from validation F1. Neither model selection nor threshold optimisation uses the test set.

MolTrans (A1) serves as the primary within-framework baseline because it uses the same FCS-derived substructure tokenisation, sequence encoders, fixed data partitions, matched seeds, and evaluation procedure as the remaining configurations. This enables direct assessment of the effects of cross-attention, global aggregation, classifier design, and optimisation under a controlled protocol.

Results reported for methods from previous studies are considered separately because differences in preprocessing, negative sampling, data partitioning, input representations, and metric definitions may prevent strictly protocol-matched comparisons. The reproduced CPI-GGS experiment is therefore treated as an external same-split reference rather than a fully protocol-identical comparison.

External CPI-GGS comparison

To provide a recent interaction-aware external comparison, CPI-GGS [16] is evaluated using the same fixed BIOSNAP training, validation, and test partitions adopted for MolTrans and BCAG-DTI. The reproduction retains the published graph- and sequence-based model components together with its core optimisation schedule. The reproduced CPI-GGS model is trained for a maximum of 20 epochs with a batch size of 32 and early-stopping patience of 5.

For compatibility with the evaluation protocol used in this study, the internal random split of CPI-GGS is replaced by the fixed BIOSNAP partitions. Five deterministic runs are performed using seeds 1–5. For each run, the checkpoint with the highest validation AUROC is selected. The F1 classification threshold is determined using the validation set and then applied unchanged to the corresponding test set.

CPI-GGS represents compounds as RDKit molecular graphs using radius-2 Weisfeiler–Lehman fingerprints and encodes proteins using overlapping 3-grams. In contrast, MolTrans and BCAG-DTI use FCS-derived substructure tokenisation implemented with the vocabulary assets released with MolTrans.

Consequently, although all three models are evaluated using the same BIOSNAP partitions, their input preprocessing pipelines are not identical. The CPI-GGS results are therefore used as an external same-split reference rather than as a fully protocol-identical comparison.

Results

Repeated-seed comparison with MolTrans

Table 5 reports the test performance of the MolTrans baseline (A1) and the complete BCAG-DTI configuration (A8) across five matched random seeds. A8 achieves higher mean AUROC, AUPRC, and F1-score than A1 on all three datasets.

On BindingDB, mean AUROC increases from 0.8815 to 0.9063, corresponding to an improvement of 0.0248. Mean AUPRC increases from 0.5231 to 0.6062, while F1-score increases from 0.5571 to 0.6015, representing gains of 0.0831 and 0.0444, respectively.

On BIOSNAP, A8 improves mean AUROC from 0.8631 to 0.8891, AUPRC from 0.8646 to 0.8992, and F1-score from 0.7988 to 0.8178. The corresponding absolute improvements are 0.0260, 0.0346, and 0.0190.

On DAVIS, mean AUROC increases from 0.8808 to 0.8955, while AUPRC improves from 0.2926 to 0.3545 and F1-score from 0.3913 to 0.4176. These correspond to improvements of 0.0146, 0.0619, and 0.0263, respectively.

Table 5 Five-seed test performance of the MolTrans baseline (A1) and the complete BCAG-DTI configuration (A8). Values are reported as mean ± standard deviation

Dataset	Model	AUROC	AUPRC	F1
BindingDB	A1 MolTrans	0.8815 ± 0.0028	0.5231 ± 0.0074	0.5571 ± 0.0089
	A8 BCAG-DTI	0.9063 ± 0.0011	0.6062 ± 0.0141	0.6015 ± 0.0056
BIOSNAP	A1 MolTrans	0.8631 ± 0.0053	0.8646 ± 0.0139	0.7988 ± 0.0061
	A8 BCAG-DTI	0.8891 ± 0.0078	0.8992 ± 0.0067	0.8178 ± 0.0094
DAVIS	A1 MolTrans	0.8808 ± 0.0128	0.2926 ± 0.0512	0.3913 ± 0.0302
	A8 BCAG-DTI	0.8955 ± 0.0060	0.3545 ± 0.0303	0.4176 ± 0.0233

The larger standard deviations observed for DAVIS, particularly for AUPRC, indicate greater variation across seeds than on BindingDB and BIOSNAP.

Figure 2 provides a visual comparison of the five-seed mean performance and variability of A1 and A8 across the three datasets.

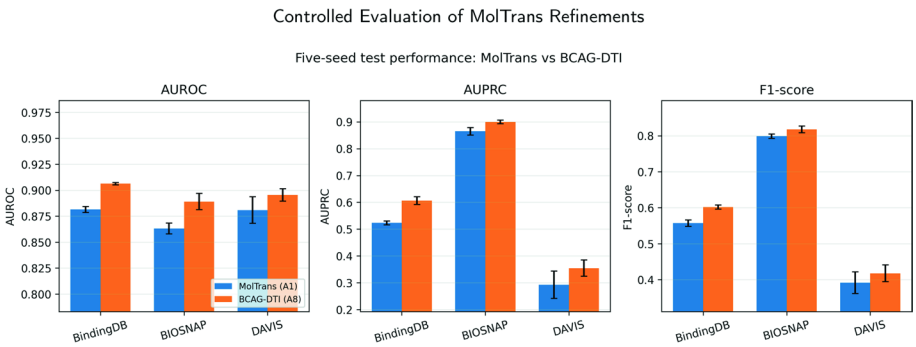

Fig. 2 Mean test AUROC, AUPRC, and F1-score of MolTrans (A1) and BCAG-DTI (A8) across five matched random seeds on BindingDB, BIOSNAP, and DAVIS. Error bars represent one standard deviation

