Supplementary Information

AlphaPPIMI: A Comprehensive Deep Learning Framework for Predicting PPI-Modulator Interactions

Dayan Liu[1,2], Tao Song[1,2], Shuang Wang[1,2], Xue Li[1,2], Peifu Han[1,2], Jianmin Wang[3*], Shudong Wang[1,2,4*]

[1]College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, Shandong, China.
[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao, 266580, Shandong, China.
[3]The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon, 21983, Korea.
[4]State Key Laboratory of Chemical Safety, Qingdao, 266580, Shandong, China.

*Corresponding author(s). E-mail(s): jmwang113@hotmail.com; wangsd@upc.edu.cn;

S1. Fold-by-Fold Performance Analysis of Baseline Models in the Cold-Pair Setting

To provide a detailed quantitative understanding of baseline performance under the cold-pair evaluation protocol, we report fold-level confusion matrices and derived metrics for two representative models: SVM [1] and MultiPPIMI [2]. This analysis highlights model-specific biases and cross-validation instability that are otherwise obscured by aggregate metrics.

Table S1 5-fold confusion matrix breakdown for the SVM model.

Fold	P / N	TP / FN / FP / TN	Accuracy	Sensitivity	Specificity
1	893 / 586	875 / 18 / 439 / 147	0.691	0.980	0.251
2	325 / 217	309 / 16 / 195 / 22	0.611	0.951	0.101
3	214 / 257	212 / 2 / 180 / 77	0.614	0.991	0.299
4	653 / 895	522 / 131 / 850 / 45	0.366	0.800	0.050
5	768 / 711	653 / 115 / 654 / 57	0.473	0.850	0.080
Mean±SD	–	–	**0.551±0.129**	**0.914±0.081**	**0.156±0.111**

Table S2 5-fold confusion matrix breakdown for the MultiPPIMI model.

Fold	P / N	TP / FN / FP / TN	Accuracy	Sensitivity	Specificity
1	893 / 586	813 / 80 / 527 / 59	0.590	0.910	0.101
2	325 / 217	267 / 58 / 174 / 43	0.572	0.822	0.198
3	214 / 257	190 / 24 / 167 / 90	0.616	0.888	0.350
4	653 / 895	555 / 98 / 447 / 448	0.648	0.850	0.500
5	768 / 711	660 / 108 / 284 / 427	0.735	0.859	0.601
Mean±SD	–	–	**0.632±0.064**	**0.866±0.033**	**0.350±0.206**

S2. Detailed Analysis of Molecular and Protein Properties Across Datasets

We performed comprehensive characterization of molecular diversity and distribution patterns within our benchmark datasets (**Fig. S1**). The analysis of key molecular descriptors revealed distinct distribution patterns between DiPPI[3] and iPPDB[4] datasets, as shown in **Fig. S1A**. Molecular weight, LogP, and rotatable bonds distributions highlight the diverse chemical properties of interface-targeting modulators in these datasets.

Chemical space analysis through t-SNE visualization of molecular fingerprints demonstrated substantial diversity in compound distributions (**Fig. S1B**). The scattered pattern observed in the chemical space suggests significant challenges in developing generalizable prediction models across these datasets.

Comparative analysis of specific protein properties (**Fig. S1C**) further revealed dataset-specific characteristics in terms of sequence length, isoelectric point, and hydrophobicity. These distinct patterns reflect the varying strategies employed in interface-targeting across different protein-protein interactions, emphasizing the importance of robust cross-domain learning approaches in our prediction framework.

