Generalizable Drug-Target Interaction Prediction via ESM-2 Representations and Progressive Contrastive Curriculum Learning

Qianyang Wu[1], **Jingwei Lv**[1], **Zilong Zhang**[1], **Feifei Cui**[1*]

[1]School of Computer Science and Technology, Hainan University, Haikou 570228, China
{qianyangwu, jingweilv, zhangzilong, feifeicui}@hainanu.edu.cn

Abstract

Predicting drug–target interactions (DTIs) is a fundamental task in computational drug discovery, yet it remains challenging under distribution shifts and limited training data. Existing approaches often suffer from poor generalization, weak cross-modal alignment between molecular and protein representations, and vulnerability to noisy supervision. We propose ESP-DTI, a unified framework designed to enhance generalization by integrating large-scale protein language models with curriculum learning and cross-modal contrastive alignment. Specifically, we leverage ESM-2 to encode context-aware protein representations and adopt a CLIP-style contrastive objective to align drug and protein embeddings in a shared latent space. To further improve learning robustness, we introduce a progressive curriculum sampling strategy that dynamically schedules training instances based on model confidence, enabling a gradual shift from easy to hard examples. Experimental results on four benchmark datasets demonstrate that ESP-DTI consistently outperforms state-of-the-art baselines, achieving a +3.1% improvement in average accuracy. Ablation studies confirm the complementary benefits of each component, validating their collective contribution to robust and generalizable DTI prediction. Our work underscores the effectiveness of combining pretrained protein language models with structured training curricula and cross-modal contrastive learning for reliable DTI prediction under real-world, distribution-shifted conditions.

Code — https://github.com/qianwindfeng/ESP-DTI/

Introduction

Predicting drug–target interactions (DTIs) is foundational to modern drug discovery (Abdul Raheem and Dhannoon 2024; Abbasi et al. 2021), allowing vast chemical spaces (Carlsson and Luttens 2024) to be pruned in silico before costly wet-lab validation. Yet models that excel on conventional random splits frequently collapse when the data distribution shifts—e.g. (Liu et al. 2024; Zeng, Chen, and Lei 2024), to unseen molecular scaffolds, sparsely annotated or novel protein families, or generally label-scarce regimes (Liu et al. 2024). These out-of-distribution (OOD) scenarios are the rule rather than the exception in real pipelines: the most therapeutically promising compounds and targets often lie far from the training manifold. Thus, achieving robustness and reliable generalization under distribution shift remains a core open problem in DTIs prediction.

Recent progress in biomolecular representation learning offers orthogonal advances for drug–target interaction (DTI) modeling, yet these remain largely unintegrated. First, protein language models (PLMs) like ESM-2 and ProtT5 provide high-capacity, structure-aware protein encoders trained on vast sequence data, capturing intricate dependencies without manual features or alignments (Rives et al. 2021; Vieira, Handojo, and Wilke 2025), yielding transferable embeddings effective in low-data regimes (Wu et al. 2023). Second, cross-modal contrastive learning (Luo et al. 2024), inspired by CLIP (Radford et al. 2021), aligns drugs and proteins in a shared latent space (Singh et al. 2023), enhancing semantic consistency and generalization to unseen pairs, crucial for capturing functional similarities in DTIs (Yang, Xu, and Zeng 2014). Third, curriculum learning (Tian et al. 2024) addresses prevalent label noise and optimization instability in biochemical data by staging training from high-confidence to harder samples (Tian et al. 2024), promoting stable convergence and generalization under shift, which is critical given data heterogeneity and false negatives in DTI (Playe and Stoven 2020). Despite their individual promise for enhancing representation power, alignment, and training stability, these techniques are typically explored in isolation within DTI research (Luo et al. 2024). Consequently, the synergistic potential of combining pretrained PLM representations, contrastive alignment, and progressive curriculum training remains underexplored. A unified framework leveraging all three could yield significantly more robust and generalizable DTI models, especially under cold-start and domain-shift conditions mirroring real-world drug discovery.

We present ESP-DTI, a unified framework that interleaves these three strands to directly target OOD robustness (Singh et al. 2023). (i) ESM-2–guided protein representation supplies high-capacity embeddings that capture residue-level dependencies and functional semantics without external alignments. (ii) A progressive, confidence-driven curriculum filters out high-confidence ("easy") samples after each epoch to denoise supervision, and periodically recalls a subset to prevent catastrophic forgetting and preserve diversity. (iii) A

*Corresponding author.

bidirectional CLIP-style contrastive objective with a learnable temperature aligns molecular and protein embeddings in a shared latent space, explicitly enhancing cross-modal interaction awareness. The components are mutually reinforcing: richer protein features enable more reliable confidence estimation; curriculum-based denoising focuses contrastive learning on informative, hard cases; improved cross-modal alignment feeds back into stronger downstream classification.

Across four public benchmarks—evaluated under random splits, drug-/target-cold-start conditions, and cluster-based shifts—ESP-DTI consistently surpasses state-of-the-art baselines on AUC, AUPR, and F1. Gains are most pronounced in stringent cold-start scenarios where both drugs and targets are unseen, underscoring the framework's practical relevance for early discovery. Ablation studies further reveal clear complementarity: removing ESM-2 degrades protein understanding; disabling the curriculum impairs convergence and noise tolerance; omitting contrastive alignment weakens modality integration and OOD generalization. These results highlight that encoder choice, training schedule, and objective design should be treated as interdependent design axes rather than isolated tweaks.