Table 6 Five-seed BIOSNAP performance of MolTrans, CPI-GGS, and BCAG-DTI using the same fixed training, validation, and test partitions

Model	AUROC	AUPRC	F1
A1 MolTrans	0.8631 ± 0.0053	0.8646 ± 0.0139	0.7988 ± 0.0061
CPI-GGS [16]	0.8619 ± 0.0023	0.8645 ± 0.0042	0.7938 ± 0.0028
A8 BCAG-DTI	0.8891 ± 0.0078	0.8992 ± 0.0067	0.8178 ± 0.0094

CPI-GGS is included as an external same-split reference and uses different input representations and preprocessing, as well as substantially lower model capacity (0.73M trainable parameters versus 106.26M for A8). Values are reported as mean ± standard deviation

External CPI-GGS comparison

To complement the within-framework comparison, CPI-GGS [16] is evaluated using the same fixed BIOSNAP training, validation, and test partitions. Five deterministic runs are performed using seeds 1–5. Table 6 compares its performance with MolTrans (A1) and BCAG-DTI (A8).

CPI-GGS achieves a mean AUROC of 0.8619 ± 0.0023, AUPRC of 0.8645 ± 0.0042, and F1-score of 0.7938 ± 0.0028. These values are close to those obtained by MolTrans, which achieves 0.8631 ± 0.0053 AUROC, 0.8646 ± 0.0139 AUPRC, and 0.7988 ± 0.0061 F1-score.

Under the same fixed BIOSNAP partitions, BCAG-DTI achieves the highest mean performance among the three models, with 0.8891 ± 0.0078 AUROC, 0.8992 ± 0.0067 AUPRC, and 0.8178 ± 0.0094 F1-score.

The input representations used by CPI-GGS differ from those used by the MolTrans-based models. CPI-GGS represents compounds as molecular graphs and proteins using overlapping 3-grams, whereas MolTrans

and BCAG-DTI use FCS-derived substructure tokenisation. The CPI-GGS results therefore provide an external same-split reference rather than a fully protocol-identical comparison.

Model capacity also differs substantially. The reproduced CPI-GGS model contains 725,260 trainable parameters, whereas A8 contains 106,261,023 trainable parameters. The higher predictive performance of A8 is therefore accompanied by substantially greater model capacity. This capacity–performance trade-off is considered further in the Discussion.

Controlled ablation: architecture and optimisation

Table 7 reports the controlled A1–A8 ablation across BindingDB, BIOSNAP, and DAVIS. The results reveal substantial differences in the contributions of individual architectural, classifier, and optimisation components.

The enhanced training strategy alone (A2) improves the MolTrans baseline on all three datasets. On BindingDB, A2 increases mean AUROC from 0.8815 to 0.9070 and AUPRC from 0.5231 to 0.6128. Similar gains are observed on BIOSNAP and DAVIS, indicating that optimisation contributes substantially even without changes to the representation architecture.

The enhanced classifier alone (A5) provides particularly strong performance. On BindingDB, A5 achieves the highest mean AUROC (0.9110), AUPRC (0.6220), and F1-score (0.6080) among the eight configurations. On BIOSNAP, A5 again achieves the highest mean AUROC (0.8927), AUPRC (0.9037), and F1-score (0.8199). On DAVIS, A5 achieves the highest mean AUPRC (0.3699) and F1-score (0.4372), whereas A7 achieves the highest mean AUROC (0.9027).

In contrast, cross-attention alone (A3) reduces performance substantially on all three datasets. On BindingDB, mean AUROC decreases from 0.8815 for A1 to 0.8059 for A3. Similar reductions occur on BIOSNAP and