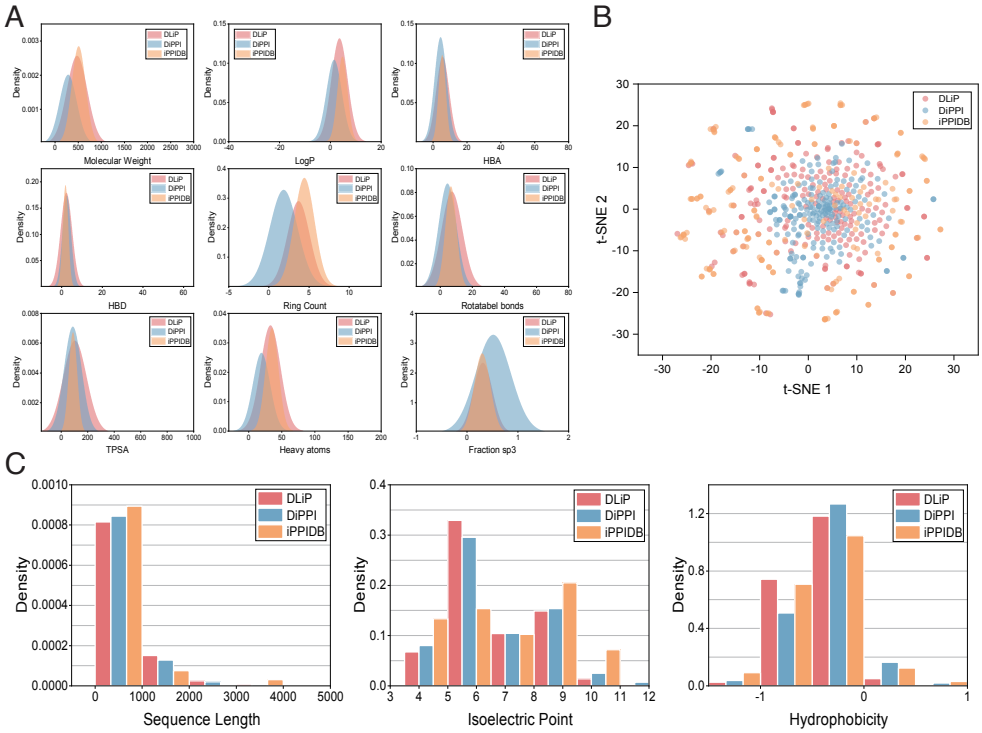

Fig. S1 Characterization of molecular physicochemical properties across distinct datasets. (**A**) Distribution of key molecular descriptors in DLiP (blue), DiPPI (orange), and iPPIDB (gray) datasets. (**B**) t-SNE visualisation of the molecular feature space for the three datasets. (**C**) Distribution of protein sequence properties: sequence length (left), isoelectric point (middle), and hydrophobicity (right).

To futher quantitatively assess the structural diversity of compounds in our datasets, we performed a comprehensive similarity analysis using Extended-Connectivity Fingerprints (ECFP4)[5]. For each dataset, we calculated pairwise Tanimoto coefficients[6] between all compounds. As shown in **Fig. S2**, both DiPPI[3] and iPPIDB[4] datasets exhibit high structural diversity. For the DiPPI dataset (**Fig. S2A**), we computed the Tanimoto similarity matrix for all 201 distinct modulators, revealing a notably low average similarity (mean = 0.078, standard deviation = 0.075) among the compounds. The majority of compound pairs exhibit Tanimoto coefficients below 0.2, with the highest frequency occurring around 0.05-0.10. Similarly, the iPPIDB dataset (**Fig. S2B**) demonstrates comparable structural diversity patterns. The hierarchical clustering dendrograms further illustrate the distinct structural clusters within both datasets, with clear separation between different compound families. These results demonstrate that our curated datasets encompass a broad and diverse chemical space, which is essential for developing robust and generalizable prediction models.