Our contributions are three-fold :

- We present ESP-DTI, to our knowledge the first DTI framework that jointly combines advanced PLM-based protein encoding, adaptive curriculum learning, and CLIP-style cross-modal contrastive alignment.
- We propose a confidence-driven filtering–recall curriculum that dynamically adjusts training complexity to mitigate label noise without sacrificing coverage.
- We provide extensive empirical validation under realistic evaluation protocols, demonstrating substantial improvements in robustness and generalization.

Related Work

Deep Learning for DTI Prediction

Deep models have become the dominant paradigm for drug–target interaction (DTI) prediction, improving upon early approaches that relied on fixed molecular fingerprints (Öztürk, Özgür, and Ozkirimli 2018), shallow protein sequence descriptors, or network topology features. Sequence-based CNN/Transformer architectures (Öztürk, Özgür, and Ozkirimli 2018; Chen et al. 2020) and graph-based encoders for small molecules (Nguyen et al. 2021) learn end-to-end representations that capture local substructure and sequence patterns, while multimodal fusion models (Huang et al. 2021) introduce cross-attention or hierarchical attention to integrate drug and protein views (Zhao et al. 2022). Despite these advances, many systems still depend on shallow protein encoders or handcrafted descriptors, limiting access to deeper, context-dependent protein semantics. Moreover, fusion is often performed late or with limited inductive bias toward interaction-relevant alignment, which can encourage shortcut learning on entity identity instead of genuine cross-modal reasoning, particularly under distribution shift.

Protein Language Models for Molecular Interaction Modeling

Large protein language models (PLMs), notably ESM-2, provide context aware residue-level embeddings directly from primary sequences, obviating multiple-sequence alignments and manual feature engineering. Emerging work has begun to transfer PLM features into DTI/CPI pipelines (Lin et al. 2023; Elnaggar et al. 2021); however, usage is frequently confined to frozen, globally pooled embeddings with minimal adaptation, and integration with the small molecule branch is commonly heuristic. As a result, the potential of PLMs to strengthen cross-modal interaction modeling and improve out-of-distribution (OOD) generalization (Stärk et al. 2022) remains underexploited.

Contrastive Learning and Cross-Modal Alignment

Contrastive objectives—popularized in multimodal representation learning by CLIP—optimize a shared latent space in which matched pairs are close and mismatched pairs are apart (Radford et al. 2021). In DTI, contrastive losses have been used to sharpen the separability of drug and protein embeddings and to reduce modality mismatch. Yet most instances operate with shallow encoders (Wang and Qi 2022), single-stage training, or random negative sampling that ignores biochemical confounders (Lin et al. 2024) (e.g., scaffold or homology clusters). Without stronger protein representations (Hua et al. 2025) and principled curriculum over pairs, contrastive alignment risks reinforcing spurious correlations (Li et al. 2025) or suffering from false negatives, which can blunt robustness under cold-start and cluster-shift evaluations.

Curriculum Learning and Self-Paced Sample Selection

Curriculum learning and self-paced schemes improve optimization and generalization by ordering or reweighting examples from easy to hard, or by confidence (Wang, Chen, and Zhu 2021; Zhou et al. 2018). In bioinformatics, curricula have been adopted to mitigate label noise and class imbalance (Wang et al. 2020), but DTI pipelines still predominantly rely on static training sets that do not adapt to evolving model competence or to redundancy in chemical space. Few works couple curriculum mechanisms with cross-modal contrastive training (Srinivasan, Ren, and Thomason 2023), and fewer still do so in conjunction with PLM-based protein encoders. Consequently, existing methods often struggle to suppress shortcut cues in dense scaffold neighborhoods and to transfer reliably across disjoint cluster partitions.

Method

Problem Definition

The goal of drug–target interaction (DTI) prediction is to determine whether a given drug compound interacts with a specific target protein. In current computational drug discovery research, most deep learning methods represent drug compounds using SMILES (Simplified Molecular Input Line Entry System) strings. Specifically, a drug molecule

can be denoted as $D = (d_1, \ldots d_m)$, where each d_i represents a chemically meaningful SMILES token, and m is the sequence length. Similarly, a target protein is represented by its amino acid sequence $T = (t_1, \ldots t_n)$, where each t_j corresponds to one of the 20 standard amino acids, and n denotes the sequence length of the protein. Given a drug SMILES sequence D and a protein sequence T, the objective is to train a predictive model that estimates the probability $P \in [0, 1]$ of interaction between any given drug–target pair.

Proposed Framework

The overall architecture of the ESP-DTI model is illustrated in Figure 1. Our model begins by extracting drug representations using RDKit (Landrum 2016), a widely adopted cheminformatics toolkit. For target proteins, we leverage ESM-2 (Lin et al. 2023), a large-scale pretrained protein language model, to obtain rich, context-aware protein embeddings. Next, a CLIP-based module is employed to project the drug and protein features into a shared latent space, aligning their representations through cross-modal contrastive learning. These aligned embeddings are then concatenated to form a unified representation for each drug–target pair. To enhance training efficiency and generalization, we introduce a progressive sample selection strategy. During each training epoch, if a sample's prediction confidence exceeds a predefined threshold A=90%, it is categorized as an "easy" sample and temporarily removed from the training set. The remaining samples are treated as "hard" examples and are retained for continued training in the subsequent epochs. As training progresses, the number of samples in the training set naturally decreases. To prevent overfitting due to insufficient data, once the training set size falls below a threshold B=50%, we dynamically recall a subset of previously filtered "easy" samples to maintain the training volume.