Table 7 Controlled five-seed ablation across BindingDB, BIOSNAP, and DAVIS

Dataset	Configuration	AUROC	AUPRC	F1
BindingDB	A1 MolTrans baseline	0.8815 ± 0.0028	0.5231 ± 0.0074	0.5571 ± 0.0089
	A2 Training strategy	0.9070 ± 0.0075	0.6128 ± 0.0337	0.6065 ± 0.0181
	A3 Cross-attention	0.8059 ± 0.0395	0.3900 ± 0.0417	0.4672 ± 0.0452
	A4 Global pooling	0.8824 ± 0.0095	0.5295 ± 0.0253	0.5665 ± 0.0115
	A5 Enhanced classifier	0.9110 ± 0.0020	0.6220 ± 0.0098	0.6080 ± 0.0034
	A6 Cross-attention + pooling	0.7808 ± 0.0462	0.3733 ± 0.0492	0.4395 ± 0.0470
	A7 Cross-attention + pooling + training	0.9017 ± 0.0015	0.5898 ± 0.0166	0.6035 ± 0.0033
	A8 BCAG-DTI	0.9063 ± 0.0011	0.6062 ± 0.0141	0.6015 ± 0.0056
BIOSNAP	A1 MolTrans baseline	0.8631 ± 0.0053	0.8646 ± 0.0139	0.7988 ± 0.0061
	A2 Training strategy	0.8829 ± 0.0050	0.8809 ± 0.0112	0.8179 ± 0.0045
	A3 Cross-attention	0.7406 ± 0.0344	0.7402 ± 0.0412	0.7023 ± 0.0171
	A4 Global pooling	0.8491 ± 0.0125	0.8530 ± 0.0167	0.7839 ± 0.0156
	A5 Enhanced classifier	0.8927 ± 0.0020	0.9037 ± 0.0042	0.8199 ± 0.0053
	A6 Cross-attention + pooling	0.7328 ± 0.0259	0.7502 ± 0.0386	0.6962 ± 0.0058
	A7 Cross-attention + pooling + training	0.8722 ± 0.0139	0.8649 ± 0.0161	0.8097 ± 0.0110
	A8 BCAG-DTI	0.8891 ± 0.0078	0.8992 ± 0.0067	0.8178 ± 0.0094
DAVIS	A1 MolTrans baseline	0.8808 ± 0.0128	0.2926 ± 0.0512	0.3913 ± 0.0302
	A2 Training strategy	0.8997 ± 0.0119	0.3415 ± 0.0449	0.4309 ± 0.0362
	A3 Cross-attention	0.8247 ± 0.0374	0.2122 ± 0.0583	0.3313 ± 0.0652
	A4 Global pooling	0.8717 ± 0.0106	0.2781 ± 0.0344	0.3773 ± 0.0264
	A5 Enhanced classifier	0.9021 ± 0.0077	0.3699 ± 0.0212	0.4372 ± 0.0170
	A6 Cross-attention + pooling	0.8244 ± 0.0152	0.2063 ± 0.0338	0.3205 ± 0.0169
	A7 Cross-attention + pooling + training	0.9027 ± 0.0049	0.3492 ± 0.0365	0.4294 ± 0.0192
	A8 BCAG-DTI	0.8955 ± 0.0060	0.3545 ± 0.0303	0.4176 ± 0.0233

Values are reported as mean ± standard deviation. Bold values indicate the highest mean performance for each metric within a dataset

DAVIS. The combination of cross-attention and global pooling under the baseline training strategy (A6) also performs below the MolTrans baseline.

Global pooling alone (A4) has a more limited effect. Its performance remains close to A1 on BindingDB, but both AUROC and AUPRC are lower than the baseline on BIOSNAP and DAVIS.

The interaction between architecture and optimisation becomes apparent when comparing A6 and A7. Both configurations contain cross-attention and global pooling, but A7 additionally uses the enhanced training strategy. On DAVIS, for example, mean AUROC increases from 0.8244 for A6 to 0.9027 for A7. Similar recovery is observed on BindingDB and BIOSNAP. This suggests that the effectiveness of the cross-modal architectural components depends strongly on the optimisation setting.

The complete BCAG-DTI configuration (A8) consistently outperforms the MolTrans baseline across all three datasets and metrics. However, it does not dominate the controlled ablation. A5 achieves higher AUPRC and F1-score than A8 on all three datasets, while A7 achieves the highest AUROC on DAVIS. These results indicate that classifier design and optimisation contribute substantially to the overall performance gains, while cross-attention and global aggregation exhibit non-trivial interactions with the training strategy.

Validation trajectories

Figure 3 presents representative per-epoch validation AUROC and AUPRC trajectories obtained from recorded training histories for the baseline and selected component configurations. The curves provide a qualitative view of the optimisation behaviour of the different configurations during training.

The trajectories should not be interpreted as repeated-run estimates of convergence speed or training stability because they correspond to individual recorded runs and have unequal training lengths. Quantitative comparisons of predictive performance and variability are therefore based on the five-seed test summaries reported in Tables 5 and 7.

