Bridging Sequence and Structure with Unified Domain Adaptation for Drug-Target Interaction Prediction

Mingcan Yuan[1,*], **He Li**[2,*], **Zhiyi Ju**[2], **Mang Ye**[2], **Qingxiong Tan**[1,2,3,†]
[1] National Center for Applied Mathematics in Hubei, Wuhan University
[2] School of Computer Science, Wuhan University
[3] Institute for Math and AI, Wuhan University

Abstract

Drug-target interaction (DTI) prediction is pivotal for accelerating drug discovery, yet existing methods struggle with generalization under cross-domain distribution shifts. Conventional approaches typically rely on static protein representations and assume distributional consistency between source and target domains, often resulting in poor calibration when encountering novel protein families or scaffolds. To address this issue, we present **Bi**-view **Fold**-aware prediction (BiFold), a novel unified interaction-adaptive framework that enhances generalization via dual-view representation learning and calibration-aware adaptation. BiFold synergistically models sequence and structural evidence, introducing a routing module to dynamically fuse the two views at the feature-channel level and exploit complementary information. Furthermore, we design a calibration-aware training principle that aligns domain statistics and encourages flatter source solutions. Crucially, target supervision is introduced only after a forward-only warm-up phase to prevent reinforcement of miscalibrated early predictions, effectively mitigating confirmation bias in pseudo-label learning. Extensive experiments demonstrate that BiFold consistently outperforms the state-of-the-art methods across diverse datasets in both in-domain and cross-domain settings. Additionally, interpretability analysis reveals that the routing module partitions the feature space into view-dominant and view-neutral subspaces, providing concrete evidence of adaptive multi-modal complementarity.

1 Introduction

Drug-target interaction (DTI) prediction serves as a critical computational step for prioritizing candidates in drug discovery and repositioning pipelines [41, 65]. Given the high attrition rate and prohibitive costs of therapeutic development [42, 66], in silico screening is widely used to reduce the experimental search space prior to wet-lab validation [35]. However, the core challenge lies not merely in statistical fitting but in generalizing across structural domains. Binding specificity is governed by sparse residue-level contacts that sequence alone cannot fully capture, while practical screening frequently involves novel chemical scaffolds and protein families absent from training distributions.

While experimental assays provide the most authoritative evidence, they remain costly across large candidate spaces [15]. Techniques such as surface plasmon resonance and cryo-electron microscopy quantify binding or structure but require substantial validation effort [5, 6]. Consequently, computational prediction is essential, yet current models struggle to extrapolate beyond observed data. Early approaches rely on fixed feature spaces, such as hand-crafted descriptors or similarity-based

[1]* Both authors contributed equally to this work.
[2]† Corresponding author: Qingxiong Tan <tanqingxiong@whu.edu.cn>

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

kernels [27, 13]. Although competitive on small benchmarks, these methods are difficult to scale to unseen chemical scaffolds or protein families, motivating a shift to learned representations.

Deep learning methods move beyond fixed descriptors by learning drug and protein representations directly from structured or sequential inputs. Graph-based and transformer-based models improve end-to-end interaction modeling from molecular graphs, SMILES strings, and protein sequences [43, 26]. Recent models further incorporate attention-based interaction modeling, adversarial transfer, biomedical pretraining, topological descriptors, and structure-aware binding representations [3, 49, 44, 33]. Despite these advances, two key limitations remain. First, **single-view representation** remains common: many models encode proteins from one dominant evidence source or fuse heterogeneous protein evidence before drug-specific interaction, weakening the complementarity between evolutionary sequence context and folded geometry. Second, **training is often decoupled**: domain adaptation is introduced after representation learning, leaving source-target statistics, sharp source-domain minima, and pseudo-label noise only weakly coordinated. Critically, **these two limitations are coupled rather than independent**: when the underlying representation has already discarded part of the binding-relevant evidence, post-hoc adaptation can only operate on a partially observed signal, and the resulting calibration error is amplified under distribution shift. This coupling explains why strong in-domain performance does not reliably transfer to unseen chemical scaffolds or protein families, and motivates a unified framework that addresses representation completeness and calibration consistency.

To tackle these issues, we present BiFold, a unified framework that bridges protein-view complementarity and source-target calibration for cross-domain DTI prediction. To address the single-view representation bottleneck, BiFold represents each protein through coordinated sequence- and structure-aware views. Each view first establishes drug-conditioned interaction evidence, preserving sequence-derived functional context and structure-derived geometric cues before fusion. BiFold then performs dimension-wise routing to allocate cross-view evidence at the feature-channel level, rather than reducing fusion to a single view-level weight for each sample. To address the decoupled training problem, BiFold introduces a calibration-aware adaptation process that turns domain adaptation from post-hoc correction into staged refinement. It first warms up target-domain statistics through forward-only inference, allowing normalization to adapt before pseudo-label supervision is introduced. Meanwhile, a new pseudo-labeled target supervision is introduced after the source classifier has reached a stable decision regime, allowing target learning to build on calibrated predictions rather than noisy early signals. This staged process progressively bridges source-domain discrimination and target-domain calibration, turning self-training into a refinement mechanism in which normalization behavior, decision smoothness, and pseudo-label reliability are improved together. By bridging sequence-structure evidence modeling with calibrated cross-domain learning, BiFold preserves binding-relevant protein evidence while improving prediction reliability under scaffold or protein-family shift. The contributions of this paper are as follows:

- We identify two distinct yet coupled sources of the persistent cross-domain gap in drug-target interaction (DTI) prediction, single-view protein representation and domain-agnostic training, and tackle both jointly within a unified BiFold framework.
- We introduce a dual-view interaction mechanism that preserves sequence-derived and structure-derived protein evidence before dimension-wise fusion, enabling cross-view complementarity while mitigating noisy early pseudo-labels through calibration-aware training.
- We theoretically establish the MSE-optimality of dimension-wise routing under heterogeneous channel-wise noise profiles, and empirically show that BiFold learns a stable partition into structure-dominant, sequence-dominant, and neutral feature subspaces.
- Extensive experiments on multiple datasets demonstrate that BiFold consistently outperforms the state-of-the-art under both in-domain and cross-domain evaluation. Case studies provide qualitative evidence that BiFold identifies residues consistent with documented binding contacts while suggesting additional candidate regions for further validation.

2 Related Work

Drug–Target Interaction Prediction. DTI prediction has evolved from descriptor-based learning to neural models that learn from molecular graphs, SMILES, protein sequences, and structures [13, 43, 26, 3, 41]. Recent models improve cold-start generalization, uncertainty estimation, semi-inductive evaluation, target-aware generation, or few-shot affinity prediction through attention, pretraining,

evidential learning, graph induction, multitask learning, and meta-learning [64, 65, 21, 9, 46, 52]. Other studies refine interaction modeling with second-order relevance, dual-granularity representation, dynamic cross-graph attention, or relational-sequential contrastive alignment [25, 14, 1, 58]. These advances reflect a central DTI property. Binding depends on local drug-protein complementarity, but the evidence is distributed across molecular substructures, residue contexts, and protein geometry. Existing methods often improve representation learning or transfer robustness separately, which can weaken residue-level binding cues under shifts in chemical scaffolds or protein families. BiFold instead treats dual-view protein evidence and cross-domain calibration as coupled requirements.

Protein Representation Learning. Pre-trained protein encoders provide transferable residue-level features from complementary sequence and structure evidence. Sequence protein language models capture evolutionary constraints, conserved motifs, and functional context from large unlabeled sequence corpora [37, 23, 8, 53, 56]. Structure-aware encoders incorporate folded geometry through predicted coordinates, structural tokens, residue-level graphs, and geometry-conditioned representations [2, 32, 48, 24, 63]. These views preserve distinct binding-relevant signals. Sequence features retain evolutionary coupling and functional conservation, whereas structural features expose spatial residue arrangements, pocket geometry, and ligand accessibility [30, 22, 55, 60]. For DTI prediction, the key issue is not the strength of individual protein features alone, but the preservation of view-specific evidence until drug-specific interaction is formed. Models that rely on one dominant protein view or fuse heterogeneous protein evidence too early can collapse cues tied to conserved motifs or local pocket geometry. BiFold keeps sequence and structure evidence separate during cross-attention and fuses the resulting interaction representations through dimension-wise routing.

Domain Adaptation and Robust Optimization. Cross-domain DTI prediction must handle shifts in chemical scaffolds, protein families, and dataset biases [36, 57, 54]. Prior methods use adversarial alignment, domain-specific normalization, sharpness-aware training, or confidence-based self-training to reduce transfer error [18, 40, 59, 34, 10]. While these strategies are effective in their original domains, DTI transfer poses a distinct challenge. The transferable signal is concentrated in a small set of residue-drug contacts, so scaffold or protein-family shifts can alter both marginal feature statistics and local interaction patterns simultaneously. Applying adaptation after representation learning therefore risks leaving the interaction space poorly calibrated. BiFold addresses this issue with a staged calibration-aware schedule that separates domain statistics, discourages sharp source solutions, and introduces pseudo-labeled target supervision after a forward-only warm-up [7, 17, 47, 50].

3 Methods

3.1 Problem Formulation

Given a protein p and a drug molecule m, DTI prediction learns a binary classifier $f_\theta(p, m) \in [0, 1]$ that estimates the interaction probability. We adopt an unsupervised domain adaptation setting with labeled source data and unlabeled target data, in which no target label is used at any stage of training or model selection:

$$\mathcal{D}_s = \{(p_i^s, m_i^s, y_i^s)\}_{i=1}^{N_s} \sim P_s, \qquad \mathcal{D}_t = \{(p_j^t, m_j^t)\}_{j=1}^{N_t} \sim P_t, \qquad P_s \neq P_t, \tag{1}$$

where $y_i^s \in \{0, 1\}$ denotes the source interaction label, and P_s and P_t denote source and target distributions. The two domains are separated by non-overlapping drug clusters for BindingDB and by non-overlapping protein families for BioSNAP and Human. A held-out labeled validation split is used only for checkpoint selection and threshold calibration. It is never used for gradient updates, pseudo-label construction, DSBN statistic updates, or test-time adaptation. The goal is to improve target-domain generalization without accessing target test labels.

Figure 1 summarizes the overall architecture of BiFold. Given a protein-drug pair, BiFold constructs sequence-aware and structure-aware protein views, forms drug-conditioned interaction representations for both views, integrates them through dimension-wise routing, and optimizes the predictor under a calibration-aware source-target training schedule.

3.2 Dual-View Drug–Protein Interaction Encoder

A single protein view does not reliably preserve all binding-relevant evidence under domain shift. Sequence encoders retain evolutionary constraints and functional context, whereas structure-aware

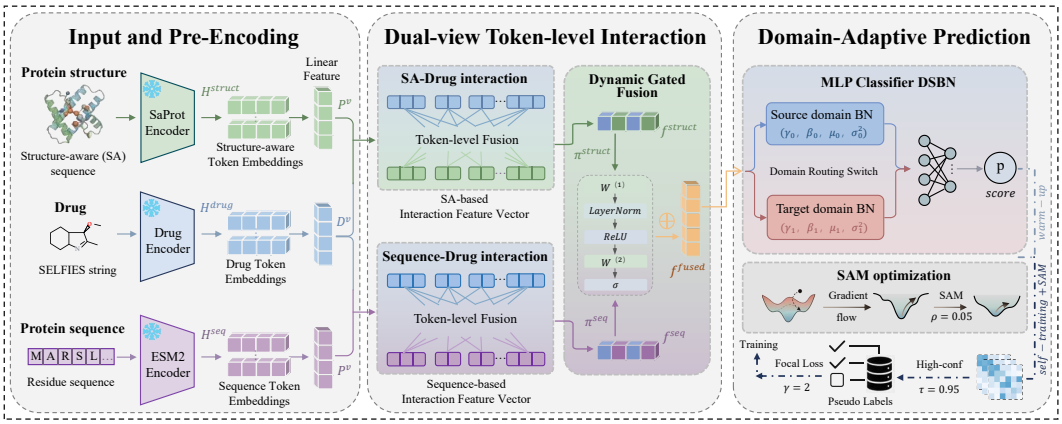


Figure 1: Overview of BiFold. Two parallel protein encoders capture co-evolutionary and structural signals, each interacting with the drug via an independent cross-attention network (CAN) to preserve view-specific binding cues. A dimension-wise routing module merges the two streams along feature channels, recovering complementarity that a scalar gate would discard. The fused representation feeds a DSBN-equipped MLP head trained under a calibration-aware schedule, jointly handling distribution shift, loss-surface sharpness, and pseudo-label noise after a forward-only warm-up.

encoders provide structure-derived residue tokens that encode folded-geometry information. BiFold therefore represents each protein through two parallel branches. The sequence branch uses ESM-2 [37], and the structure branch uses SaProt with Foldseek structural tokens [48, 51]. The drug is converted into a SELFIES string [31] and encoded by SELFormer [61]. All pretrained encoders are kept frozen, and BiFold trains the projection layers, cross-attention modules, routing network, and classifier. Let $\mathcal{V} = \{\text{seq}, \text{struct}\}$ denote the two protein views. The view-specific protein and drug tokens are obtained as follows:

$$\mathbf{H}^v = \mathcal{E}_v(p), \qquad \mathbf{H}^d = \mathcal{E}_d(m), \qquad \mathbf{P}^v = \mathbf{H}^v\mathbf{W}_p^v, \qquad \mathbf{D}^v = \mathbf{H}^d\mathbf{W}_d^v, \qquad v \in \mathcal{V}, \tag{2}$$

where $\mathcal{E}_{\text{seq}}$, $\mathcal{E}_{\text{struct}}$, and $\mathcal{E}_d$ denote ESM-2, SaProt, and SELFormer, respectively. The view-specific projections align protein and drug tokens to the same hidden dimension d, while allowing each protein view to form its own drug-conditioned interaction pattern.