In parallel, we enable a contrastive learning mechanism to continuously refine the quality of learned representations throughout training. Finally, the fused representation of each drug–target pair is fed into a two-layer MLP (Multi-Layer Perceptron) to produce the final interaction prediction.

Drug Feature Extraction

To obtain rich and informative drug representations, we extract three types of features for each drug using the widely-used cheminformatics toolkit RDKit: (i)MORGAN fingerprints, (ii)MACCS keys,and (iii)Molecular descriptors.

These three feature types are computed independently and then concatenated to form the final comprehensive representation of the drug:

$$d_f = d_{morgan} \parallel d_{maccs} \parallel d_{desc} \tag{1}$$

Where, $\parallel$ denotes the concatenation operation. This multi-view fusion ensures that both structural and physicochemical characteristics of the drug are comprehensively captured in the final embedding.

Protein Feature Extraction

To obtain semantically rich and biologically meaningful protein representations, we utilize ESM-2, a large-scale pretrained protein language model based on the Transformer architecture. ESM-2 has been pretrained on extensive protein sequence databases, enabling it to capture deep contextual and evolutionary features from raw amino acid sequences.

Given a protein sequence , we input it into the ESM-2 model, which generates a high-dimensional embedding for each individual residue (i.e., amino acid position):

$$E_T = ESM2(T) = [e_1, e_2, \ldots, e_j], e_j \in \mathbb{R}^{\mathrm{d}} \tag{2}$$

To obtain a global fixed-length representation of the protein, we apply mean pooling across the sequence dimension:

$$t_f = \frac{1}{L} \sum_{j=1}^{L} e_j \tag{3}$$

This results in a single vector $\mathrm{t_f} \in \mathrm{R^d}$ that summarizes the overall structural and functional information of the protein. By leveraging the pretrained knowledge of ESM-2, our model benefits from powerful protein embeddings that improve downstream drug–target interaction prediction.

Fusion for Prediction

The aligned embeddings are concatenated to form the joint pair representation $[d_f \parallel t_f]$,which is fed to a two-layer MLP for interaction prediction. Training minimizes the combined objective, $\mathcal{L}_{cls}$ is the supervised binary cross-entropy on interaction labels.

CLIP-Style Cross-Modal Feature Alignment

Most existing DTI models perform drug–target interaction prediction by simply concatenating the extracted features of drug molecules and protein targets. However, Naïve fusion by concatenating drug and protein features often leaves a modality gap: drug embeddings and protein embeddings may occupy disparate regions of feature space, weakening interaction reasoning. To address this, our model incorporates a CLIP (Contrastive Language–Image Pretraining)-inspired module to align and integrate the features derived from SMILES strings and amino acid sequences into a shared semantic space.

The core idea of this fusion strategy is to project drug and protein features into a common latent embedding space where their semantic similarity can be directly measured. This is achieved by training two modality-specific encoders—one for drugs and one for proteins—using a contrastive loss that encourages correct drug–target pairs to have similar representations while pushing apart the representations of mismatched pairs.

Projection and Normalization

Let u and v be the drug and protein features produced by the modality-specific encoders (see Section Drug/Protein Features). We apply projection heads $h_d, h_p : \mathbb{R}^d \to \mathbb{R}^k$ and ℓ_2-normalize the outputs:

$$z_d = \frac{h_d(u)}{\|h_d(u)\|_2} \quad , \quad z_p = \frac{h_p(v)}{\|h_p(v)\|_2} \tag{4}$$