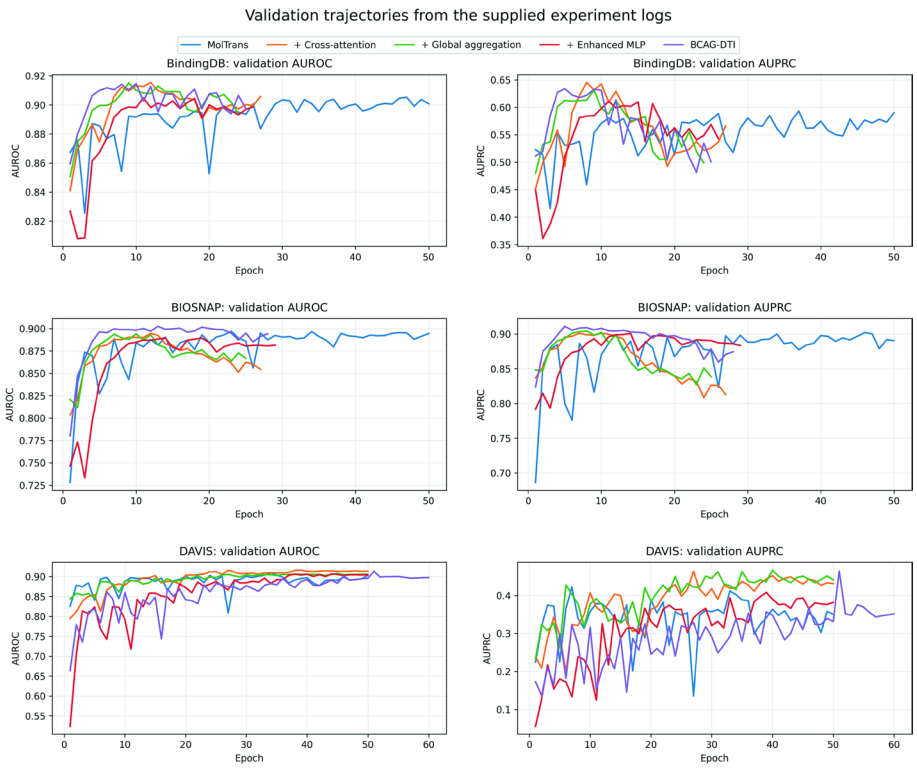

Fig. 3 Representative per-epoch validation AUROC and AUPRC trajectories for selected model configurations. The trajectories illustrate optimisation behaviour for individual runs and are not used as repeated-run estimates of training stability

Unseen-entity and missing-data robustness

Table 8 compares MolTrans (A1) and BCAG-DTI (A8) under unseen-entity and missing-interaction conditions on BIOSNAP. Both models are evaluated using identical partitions and five matched random seeds.

For unseen drugs, A8 improves mean AUROC from 0.8245 to 0.8349, corresponding to an absolute gain of 0.0105. Mean AUPRC increases from 0.8468 to 0.8648, while F1-score increases slightly from 0.7563 to 0.7582.

The unseen-protein setting is more challenging for both models. MolTrans achieves a mean AUROC of 0.6482, whereas A8 reaches 0.6639, an improvement of 0.0157. Mean AUPRC increases from 0.6566 to 0.6820, and F1-score increases from 0.4911 to 0.5029. The lower absolute performance in the unseen-protein setting indicates that generalisation to previously unseen targets remains more difficult than generalisation to unseen drugs.

Under the missing-interaction conditions, A8 improves mean AUROC by 0.0194 at 70% missing data, 0.0110 at 80%, and 0.0096 at 90%. Mean AUPRC is also higher for A8 at each of these levels. The largest improvement occurs at the 70% missing-interaction level, where AUROC increases from 0.8349 to 0.8543 and AUPRC from 0.8431 to 0.8672.

At 95% missing data, the advantage is no longer consistent across all metrics. A8 obtains slightly lower mean AUROC (0.7491 versus 0.7510) and F1-score (0.6933 versus 0.6986), while maintaining a higher mean AUPRC (0.7755 versus 0.7658). The results therefore indicate improved performance across several unseen-entity and data-scarcity settings, but not across every metric under the most extreme missing-interaction condition.

Figure 4 provides a complementary visual summary of the robustness experiments. The unseen-entity comparison includes the full BIOSNAP setting as a reference and shows the larger performance reduction associated with unseen proteins than with unseen drugs. The missing-interaction curves show progressive degradation for both models as the proportion of missing interaction data increases. A8 maintains higher mean AUROC and AUPRC through the 90% condition, whereas the AUROC advantage disappears at 95% missing data.

Attention case analysis

To examine the representations learned by the bidirectional cross-attention module, we analyse representative predictions from the BIOSNAP test set. The analysis uses the saved A8 checkpoint from seed 3 and includes true-positive, false-positive, and false-negative cases.

For each example, attention weights are first averaged across the eight cross-attention heads. Drug-to-protein and protein-to-drug attention are normalised separately by their total attention mass. The protein-to-drug matrix is then transposed to obtain a common drug–protein orientation, after which the two directions are equally averaged. Padded positions are removed before visualisation. Protein-token attention is projected onto 1-based residue positions using a length-weighted, mass-preserving mapping. The resulting maps therefore describe model-internal attention allocation rather than three-dimensional atom–residue contacts.

Figure 5 presents three representative cases. For the true-positive corticosterone–HSD11B1 interaction, A8 assigns an interaction probability of 0.999996. Structural context is available from the 3.0 Å X-ray crystal structure

Table 8 BIOSNAP robustness comparison between MolTrans (A1) and BCAG-DTI (A8)

Setting	A1 AUROC	A8 AUROC	A1 AUPRC	A8 AUPRC	A1 F1	A8 F1
Unseen drug	0.8245 ± 0.0132	0.8349 ± 0.0074	0.8468 ± 0.0069	0.8648 ± 0.0051	0.7563 ± 0.0198	0.7582 ± 0.0085
Unseen protein	0.6482 ± 0.0142	0.6639 ± 0.0165	0.6566 ± 0.0147	0.6820 ± 0.0139	0.4911 ± 0.0285	0.5029 ± 0.0344
70% missing	0.8349 ± 0.0052	0.8543 ± 0.0063	0.8431 ± 0.0065	0.8672 ± 0.0050	0.7668 ± 0.0051	0.7824 ± 0.0050
80% missing	0.8212 ± 0.0047	0.8322 ± 0.0055	0.8313 ± 0.0059	0.8506 ± 0.0052	0.7558 ± 0.0060	0.7576 ± 0.0063
90% missing	0.7809 ± 0.0040	0.7905 ± 0.0068	0.7900 ± 0.0227	0.8119 ± 0.0069	0.7221 ± 0.0047	0.7235 ± 0.0057
95% missing	0.7510 ± 0.0069	0.7491 ± 0.0065	0.7658 ± 0.0152	0.7755 ± 0.0082	0.6986 ± 0.0031	0.6933 ± 0.0041