The relevance of a residue depends on the drug being evaluated. BiFold therefore avoids fusing the two protein views before interaction modeling. Instead, each view uses an independent cross-attention network. We first define the attention operator:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{3}$$

For a view v, the drug-conditioned update is written compactly as follows:

$$\begin{aligned}
\widetilde{\mathbf{P}}^v &= \frac{1}{2}\left[\text{Attn}(\mathbf{Q}_p^v, \mathbf{K}_p^v, \mathbf{V}_p^v) + \text{Attn}(\mathbf{Q}_p^v, \mathbf{K}_d^v, \mathbf{V}_d^v)\right], \\
\widetilde{\mathbf{D}}^v &= \frac{1}{2}\left[\text{Attn}(\mathbf{Q}_d^v, \mathbf{K}_d^v, \mathbf{V}_d^v) + \text{Attn}(\mathbf{Q}_d^v, \mathbf{K}_p^v, \mathbf{V}_p^v)\right], \\
\mathbf{f}^v &= \left[\text{Pool}(\widetilde{\mathbf{P}}^v); \text{Pool}(\widetilde{\mathbf{D}}^v)\right] \in \mathbb{R}^{2d},
\end{aligned} \tag{4}$$

where subscripts p and d denote protein and drug projections. The self-attention terms preserve within-modality context, while the cross-attention terms identify drug-conditioned features. We use an equal-weighted symmetric update as a lightweight design to balance within-view context and cross-modal interaction without introducing additional routing parameters before view fusion. The resulting vector $\mathbf{f}^v$ summarizes the interaction evidence from protein view v.

3.3 Dimension-Wise Routing for View Fusion

The two interaction representations can differ across feature channels. Some channels retain stronger sequence-derived functional context, whereas others encode more reliable structure-derived evidence. A scalar gate assigns one weight per view and expresses only a sample-level preference. BiFold

instead uses a dimension-wise routing vector to select evidence at the feature-channel level:

$$\begin{aligned} \hat{\mathbf{f}}^v &= \tanh(\mathbf{W}_t^v \mathbf{f}^v), \qquad v \in \mathcal{V}, \\ \boldsymbol{\pi}^{\text{struct}} &= \Phi_r([\mathbf{f}^{\text{struct}}; \mathbf{f}^{\text{seq}}]), \qquad \boldsymbol{\pi}^{\text{seq}} = \mathbf{1} - \boldsymbol{\pi}^{\text{struct}}, \\ \mathbf{f}^{\text{fused}} &= \boldsymbol{\pi}^{\text{struct}} \odot \hat{\mathbf{f}}^{\text{struct}} + \boldsymbol{\pi}^{\text{seq}} \odot \hat{\mathbf{f}}^{\text{seq}}, \end{aligned} \tag{5}$$

where Φ_r denotes an MLP followed by a sigmoid function and outputs a vector in $[0,1]^{2d}$, and $\odot$ denotes element-wise multiplication. This routing expression transforms each view, assigns complementary channel-level weights, and produces a convex fusion that preserves a division of evidence between sequence and structure views.

The following proposition formalizes, under a controlled signal-plus-noise model, the conditions under which dimension-wise routing strictly dominates scalar gating. While the analysis is restricted to a surrogate reconstruction objective and does not assign biological meaning to individual feature channels, it identifies the heterogeneous channel-wise reliability assumption that the empirical gate analysis in Section 4.4 later confirms. The proof is provided in Appendix C.4.

Proposition 1 (Channel-wise routing)**.** *Let $\mathbf{z} \in \mathbb{R}^h$ be a latent interaction signal. Suppose two views observe z_k through $f_k^{seq} = z_k + \epsilon_k^{seq}$ and $f_k^{struct} = z_k + \epsilon_k^{struct}$, where the noises are unbiased, independent, and have variances $(\sigma_k^{seq})^2$ and $(\sigma_k^{struct})^2$. Among convex estimators $\hat{z}_k = \alpha_k f_k^{struct} + (1-\alpha_k) f_k^{seq}$, the MSE-optimal coefficient is:*

$$\alpha_k^\star = \frac{(\sigma_k^{seq})^2}{(\sigma_k^{seq})^2 + (\sigma_k^{struct})^2}. \tag{6}$$

Scalar gating is optimal across channels only when $(\sigma_k^{seq})^2/(\sigma_k^{struct})^2$ is channel-invariant.

3.4 Calibration-Aware Cross-Domain Training

The fused interaction representation must remain calibrated when the target domain contains unseen scaffolds or protein families. The key idea is to control when and how unlabeled target data enter training. Target samples first calibrate domain-specific statistics without labels, then the source classifier is regularized toward a flatter solution, and only then are confident target predictions converted into pseudo-labeled supervision. This staged process treats unlabeled target samples as distributional evidence that should be calibrated before being used as supervision.

The classifier is a three-layer MLP with domain-specific batch normalization (DSBN) [7]. For a fused feature $\mathbf{f}^{\text{fused}}$ and an intermediate activation $\mathbf{h}$ in domain $r \in \{s,t\}$, prediction and normalization are:

$$\hat{y} = f_\theta(p,m) = \sigma\big(\text{MLP}_{\text{DSBN}}(\mathbf{f}^{\text{fused}})\big), \qquad \text{DSBN}_r(\mathbf{h}) = \boldsymbol{\gamma}_r \odot \frac{\mathbf{h} - \boldsymbol{\mu}_r}{\sqrt{\boldsymbol{\sigma}_r^2 + \varepsilon_{\text{BN}}}} + \boldsymbol{\beta}_r, \tag{7}$$

where $(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r^2)$ are the running statistics, $(\boldsymbol{\gamma}_r, \boldsymbol{\beta}_r)$ are the affine parameters for domain r, and ε_{BN} is a numerical stability constant. The source and target branches have separate normalization parameters, but classifier weights are shared. This lets the classifier learn common interaction patterns while normalization absorbs domain-specific activation shifts.

Sharpness-aware minimization (SAM) [17] is applied to the source-domain objective to reduce sensitivity to source-specific sharp minima. Let $\mathcal{L}_s(\theta)$ denote the supervised source loss. The SAM loss and its first-order perturbation are:

$$\mathcal{L}_s^{\text{SAM}}(\theta) = \max_{\|\boldsymbol{\delta}\|_2 \le \rho} \mathcal{L}_s(\theta + \boldsymbol{\delta}), \qquad \boldsymbol{\delta}^\star = \rho \frac{\nabla_\theta \mathcal{L}_s(\theta)}{\|\nabla_\theta \mathcal{L}_s(\theta)\|_2}, \tag{8}$$

where ρ controls the perturbation radius. This objective encourages source solutions that are less sensitive to small weight perturbations and more stable under target-domain shifts.

Before self-training begins, BiFold uses a forward-only warm-up. During warm-up, target samples pass through the target DSBN branch to update running statistics, but gradients are computed only from the source loss. After warm-up, target-domain supervision is introduced through confidence-thresholded pseudo-labeling, following the common self-training principle [47]:

$$\tilde{q}_j = f_\theta(x_j^t), \qquad \tilde{y}_j = \mathbb{1}[\tilde{q}_j > \tfrac{1}{2}], \qquad c_j = \mathbb{1}[\tilde{q}_j > \tau \text{ or } \tilde{q}_j < 1-\tau], \tag{9}$$

where $x_j^t = (p_j^t, m_j^t)$ is a target sample, $\tau \in (\frac{1}{2}, 1)$ is a confidence threshold, and c_j removes target samples near the decision boundary. The pseudo-label $\tilde{y}_j$ and mask c_j are detached from the computation graph when computing the target loss.

We use focal loss for source labels and target pseudo-labels so that training is not dominated by easy examples or high-confidence pseudo-labels. With $q_i^s = f_\theta(p_i^s, m_i^s)$, the true-class probabilities are:

$$q_{t,i}^s = y_i^s q_i^s + (1 - y_i^s)(1 - q_i^s), \qquad \tilde{q}_{t,j} = \tilde{y}_j \tilde{q}_j + (1 - \tilde{y}_j)(1 - \tilde{q}_j). \tag{10}$$

The supervised source loss, masked pseudo-label loss, and final training objective are:

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{N_s} \sum_{i=1}^{N_s} (1 - q_{t,i}^s)^\gamma \, \mathrm{BCE}(q_i^s, y_i^s), \\ \mathcal{L}_p &= \frac{1}{\max(1, \sum_j c_j)} \sum_{j=1}^{N_t} c_j (1 - \tilde{q}_{t,j})^\gamma \, \mathrm{BCE}(\tilde{q}_j, \tilde{y}_j), \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_s^{\text{SAM}} + \lambda(t) \mathcal{L}_p, \end{aligned} \tag{11}$$

where $\lambda(t)$ increases linearly from 0 to 1 after the warm-up stage. At inference time, target samples use the target DSBN branch accumulated during training.

4 Experiments

4.1 Experimental Setup

Datasets, baselines, and evaluation metrics. We evaluate BiFold on three DTI benchmark datasets, BindingDB [39], BioSNAP [67], and Human [38]. For the in-domain setting, each dataset uses the training, validation, and test splits established by DrugBAN [3]. For the cross-domain setting, we adopt the source-target split protocol of DrugBAN, as formalized in Section 3.1. BindingDB is split by non-overlapping drug clusters obtained through single-linkage clustering on Tanimoto similarity, while BioSNAP and Human are split by non-overlapping protein families. Dataset statistics are provided in Appendix A. We compare BiFold with seven state-of-the-art methods, including five deep DTI models (GraphDTA [43], MolTrans [26], DrugBAN [3], SaBAN [33], and SiamDTI [62]) and two pre-trained language-model methods (BioT5 [44] and Top-DTI [49]). We use the same benchmark splits for all methods. When published results exactly match the split, metric, and model-selection protocol, we report them directly. Otherwise, we retrain the method under our protocol, as detailed in Appendix E. Main comparisons use AUROC, AUPRC, and accuracy following Bai et al. [3]. Ablations and detailed analyses additionally report threshold-dependent metrics, including F1, sensitivity, specificity, and MCC.

Implementation details. The three pre-trained encoders, SaProt 650M, ESM-2 650M, and SELFormer, are frozen. Maximum token lengths are 1024 for proteins and 512 for drugs. The CAN layers use 8 attention heads with a hidden dimension of 768. The model is trained for up to 200 epochs with a batch size of 64, using SAM with AdamW as the base optimizer, a learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, $\rho = 0.05$, and cosine annealing scheduling. The self-training warm-up lasts 6 epochs, and the confidence threshold is set to $\tau = 0.95$. Early stopping is performed with a patience of 15 epochs using the benchmark validation split. In cross-domain runs, validation labels are used only for checkpoint selection and threshold calibration. They are not used to compute training gradients, construct pseudo-labels, update DSBN statistics, or fit trainable model weights. The final test results are reported once on the held-out target test set. The same validation-based threshold selection protocol is applied to all retrained methods.

4.2 Comparison with State-of-the-Art Methods

In-domain results. Table 1 reports the in-domain performance on BindingDB and BioSNAP, while Figure 2(a) visualizes the corresponding results on the Human dataset. The complete metric breakdown of BiFold on the Human dataset, with all seven evaluation metrics and standard deviations, is provided in Tables 9–10 of Appendix C. BiFold achieves the best mean performance across the three datasets and the three primary metrics. On BindingDB, most baselines already operate in a high-AUROC regime, which leaves limited room for further AUROC improvement. On BioSNAP, BiFold improves AUROC by 2.1 percentage points over the strongest baseline, indicating a clearer margin on this more diverse benchmark. Combined with the ablation results in Section 4.3, these findings are consistent with the benefit of preserving complementary sequence and structure evidence even under in-domain evaluation.

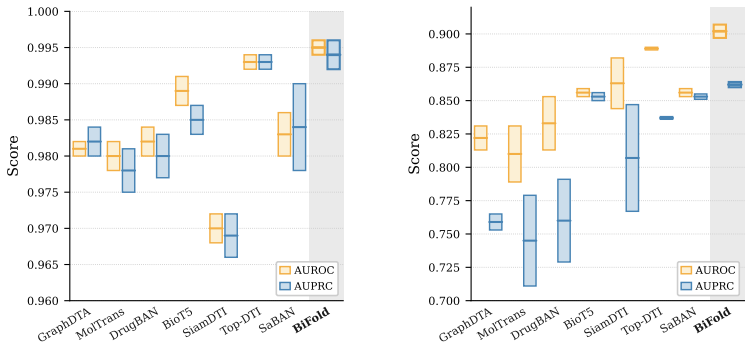
Table 1: In-domain performance comparison of BiFold and the state-of-the-art methods on the BindingDB and BioSNAP datasets. The best and the second best performances are respectively indicated in **bold** and underlined.