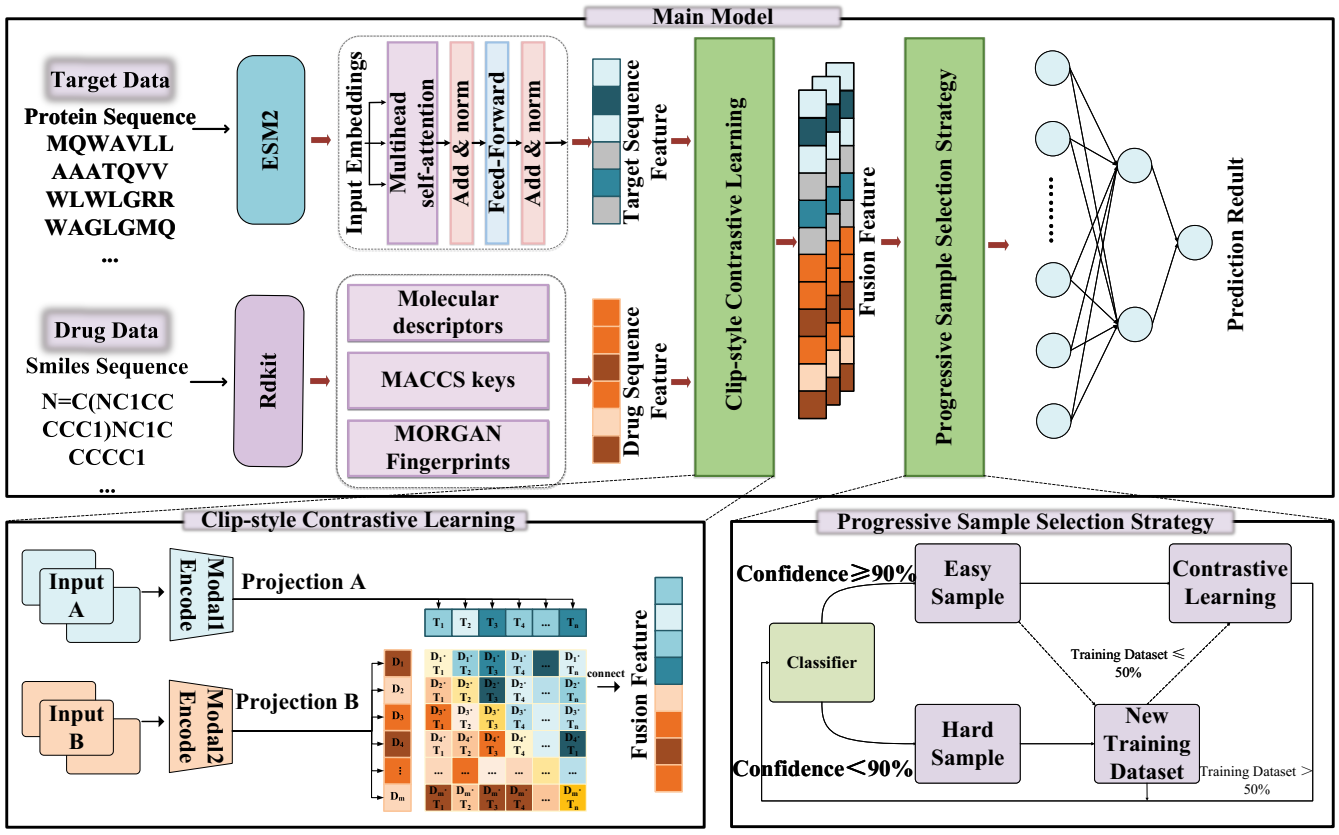

Figure 1: (A). Overall framework of ESP-DTI. The inputs to the model are the SMILES and Amino acid sequence. The output is a predicted result. The model first uses RDKit and ESM-2 to extract drug and target features respectively, then concatenates the extracted features after the Clip-style Contrastive Learning module, and then dynamically trains them through the Progressive Sample Selection Strategy module proposed by our model, and finally outputs the predicted interaction relationship. (B). The Clip-style Contrastive Learning module. (C). The Progressive Sample Selection Strategy module.

Normalization makes cosine similarity a proper metric and stabilizes contrastive optimization.

Symmetric InfoNCE Loss

For a mini-batch of B matched drug–protein pairs $\left\{\left(z_d^{(i)}, z_p^{(i)}\right)\right\}_{i=1}^{B}$, we compute a similarity matrix

$$S_{ij} = \frac{z_d^{(i)\mathrm{T}} z_p^{(j)}}{\tau} \tag{5}$$

where $\tau > 0$, is a learnable temperature.

The CLIP-style objective applies cross-entropy in both directions (drugs→proteins and proteins→drugs):

$$\mathcal{L}_{\text{clip}} = \frac{1}{2}\left[\frac{1}{B}\sum_{i=1}^{B}\mathrm{CE}\big(\mathrm{softmax}(S_{i,:}), i\big) + \frac{1}{B}\sum_{j=1}^{B}\mathrm{CE}\big(\mathrm{softmax}(S_{:,j}), j\big)\right] \tag{6}$$

This bi-directional contrast maximizes similarity for the true pair $(i = i)$ while treating the remaining $B - 1$ batch items as in-batch negatives, thereby tightening cross-modal alignment.

This bi-directional contrastive objective not only aligns drug and protein features more effectively but also improves the model's capacity to distinguish interacting from non-interacting pairs. After training, the aligned embeddings are fused (e.g., concatenated) to form the final joint representation of the drug–target pair, which is then passed to downstream prediction modules.

This CLIP-style alignment module allows our model to explicitly capture inter-modality dependencies and enhances its generalization ability in DTI prediction tasks.

Progressive Sample Selection Strategy

To address class imbalance, label noise, and mixed example difficulty in DTI prediction, we introduce a progressive, confidence-aware sample selection strategy that adapts the active training set to the model's evolving competence. Unlike classic curricula that proceed easy-to-hard, our schedule emphasizes hard-first learning with controlled reintroduction of easy samples, which we find better suited to fine-grained molecular discrimination and redundancy-rich

chemical spaces.

Let $\mathcal{S}_0$ be the initial training set and $c_\theta(x) = \max_y P_\theta(y \mid x)$ be the model confidence for pair x. At epoch t, we use a threshold $\alpha_t \in (0, 1)$ to partition the current set $\mathcal{S}_t$ into

$$\mathcal{E}_t = \{x \in \mathcal{S}_t : c_\theta(x) \geq \alpha_t\} \quad \text{(easy)} \tag{7}$$

$$\mathcal{H}_t = \mathcal{S}_t - \mathcal{E}_t \quad \text{(hard)} \tag{8}$$

We then train the next epoch primarily on $\mathcal{H}_t$, temporarily filtering $\mathcal{E}_t$ to reduce redundancy and noise amplification. The threshold follows a monotone schedule $\alpha_{t+1} = \max(\alpha_{\min}, \alpha_t - \Delta)$, with $\alpha_0 = 0.90$ by default, gradually lowering the bar for easy so that progressively simpler examples are reincorporated as training proceeds.