Values are reported as mean ± standard deviation over five matched random seeds. Bold values indicate the higher mean performance within each evaluation setting

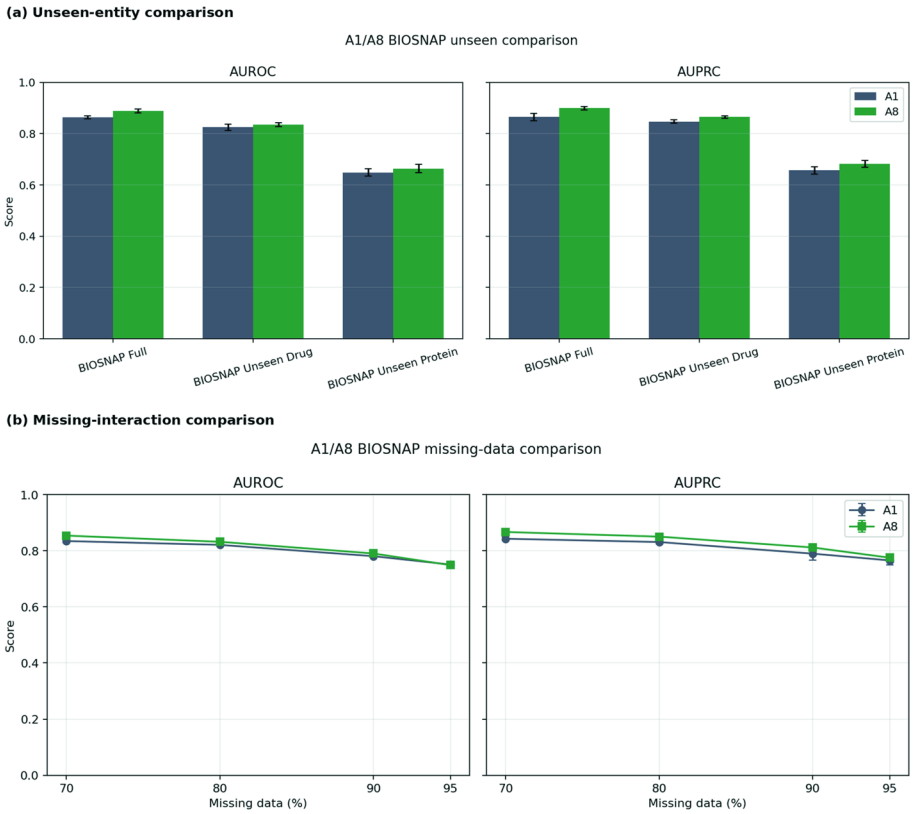

Fig. 4 Robustness comparison between MolTrans (A1) and BCAG-DTI (A8) on BIOSNAP. **a** AUROC and AUPRC under the full-data, unseen-drug, and unseen-protein settings. **b** AUROC and AUPRC under the 70%, 80%, 90%, and 95% missing-interaction conditions. Values represent means across five matched random seeds, with error bars denoting one standard deviation

of murine HSD11B1 in complex with corticosterone and NADP(H) (PDB 1Y5R) [23]. However, because this structure is from the murine ortholog, the highest-ranked residues in the human HSD11B1 attention analysis cannot be directly interpreted as experimentally validated corticosterone-contact residues. The example therefore provides structural context rather than residue-level validation of the attention map.

An additional true-positive case involving itopride–DRD2 provides a similar illustration. A8 assigns an interaction probability of 0.999996, and residue D114 appears among the five highest-ranked protein residues. In the crystal structure of human DRD2 bound to risperidone (PDB 6CM4), Asp114$^{3.32}$ forms a salt bridge with the tertiary amine of risperidone [24]. Thus, the high attention assigned to D114 coincides with a residue known to participate directly in ligand recognition in DRD2. However, PDB 6CM4 contains risperidone rather than itopride, and therefore does not validate D114 as an itopride-contact residue. This correspondence is consequently treated as structural context rather than evidence that the attention mechanism has recovered an experimentally validated itopride-binding determinant.

The false-positive diphenhydramine–OPRK1 case demonstrates a different limitation. A8 assigns an interaction probability of 0.999996 despite the negative reference label. The corresponding attention map contains concentrated regions and identifiable high-ranking fragments and residues. However, attention concentration does not establish that the predicted interaction is biologically correct. No reliable direct diphenhydramine–OPRK1 binding evidence was identified in the accompanying evidence review; importantly, absence of such evidence is not evidence of non-binding.

The false-negative paclitaxel–BCL2 example provides the complementary failure mode. A8 assigns an interaction probability of only 0.000061 even though direct paclitaxel–BCL2 binding has been experimentally reported [25]. For the same test pair, the MolTrans baseline (A1) assigns a probability of 0.789312 and produces the correct positive prediction. This example demonstrates that A8 can fail on individual interactions that are correctly classified by A1 and further cautions against interpreting attention magnitude as evidence for or against physical binding.