Method	BindingDB			BioSNAP		
	AUROC	AUPRC	Accuracy	AUROC	AUPRC	Accuracy
GraphDTA (Bioinform.'21)	0.951 ± 0.002	0.934 ± 0.002	0.888 ± 0.005	0.887 ± 0.008	0.890 ± 0.007	0.800 ± 0.007
MolTrans (Bioinform.'21)	0.952 ± 0.002	0.936 ± 0.001	0.887 ± 0.006	0.895 ± 0.004	0.897 ± 0.005	0.825 ± 0.010
DrugBAN (NMI'23)	0.960 ± 0.001	0.948 ± 0.002	0.904 ± 0.004	0.903 ± 0.005	0.902 ± 0.004	0.834 ± 0.008
BioT5 (EMNLP'23)	0.963 ± 0.001	0.952 ± 0.001	0.907 ± 0.003	0.937 ± 0.002	0.937 ± 0.002	0.874 ± 0.002
SiamDTI (BIB'24)	0.961 ± 0.002	0.945 ± 0.002	0.890 ± 0.006	0.912 ± 0.005	0.910 ± 0.003	0.855 ± 0.004
Top-DTI (Bioinform.'25)	0.965 ± 0.002	0.958 ± 0.002	0.898 ± 0.005	0.939 ± 0.002	0.941 ± 0.002	0.866 ± 0.003
SaBAN (ICASSP'26)	0.957 ± 0.002	0.948 ± 0.001	0.902 ± 0.003	0.930 ± 0.003	0.929 ± 0.002	0.858 ± 0.006
BiFold (Ours)	**0.982** ± 0.002	**0.967** ± 0.002	**0.965** ± 0.002	**0.960** ± 0.002	**0.962** ± 0.002	**0.884** ± 0.002

Cross-domain results. Table 2 presents the cross-domain evaluation on BindingDB and BioSNAP, and Figure 2(b) visualizes the corresponding results on the Human benchmark. BiFold achieves the best cross-domain performance across the reported benchmarks, with especially clear gains on BioSNAP and on AUPRC under scaffold shift. On BioSNAP, BiFold improves AUROC over Top-DTI by 3.6 percentage points and AUPRC by 7.0 percentage points, compared with the 2.1-point in-domain AUROC margin. As shown in Figure 2(b), the uncertainty intervals of BiFold are tighter than those of high-variance baselines such as SiamDTI and DrugBAN. These results, together with the ablation analysis, support the intended role of BiFold under distribution shift: preserving complementary protein evidence and improving calibration when representation errors and source-target mismatch appear together.

Table 2: Performance comparison of BiFold and the state-of-the-art methods on the BindingDB and BioSNAP datasets under the challenging cross-domain setting. The best and the second best performances are respectively indicated in **bold** and underlined.

Method	BindingDB			BioSNAP		
	AUROC	AUPRC	Accuracy	AUROC	AUPRC	Accuracy
GraphDTA (Bioinform.'21)	0.536 ± 0.015	0.496 ± 0.029	0.472 ± 0.009	0.618 ± 0.005	0.618 ± 0.008	0.535 ± 0.024
MolTrans (Bioinform.'21)	0.554 ± 0.024	0.511 ± 0.025	0.470 ± 0.004	0.621 ± 0.015	0.608 ± 0.022	0.546 ± 0.032
DrugBAN (NMI'23)	0.604 ± 0.027	0.570 ± 0.047	0.509 ± 0.021	0.685 ± 0.044	0.713 ± 0.041	0.565 ± 0.056
BioT5 (EMNLP'23)	0.651 ± 0.002	0.653 ± 0.003	0.621 ± 0.005	0.720 ± 0.008	0.718 ± 0.004	0.715 ± 0.009
SiamDTI (BIB'24)	0.627 ± 0.027	0.571 ± 0.024	0.563 ± 0.033	0.718 ± 0.055	0.725 ± 0.054	0.623 ± 0.070
Top-DTI (Bioinform.'25)	0.682 ± 0.002	0.638 ± 0.002	0.633 ± 0.003	0.750 ± 0.002	0.727 ± 0.002	0.722 ± 0.003
SaBAN (ICASSP'26)	0.637 ± 0.003	0.633 ± 0.002	0.621 ± 0.005	0.720 ± 0.002	0.718 ± 0.002	0.715 ± 0.005
BiFold (Ours)	**0.697** ± 0.002	**0.680** ± 0.003	**0.660** ± 0.003	**0.786** ± 0.003	**0.797** ± 0.003	**0.736** ± 0.004

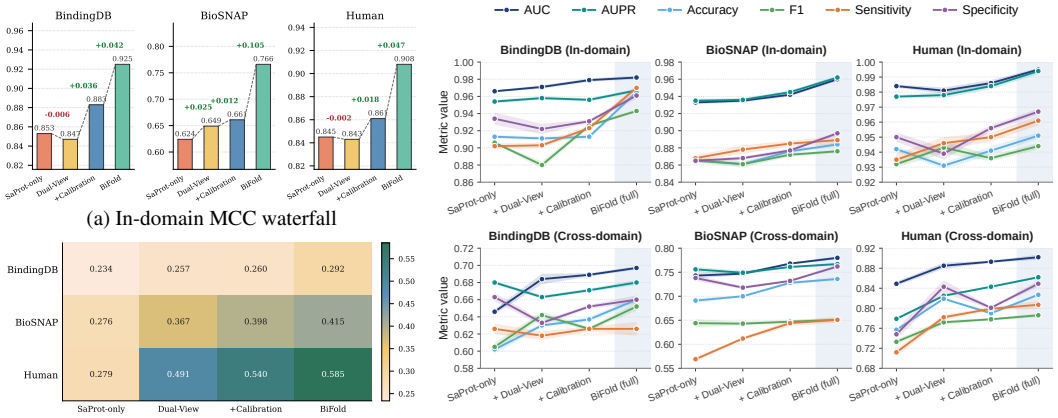


(a) In-domain results (b) Cross-domain results

Figure 2: Performance comparison of BiFold and the state-of-the-art methods on the Human benchmark under both in-domain and cross-domain settings, which shows that BiFold achieves the best mean performance with tighter uncertainty intervals than high-variance baselines. The horizontal line and box edges respectively denote the mean and mean ± std over three independent runs.

4.3 Ablation Studies

To evaluate the contribution of each design component, we apply an identical experimental protocol to a series of progressively richer model variants. Building on a single-view SaProt baseline, the model is first enriched with dual-view cross-attention and gated fusion (w/ Dual-View), complementing the structure-aware representation with a sequence-derived stream. Cross-domain training is subsequently stabilized through domain-specific normalization combined with sharpness-aware minimization (+ Calibration DSBN & SAM). Confidence-thresholded self-training is then introduced to leverage unlabeled target samples, yielding BiFold (full). With the benchmark split and model-selection protocol held fixed, the stepwise comparison isolates the primary effect associated with each newly introduced component, while the full model captures their joint behavior. Complete metrics with standard deviations over three independent runs are reported in Tables 9–10 of Appendix C.



(a) In-domain MCC waterfall

(b) Cross-domain MCC heatmap

(c) Per-metric trajectories with one-σ bands

Figure 3: Ablation analysis on three datasets. (a) In-domain MCC waterfall, with labels marking the marginal contribution of each component. (b) Cross-domain MCC heatmap across datasets and variants. (c) Per-metric trajectories of six metrics across the four design variants on three datasets, where the first and the second rows respectively denote the in-domain and cross-domain tasks.

Three observations emerge from Figure 3. First, the added components produce markedly larger cross-domain than in-domain improvements. No individual in-domain step yields more than $+0.105$ MCC, while the cumulative cross-domain MCC gain reaches $+0.306$ on Human and remains positive on BindingDB and BioSNAP (Figure 3(b)). Second, the dominant contributor to this gain depends on the dataset and the metric: under cross-domain MCC (Figure 3(b)), dual-view encoding leads on Human and BioSNAP, while self-training leads on BindingDB; under cross-domain AUROC, calibration leads on BioSNAP, indicating that domain-specific normalization and sharpness-aware minimization are most beneficial when source and target protein families differ. Third, BiFold (full) achieves the highest or tied-highest mean on at least five of the six metrics across the evaluated settings. Its improvements over the SaProt-only baseline are generally substantial, although some uncertainty intervals overlap on the BindingDB cross-domain split, particularly for AUPRC, sensitivity, and specificity. These results support improved overall predictive performance while highlighting the dependence of threshold-based metrics on the evaluation setting.

Table 3: BioSNAP cross-domain fusion ablation (seed = 26). Post-int. denotes fusion after drug-conditioned interaction.

Seq.	Str.	Post-int.	Fusion	AUROC ↑	AUPRC ↑	Acc. ↑
✓	×	–	–	0.7264	0.7554	0.6421
×	✓	–	–	0.7340	0.7550	0.6551
✓	✓	✓	Mean	0.7422	0.7540	0.6527
✓	✓	✓	Concat	0.7510	0.7700	0.6663
✓	✓	✓	Global	0.7560	0.7679	0.6710
✓	✓	✓	Sample	0.7600	0.7812	0.6886
✓	✓	×	Token	0.7636	0.7800	0.6968
✓	✓	✓	Channel	**0.7891**	**0.7993**	**0.7361**

Controlled fusion and calibration analysis. Table 3 shows that BiFold consistently outperforms the alternative fusion strategies. The comparisons with single-view and fixed-fusion variants highlight the complementary contributions of sequence and structure representations. The improvements over scalar gating and early token fusion further demonstrate the benefits of channel-wise routing

and drug-conditioned interaction before fusion. Additional calibration results and component-level ablations are provided in Appendix C.2 and Appendix C.3.

4.4 The Gate Learns a Stable Subspace Partition

A learned routing module could in principle act as a per-sample selector, weighting the structure stream more heavily for some drug-target pairs and the sequence stream more heavily for others. The trained gate of BiFold instead exhibits a markedly different behavior. At the sample level, the aggregate routing weight assigned to each view remains nearly constant across input pairs. At the dimension level, the routing weights are strongly polarized, varying within a routing vector substantially more than across samples for any fixed dimension. Sorting feature dimensions by their mean routing weight produces a smooth S-shaped profile (Figure 4) that partitions the fused representation into three regions of comparable size: a sequence-dominant region, a structure-dominant region, and a neutral region in between. The gate statistics across datasets (Appendix C.5) support the stability of this partition, while the subspace-knockout experiments on Human and BindingDB (Appendix C.6) provide evidence of its functional relevance.

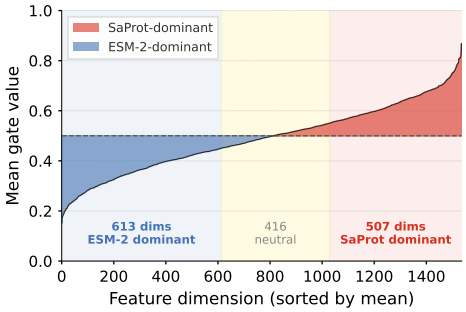


Figure 4: Per-dimension mean of $\boldsymbol{\pi}^{\text{struct}}$ on the in-domain Human test set, sorted ascending.

This finding constitutes a structural observation about sequence-structure complementarity in the encoder pair and the datasets considered here. Sequence and structure contribute complementary, additive evidence through largely disjoint feature channels that are consistently reused across inputs. This observation supports the design choice of preserving view-specific subspaces until drug-conditioned interaction representations are formed, and is consistent with the heterogeneous channel-wise reliability assumption underlying Section 3.3.

4.5 Case Study

The dual-stream cross-attention of BiFold provides a token-level correspondence between drug tokens and protein residues without residue-level supervision. We examine two co-crystallized drug-target pairs from the Protein Data Bank [4], EZL-6QL2 [29] and 9YA-5W8L [45], selected under four pre-specified criteria. Both pairs are absent from BindingDB, BioSNAP, and Human, which are the benchmark datasets used to train and evaluate BiFold in this work. They are also resolved at better than 2.5 Å, target a human protein, and report inhibitor potency $\text{IC}_{50} \leqslant 100$ nM. For each pair, drug-to-protein attention is averaged across heads and streams to obtain one per-residue importance score.

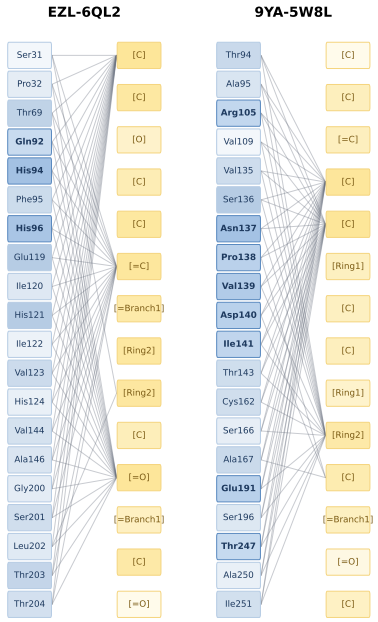


Figure 5: Drug-protein attention on EZL-6QL2 (left) and 9YA-5W8L (right). **Bold labels** mark documented contacts.

Quantitatively, the top-20 residues recover $3/11$ and $8/14$ documented contacts, respectively, representing approximately $4.0\times$ and $9.5\times$ enrichment over random selection (hypergeometric $p = 0.028$ and $p = 9.2 \times 10^{-8}$). The recovered residues also form spatially coherent regions around the binding pocket. These case studies provide qualitative evidence that the cross-attention maps can identify residues consistent with known binding contacts while suggesting additional candidate regions for further validation. We interpret these maps as correspondence patterns rather than causal mechanisms, since attention reflects where the model looks rather than the binding-energy contribution of each residue. Full residue lists, random-baseline comparison, significance derivation, and biochemical discussion are provided in Appendix D.

5 Conclusion

We present BiFold, a novel unified interaction-adaptive framework for drug-target interaction prediction that enhances generalization across in-domain and cross-domain settings. BiFold preserves sequence-derived and structure-derived protein evidence through independent drug-conditioned interaction streams and integrates the two views with dimension-wise routing at the feature-channel level. Furthermore, a calibration-aware adaptation principle is designed to align domain statistics, encourage flatter source solutions, and admit target supervision only after a forward-only warm-up to mitigate confirmation bias in pseudo-label learning. Extensive experiments and interpretability analysis demonstrate that BiFold consistently outperforms the state-of-the-art methods and reveals an adaptive partition of the feature space into view-dominant and view-neutral subspaces, providing concrete evidence of multi-modal complementarity.

Limitations. BiFold is developed and evaluated under the standard binary DTI prediction protocol. Extending it to continuous affinity modeling or multi-task endpoint prediction would require task-specific labels, objectives, and evaluation criteria. The case studies provide qualitative evidence for routing behavior and token-level correspondence, but they should be interpreted as model-level hypotheses rather than causal biochemical explanations. These extensions are left for future research.