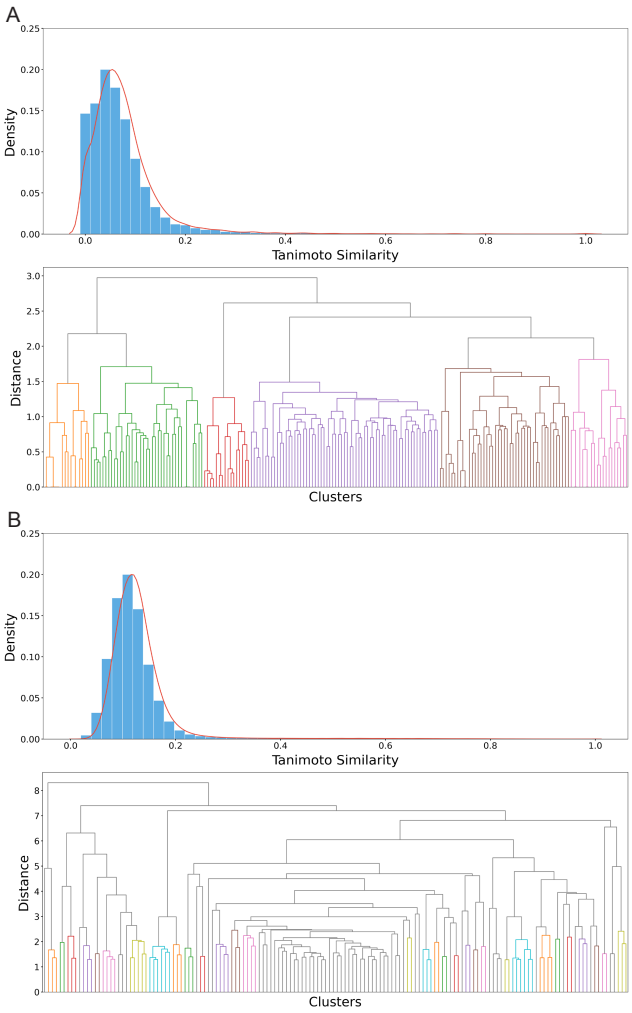

Fig. S2 Structural diversity analysis of PPI modulators. (**A**) Tanimoto similarity distribution and hierarchical clustering analysis of DiPPI dataset compounds. (**B**) Corresponding analysis for iPPIDB dataset compounds. Both datasets show high structural diversity with predominantly low pairwise similarities (Tanimoto coefficient < 0.2).

To better understand the impact of data partitioning on evaluation difficulty, we analyzed entity-level overlap statistics between training and test sets under both Random and Cold-pair split settings. For each of the five folds, we computed the percentage of overlapping proteins and modulators. As shown in **Fig. S3**, the Random split exhibits near-complete overlap at the protein level and substantial overlap at the modulator level, indicating high potential for entity-level information leakage. In contrast, the Cold-pair split shows significantly reduced overlap. Notably, we observed that Fold 1 in the cold-pair split exhibited a relatively high degree of protein and

modulator overlap. This is attributable to the fact that the cold-pair strategy only constrains interaction-level separation, and does not enforce disjoint protein or modulator sets across folds. Due to the inherent skew in the dataset—where certain proteins or compounds appear in a large number of interactions—some folds may contain higher entity-level overlap despite strict pair separation. We acknowledge this imbalance and emphasize that folds 2–5 represent more stringent generalization settings.

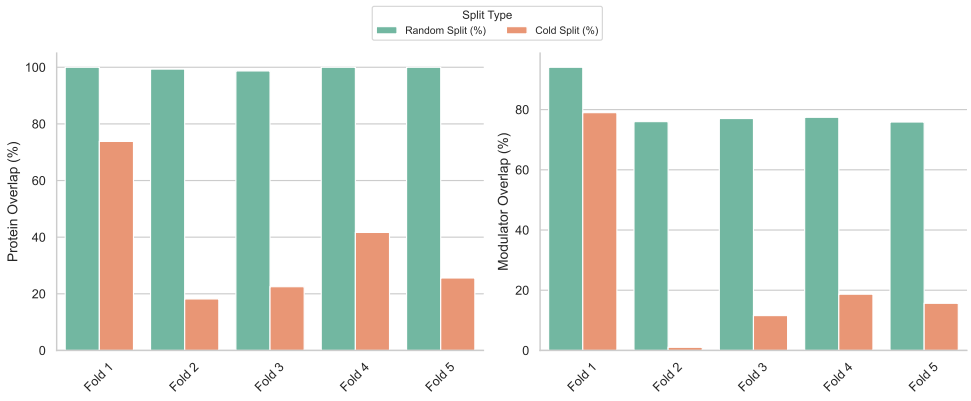

Fig. S3 Comparison of protein and modulator overlap rates under Random and Cold-pair split strategies across five folds.