Recall and stability control. Because $|\mathcal{H}_t|$ may shrink as the model improves, we prevent data starvation via a size guard: when $|\mathcal{H}_t|/|\mathcal{S}_0| < \beta$ (default $\beta = 0.50$), we recall a random subset $\mathcal{R}_t \subset \mathcal{E}_t$ to form the next active set $\mathcal{S}_{t+1} = \mathcal{H}_t \cup \mathcal{R}_t$. This filter–recall cycle preserves diversity, mitigates catastrophic forgetting, and stabilizes optimization without manual per-dataset heuristics.

Contrastive refinement on mixed batches. When the active set becomes small or after a warm-up period, we activate cross-modal contrastive refinement to sharpen representation geometry. Mini-batches are constructed by mixing hard examples with recalled easy ones.

$$\mathcal{L}_{\text{contrast}} = \text{InfoNCE}(\text{hard} \cup \text{easy}) \tag{9}$$

This step emphasizes separation of confusing negatives while anchoring positives with reliable recalled pairs, improving robustness under noisy supervision and distribution shift.

Classifier and Overall Training Objective

In the final stage of the ESP-DTI model, we utilize a Multi-Layer Perceptron (MLP) as the classifier to perform binary interaction prediction. The input to the classifier is the concatenated embedding of the drug and protein features, which have been aligned and fused through the CLIP-based module.

The classifier is designed as a lightweight feed-forward neural network comprising one or more fully connected layers with non-linear activation functions (e.g., ReLU), followed by a sigmoid output layer to produce the final interaction probability $P \in [0, 1]$.

The training objective of the model integrates three key components to jointly optimize classification accuracy, cross-modal representation alignment, and contrastive discrimination on filtered samples:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}} \tag{10}$$

Experiments

Datasets and Splits

We evaluate ESP-DTI on four benchmarks: human, Caenorhabditis elegans (Tsubaki, Tomii, and Sese 2018), BindingDB (Liu et al. 2006), and BioSNAP (Huang et al. 2021). For BindingDB and BioSNAP, we consider both in-domain and cross-domain (distribution-shifted) settings. In-domain splits. For the smaller Human and C. elegans datasets, we randomly partition interactions into train: validation: test with an 80:10:10 split. For the larger BindingDB and BioSNAP datasets, we use a 70:10:20 split. We also construct cold-start (CS) evaluations on BindingDB and BioSNAP: we randomly select 70% of drugs (or proteins) and all incident drug–target pairs for training; the remaining 30% of entities and their pairs are held out and further divided 30:70 into validation:test. This protocol ensures that every drug and protein in the test set is unseen during training.

Cross-domain evaluation (cluster-based). Following the cluster-based partitioning strategy popularized by DrugBAN (Bai et al. 2023), we form disjoint source and target domains by clustering drugs using ECFP4 and proteins using PSC. We then sample 60% of drug and protein clusters as the source domain and use all associated pairs for training; the remaining 40% of clusters define the target domain. To emulate realistic transfer, we augment training with 80% of the unlabeled target-domain data (for adaptation) while withholding target labels. Model selection is performed on 80% of the labeled target-domain pairs (validation), and final generalization is reported on the remaining 20% of labeled target-domain pairs (test). This design enforces non-overlapping cluster support between domains, yielding a challenging and faithful assessment of performance under distribution shift.

Baselines and Implementation Details

We compare ESP-DTI against nine representative methods spanning classical learners and recent neural architectures: Support Vector Machine (SVM) (Cortes and Vapnik 1995), Random Forest (RF) (Ho 1995), GraphDTA (Nguyen et al. 2021), DeepConvDTI (Lee, Keum, and Nam 2019) , MolTrans (Huang et al. 2021), TransformerCPI (Chen et al. 2020), HyperAttentionDTI (Zhao et al. 2022), DrugBAN (Bai et al. 2023), and MlanDTI (Xie, Tu, and Xu 2024). Together, these baselines cover feature-based classifiers, sequence/graph encoders, and cross-modal fusion models, enabling a comprehensive assessment. ESP-DTI is implemented in PyTorch and optimized with Adam (initial learning rate 0.001). For deep learning baselines, including MlanDTI, we directly report the experimental results presented in the respective original papers. When available, we also used authors' official implementations and default configurations; otherwise, we followed the hyperparameters reported in the original papers. Unless stated otherwise, all methods use the same data partitions, preprocessing pipeline, and evaluation metrics described in MlanDTI Datasets.