Acknowledgments and Disclosure of Funding

Acknowledgments. This work was supported by the Hubei Provincial Natural Science Foundation of China under Grant (2026AFB556) and the National Natural Science Foundation of China under Grant (62606374, 62361166629).

Competing Interests. The authors declare no competing financial or non-financial interests.

References

[1] Abrar Rahman Abir, Muhtasim Noor Alif, Wencai Zhang, Khandakar Tanvir Ahmed, and Wei Zhang. DCGAT-DTI: Dynamic cross-graph attention network for drug–target interaction prediction. *Bioinformatics Advances*, 6(1):vbaf306, 2026.

[2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alex Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

[3] Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023.

[4] Stephen K. Burley, Rusham Bhatt, Charmi Bhikadiya, Chunxiao Bi, Alison Biester, Pratyoy Biswas, Sebastian Bittrich, Santiago Blaumann, Ronald Brown, Henry Chao, et al. Updated resources for exploring experimentally-determined PDB structures and computed structure models at the RCSB Protein Data Bank. *Nucleic Acids Research*, 53(D1):D564–D574, 2025.

[5] Davide Capelli, Viviana Scognamiglio, and Roberta Montanari. Surface plasmon resonance technology: Recent advances, applications and experimental cases. *TrAC Trends in Analytical Chemistry*, 163:117079, 2023.

[6] Ecenur Cebi, Joohyun Lee, Vinod Kumar Subramani, Nayeon Bak, Changsuk Oh, and Kyeong Kyu Kim. Cryo-electron microscopy-based drug design. *Frontiers in Molecular Biosciences*, 11:1342179, 2024.

[7] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7354–7362, 2019.

[8] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, Chiming Liu, Aohan Zeng, Yuxiao Dong, Jie Tang, and Le Song. xTrimoPGLM: Unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nature Methods*, 22:1028–1039, 2025.

[9] Yigang Chen, Xiang Ji, Ziyue Zhang, Zihao Zhu, Yuming Zhou, Chang Su, Yang-Chi-Dung Lin, Hsi-Yuan Huang, Kangping Wei, Yi Lai, Ke Chen, Xingqiao Lin, Yangyi Zhang, Jiehui Fu, Yixian Huang, Shidong Cui, Shih-Chung Yen, Tao Zhang, Arieh Warshel, and Hsien-Da Huang. Semi-inductive dataset construction and framework optimization for practical drug target interaction prediction with ScopeDTI. *Nature Communications*, 16(1):11509, 2025.

[10] Bo Cheng, Jueqing Lu, Yuan Tian, Haifeng Zhao, Yi Chang, and Lan Du. CGMatch: A different perspective of semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15381–15391, 2025.

[11] Giuseppina Claps, Sara Faouzi, Virginie Quidville, Feras Chehade, Shensi Shen, Stéphan Vagner, and Caroline Robert. The multiple roles of LDH in cancer. *Nature Reviews Clinical Oncology*, 19(12):749–762, 2022.

[12] Katia D'Ambrosio, Anna Di Fiore, Vincenzo Alterio, Emma Langella, Simona Maria Monti, Claudiu T Supuran, and Giuseppina De Simone. Multiple binding modes of inhibitors to human carbonic anhydrases: an update on the design of isoform-specific modulators of activity. *Chemical Reviews*, 125(1):150–222, 2025.

[13] Kusal Debnath, Pratip Rana, and Preetam Ghosh. A survey on deep learning for drug-target binding prediction: models, benchmarks, evaluation, and case studies. *Briefings in Bioinformatics*, 26(5):bbaf491, 2025.

[14] Zhiyuan Dong, Yijia Zhang, Yang Shi, Chao Zhang, Chaoyang Han, Hongfeng Li, Yajuan Gao, Wanyi Fu, Yanye Lu, Ping Ren, Lei Liang, and Zhiyu Liu. Decoupled dual-granularity rebalanced pyramid network for drug–target interaction prediction. *Briefings in Bioinformatics*, 26(6):bbaf670, 2025.

[15] Maria Emilia Dueñas, Rachel E Peltier-Heap, Melanie Leveridge, Roland S Annan, Frank H Büttner, and Matthias Trost. Advances in high-throughput mass spectrometry in drug discovery. *EMBO Molecular Medicine*, 15(1):e14850, 2023.

[16] Marianna Fontana, Alberto Aimo, Michele Emdin, Aldostefano Porcari, Scott D. Solomon, Philip N. Hawkins, and Julian D. Gillmore. Transthyretin amyloid cardiomyopathy: from cause to novel treatments. *European Heart Journal*, 47(1):54–63, 2026.

[17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[19] Julian D Gillmore, Daniel P Judge, Francesco Cappelli, Marianna Fontana, Pablo Garcia-Pavia, Simon Gibbs, Martha Grogan, Mazen Hanna, James Hoffman, Ahmad Masri, et al. Efficacy and safety of acoramidis in transthyretin amyloid cardiomyopathy. *New England Journal of Medicine*, 390(2):132–142, 2024.

[20] Simone Giovannuzzi and Claudiu T Supuran. Human carbonic anhydrase modulators: the past, present, and future. *Trends in Pharmacological Sciences*, 46(9):836–847, 2025.

[21] Hamid Hadipour, Yan Yi Li, Yan Sun, Chutong Deng, Leann Lac, Rebecca Davis, Silvia T. Cardona, and Pingzhao Hu. GraphBAN: An inductive graph-based approach for enhanced prediction of compound-protein interactions. *Nature Communications*, 16(1):2541, 2025.

[22] Jiyun Han, Shizhuo Zhang, Mingming Guan, Qiuyu Li, Xin Gao, and Juntao Liu. GeoNet enables the accurate prediction of protein-ligand binding sites through interpretable geometric deep learning. *Structure*, 32(12):2435–2448.e5, 2024.

[23] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S Molina, Neil Thomas, Yousuf A Khan, Chetan Mishra, Carolyn Kim, Liam J Bartie, Matthew Nemeth, Patrick D Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[24] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[25] Yang Hua, Tianyang Xu, Xiaoning Song, Zhenhua Feng, Rui Wang, Wenjie Zhang, and Xiaojun Wu. R-DTI: Drug target interaction prediction based on second-order relevance exploration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17368–17376, 2025.

[26] 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.

[27] Tiago Janela and Jürgen Bajorath. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. *Nature Machine Intelligence*, 4(12):1246–1255, 2022.

[28] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[29] Justina Kazokaitė, Visvaldas Kairys, Joana Smirnovienė, Alexey Smirnov, Elena Manakova, Martti Tolvanen, Seppo Parkkila, and Daumantas Matulis. Engineered carbonic anhydrase VI-mimic enzyme switched the structure and affinities of inhibitors. *Scientific Reports*, 9(1): 12710, 2019.

[30] Lucien F. Krapp, Luciano A. Abriata, Fabio Cortés Rodriguez, and Matteo Dal Peraro. PeSTo: Parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications*, 14(1):2175, 2023.

[31] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[32] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science*, 384(6693):eadl2528, 2024.

[33] Jing Lan, Hexiao Ding, Hongzhao Chen, Yufeng Jiang, Nga-Chun Ng, Gwing Kei Yip, Gerald W.Y. Cheng, Yunlin Mao, Jing Cai, Liang-ting Lin, and Jung Sun Yoo. Structure-aware contrastive learning with fine-grained binding representations for drug discovery. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[34] Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. Seeking consistent flat minima for better domain generalization via refining loss landscapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15349–15359, 2025.

[35] Qian Liao, Yu Zhang, Ying Chu, Yi Ding, Zhen Liu, Xianyi Zhao, Yizheng Wang, Jie Wan, Yijie Ding, Prayag Tiwari, Quan Zou, and Ke Han. Application of Artificial Intelligence in drug-target interactions prediction: A review. *npj Biomedical Innovations*, 2(1):1, 2025.

[36] Gaoming Lin, Xin Zhang, Zhonghao Ren, Quan Zou, Prayag Tiwari, Changjun Zhou, and Yijie Ding. TAPB: An interventional debiasing framework for alleviating target prior bias in drug-target interaction prediction. *Nature Communications*, 16(1):10867, 2025.

[37] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[38] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229, 2015.

[39] Tiqing Liu, Linda Hwang, Stephen K. Burley, Carmen I. Nitsche, Christopher Southan, W. Patrick Walters, and Michael K. Gilson. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1):D1633–D1644, 2025.

[40] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, pages 1645–1655, 2018.

[41] Zhangli Lu, Guoqiang Song, Huimin Zhu, Chuqi Lei, Xinliang Sun, Kaili Wang, Libo Qin, Yafei Chen, Jing Tang, and Min Li. DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nature Communications*, 16(1):2548, 2025.

[42] Andrew W. Mulcahy, Stephanie Rennane, Daniel Schwam, Reid Dickerson, Lawrence Baker, and Kanaka Shetty. Use of clinical trial characteristics to estimate costs of new drug development. *JAMA Network Open*, 8(1):e2453275, 2025.

[43] 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.

[44] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, 2023.

[45] Ganesha Rai, Kyle R Brimacombe, Bryan T Mott, Daniel J Urban, Xin Hu, Shyh-Ming Yang, Tobie D Lee, Dorian M Cheff, Jennifer Kouznetsova, Gloria A Benavides, et al. Discovery and optimization of potent, cell-active pyrazole-based inhibitors of lactate dehydrogenase (LDH). *Journal of Medicinal Chemistry*, 60(22):9184–9204, 2017.

[46] Pir Masoom Shah, Huimin Zhu, Zhangli Lu, Kaili Wang, Jing Tang, and Min Li. DeepDTAGen: A multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nature Communications*, 16(1):5021, 2025.

[47] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 596–608, 2020.

[48] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[49] Muhammed Talo and Serdar Bozdag. Top-dti: integrating topological deep learning and large language models for drug–target interaction prediction. *Bioinformatics*, 41(Supplement_1): i133–i141, 2025.

[50] Linwei Tao, Minjing Dong, and Chang Xu. Dual focal loss for calibration. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 33833–33849, 2023.

[51] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

[52] Mengxuan Wan, Yanpeng Zhao, Yixin Zhang, Huiyan Xu, Duoyun Yi, Peng Zan, Song He, and Xiaochen Bo. A meta learning and task adaptive approach for drug target affinity prediction. *Nature Communications*, 17(1):3734, 2026.

[53] Duolin Wang, Mahdi Pourmirzaei, Usman L. Abbas, Shuai Zeng, Negin Manshour, Farzaneh Esmaili, Biplab Poudel, Yuexu Jiang, Qing Shao, Jin Chen, and Dong Xu. S-PLM: Structure-aware protein language model via contrastive learning between sequence and structure. *Advanced Science*, 12(5):2404212, 2025.

[54] Meng-zhu Wang. SimProF: A simple probabilistic framework for unsupervised domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21153–21161, 2025.

[55] Tianci Wang, Guanyu Qiao, Guohua Wang, and Yang Li. PLiSAGE: Enhancing protein-ligand interaction prediction with multimodal surface and geometry encoding. *Bioinformatics*, 41(12): btaf608, 2025.

[56] Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[57] Donghao Xu, Bibo Wang, Dexing Zhu, Chenzhi Zheng, Haitao Wu, Ruhan Wang, Jinxing Sun, Mingchen Xie, Xu Li, Liang Wu, and Chao Zhang. MoleProLink-RL: Geometric transport for domain-policy reinforcement learning in drug-target interaction prediction. *npj Digital Medicine*, 8(1):794, 2025.

[58] Zhenxiang Xu, Jiayi Que, Yue Hong, Lantian Yao, Juan Liu, and Xiangrong Liu. CrossLinker: Aligning relational and sequential contexts for drug–target interaction prediction in cold-start and few-shot scenarios. *Journal of Chemical Information and Modeling*, 2026.

[59] Ellen Yi-Ge, Mingjing Wu, and Zhenghan Chen. Reducing divergence in batch normalization for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22155–22163, 2025.

[60] Qinze Yu, Chang Zhou, Jiyue Jiang, Xiangyu Shi, and Yu Li. GS-DTI: A graph-structure-aware framework leveraging large language models for drug–target interaction prediction. *Bioinformatics*, 41(8):btaf445, 2025.

[61] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, Gamze Deniz, and Tunca Doğan. SELFormer: molecular representation learning via SELFIES language models. *Machine Learning: Science and Technology*, 4(2):025035, 2023.

[62] Hongzhi Zhang, Xiuwen Gong, Shirui Pan, Jia Wu, Bo Du, and Wenbin Hu. A cross-field fusion strategy for drug-target interaction prediction. *arXiv preprint arXiv:2405.14545*, 2024.

[63] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations (ICLR)*, 2023.

[64] Qichang Zhao, Haochen Zhao, Linyuan Guo, Kai Zheng, Yajie Li, Qiao Ling, Jing Tang, Yaohang Li, and Jianxin Wang. ColdstartCPI: Induced-fit theory-guided DTI predictive model with improved generalization performance. *Nature Communications*, 16(1):6436, 2025.

[65] Yanpeng Zhao, Yuting Xing, Yixin Zhang, Yifei Wang, Mengxuan Wan, Duoyun Yi, Chengkun Wu, Shangze Li, Huiyan Xu, Hongyang Zhang, et al. Evidential deep learning-based drug-target interaction prediction. *Nature Communications*, 16(1):6915, 2025.