S3. Analysis of ChemDiv PPI-targeted Library

The ChemDiv[7] PPI-targeted library was analyzed to understand its molecular diversity and chemical space coverage (**Fig. S4**). Distribution analysis of key molecular descriptors revealed properties specifically optimized for PPI targeting (**Fig. S4A**). The library shows appropriate ranges of molecular weight, LogP, and other physicochemical parameters crucial for interface binding.

Scaffold analysis identified several privileged core structures frequently represented in the library (**Fig. S4B**). These scaffolds demonstrate strategic design considerations for targeting protein-protein interfaces. The frequency distribution of scaffold types (**Fig. S4C**) indicates a balanced representation of structural features, suggesting comprehensive coverage of potential PPI-targeting chemical space.

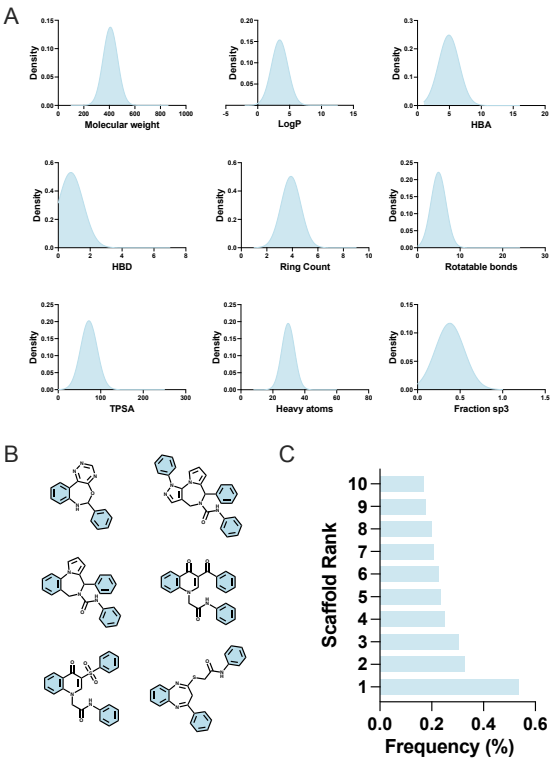

Fig. S4 Analysis of molecular properties and scaffold diversity of the ChemDiv library. (**A**) Distribution of key molecular descriptors. (**B**) Representative core scaffolds of the most frequent molecular frameworks in the library. (**C**) Frequency distribution of the top 10 scaffolds, showing the structural diversity of the compound collection.

S4. Hyperparameter setting and sensitivity analysis

Table S3 shows a list of model hyperparameters and values used in experiment. Furthermore, we conducted comprehensive hyperparameter sensitivity analysis on the DLiP dataset. **Fig. S5** shows the learning curves with different choices of key hyperparameters, including Learning rate, batch size, number of cross-attention heads and hidden dimension size. The results demonstrate that our model maintains stable performance across different hyperparameter settings, with optimal convergence achieved within 200 epochs. The most significant impact comes from the learning rate selection, while the model shows robustness to variations in other parameters within reasonable ranges.

Table S3 Notations and descriptions used in AlphaPPIMI

Notations	Description
$\mathbf{E_p} \in \mathbb{R}^{D_p \times N}$	Protein sequence embedding matrix from language models
$\mathbf{E_m} \in \mathbb{R}^{D_m \times N}$	Molecular representation matrix from Uni-Mol2
$F(\cdot), G(\cdot), D(\cdot)$	Feature extractor, decoder and domain discriminator in CDAN
$\mathbf{H^{(p)}}, \mathbf{H^{(m)}}$	Hidden representations for protein and molecule
$\mathbf{I} \in \mathbb{R}^{N \times M}$	Cross-attention interaction matrix
$\mathbf{g} \in \mathbb{R}^2$	Output modulation probability by softmax
$p \in \mathbb{R}^1$	Output interaction probability by Sigmoid
$\mathbf{W_p}, \mathbf{b_p}$	Weight matrix and bias for protein encoder
$\mathbf{W_m}, \mathbf{b_m}$	Weight matrix and bias for molecule encoder
$\mathbf{W_a}, \mathbf{b_a}$	Weight matrix and bias for cross-attention
$\mathbf{W_d}, \mathbf{b_d}$	Weight matrix and bias for decoder