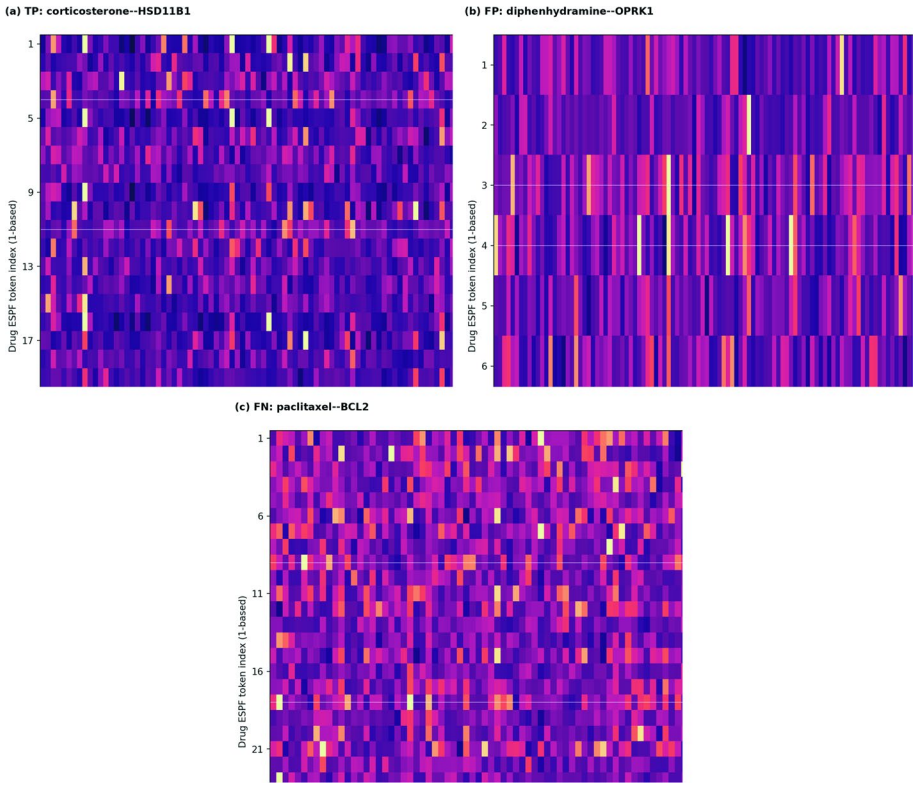

Fig. 5 Representative BCAG-DTI bidirectional cross-attention maps from the BIOSNAP test set: (**a**) a true-positive corticosterone–HSD11B1 interaction, (**b**) a false-positive diphenhydramine–OPRK1 interaction, and (**c**) a false-negative paclitaxel–BCL2 interaction. The maps are obtained from the saved A8 checkpoint for seed 3. Attention is averaged across eight heads; the two cross-attention directions are normalised separately, aligned to a common drug–protein orientation, and equally averaged after removal of padded positions. The axes show ESPF token indices corresponding to the FCS-derived substructure tokenisation used in the MolTrans implementation. These visualisations represent model-internal attention allocation and are not interpreted as experimentally validated binding contacts or causal binding determinants

Taken together, the case analyses show that cross-attention maps can be useful for inspecting where the model allocates representation weight, but they do not constitute causal explanations, experimentally validated binding-site predictions, or molecular docking evidence. Detailed attention panels containing the five highest-ranked drug fragments and protein residues, together with additional true-positive and true-negative examples, are provided in Additional file 1 (Supplementary Table S1 and Supplementary Figs. S1–S6).

Discussion

The experimental results provide several insights into the contributions of architectural and optimisation choices within the MolTrans-based framework.

First, the complete BCAG-DTI configuration consistently improves the MolTrans baseline in mean AUROC, AUPRC, and F1-score across BindingDB, BIOSNAP, and DAVIS. The largest AUPRC improvement occurs on BindingDB, where the score increases from 0.5231 to 0.6062, while substantial AUPRC improvement is also observed on the more imbalanced DAVIS dataset. These results indicate that the combined BCAG-DTI configuration provides a stronger predictive model than the baseline under the fixed evaluation protocol.

The controlled ablation, however, shows that the observed gains cannot be attributed primarily to bidirectional cross-attention or global aggregation. The enhanced classifier (A5) achieves the highest mean AUROC on BindingDB and BIOSNAP and the highest mean AUPRC and F1-score on all three datasets. Importantly, A5 retains the same MolTrans sequence encoders and interaction representation as A1 but replaces the original classifier ($512 \rightarrow 64 \rightarrow 32$) with the substantially larger enhanced classifier ($1024 \rightarrow 512 \rightarrow 256$). Its strong performance should therefore be interpreted, at least in part, as an effect of increased classifier capacity and the associated classifier design and regularisation, rather than as evidence of improved raw sequence representation quality. On DAVIS, the highest mean AUROC is obtained by A7, which combines cross-attention, global pooling, and the enhanced training strategy. The enhanced training strategy alone (A2) also improves the MolTrans

baseline across all three datasets. These results indicate that classifier capacity, regularisation, and optimisation play major roles in the overall performance improvement.