[66] Ying Zhou, Yintao Zhang, Hangwei Xu, Zhen Chen, Shijie Huang, Yinghong Li, Jianbo Fu, Hongning Zhang, Donghai Zhao, Xichen Lian, Yuan Zhou, Xinyi Shen, Kaixuan Liu, Yunqing Qiu, Yanzhong Wang, Wanqing Xie, Lianyi Han, Haibin Dai, and Feng Zhu. Dynamic clinical trial success rates for drugs in the 21st century. *Nature Communications*, 16(1):9537, 2025.

[67] Marinka Zitnik, Rok Sosič, Sagar Maheshwari, and Jure Leskovec. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection, 2018.

A Dataset Details

We evaluate BiFold on three widely used DTI benchmark datasets. Table 4 summarizes the statistics for both in-domain and cross-domain settings.

BindingDB [39] is a large-scale public database of measured binding affinities between drug-like molecules and protein targets. We extract positive drug-target pairs from the BindingDB records whose binding is characterized by an experimentally measured inhibition constant K_i, equilibrium dissociation constant K_d, or half-maximal inhibitory concentration IC_{50}, and construct negative pairs by random sampling from unobserved drug-target combinations under a matched one-to-one positive-to-negative ratio. For the cross-domain setting, drug clusters obtained through single-linkage clustering on Tanimoto similarity are used to split the source and target domains following DrugBAN [3], ensuring that no drug cluster appears in both domains.

BioSNAP [67] is derived from the DrugBank database and includes genes targeted by drugs available on the US market. The dataset contains verified positive interactions and randomly sampled negative interactions. The cross-domain split separates target protein families between source and target domains.

Human [38] contains positive interactions and high-confidence negative interactions obtained through an in silico screening process. The cross-domain split follows the same protein family separation strategy as BioSNAP.

Table 4: Numbers of unique drugs, unique proteins, and drug-target pairs for the three benchmark datasets used throughout our experiments.

Dataset	Drugs	Proteins	Interactions
BindingDB	14,643	2,623	49,199
BioSNAP	4,510	2,181	27,464
Human	2,726	2,001	6,728

For each protein, we obtain two representations. The raw amino acid sequence is encoded by ESM-2 (650M parameters) [37] with a maximum token length of 1024. The structure-aware sequence is generated by running Foldseek [51] on AlphaFold2-predicted structures [28] and encoded by SaProt (650M parameters) [48] with the same maximum length. For each drug, we convert its SMILES string to the SELFIES representation [31] and encode it with SELFormer [61] using a maximum token length of 512. All embeddings are pre-computed and stored as half-precision tensors to reduce disk I/O during training.

B Additional Implementation Details

Model architecture. The CAN layers use 8 attention heads with a per-head dimension of 96 (hidden dimension 768). The dimension-wise routing module consists of two linear transform layers with tanh activation and a two-layer routing network with layer normalization between layers. The MLP prediction head has three fully connected layers with dimensions 1536, 768, and 384, each followed by DSBN and ReLU activation, and a final output layer with dimension 1.

Training details. The model is trained using SAM with AdamW as the base optimizer. The learning rate is set to 1×10^{-4} with a weight decay of 1×10^{-4}. The SAM perturbation radius is $\rho = 0.05$. We use cosine annealing scheduling with $T_{\max}$ equal to the total number of epochs. The batch size is 64 for both source and target domains. Training runs for up to 200 epochs with early stopping based on the benchmark validation AUROC and a patience of 15 epochs. At validation time, the classification threshold is selected by maximizing the G-Mean, the geometric mean of sensitivity and specificity, on the validation split rather than using a fixed threshold of 0.5. Validation labels are used only for checkpoint selection and threshold calibration, and are not used for gradient updates, pseudo-label construction, DSBN statistic updates, or test-time adaptation. Following the cross-domain protocol of DrugBAN [3], the labeled validation split used for checkpoint selection and threshold calibration is held out from the source domain prior to training, and is therefore disjoint from both the source training set and the target test set. The target domain is partitioned only into 80% unlabeled training

samples and 20% labeled test samples, with no labeled target subset reserved for validation. As a result, no target label is observed at any stage of training, validation, or model selection, and the cross-domain evaluation follows a strict unsupervised domain adaptation protocol with respect to the target distribution. The same validation construction is applied uniformly to BiFold and to all retrained baselines.

Self-training details. The warm-up phase lasts $T_w = 6$ epochs, during which only source-domain focal loss is applied ($\gamma = 2$, no class-balancing α). Target-domain data is passed through the model in forward-only mode during warm-up so that DSBN running statistics for the target domain are properly initialized. After warm-up, pseudo-labels are generated for target-domain samples with confidence exceeding $\tau = 0.95$ (positive) or below $1 - \tau = 0.05$ (negative). The pseudo-label loss weight $\lambda(t)$ increases linearly from 0 to 1.0 over the remaining epochs.

Computational resources and runtime. All experiments are implemented in Python 3.12 with PyTorch 2.11.0 (CUDA 12.8) on a single NVIDIA GeForce RTX 5090 GPU with 32 GB of memory. The three pre-trained encoders (SaProt, ESM-2, and SELFormer) are used only for one-time embedding extraction and are not loaded during BiFold training. After the embeddings are pre-computed, each training-and-evaluation run completes within 3 hours on this single-GPU setup.

Reproducibility. The main benchmark comparisons and stepwise ablations are repeated three times with independent random initializations, and their results are reported as the mean and standard deviation across runs. The additional controlled fusion, calibration, and component-level ablations on the BioSNAP cross-domain split are evaluated using one fixed random seed (26), as indicated in the corresponding table captions. Code and pre-computed embeddings are not yet publicly available; their public release is planned.

Hyperparameter search. Table 5 summarizes the hyperparameter search space and the final selected values. The values in parentheses indicate the candidate range explored during preliminary experiments on the BindingDB cross-domain validation set.

Table 5: Hyperparameter configuration. The selected value is listed first, with the search range in parentheses.

Module	Hyperparameter	Value (Search Range)
Encoders	Drug encoder	HUBioDataLab/SELFormer
	Protein encoder (SA)	westlake-repl/SaProt_650M_AF2
	Protein encoder (Seq)	facebook/esm2_t33_650M_UR50D
CAN	Hidden dimension	768 (32, 64, 128, 256, 512, 768)
	Attention heads	8
	Aggregation mode	mean_all_tok (mean_all_tok, CLS)
MLP	Hidden sizes	1536, 768, 384
	Activation	ReLU (Tanh, ReLU)
	Batch normalization	DSBN
Training	Batch size	64 (64, 128)
	Optimizer	SAM / AdamW
	Learning rate	1×10^{-4} (1×10^{-3}–1×10^{-6})
	LR scheduler	CosineAnnealingLR
	Max epochs	200
	SAM perturbation ρ	0.05 (0.05, 0.10, 0.15)
	Focal loss γ	2
	Pseudo-label threshold τ	0.95 / 0.05

C Additional Experimental Results

C.1 Complete Ablation Results

Tables 9 and 10 provide the complete ablation results with all seven evaluation metrics and standard deviations across three independent runs. Figure 3 in the main text visualizes the key trends from these tables from three complementary perspectives (waterfall, heatmap, and per-metric trajectories).

C.2 Controlled Fusion and Calibration Comparisons

To assess fusion timing and routing granularity, we compare BiFold with seven controlled variants on the BioSNAP cross-domain split. All variants use the same data split, frozen encoders, interaction backbone, optimization settings, source-domain validation protocol, and fixed random seed (26). These controlled results correspond to the rebuttal experiments and are reported separately from the three-run mean and standard deviation of the main benchmark comparisons.

Table 6: BioSNAP cross-domain fusion and calibration ablation (one run; seed = 26). Post-int. indicates that cross-view fusion follows drug-conditioned interaction. The arrows indicate the preferred direction for each metric.

Seq.	Str.	Post-int.	Fusion	AUROC ↑	AUPRC ↑	Accuracy ↑	ECE ↓	Brier ↓	NLL ↓
✓	×	–	–	0.7264	0.7554	0.6421	0.1250	0.2051	0.4731
×	✓	–	–	0.7340	0.7550	0.6551	0.1144	0.2051	0.4833
✓	✓	✓	Mean	0.7422	0.7540	0.6527	0.1206	0.2065	0.4725
✓	✓	✓	Concat	0.7510	0.7700	0.6663	0.1207	0.2028	0.4610
✓	✓	✓	Global	0.7560	0.7679	0.6710	0.1255	0.1951	0.4627
✓	✓	✓	Sample	0.7600	0.7812	0.6886	0.1192	0.1820	0.4416
✓	✓	×	Token	0.7636	0.7800	0.6968	0.1093	0.1734	0.4365
✓	✓	✓	Channel	**0.7891**	**0.7993**	**0.7361**	**0.0520**	**0.1573**	**0.4062**

In the Fusion column, Mean and Concat represent fixed fusion, Global and Sample denote the corresponding scalar gates, Token denotes early token fusion, and Channel denotes the dimension-wise routing of BiFold. A dash indicates that cross-view fusion is inapplicable to a single-view variant.

Relative to the stronger single-view result on each metric, BiFold improves AUROC by 0.0551, AUPRC by 0.0439, and accuracy by 0.0810. Compared with early token fusion, it improves AUROC by 0.0255 and AUPRC by 0.0193. Under this controlled setup, BiFold has $11{,}126{,}785$ trainable parameters, compared with $18{,}796{,}033$ for early token fusion and $21{,}155{,}329$ for concatenation. The sample-wise scalar gate further confirms that fine-grained channel allocation benefits predictive performance.

Table 7: Paired-bootstrap AUPRC differences on the same BioSNAP cross-domain test samples (one run; seed = 26).

Comparison	AUPRC difference	95% paired-bootstrap CI
BiFold vs. early token fusion	$+0.0193$	$[0.0126, 0.0372]$
BiFold vs. sample-wise scalar gate	$+0.0181$	$[0.0114, 0.0355]$

The paired-bootstrap confidence intervals quantify sampling variability across test pairs in the fixed-seed evaluation. They do not capture variability across independent training runs.

C.3 Component-Level Adaptation and Calibration Ablations

We retain the dimension-wise routing architecture and separately remove DSBN, SAM, forward-only warm-up, and pseudo-labeling on the BioSNAP cross-domain split (one run; seed = 26). The source-only variant does not use the target-domain adaptation procedure.

Table 8: BioSNAP cross-domain component ablations for prediction and probability calibration (one run; seed = 26).

Variant	AUROC ↑	AUPRC ↑	Accuracy ↑	ECE ↓	Brier ↓	NLL ↓
Routing only (source-only)	0.7574	0.7387	0.6592	0.2601	0.2750	1.3385
Without DSBN	0.7664	0.7737	0.6757	0.1242	0.2121	0.6523
Without SAM	0.7575	0.7527	0.6886	0.1269	0.2188	0.6913
Without forward-only warm-up	0.7598	0.7566	0.5088	**0.0162**	0.2483	0.6897
Without pseudo-labeling	0.7572	0.7628	0.6898	0.1337	0.2163	0.6707
Full BiFold	**0.7891**	**0.7993**	**0.7361**	0.0520	**0.1573**	**0.4062**

Removing any individual component decreases AUROC and AUPRC relative to the full model. The full configuration also yields the lowest Brier score and NLL among the variants. The variant without forward-only warm-up has the lowest ECE (0.0162) but substantially lower accuracy (0.5088) and worse Brier score (0.2483) and NLL (0.6897), indicating that ECE alone does not capture the quality of the resulting predictions. Together, these results support the contributions of staged domain-statistics adaptation, flat-minima optimization, and confidence-thresholded self-training.

C.4 Optimality of Dimension-wise Routing

This appendix provides the formal proof of the channel-wise routing result stated as Proposition 1 in Section 3.3, restated below as Proposition 2 with an explicit uniqueness claim. An extension to correlated view noise is given at the end of the appendix.

Setup. Let $\mathbf{z} \in \mathbb{R}^h$ denote a latent interaction signal, and let two transformed view-specific features observe this signal coordinate-wise:

$$f_k^{\text{seq}} = z_k + \epsilon_k^{\text{seq}}, \qquad f_k^{\text{struct}} = z_k + \epsilon_k^{\text{struct}}, \qquad k = 1, \dots, h. \tag{12}$$

We assume that ϵ_k^{seq} and $\epsilon_k^{\text{struct}}$ are zero-mean, mutually independent, independent of z_k, and have strictly positive finite variances

$$\text{Var}(\epsilon_k^{\text{seq}}) = (\sigma_k^{\text{seq}})^2, \qquad \text{Var}(\epsilon_k^{\text{struct}}) = (\sigma_k^{\text{struct}})^2. \tag{13}$$

A dimension-wise convex fusion estimates z_k by

$$\hat{z}_k(\alpha_k) = \alpha_k f_k^{\text{struct}} + (1 - \alpha_k) f_k^{\text{seq}}, \qquad \alpha_k \in [0, 1], \tag{14}$$

and its reconstruction risk is

$$R(\boldsymbol{\alpha}) = \mathbb{E}\left[\sum_{k=1}^{h} (\hat{z}_k(\alpha_k) - z_k)^2\right]. \tag{15}$$

Proposition 2 (Dimension-wise MSE optimality)**.** *Under the setup above, the following statements hold.*

(i) The unique minimizer of $R(\boldsymbol{\alpha})$ over $[0, 1]^h$ is given coordinate-wise by

$$\alpha_k^\star = \frac{(\sigma_k^{seq})^2}{(\sigma_k^{seq})^2 + (\sigma_k^{struct})^2}, \qquad k = 1, \dots, h. \tag{16}$$

(ii) Let $c^\star$ be the optimal scalar gate obtained under the restriction $\alpha_1 = \cdots = \alpha_h = c$. Then

$$R(c^\star \mathbf{1}) \geq R(\boldsymbol{\alpha}^\star), \tag{17}$$

with strict inequality if and only if the ratios $(\sigma_k^{struct})^2/(\sigma_k^{seq})^2$ are not constant across coordinates.