Intra-Domain Experiments

We initially assess ESP-DTI under in-domain (random) splits on two small, class-balanced benchmarks—Human and C. elegans—and two larger benchmarks—BindingDB and BioSNAP. As reported in Table 1 for Human and C. elegans, ESP-DTI achieves the strongest performance across AUROC and AUPRC and delivers consistent gains in F1. Because these datasets are balanced and drawn i.i.d. from

Methods	human			C.elegans			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
SVM	0.913	0.905	0.811	0.894	0.910	0.801	0.939	0.928	0.787	0.862	0.864	0.762
RF	0.940	0.930	0.850	0.902	0.920	0.832	0.942	0.921	0.858	0.860	0.886	0.808
GraphDTA	0.960	0.959	0.897	0.974	0.975	0.919	0.951	0.934	0.867	0.887	0.890	0.789
DeepConvDTI	0.967	0.964	0.922	0.983	0.985	0.944	0.945	0.925	0.859	0.886	0.890	0.797
MolTrans	0.974	0.976	0.944	0.982	0.985	0.966	0.952	0.936	0.865	0.895	0.897	0.824
TransformerCPI	0.973	0.975	0.920	0.988	0.986	0.952	0.943	0.925	0.855	0.889	0.893	0.798
HyperAttDTI	0.984	0.984	0.946	0.989	0.990	0.958	0.959	0.948	0.887	0.901	0.902	0.838
DrugBAN	0.981	0.983	0.940	0.986	0.988	0.949	0.959	0.947	0.881	0.903	0.902	0.832
MlanDTI	0.988	0.990	**0.961**	0.990	0.992	0.962	0.945	0.926	0.857	0.909	0.912	0.841
Ours	**0.989**	**0.990**	0.944	**0.993**	**0.994**	**0.966**	**0.995**	**0.992**	**0.971**	**0.943**	**0.947**	**0.879**

Table 1: The results of the proposed model and baselines on four datasets (10 random runs), Metric: AUROC (AUC), AUPRC (AUPR), F1-score (F1). Bold indicates the best performance, and underline indicates the second best for each metric.

Methods	Cold						Cross-domain					
	BindingDB			BioSNAP			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
Moltrans	0.595	0.522	0.511	0.672	0.697	0.437	0.537	0.476	0.389	0.632	0.635	0.401
TransformerCPI	0.656	0.594	0.566	0.680	0.708	0.523	0.568	0.450	0.410	0.656	0.693	0.432
HyperAttDTI	0.661	0.598	0.582	0.732	0.760	0.539	0.545	0.462	0.376	0.654	0.685	0.395
DrugBAN	0.655	0.600	0.542	0.651	0.667	0.449	0.578	0.471	0.484	0.608	0.606	0.438
MlanDTI	**0.671**	0.594	0.601	0.782	0.801	0.653	0.657	**0.537**	0.489	0.728	0.759	0.604
Ours	0.666	**0.660**	**0.607**	**0.820**	**0.841**	**0.702**	**0.701**	0.490	**0.496**	**0.763**	**0.817**	**0.665**

Table 2: In-domain (cold pair split: unseen drugs & proteins) and cross-domain (clustering-based split) comparison on the BindingDB and BioSNAP datasets (10 random runs).

a single distribution, the results primarily reflect same-distribution discrimination rather than robustness to shift; the observed improvements indicate that ESP-DTI's representation and training objectives translate into better precision–recall trade-offs even in the absence of distributional stress.

On BindingDB and BioSNAP with random partitions, ESP-DTI again attains the best results across primary metrics. Notably, the framework overcomes the historically weaker performance on BindingDB observed for several recent methods, producing substantial improvements without task-specific heuristics. This suggests that the combined effect of ESM-2 representations, cross-modal contrastive alignment, and a progressive curriculum yields benefits that scale to larger, more heterogeneous corpora.

BindingDB poses a distinctive challenge due to its entity configuration and redundancy: 14,643 drugs versus 2,623 proteins, in contrast with BioSNAP (4,510/2,181), Human (2,726/2,001), and C. elegans (1,767/1,876). In such regimes, baseline models often drift toward shortcut cues—e.g., memorizing drug identity or leveraging nearest-neighbor similarity within dense clusters—leading to a surprisingly narrow gap between classical learners and deep architectures (RF AUROC 0.942 vs. MlanDTI 0.962). Prior work (MlanDTI) further reports a drop under drug cold-start on BindingDB relative to random splits, plausibly due to many highly similar drug molecules that make fine-grained discrimination difficult when entity information is withheld. These observations raise concerns about overfitting and practical reliability when distributions shift.

ESP-DTI addresses these failure modes through three interacting mechanisms. First, a confidence-aware progressive curriculum dynamically curates the training set, tempering the influence of near-duplicate drugs and helping optimization escape local minima that arise from redundant neighborhoods. Second, ESM-2 protein embeddings supply residue-contextualized representations that strengthen the target modality and reduce reliance on drug-only cues. Third, a CLIP-style cross-modal contrastive objective explicitly aligns molecular and protein views in a shared latent space, encouraging the model to focus on interaction-relevant signals rather than entity memorization. Together, these components improve discriminative power on BindingDB while mitigating shortcut learning.