In contrast, cross-attention alone (A3) consistently reduces performance relative to MolTrans. A similar decrease is observed when cross-attention and global pooling are combined under the baseline training strategy (A6). The performance recovery observed in A7 demonstrates that the effectiveness of these architectural components depends strongly on the optimisation setting. For example, DAVIS AUROC increases from 0.8244 for A6 to 0.9027 for A7 when the enhanced training strategy is introduced. Cross-attention and global aggregation should therefore be regarded as optimisation-sensitive components rather than universally beneficial additions.

The complete A8 configuration provides consistent gains over A1 but does not dominate the individual configurations in the ablation study. In particular, A5 achieves higher AUPRC and F1-score than A8 on all three datasets. This result suggests that increasing decoder capacity and improving regularisation can contribute as much as, or more than, the additional cross-modal interaction modules. The ablation therefore highlights the importance of separating representational changes from classifier and optimisation effects when evaluating architectural improvements in DTI prediction.

The robustness experiments provide a similarly qualified picture. BCAG-DTI improves MolTrans under both unseen-drug and unseen-protein conditions. The improvement is larger for unseen proteins in AUROC and AUPRC, although absolute performance in this setting remains lower than for unseen drugs, indicating that generalisation to previously unseen protein targets remains challenging. Under missing-interaction conditions, BCAG-DTI improves AUROC and AUPRC at 70%, 80%, and 90% missing data. At 95% missing data, however, its AUROC and F1-score are slightly lower than those of MolTrans, while AUPRC remains higher. The benefit of the complete configuration therefore extends across several data-scarcity settings but is not uniform under the most extreme missing-interaction condition.

The external CPI-GGS experiment provides an additional reference outside the MolTrans family. Under the same fixed BIOSNAP partitions, BCAG-DTI achieves higher mean AUROC, AUPRC, and F1-score than both MolTrans and CPI-GGS. However, the comparison must be interpreted together with differences in input representation and model capacity. CPI-GGS represents compounds using molecular graphs and proteins using overlapping 3-grams, whereas MolTrans and BCAG-DTI use FCS-derived substructure representations. In addition, the reproduced CPI-GGS model contains approximately 0.73 million trainable parameters, compared with approximately 106.26 million for A8. The higher predictive performance of BCAG-DTI therefore comes with a substantial increase in model capacity and likely computational cost. The external experiment is consequently most informative as a same-split reference rather than as a fully protocol-identical comparison.

The attention case analysis further illustrates the limitations of interpreting cross-attention weights biologically. In the corticosterone–HSD11B1 case, structural information provides contextual evidence, but the highly attended positions cannot be directly matched to experimentally validated human ligand-contact residues. The itopride–DRD2 analysis similarly identifies a highly attended residue within relevant structural context, but the available structure is not an itopride co-crystal. The diphenhydramine–OPRK1 false-positive case shows that concentrated attention can occur even when the reference interaction label is negative. Conversely, paclitaxel–BCL2 is predicted as a false negative despite experimental evidence supporting direct binding [25]. These observations indicate that attention maps are useful for examining model-internal token allocation but do not establish causal binding determinants.

Several limitations should be considered. First, BCAG-DTI remains primarily sequence-based and does not explicitly incorporate experimentally resolved protein structures, binding pockets, molecular docking information, or atom–residue contact supervision. Structure-aware representations may provide complementary information for modelling physical drug–target interactions. Second, the evaluation relies on established benchmark datasets and fixed partitions. Additional testing across independent datasets and prospective interaction data would provide stronger evidence of generalisation. Third, the enhanced classifier substantially increases model capacity, making it difficult to separate representation quality from the effect of parameter count without additional parameter-matched experiments. Finally, although the five-seed evaluation provides a more reliable estimate of

variability than single-run experiments, broader statistical evaluation across additional datasets and experimental protocols would further strengthen the conclusions.

Conclusion

This study presents a controlled evaluation of cross-attention, global aggregation, classifier design, and optimisation within a MolTrans-based drug–target interaction prediction framework. Across five matched random seeds, the complete BCAG-DTI configuration consistently improves the MolTrans baseline in mean AUROC, AUPRC, and F1-score on BindingDB, BIOSNAP, and DAVIS. It also improves performance for unseen drugs, unseen proteins, and most missing-interaction conditions evaluated on BIOSNAP.

The controlled ablation demonstrates that these improvements do not arise uniformly from the added cross-modal components. The enhanced classifier achieves the strongest AUPRC and F1-score performance across all three datasets and the highest AUROC on BindingDB and BIOSNAP, while the combination of cross-attention, global pooling, and enhanced training achieves the highest AUROC on DAVIS. Cross-attention alone reduces performance under the baseline optimisation strategy, whereas much of this loss is recovered when enhanced training is introduced. These findings show that classifier design and optimisation contribute substantially to the overall gains and that the effectiveness of cross-attention and global aggregation depends on the training configuration.

On the fixed BIOSNAP partitions, BCAG-DTI also achieves higher mean AUROC, AUPRC, and F1-score than the reproduced CPI-GGS external reference, although this comparison involves different preprocessing pipelines and substantially different model capacities. The attention case analysis further shows that learned cross-attention patterns should be interpreted as internal model representations rather than direct evidence of physical binding contacts.