Proof. For each coordinate, the estimation error is

$$\hat{z}_k(\alpha_k) - z_k = \alpha_k \epsilon_k^{\text{struct}} + (1 - \alpha_k)\epsilon_k^{\text{seq}}. \tag{18}$$

Because the two noise terms are zero-mean and independent, the coordinate-wise risk is

$$R_k(\alpha_k) = \mathbb{E}\big[(\hat{z}_k(\alpha_k) - z_k)^2\big] = \alpha_k^2(\sigma_k^{\text{struct}})^2 + (1-\alpha_k)^2(\sigma_k^{\text{seq}})^2. \tag{19}$$

Thus $R(\boldsymbol{\alpha}) = \sum_{k=1}^{h} R_k(\alpha_k)$, so the minimization separates across coordinates. Differentiating R_k gives

$$\frac{\partial R_k}{\partial \alpha_k} = 2\alpha_k(\sigma_k^{\text{struct}})^2 - 2(1-\alpha_k)(\sigma_k^{\text{seq}})^2. \tag{20}$$

Setting the derivative to zero yields Equation 16. Since both variances are strictly positive, the solution lies in $(0, 1)$. Moreover,

$$\frac{\partial^2 R_k}{\partial \alpha_k^2} = 2\left((\sigma_k^{\text{struct}})^2 + (\sigma_k^{\text{seq}})^2\right) > 0, \tag{21}$$

so each R_k is strictly convex and the coordinate-wise minimizer is unique. This proves (i).

For (ii), define

$$a_k = (\sigma_k^{\text{struct}})^2, \qquad b_k = (\sigma_k^{\text{seq}})^2, \qquad w_k = a_k + b_k. \tag{22}$$

Using $\alpha_k^\star = b_k/w_k$, direct expansion gives the exact excess-risk identity

$$R_k(\alpha_k) = R_k(\alpha_k^\star) + w_k(\alpha_k - \alpha_k^\star)^2. \tag{23}$$

Summing over coordinates yields

$$R(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha}^\star) + \sum_{k=1}^{h} w_k(\alpha_k - \alpha_k^\star)^2. \tag{24}$$

Under the scalar restriction $\boldsymbol{\alpha} = c\mathbf{1}$, the risk becomes

$$R(c\mathbf{1}) = R(\boldsymbol{\alpha}^\star) + \sum_{k=1}^{h} w_k(c - \alpha_k^\star)^2. \tag{25}$$

The optimal scalar coefficient is therefore the weighted average

$$c^\star = \frac{\sum_{k=1}^{h} w_k\alpha_k^\star}{\sum_{k=1}^{h} w_k} = \frac{\sum_{k=1}^{h} b_k}{\sum_{k=1}^{h}(a_k + b_k)}. \tag{26}$$

Substituting $c^\star$ into the previous display gives

$$R(c^\star\mathbf{1}) - R(\boldsymbol{\alpha}^\star) = \sum_{k=1}^{h} w_k(c^\star - \alpha_k^\star)^2 \geq 0. \tag{27}$$

Equality holds if and only if $\alpha_k^\star = c^\star$ for all k. By Equation 16, this is equivalent to the ratio $a_k/b_k = (\sigma_k^{\text{struct}})^2/(\sigma_k^{\text{seq}})^2$ being constant across k. If the ratio varies, at least one $\alpha_k^\star$ differs from $c^\star$, and Equation 27 is strictly positive. This proves (ii). □

Consequences for routing. The proof yields two useful consequences. First, Equation 16 shows that the more reliable view at a coordinate receives a larger weight. If the structure-view noise is small relative to the sequence-view noise, then $\alpha_k^\star$ is close to 1, and the structure view dominates that coordinate. If the reverse is true, then $\alpha_k^\star$ is close to 0, and the sequence view dominates. If the two variances are comparable, then $\alpha_k^\star$ is near 0.5, producing a neutral coordinate. Second, Equation 27 shows that the performance gap between scalar and dimension-wise fusion is exactly a weighted dispersion of the coordinate-wise optimal weights. The gap is zero only when all coordinates prefer the same mixing ratio.

Connection to the learned gate. The proposition should be interpreted as a structural motivation rather than a proof of classification improvement. It analyzes a surrogate reconstruction problem under an idealized noise model, whereas BiFold is trained with a classification loss. Nevertheless, the result motivates the dimension-wise design: if the two protein views have heterogeneous reliability across feature channels, a scalar gate cannot represent the coordinate-wise optimum of the surrogate. The gate analysis in Appendix C.5 checks whether the trained model uses this extra degree of freedom. The observed stable partition into sequence-dominant, neutral, and structure-dominant coordinates is consistent with the channel-wise reliability view, but it does not imply that individual channels have direct biological meanings or that the noise model is the only explanation for the learned partition.

Remark on correlated noise. The independence assumption is made for clarity. If the two noise terms at coordinate k have covariance $c_k = \mathrm{Cov}(\epsilon_k^{\text{struct}}, \epsilon_k^{\text{seq}})$, the same calculation gives

$$R_k(\alpha_k) = \alpha_k^2(\sigma_k^{\text{struct}})^2 + (1-\alpha_k)^2(\sigma_k^{\text{seq}})^2 + 2\alpha_k(1-\alpha_k)c_k. \tag{28}$$

When $(\sigma_k^{\text{struct}})^2 + (\sigma_k^{\text{seq}})^2 - 2c_k > 0$, the unconstrained optimum is

$$\alpha_k^\star = \frac{(\sigma_k^{\text{seq}})^2 - c_k}{(\sigma_k^{\text{struct}})^2 + (\sigma_k^{\text{seq}})^2 - 2c_k}, \tag{29}$$

followed by projection to $[0,1]$ if needed. Thus the qualitative conclusion remains the same: when the effective reliability profile varies across coordinates, a dimension-wise gate is more expressive than a scalar gate.

Table 9: Ablation study under the in-domain setting on BindingDB, BioSNAP, and Human datasets. Best values are in **bold**; second-best are underlined. Colored arrows below each value indicate the incremental change in percentage points (pp) relative to the preceding row: ↑ denotes an increase and ↓ a decrease.

Model	AUC	AUPR	Acc.	F1	Sens.	Spec.	MCC
BindingDB							
SaProt-only	0.966±0.001	0.954±0.002	0.913±0.002	0.906±0.001	0.902±0.003	0.934±0.007	0.853±0.005
w/ Dual-View only	0.971±0.002 ↑0.5	0.958±0.002 ↑0.4	0.911±0.002 ↓0.2	0.880±0.004 ↓2.6	0.903±0.003 ↑0.1	0.922±0.006 ↓1.2	0.847±0.002 ↓0.6
+ DSBN & SAM	0.979±0.002 ↑0.8	0.956±0.002 ↓0.2	0.913±0.003 ↑0.2	0.926±0.002 ↑4.6	0.923±0.003 ↑2.0	0.931±0.002 ↓0.3	0.883±0.003 ↑3.6
BiFold (full)	**0.982±0.002** ↑0.3	**0.967±0.002** ↑1.1	**0.965±0.002** ↑5.2	**0.943±0.002** ↑1.7	**0.970±0.003** ↑4.7	**0.961±0.004** ↑3.0	**0.925±0.005** ↑4.2
BioSNAP							
SaProt-only	0.933±0.002	0.935±0.002	0.866±0.002	0.865±0.002	0.868±0.003	0.865±0.002	0.624±0.002
w/ Dual-View only	0.935±0.002 ↑0.2	0.936±0.002 ↑0.1	0.861±0.002 ↓0.5	0.861±0.004 ↓0.4	0.878±0.005 ↑1.0	0.868±0.002 ↑0.3	0.649±0.002 ↑2.5
+ DSBN & SAM	0.942±0.002 ↑0.7	0.945±0.002 ↑0.9	0.876±0.003 ↑1.5	0.872±0.002 ↑1.1	0.885±0.003 ↑0.7	0.877±0.002 ↑0.9	0.661±0.003 ↑1.2
BiFold (full)	**0.960±0.002** ↑1.8	**0.962±0.002** ↑1.7	**0.884±0.002** ↑0.8	**0.876±0.002** ↑0.4	**0.889±0.005** ↑0.4	**0.897±0.002** ↑2.0	**0.766±0.004** ↑10.5
Human							
SaProt-only	0.984±0.001	0.977±0.001	0.942±0.002	0.932±0.002	0.935±0.004	0.950±0.003	0.845±0.003
w/ Dual-View only	0.981±0.002 ↓0.3	0.978±0.002 ↑0.1	0.931±0.002 ↓1.1	0.943±0.004 ↑1.1	0.946±0.004 ↑1.1	0.939±0.004 ↓1.1	0.843±0.003 ↓0.2
+ DSBN & SAM	0.986±0.002 ↑0.5	0.984±0.002 ↑0.6	0.941±0.003 ↑1.0	0.936±0.002 ↓0.7	0.950±0.003 ↑0.4	0.956±0.002 ↑1.7	0.861±0.003 ↑1.8
BiFold (full)	**0.995±0.001** ↑0.9	**0.994±0.002** ↑1.0	**0.951±0.002** ↑1.0	**0.944±0.002** ↑0.8	**0.961±0.003** ↑1.1	**0.967±0.002** ↑1.1	**0.908±0.005** ↑4.7

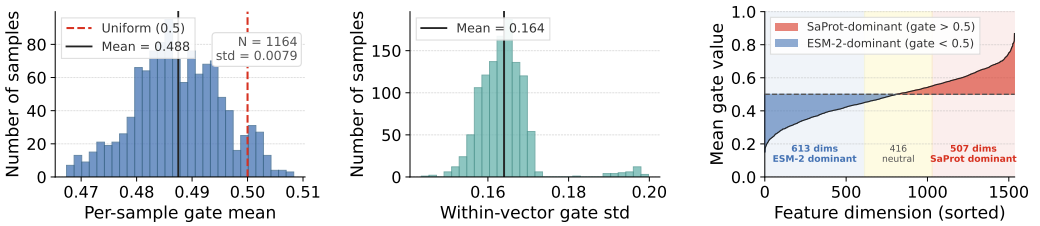
C.5 Routing Gate Analysis

The dimension-wise routing module of Section 3.3 produces, for each drug-target pair, a routing vector $\boldsymbol{\pi}^{\text{struct}} \in [0,1]^{2d}$ with $2d = 1536$, together with its complement $\boldsymbol{\pi}^{\text{seq}} = \mathbf{1} - \boldsymbol{\pi}^{\text{struct}}$. Coordinates of $\boldsymbol{\pi}^{\text{struct}}$ closer to 1 indicate higher reliance on the structure-aware stream (instantiated as SaProt in this work), while values closer to 0 indicate higher reliance on the sequence-based stream (instantiated as ESM-2). To understand what the routing module actually learns, we extract $\boldsymbol{\pi}^{\text{struct}}$ for all 1,164 samples in the in-domain Human test set and analyze its distribution along three complementary angles.

Figure 6(a) shows the distribution of the per-sample mean of $\boldsymbol{\pi}^{\text{struct}}$, computed as the average of its 1,536 coordinates for each sample. The values are tightly concentrated around 0.488 with a standard deviation of only 0.008, indicating that at the sample level, the model assigns nearly equal aggregate weight to the structure-aware and sequence-based streams regardless of the input pair. Figure 6(b) shows the within-vector standard deviation, which captures how much the routing weights vary across the 1,536 feature dimensions of a single vector. This quantity is approximately 0.164 for every sample, indicating that the routing is far from uniform within a vector.

Table 10: Ablation study under the cross-domain setting on BindingDB, BioSNAP, and Human datasets. Best values are in **bold**; second-best are underlined. Colored arrows follow the same convention as Table 9.

Model	AUC	AUPR	Acc.	F1	Sens.	Spec.	MCC
BindingDB							
SaProt-only	0.646±0.003	0.680±0.002	0.602±0.005	0.605±0.005	0.626±0.006	**0.663±0.005**	0.234±0.002
w/ Dual-View only	0.684±0.006 ↑3.8	0.663±0.002 ↓1.7	0.630±0.002 ↑2.8	0.642±0.004 ↑3.7	0.618±0.005 ↓0.8	0.633±0.003 ↓3.0	0.257±0.002 ↑2.3
+ DSBN & SAM	0.689±0.002 ↑0.5	0.671±0.002 ↑0.8	0.637±0.003 ↑0.7	0.626±0.002 ↓1.6	0.626±0.003 ↑0.8	0.652±0.002 ↑1.9	0.260±0.003 ↑0.3
BiFold (full)	**0.697±0.002** ↑0.8	**0.680±0.003** ↑0.9	**0.660±0.003** ↑2.3	**0.652±0.005** ↑2.6	**0.626±0.008** ↑0.0	0.660±0.005 ↑0.8	**0.292±0.003** ↑3.2
BioSNAP							
SaProt-only	0.743±0.006	0.756±0.005	0.691±0.005	0.644±0.008	0.569±0.004	0.738±0.006	0.276±0.015
w/ Dual-View only	0.747±0.003 ↑0.4	0.749±0.004 ↓0.7	0.700±0.002 ↑0.9	0.643±0.004 ↓0.1	0.612±0.003 ↑4.3	0.718±0.003 ↓2.0	0.367±0.002 ↑9.1
+ DSBN & SAM	0.768±0.002 ↑2.1	0.761±0.002 ↑1.2	0.728±0.003 ↑2.8	0.647±0.003 ↑0.4	0.644±0.006 ↑3.2	0.732±0.002 ↑1.4	0.398±0.003 ↑3.1
BiFold (full)	**0.786±0.003** ↑1.8	**0.797±0.003** ↑3.6	**0.736±0.004** ↑0.8	**0.651±0.004** ↑0.4	**0.651±0.006** ↑0.7	**0.762±0.003** ↑3.0	**0.415±0.005** ↑1.7
Human							
SaProt-only	0.849±0.004	0.779±0.003	0.757±0.002	0.733±0.005	0.712±0.007	0.748±0.010	0.279±0.013
w/ Dual-View only	0.885±0.006 ↑3.6	0.825±0.002 ↑4.6	0.819±0.002 ↑6.2	0.772±0.004 ↑3.9	0.782±0.005 ↑7.0	0.843±0.013 ↑9.5	0.491±0.002 ↑21.2
+ DSBN & SAM	0.893±0.002 ↑0.8	0.843±0.002 ↑1.8	0.790±0.003 ↓2.9	0.778±0.002 ↑0.6	0.799±0.003 ↑1.7	0.801±0.002 ↓4.2	0.540±0.003 ↑4.9
BiFold (full)	**0.902±0.005** ↑0.9	**0.862±0.002** ↑1.9	**0.827±0.002** ↑3.7	**0.786±0.004** ↑0.8	**0.807±0.008** ↑0.8	**0.849±0.006** ↑4.8	**0.585±0.005** ↑4.5



(a) Sample-level gate mean (b) Dimension-level diversity (c) Per-dimension gate routing

Figure 6: Routing gate analysis on the in-domain Human test set (1,164 samples, 1,536-dimensional routing vectors). (a) Distribution of the per-sample mean of π^{struct}. (b) Distribution of the within-vector standard deviation across the 1,536 dimensions. (c) Per-dimension mean of π^{struct}, sorted in ascending order. Quantitative summaries are reported in the prose below.