Further we next evaluate cold-start generalization to unseen entities (drug-CS and protein-CS). In this setting, close analogs are scarce, and methods that depend on entity memorization typically degrade. ESP-DTI remains superior on BindingDB cold-start and also outperforms baselines on the more balanced BioSNAP split, indicating that the learned cross-modal alignment and curriculum-driven denoising transfer to genuinely novel drugs and proteins rather than only to previously observed neighborhoods. Overall, the BindingDB results highlight the importance of jointly (i) dynamically shaping the training distribution, (ii) leveraging pretrained protein language models, and (iii) enforcing cross-modal alignment. ESP-DTI's integrated design yields reliable gains where baselines struggle most—under entity

Ablation	BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1
w/o CL	0.637	0.422	0.442	0.755	0.791	0.649
w/o ESM	0.584	0.369	0.450	0.602	0.634	0.432
w/o CLIP	0.640	0.426	0.463	0.758	0.797	0.648
w/o PSS	0.639	0.427	0.442	0.749	0.789	0.647
Full	**0.701**	**0.490**	**0.496**	**0.763**	**0.817**	**0.665**

Table 3: Ablation study on BindingDB and BioSNAP datasets (AUC, AUPR, F1)

imbalance, redundancy, and distribution shift—while preserving strong performance in standard in-domain regimes.

Cross-Domain Experiments

Table 2 reports performance on BindingDB and BioSNAP under the cluster-based cross-domain protocol, which enforces disjoint drug scaffolds and protein clusters between training and test sets. Relative to in-domain splits, most baselines degrade markedly, reflecting the removal of scaffold/homology leakage and the resulting distribution shift. The effect is especially pronounced on BindingDB: by eliminating close analogs across domains, the protocol prevents reliance on drug-pattern memorization, and several methods regress toward chance (AUROC$\approx$0.5). Among prior approaches, MlanDTI (with multi-level attention) is the strongest baseline; however, ESP-DTI surpasses it by a clear margin on both datasets,+4.4% and +3.5% AUROC points on BindingDB and BioSNAP, respectively,under the identical cross-domain setting. These gains indicate that ESP-DTI captures interaction-relevant signals that transfer to novel drug/protein clusters, rather than exploiting distributional overlap.

Ablation Studies

We perform ablations on BindingDB and BioSNAP under the cross-domain protocol to quantify the contribution of each component in ESP-DTI (Table 3). In each variant, a single module is removed or replaced while keeping all other settings fixed.

Contrastive Refinement over Easy/Hard Partitions We disable the CLIP-style contrastive optimization applied after the dynamic easy/hard partitioning (i.e., no representation refinement on the evolving active set). Performance drops markedly, indicating that contrastive refinement is critical for sharpening the geometry of the joint embedding space once the curriculum focuses training on uncertain (hard) pairs. Without this step, the model underutilizes the informative structure of the hard pool and struggles to separate challenging negatives.

Protein Representations without ESM-2 We replace ESM-2 embeddings with handcrafted protein features (amino-acid composition and physicochemical descriptors). This yields the largest degradation across both datasets, underscoring the importance of context-aware PLM representations for transfer to unseen protein clusters. Compared with shallow descriptors, ESM-2 supplies residue-level semantics and long-range context that are essential for robust cross-modal alignment.

Removal of CLIP-Style Cross-Modal Alignment We assess the effect of removing the CLIP-style contrastive alignment by eliminating the contrastive loss without introducing any replacement mechanism. This results in a notable drop in AUROC, AUPRC, and F1 across both datasets, highlighting the importance of explicit cross-modal alignment. Without this objective, the model struggles to reconcile modality-specific representations, leading to overfitting on shallow cues like scaffold or sequence identity. The joint embedding space becomes less structured and less transferable under distribution shifts. These findings confirm that CLIP-style alignment plays a crucial role in enforcing semantic consistency and enhancing generalization.

Curriculum Disabled (No Progressive Partitioning/Recall) We train on a static dataset without confidence-aware filtering or recall. Performance declines on both datasets, with a larger drop on BindingDB. As discussed earlier, BindingDB exhibits substantial scaffold redundancy and entity imbalance; without the curriculum, optimization gravitates toward redundant neighborhoods, increasing overfitting and hindering transfer. The progressive filter–recall schedule counteracts this by concentrating updates on informative pairs while preserving diversity. Taken together, the ablations show that all three pillars—ESM-2 protein semantics, CLIP-style cross-modal alignment, and the progressive confidence-aware curriculum—are necessary and complementary: removing any one significantly degrades AUROC/AUPRC/F1, with the strongest impact observed when discarding ESM-2 features.

Conclusion

In this work, we presented ESP-DTI, a unified framework for drug–target interaction prediction that combines ESM-2–guided protein representations, a progressive confidence-aware curriculum, and CLIP-style cross-modal contrastive alignment. The design explicitly targets failure modes common in DTI modeling—overfitting to training distributions, degradation under cold-start and cluster-shift conditions, and insensitivity to example difficulty—by denoising supervision, aligning molecular and protein views in a shared latent space, and emphasizing interaction-relevant signals over entity memorization. Extensive experiments on four benchmarks under multiple protocols demonstrate state-of-the-art performance and robust transfer, with especially strong gains when both drugs and targets are unseen. Ablation studies validate the complementary roles of each component: the curriculum curbs redundancy-induced overfitting, ESM-2 enriches protein semantics, and the contrastive objective improves cross-modal alignment, together yielding consistent improvements in AUROC, AUPRC, and F1. Future directions involve integrating 3D structural signals and domain adaptation. We also aim to extend ESP-DTI to few-shot regimes and generative molecular design, establishing a versatile closed-loop framework for drug discovery.