Overall, the results emphasise the importance of controlled component-level evaluation when modifying established DTI architectures. Within the MolTrans framework, improvements arise from interactions among representation learning, classifier capacity, and optimisation rather than from a single architectural component. Future work should investigate parameter-efficient variants, structure-aware representations, and broader out-of-distribution evaluation to determine whether these improvements generalise beyond the benchmark settings considered here.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06634-6.

Acknowledgements NA

Author contributions H.P: Formal analysis, Methodology, Software, Visualization, Writing– original draft. F.B.T: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing–review andediting .T.C: Writing–review and editing.Y.C: Writing– Funding acquisition, review and editing.Y.M: Funding acquisition, review and editing.

Funding This work was partially supported by a Ningbo Municipal Contract Award (Project No. E11260700005), funded by the State Grid Corporation of China.

Data Availability All datasets used in this study are publicly available. BIOSNAP was obtained from the Stanford Network Analysis Project (SNAP) Biomedical Dataset Collection: https://snap.stanford.edu/biodata. DAVIS was obtained from the benchmark dataset introduced by Davis et al. and is publicly available at: http://staff.cs.utu.fi/~aatapa/data/DrugTarget/. BindingDB was obtained from the BindingDB database: https://www.bindingdb.org.The source code, configuration files, and data processing scripts used in this study are available in the BCAG-DTI repository: https://github.com/Falmi/BCAG-DTI.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162:1239–49.
2. Broach JR, Thorner J, et al. High-throughput screening for drug discovery. Nature. 1996;384:14–16.
3. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. 2013;9:232–40.
4. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner JK, Ceulemans H, et al. Deep learning as an opportunity in virtual screening. In Proceedings of the deep learning workshop at NIPS, MA, United States: MIT Press Cambridge, pp. 1–9; 2014.
5. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on chembl. Chemical Science. 2018;9:5441–51.
6. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug-target interaction prediction. J Proteome Res. 2017;16:1401–9.
7. Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug–target binding affinity prediction. Bioinformatics. 2018;34:i821–9.
8. Lee I, Keum J, Nam H. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15:e1007129.
9. Qian Y, Li X, Wu J, Zhang Q. Mcl-dti: using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction. BMC Bioinform. 2023;24:323.
10. Zeng X, Chen W, Lei B. Cat-dti: cross-attention and transformer network with domain adaptation for drug-target interaction prediction. BMC Bioinform. 2024;25:141.
11. Zhang Q, Zuo L, Ren Y, Wang S, Wang W, Ma L, et al. Fmca-dti: a fragment-oriented method based on a multihead cross attention mechanism to improve drug-target interaction prediction. Bioinformatics. 2024;40:btae347.
12. Gao M, Zhang D, Chen Y, Zhang Y, Wang Z, Wang X, et al. Graphormerdti: a graph transformer-based approach for drug-target interaction prediction. Comput Biol Med. 2024;173:108339.
13. Sun Y, Li YY, Leung CK, Hu P. ingnn-dti: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models. Bioinformatics. 2024;40:btae135.
14. Gu Y, Xia S, Ouyang Q, Zhang Y. Bioactivity deep learning for complex structure-free compound-protein interaction prediction. J Chem Inf Model. 2025;65:9910.
15. Huang K, Xiao C, Glass LM, Sun J. Moltrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics. 2021;37:830–6.
16. Hou Z, Xu Z, Yan C, Luo H, Luo J. Cpi-ggs: A deep learning model for predicting compound-protein interaction based on graphs and sequences. Comput Biol Chem. 2025;115:108326.
17. Öztürk H, Ozkirimli E, Özgür, A. Widedta: Prediction of drug-target binding affinity. arxiv 2019. arXiv preprint arXiv:1902.04166;2019.
18. Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics. 2019;35:309–18.
19. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. Graphdta: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37:1140–7.
20. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res. 2007;35:D198–D201.

21. Zitnik M, Sosic R, Leskovec J. Biosnap datasets: Stanford biomedical network dataset collection. Note: http://snap.stanford.edu/biodata Cited by 5; 2018.
22. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011;29:1046–51.
23. Zhang J, Osslund TD, Plant MH, Clogston CL, Nybo RE, Xiong F, et al. Crystal structure of murine 11β-hydroxysteroid dehydrogenase 1: an important therapeutic target for diabetes. Biochemistry. 2005;44:6948–57.
24. Wang S, Che T, Levit A, Shoichet BK, Wacker D, Roth BL. Structure of the d2 dopamine receptor bound to the atypical antipsychotic drug risperidone. Nature. 2018;555:269–73.
25. Ferlini C, Cicchillitti L, Raspaglio G, Bartollino S, Cimitan S, Bertucci C, et al. Paclitaxel directly binds to bcl-2 and functionally mimics activity of nur77. Cancer Res. 2009;69:6906–14.

Authors and Affiliations

Hao Pang[1] · Fiseha Berhanu Tesema[1] · Tianxiang Cui[1] · Yuan Cheng[1] · Yanwen Mao[1]

✉ Fiseha Berhanu Tesema
Fiseha-Berhanu.Tesema@nottingham.edu.cn

[1] School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo 315100, Zhejiang, China