Figure 6(c) provides the most direct visualization of the learned routing. We sort the 1,536 feature dimensions by the mean of π^{struct} at that dimension (averaged across all 1,164 samples) and plot the resulting profile. The profile exhibits a smooth S-shaped curve spanning a wide range of $[0.151, 0.868]$, indicating that individual dimensions take strongly polarized values rather than clustering around 0.5. Of the 1,536 dimensions, 613 (40%) consistently satisfy mean routing weight < 0.45 and can be classified as sequence-dominant, 507 (33%) consistently satisfy mean routing weight > 0.55 and are structure-dominant, while only 416 (27%) fall in the neutral range $[0.45, 0.55]$. Combined with the stability across samples (Figure 6(a), std = 0.008), this pattern shows that the routing module learns a consistent dimension-level partition. Specific feature channels are systematically assigned to one view while others are assigned to the complementary view, and this assignment is largely shared across input pairs.

We interpret this as the routing module partitioning the joint 1,536-dimensional feature space into a structure-dominant subspace and a sequence-dominant subspace at training time, then using both subspaces for every input rather than switching between views per sample. This finding has two

implications. First, it confirms that simple concatenation or averaging would not be equivalent to the learned routing. A uniform 0.5 routing weight across all dimensions would lose the dimension-level specialization that the model has learned. Second, it suggests that for the present datasets and the 650M-parameter encoder pair used in this work (SaProt and ESM-2, which are of comparable capacity), the two views provide information that is best treated as additive rather than alternative. A truly input-dependent routing mechanism may become more useful when the two encoders differ substantially in capacity or specialization, or when target families with markedly different binding modes are present in the same training set.

A note on generality. The analysis above is reported on the in-domain Human test set for concreteness. The dimension-level partition is a property of the gate learned by BiFold rather than of any single dataset, and we observe the same three-region structure (tightly concentrated per-sample mean, within-vector standard deviation around 0.16, roughly one-third sequence-dominant and one-third structure-dominant dimensions) on the remaining dataset–setting combinations. The extraction and plotting scripts are planned for release with the code to support reproduction of the corresponding profiles on trained checkpoints.

C.6 Subspace Knockout: How Strongly Does Each Subspace Contribute?

The dimension-level partition reported in Appendix C.5 divides the fused interaction representation into three regions: sequence-dominant, structure-dominant, and neutral. To examine the functional contribution of these regions, we evaluate subspace knockout on two representative trained BiFold checkpoints, corresponding to in-domain Human and cross-domain BindingDB. For each checkpoint, the subspaces are identified using the thresholds in Appendix C.5. We consider two interventions on the fused representation $\mathbf{f}^{\text{fused}}$ before the DSBN MLP head. The *masked* intervention sets one subspace to zero while retaining the other two; the *only* intervention retains one subspace and masks the other two. All interventions are applied at inference time without further training.

Table 11: Subspace knockout on a single representative trained checkpoint for each reported setting. The main benchmark results in Tables 1 and 2 are three-run means. Masked rows zero one subspace; only rows retain one subspace. Parentheses indicate the AUROC change in percentage points relative to the unmasked prediction of the same checkpoint.

Intervention	Human (in-domain)		BindingDB (cross)	
	AUROC	AUPRC	AUROC	AUPRC
None (full)	0.995	0.987	0.697	0.680
Sequence-dominant masked	0.977 (-1.8)	0.963	0.700 ($+0.3$)	0.686
Structure-dominant masked	0.982 (-1.3)	0.969	0.700 ($+0.3$)	0.691
Neutral masked	0.991 (-0.4)	0.982	0.705 ($+0.8$)	0.699
Only sequence-dominant	0.950 (-4.5)	0.928	0.701 ($+0.4$)	0.657
Only structure-dominant	0.944 (-5.1)	0.924	0.688 (-0.9)	0.679
Only neutral	0.922 (-7.3)	0.910	0.694 (-0.3)	0.657

Table 11 shows that masking any single subspace produces modest AUROC changes on the two evaluated checkpoints. On Human, retaining only one subspace yields larger AUROC decreases of 4.5–7.3 percentage points, indicating that the three regions contribute complementary predictive information. On BindingDB, the changes are smaller and some masked variants slightly improve AUROC, suggesting that the effect of subspace removal depends on the evaluation setting and the behavior of the downstream classifier.

These interventions support the functional relevance of the learned feature partition while also showing that the contribution of an individual region is context-dependent. The unchanged DSBN statistics and classifier parameters may further influence the response to masking. The updated BioSNAP cross-domain benchmark reports 0.786 ± 0.003 AUROC and 0.797 ± 0.003 AUPRC in Table 2. The knockout outputs from the earlier BioSNAP checkpoint are omitted because corresponding masked and only-subspace evaluations are not available for the updated model.

C.7 Per-Dataset Breakdown of Ablation Results

In the in-domain setting, starting from the SaProt-only baseline, dual-view encoding, DSBN and SAM, and self-training generally contribute improvements in AUC (0.966 / 0.971 / 0.979 / 0.982 on BindingDB; 0.933 / 0.935 / 0.942 / 0.960 on BioSNAP; 0.984 / 0.981 / 0.986 / 0.995 on Human). The MCC improvements are particularly notable on BioSNAP, rising from 0.624 (SaProt-only) to 0.766 (BiFold full), a gain of 14.2 percentage points, and on Human, rising from 0.845 to 0.908 (+6.3 pp). These substantial MCC improvements indicate that the combined model produces more balanced predictions even under the relatively stable in-domain conditions.

In the cross-domain setting, the improvements from each component are proportionally larger. On BindingDB, adding DSBN and SAM to the dual-view architecture improves AUC from 0.684 to 0.689 (+0.5 pp) and MCC from 0.257 to 0.260 (+0.3 pp), and self-training further raises AUC to 0.697 (+0.8 pp) and MCC to 0.292 (+3.2 pp). The especially large MCC improvement from self-training suggests that pseudo-labels particularly help with reducing false positives and false negatives in a balanced manner.

On BioSNAP, the largest single-component improvement comes from adding DSBN and SAM, which raises AUC from 0.747 to 0.768 (+2.1 pp). This is consistent with the observation that BioSNAP has the most pronounced distributional mismatch between source and target protein families.

On Human, all three components contribute sizably, with the full model achieving 0.902 AUC and 0.585 MCC, compared to 0.885 AUC and 0.491 MCC for the dual-view only baseline. The 9.4 pp improvement in MCC is the largest across all datasets, indicating that the domain-adaptive training is most effective when the baseline already has a reasonable AUC but struggles with balanced classification.

D Interpretability Analysis

This appendix expands on the case study of Section 4.5. We first describe the selection protocol and the attention aggregation procedure, then provide the full list of residues recovered for each of the two drug-target pairs, and conclude with a detailed biochemical discussion of the recovered patterns.

Case selection. We select drug-target pairs from the Protein Data Bank [4] under four criteria. First, the co-crystal structure has a resolution better than 2.5 Å. Second, the target is a human protein. Third, the reported inhibitor potency satisfies $\mathrm{IC}_{50} \leqslant 100$ nM, so the bound ligand is a genuine high-affinity binder rather than a weak hit. Fourth, neither the drug nor the target appears in the training data of BiFold. The two pairs studied here, EZL-6QL2 [29] and 9YA-5W8L [45], satisfy all four criteria and offer contrasting binding chemistries, a zinc-coordinated sulfonamide inhibitor on one side and a pyrazole-based inhibitor of a dehydrogenase on the other.

Attention aggregation. For a protein of length L_p and a drug of length L_d, the cross-attention module of each stream in BiFold produces a $(L_d \times L_p)$ matrix $A^{d\to p}$, with $A^{d\to p}_{i,j}$ denoting the attention weight from drug token i to protein token j, softmax-normalized along the protein dimension. We average this matrix over the eight attention heads of the CAN layer, giving a per-head-averaged map for each of the two streams (structure-aware and sequence-based). The two streams are then averaged to obtain a single matrix $\bar{A}^{d\to p}$. The per-residue importance score of residue j is defined as

$$s_j \;=\; \sum_{i=1}^{L_d} \bar{A}^{d\to p}_{i,j}\,, \tag{30}$$

that is, the total attention received by protein token j summed across all drug tokens. Protein tokens corresponding to special tokens of the tokenizer are excluded, after which the remaining tokens are mapped back to residue positions on the full UniProt sequence.

PDB to UniProt numbering. Residue indices reported in PDB entries of mature proteins generally differ from positions on the full UniProt sequence because signal peptides and construct boundaries are removed during crystallization. To compare our predictions with the binding contacts documented in the reference crystal structures, we map each UniProt position p_{uni} to the corresponding PDB position $p_{\text{pdb}} = p_{\text{uni}} - \Delta$, with the offset Δ determined from the UniProt signal-peptide annotation.

For EZL-6QL2 (carbonic anhydrase VI, UniProt P23280), the signal peptide spans residues 1–17 and $\Delta = 17$; for 9YA-5W8L (lactate dehydrogenase-A, UniProt P00338), the PDB construct is offset by one residue and $\Delta = 1$. All residue labels reported below are in PDB numbering after this mapping.

Predicted binding residues. Table 12 lists the twenty residues with the largest importance score s_j for each of the two pairs. Residues documented in the reference crystal structures as binding-site contacts are shown in bold. On EZL-6QL2, three of the eleven documented contacts are recovered (Gln92, His94, His96) among the top twenty. On 9YA-5W8L, eight of the fourteen documented contacts are recovered (Arg105, Asn137, Pro138, Val139, Asp140, Ile141, Glu191, Thr247).

Random-baseline comparison and significance. As a calibration of these recovery counts, if twenty residues were drawn uniformly at random from a protein of length L with K documented contacts, the expected number of hits would be $K \cdot 20/L$. For EZL-6QL2 (CA-VI, $L = 308$, $K = 11$), this gives approximately 0.71 expected hits, so the observed 3 hits is approximately $4\times$ the random rate. For 9YA-5W8L (LDH-A, $L = 332$, $K = 14$), the expected is approximately 0.84, so the observed 8 hits is approximately $9.5\times$ the random rate. Under the corresponding hypergeometric null, where 20 residues are drawn without replacement from a protein with K documented contacts, the probability of recovering at least the observed number of hits is

$$P(X \geqslant k) \;=\; \sum_{i=k}^{\min(K,20)} \frac{\binom{K}{i}\binom{L-K}{20-i}}{\binom{L}{20}} .$$

This evaluates to $p \approx 0.028$ for EZL-6QL2 (a marginal but conventional rejection of the null) and $p \approx 9.2 \times 10^{-8}$ for 9YA-5W8L. The two pairs therefore differ in the strength of the alignment evidence, with the LDH-A example providing strong rejection of random recovery and the CA-VI example providing weaker but still significant rejection.

Table 12: Top-20 residues recovered by BiFold for two co-crystallized drug-target pairs, ranked by the averaged drug-to-protein attention. Residues in **bold** coincide with binding contacts documented in the corresponding PDB entry. Numbering follows the PDB entries.