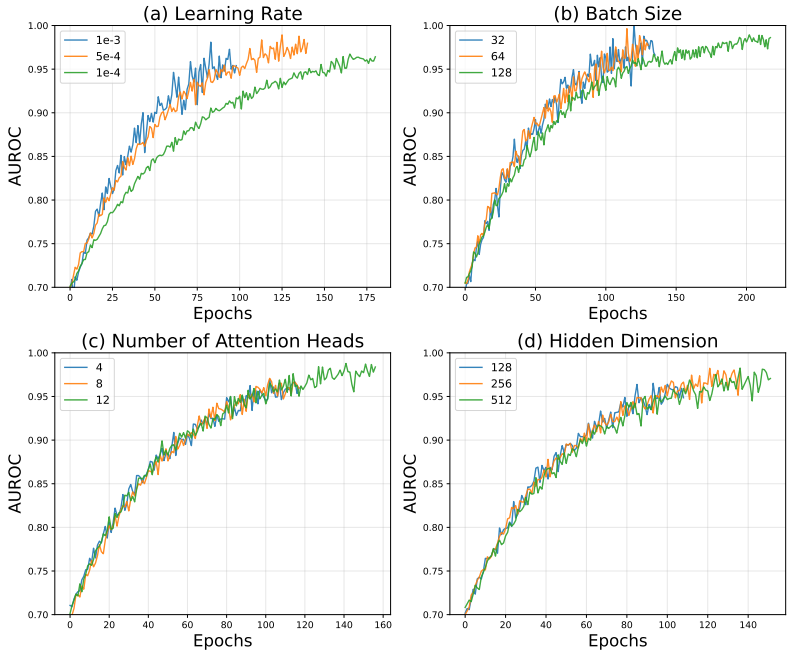

Fig. S5 Hyperparameter sensitivity analysis of AlphaPPIMI.

S5. Analysis of ESM2 Model Size Impact on Performance

To justify our selection of the ESM2-150M model in the AlphaPPIMI framework, we conducted a comparative analysis of ESM2 models across different parameter scales (**Fig. S6**). The analysis focused on evaluating the performance-efficiency tradeoff to determine the optimal model size for our application.