Acknowledgments

This work was supported by the Science and Technology Special Fund of Hainan Province (ZDYF2024GXJS018).

References

Abbasi, K.; Razzaghi, P.; Poso, A.; Ghanbari-Ara, S.; and Masoudi-Nejad, A. 2021. Deep learning in drug target interaction prediction: current and future perspectives. *Current Medicinal Chemistry*, 28(11): 2100–2113.

Abdul Raheem, A. K.; and Dhannoon, B. N. 2024. Comprehensive review on drug-target interaction prediction-latest developments and overview. *Current Drug Discovery Technologies*, 21(2): 56–67.

Bai, P.; Miljković, F.; John, B.; and Lu, H. 2023. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence*, 5: 126–136.

Carlsson, J.; and Luttens, A. 2024. Structure-based virtual screening of vast chemical space as a starting point for drug discovery. *Current Opinion in Structural Biology*, 87: 102829.

Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; and Zheng, M. 2020. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16): 4406–4414.

Cortes, C.; and Vapnik, V. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3): 273–297.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112–7127.

Ho, T. K. 1995. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, 278–282. USA: IEEE Computer Society. ISBN 0818671289.

Hua, Y.; Feng, Z.; Song, X.; Wu, X.-J.; and Kittler, J. 2025. MMDG-DTI: Drug–target interaction prediction via multimodal feature fusion and domain generalization. *Pattern Recognition*, 157: 110887.

Huang, K.; Xiao, C.; Glass, L. M.; and Sun, J. 2021. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6): 830–836.

Landrum, G. 2016. RDKit: Open-Source Cheminformatics. Available at: http://www.rdkit.org. Accessed: 2024-01-15.

Lee, I.; Keum, J.; and Nam, H. 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology*, 15(6): 1–21.

Li, C.; Zhang, L.; Sun, G.; and Su, L. 2025. Multi-view based heterogeneous graph contrastive learning for drug-target interaction prediction. *Journal of Biomedical Informatics*, 104852.

Lin, X.; Zhang, X.; Yu, Z.-G.; Long, Y.; Zeng, X.; and Yu, P. S. 2024. CSCL-DTI: predicting drug-target interaction through cross-view and self-supervised contrastive learning. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 707–712. IEEE.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Liu, S.; Yu, J.; Ni, N.; Wang, Z.; Chen, M.; Li, Y.; Xu, C.; Ding, Y.; Zhang, J.; Yao, X.; et al. 2024. Versatile Framework for Drug–Target Interaction Prediction by Considering Domain-Specific Features. *Journal of Chemical Information and Modeling*, 64(14): 5646–5656.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; and Gilson, M. K. 2006. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35(suppl$_1$) : $D198 - -D201$.

Luo, Z.; Wu, W.; Sun, Q.; and Wang, J. 2024. Accurate and transferable drug–target interaction prediction with DrugLAMP. *Bioinformatics*, 40(12): btae693.

Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; and Venkatesh, S. 2021. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8): 1140–1147.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Playe, B.; and Stoven, V. 2020. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *Journal of Cheminformatics*, 12(1): 11.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Singh, R.; Sledzieski, S.; Bryson, B.; Cowen, L.; and Berger, B. 2023. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24): e2220778120.

Srinivasan, T.; Ren, X.; and Thomason, J. 2023. Curriculum learning for data-efficient vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5619–5624.

Stärk, H.; Ganea, O.; Pattanaik, L.; Barzilay, R.; and Jaakkola, T. 2022. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, 20503–20521. PMLR.

Tian, Z.; Yu, Y.; Ni, F.; and Zou, Q. 2024. Drug-target interaction prediction with collaborative contrastive learning and adaptive self-paced sampling strategy. *BMC Biology*, 22(1): 216.

Tsubaki, M.; Tomii, K.; and Sese, J. 2018. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2): 309–318.

Vieira, L. C.; Handojo, M. L.; and Wilke, C. O. 2025. Medium-sized protein language models perform well at transfer learning on realistic datasets: LC Vieira et al. *Scientific Reports*, 15(1): 21400.

Wang, Q.; Zhou, Y.; Zhang, W.; Tang, Z.; and Chen, X. 2020. Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. *Expert Systems with Applications*, 152: 113334.

Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4555–4576.

Wang, X.; and Qi, G.-J. 2022. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5549–5560.

Wu, F.; Wu, L.; Radev, D.; Xu, J.; and Li, S. Z. 2023. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6(1): 876.

Xie, Z.; Tu, S.; and Xu, L. 2024. Multilevel attention network with semi-supervised domain adaptation for drug-target prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 329–337.

Yang, F.; Xu, J.; and Zeng, J. 2014. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. In *Biocomputing 2014*, 148–159. World Scientific.

Zeng, X.; Chen, W.; and Lei, B. 2024. CAT-DTI: cross-attention and Transformer network with domain adaptation for drug-target interaction prediction. *BMC Bioinformatics*, 25(1): 141.

Zhao, Q.; Zhao, H.; Zheng, K.; and Wang, J. 2022. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3): 655–662.

Zhou, S.; Wang, J.; Meng, D.; Xin, X.; Li, Y.; Gong, Y.; and Zheng, N. 2018. Deep self-paced learning for person re-identification. *Pattern Recognition*, 76: 739–751.