Pair	Predicted binding residues (top-20, by attention)
EZL-6QL2	Ser31, Pro32, Thr69, **Gln92**, **His94**, Phe95, **His96**, Glu119, Ile120, His121, Ile122, Val123, His124, Val144, Ala146, Gly200, Ser201, Leu202, Thr203, Thr204
9YA-5W8L	Thr94, Ala95, **Arg105**, Val109, Val135, Ser136, **Asn137**, **Pro138**, **Val139**, **Asp140**, **Ile141**, Thr143, Cys162, Ser166, Ala167, **Glu191**, Ser196, **Thr247**, Ala250, Ile251

Biochemical discussion of EZL-6QL2. Carbonic anhydrase VI (CA-VI) is a zinc metalloenzyme whose catalytic activity relies on three conserved histidines (His94, His96, His119 in PDB numbering) that coordinate the active-site zinc ion [20, 12]. We note that PDB 6QL2 itself is not pure CA-VI but a CA II mutant carrying five point mutations (A65T, N67Q, F130Y, V134Q, L203T) that reproduce the CA-VI active-site pocket [29]. We use it because it is the only high-resolution co-crystal of ethoxzolamide with a CA-VI-like sulfonamide-binding site. BiFold is fed the authentic CA-VI sequence (UniProt P23280), and the residue-level comparison below is valid at positions where the CA II mutant and CA-VI share the same amino acid. Ethoxzolamide (EZL) is a classical sulfonamide inhibitor whose sulfonamide nitrogen replaces the catalytic hydroxide and directly contacts the zinc, while the benzothiazole scaffold occupies the hydrophobic half of the binding pocket. Among the residues recovered by BiFold in the top-twenty, His94 and His96 are two of the three zinc-coordinating histidines, and Gln92 is a non-trivial active-site contact documented in the reference crystal structure [29]. Beyond these three, the top twenty contains a cluster of residues numbered 119–124 (Glu119, Ile120, His121, Ile122, Val123, His124) that line the upper rim of the zinc pocket and participate in shaping the accessible surface of the inhibitor. The position of the third zinc-binding residue is where the CA II mutant and CA-VI differ. At mature position 119 the residue is His in the 6QL2 construct but Glu in the authentic CA-VI sequence, and the two nearby

histidines of CA-VI (His121 and His124) appear in our top twenty alongside this Glu119. Given that neighbouring residues share strongly correlated attention patterns in transformer models, this distribution is consistent with the model recognizing the zinc-binding region as a whole. A second cluster at positions 200–204 (Gly200, Ser201, Leu202, Thr203, Thr204) lies along the second-shell residues surrounding the ethoxy tail of the ligand and is known from the CA-II–ethoxzolamide crystal structure to participate in the hydrogen-bond network of the inhibitor through Thr200 and Thr199 [12].

Biochemical discussion of 9YA-5W8L. Lactate dehydrogenase-A (LDH-A) catalyzes the reversible conversion of pyruvate to lactate and is a validated oncology target [11]. The pyrazole-based inhibitor 9YA [45] binds in a pocket that bridges the NADH-binding cleft and the substrate-binding loop, making contacts with both the adenosine ribose region (through Arg105) and the mobile loop that closes over the substrate during catalysis (residues 137–141 of the loop, together with Asp140 and Ile141 in its closed conformation). In the top twenty predicted residues, BiFold recovers a contiguous five-residue segment of the substrate-binding loop (Asn137, Pro138, Val139, Asp140, Ile141) together with the coenzyme-binding residue Arg105, the catalytic residue Glu191, and the conserved Thr247 in the nicotinamide-binding subpocket. Among the residues not recovered, Arg168, His192 and Asp194 lie in a secondary basic patch that interacts with the carboxylate group of the inhibitor, and Ala237, Tyr238 and Leu322 form a peripheral hydrophobic contact. The eight recovered contacts therefore span the two main functional elements of the binding pocket (the NADH cleft and the substrate loop), while the six missed contacts are concentrated in secondary contact regions that are not in the immediate catalytic neighbourhood. This distribution is consistent with the model attending to residues that are most strongly coupled to the substrate- and cofactor-binding function of LDH-A, rather than to isolated surface contacts.

Candidate regions for further validation. Beyond the documented contacts recovered among the top-ranked residues, BiFold also assigns high attention to neighbouring residues around the corresponding binding pockets. For EZL-6QL2, these residues are mainly located around the upper rim of the zinc-binding pocket and the second-shell region near the ethoxzolamide scaffold. For 9YA-5W8L, the additional high-attention residues are concentrated around the substrate-binding loop and the NADH-binding cleft. We interpret these residues as candidate regions for further validation rather than newly confirmed binding contacts, since the attention scores indicate model correspondence but do not measure the binding-energy contribution of each residue.

Summary. Across the two pairs, eleven of the twenty-five documented binding contacts are recovered in the top-twenty residues ranked by averaged cross-attention, with the recovered residues concentrated in functionally coherent clusters, namely the zinc-coordinating region of the CA-VI-like active site and the NADH-loop region of lactate dehydrogenase-A. The predicted interaction scores are 0.81 for EZL-6QL2 and 0.48 for 9YA-5W8L, while the latter exhibits stronger residue-contact recovery. These examples suggest that attention-based residue ranking and pair-level prediction scores capture different aspects of model behavior.

Additional qualitative visualization. As a supplementary check that the attention localization behaviour generalizes beyond the two co-crystallized cases above, Figure 7 shows the per-residue attention profile of BiFold on a Transthyretin–inhibitor interaction drawn from the Human in-domain test set (prediction score 0.988). Transthyretin is a homotetrameric transport protein whose T4 binding channel, comprising three halogen binding pockets (HBP1–3), accommodates thyroxine and thyroxine-mimicking inhibitors [16]. Small-molecule TTR stabilizers that occupy these pockets have recently entered clinical use, with acoramidis demonstrating efficacy in transthyretin amyloid cardiomyopathy in a phase 3 trial [19]. The highest attention values on this pair concentrate on residues within and immediately adjacent to this channel, without prior knowledge of the binding site being provided to the model. We include this example for its clarity rather than as a quantitative recall benchmark. It is intended to illustrate that the aggregation procedure described above produces interpretable residue-level signals on targets of varying size and ligand chemistry.

E Baseline Descriptions

We compare BiFold against seven baseline methods spanning the recent advances in DTI prediction.

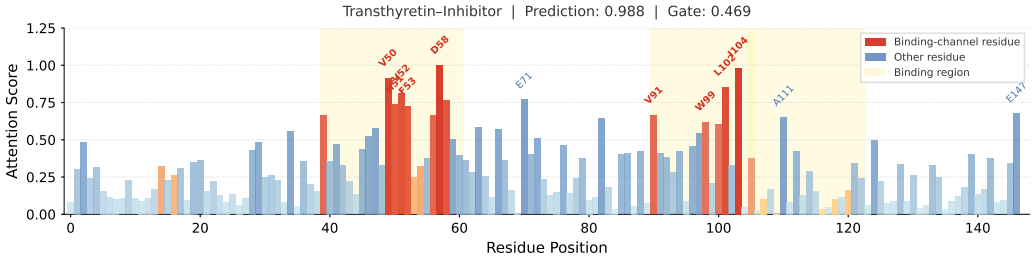


Figure 7: Per-residue cross-attention profile of BiFold on a Transthyretin–inhibitor interaction from the Human in-domain test set. The aggregation procedure follows the definition earlier in this section.

GraphDTA [43] models drug molecules as 2D molecular graphs using graph neural networks and encodes protein sequences with convolutional layers. The drug and protein representations are concatenated for prediction.

MolTrans [26] uses a transformer architecture with self-attention to capture interactions between drug and protein tokens. Substructural patterns are identified from large-scale unlabeled molecular data.

DrugBAN [3] encodes drugs with graph convolutional networks and proteins with 1D CNNs, then uses a bilinear attention network to model pairwise local interactions. It incorporates conditional domain adversarial networks for cross-domain generalization.

BioT5 [44] is a pre-trained biomedical language model that treats DTI prediction as a cross-modal sequence generation task, jointly encoding chemical and biological modalities.

SiamDTI [62] employs a dual-channel Siamese architecture to extract both local and global protein features and uses contrastive learning objectives for cross-domain transfer.

Top-DTI [49] integrates topological deep learning with large language model representations to capture higher-order structural relationships in drug and protein representations.

SaBAN [33] extends the bilinear attention framework with structure-aware protein vocabulary derived from Foldseek and incorporates contrastive learning for fine-grained binding representation.

Sources of reported numbers. Because our data splits coincide with those of Bai et al. [3], the numbers reported for GraphDTA, MolTrans, DrugBAN, and BioT5 are taken directly from the original papers, which use the identical train/validation/test partitions on BindingDB, BioSNAP, and Human. The remaining three baselines, Top-DTI, SaBAN, and SiamDTI, are trained from scratch on our machines under the same data splits and the same evaluation protocol, using the publicly available code and recommended hyperparameters of each method, because these methods either report results on a different split or do not release their evaluation pipeline in a directly comparable form. All reported standard deviations are over three independent runs.

F Broader Impacts

Computational DTI prediction can support early drug discovery by prioritizing candidate drug-target pairs for experimental validation, thereby reducing the cost of exploratory screening. Prediction scores should be interpreted as prioritization signals and verified through experimental assays before biomedical use. False-positive predictions may increase validation costs, while false negatives may exclude promising candidates from further investigation. Dataset biases and domain shifts may also affect the reliability of candidate prioritization. BiFold is intended to support experimental screening and does not generate novel chemical structures or provide clinical recommendations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that BiFold addresses two coupled challenges in DTI prediction: insufficient sequence-structure complementarity in protein representations and weak calibration under source-target shift. These claims are supported by the model design in Section 3, the state-of-the-art comparisons in Section 4.2, the block-level ablation study in Section 4.3, the routing analysis in Section 4.4, and the case study in Section 4.5. The conclusion includes a concise limitations paragraph that clarifies the task scope and natural extensions of the method. The case study is stated as qualitative evidence and is further supported by the residue lists, random-baseline comparison, significance derivation, biochemical discussion, and candidate-region discussion in Appendix D.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a concise limitations paragraph in the main text. It clarifies the task scope under the standard binary DTI prediction protocol and identifies continuous affinity modeling, multi-task endpoint prediction, and extended interpretation as future research directions.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [Yes]

   Justification: Section 3.3 states the channel-wise routing proposition and its limited role as a design motivation. Appendix C.4 provides the full signal-plus-noise assumptions, a complete proof of the MSE-optimal dimension-wise routing coefficient, and the condition under which scalar gating is strictly suboptimal for the surrogate objective. The appendix also states that the result is an idealized MSE analysis rather than a guarantee of classification performance.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper, regardless of whether the code and data are provided?

   Answer: [Yes]

   Justification: Section 4.1 specifies datasets, split protocols, baseline selection, main metrics, optimizer, learning rate, weight decay, SAM perturbation radius, batch size, number of epochs, warm-up length, confidence threshold, early stopping, and validation-based threshold calibration. Section 3 defines the architecture and objectives. Appendix A reports dataset construction and statistics, Appendix B reports implementation and hyperparameter details, Appendix E identifies the source of every baseline number, and Appendix C provides full ablation tables with standard deviations. Appendix D further specifies the case selection, attention aggregation, residue lists, random-baseline comparison, and significance derivation used in the case study. Although the code is not yet publicly released (see Question 5), the descriptions in Sections 4.1 and 3, together with the appendices, are sufficient for an independent re-implementation.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

Justification: The benchmark datasets are publicly available and the paper specifies the exact split protocols. Code and pre-computed embeddings are not yet publicly available. We plan to release them after the camera-ready submission.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the in-domain and cross-domain split protocols, optimizer, learning rate schedule, early stopping, validation-based threshold calibration, and the validation-assisted UDA boundary. It also states that validation labels are used only for checkpoint selection and threshold calibration, not for gradient computation, pseudo-label construction, DSBN statistic updates, or fitting trainable model weights. Appendix B provides architecture details, training settings, self-training settings, computational cost, reproducibility details, and the hyperparameter search space. Appendix E specifies which baseline numbers are reported from prior work and which are retrained under the same protocol.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main benchmark comparisons and stepwise ablations are repeated three times with independent random initializations, with mean and standard deviation reported in the corresponding tables. The additional BioSNAP fusion and component-level ablations use a single fixed seed (26). Paired-bootstrap confidence intervals for AUPRC differences are provided in Table 7. The interpretability analysis further includes hypergeometric significance tests for residue-contact recovery in Appendix D.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix B report that experiments are implemented in Python 3.12 with PyTorch 2.11.0 compiled for CUDA 12.8, and run on a single NVIDIA GeForce RTX 5090 GPU. They also state that the pre-trained encoders are used for one-time embedding extraction and are not loaded during training, and that each run completes within 3 hours.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The work uses publicly available biological datasets and publicly available pre-trained encoders. It does not involve human subjects, private personal data, deception, surveillance, or safety-critical deployment. The model is intended as a computational screening aid whose predictions require experimental validation before any biomedical use.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F includes a broader-impact discussion. It describes the positive impact of accelerating early drug discovery by prioritizing candidates for experimental validation, and states the standard caution that computational scores should be interpreted as prioritization signals rather than experimental evidence. The paper also clarifies that the model is intended to support experimental assays rather than provide clinical recommendations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: BiFold is a binary DTI classifier and does not generate molecules, toxic compounds, biological sequences, or actionable wet-lab protocols. The paper does not introduce a high-risk generative model or a dataset containing sensitive information. Standard scientific safeguards are provided by the requirement that predictions be experimentally validated before biomedical use.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited, and are the license and terms of use properly respected?

    Answer: [Yes]

    Justification: The paper cites all benchmark datasets and pre-trained encoders used in the study, including BindingDB, BioSNAP, Human, ESM-2, SaProt, Foldseek, and SELFormer. Appendix B lists the exact public model identifiers used for the encoders. All datasets and pre-trained encoders are publicly released research artifacts and are used in accordance with their respective licenses. In particular, the ESM-2, SaProt, Foldseek, and SELFormer model weights and source code are released under permissive open-source licenses (such as MIT or Apache-2.0) by their respective authors, and the BindingDB, BioSNAP, and Human benchmarks are released for non-commercial academic research under the terms specified by their original publishers. The work uses these assets as publicly released research resources and does not redistribute modified versions of the datasets or pre-trained models in the manuscript.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or pre-trained model. The contribution is a DTI prediction framework built from existing public datasets and frozen public encoders. Code and pre-computed embeddings are planned for release upon acceptance, as stated in Appendix B.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and details about compensation?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants and whether Institutional Review Board approvals were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: No general-purpose LLM is used as a core experimental component. BiFold uses frozen pre-trained sequence encoders for proteins and drugs, namely ESM-2, SaProt, and SELFormer, which are cited and described as standard feature extractors in Sections 3.2 and 4.1. The method does not rely on LLM prompting, generation, or interactive inference.