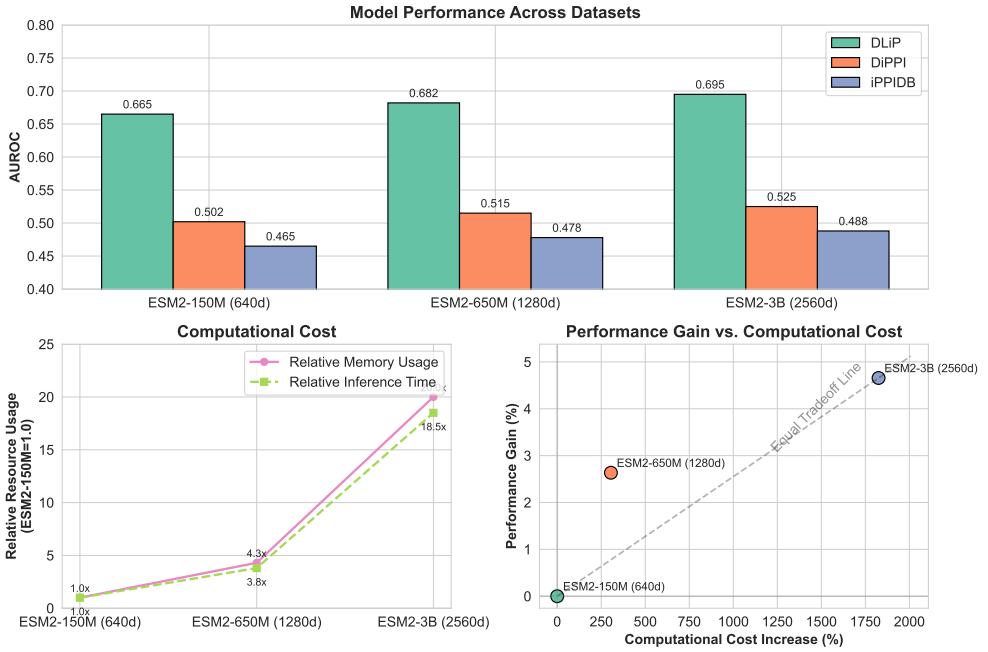

Fig. S6 Performance comparison of ESM2 models with different parameter sizes. (**A**) AUROC performance of three ESM2 variants (ESM2-150M, ESM2-650M, and ESM2-3B) across DLiP, DiPPI, and iPPIDB datasets. (**B**) Computational resource requirements normalized to ESM2-150M as baseline (1.0×). (**C**) Analysis of performance gains versus computational cost increases, with the dashed line representing the equal tradeoff threshold.

The analysis supports our decision to employ ESM2-150M in the AlphaPPIMI framework. The model achieves a favorable performance-efficiency tradeoff, enabling integration with complementary feature sources while maintaining reasonable computational requirements. This strategic choice facilitates broader applicability of our method across diverse computational environments without substantially compromising prediction accuracy.

S6. Structural Similarity Assessment and Scoring System

As mentioned in the main text, the structural similarity between compounds was evaluated using a comprehensive scoring system that combines three complementary metrics: Maximum Common Substructure (MCS)[8], Tanimoto coefficient based on Morgan fingerprints[9], and pharmacophore feature similarity. The MCS analysis identifies the largest common substructure between molecules, while Morgan fingerprints (radius=2, 2048 bits) capture the presence of specific structural fragments. The pharmacophore feature similarity considers three key aspects: aromatic ring systems, hydrogen bond patterns (donors and acceptors), and halogen atom distributions.

We developed a weighted scoring scheme that emphasizes the functional importance of different similarity metrics. Pharmacophore features were assigned the highest weight (0.5) due to their crucial role in PPI inhibition, particularly in maintaining specific interaction patterns essential for binding affinity and selectivity. The MCS similarity and Tanimoto coefficient were equally weighted (0.25 each) to provide complementary assessment of scaffold preservation and overall molecular similarity. As shown in **Fig. S7**, this weighting scheme effectively identified compounds that preserve key pharmacophoric features while maintaining reasonable structural similarity to the reference compound.

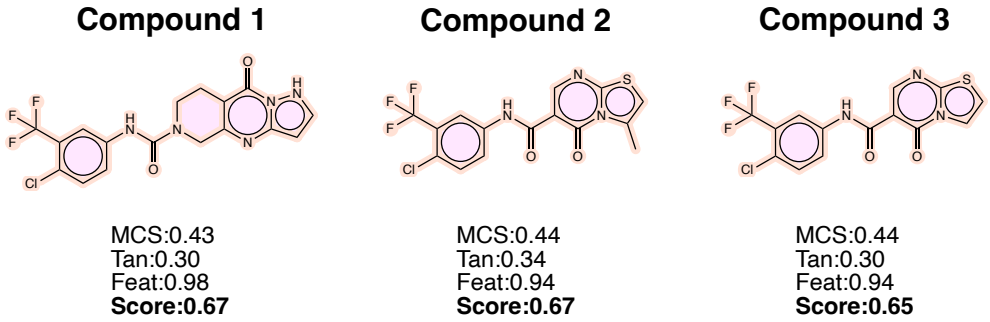

Fig. S7 Chemical structures and similarity scores of the three predicted compounds.

S7. Ablation Design and Baselines

To rigorously assess the contribution of the Conditional Domain Adversarial Network (CDAN) to cross-domain generalization, we performed a controlled ablation study involving three model variants, all evaluated on the DiPPI and iPPIDB datasets under identical data exposure. (1) AlphaPPIMI (Direct Transfer) was trained solely on the labeled source domain (DLiP) and directly tested on target domains without any adaptation. (2) AlphaPPIMI-PL, a stronger baseline, shared the same architecture as the direct transfer model but incorporated 80% of the unlabeled target-domain data via a pseudo-labeling strategy. Specifically, we first trained an initial model on DLiP, then used it to predict class probabilities for the unlabeled target samples. Only predictions with high confidence (i.e., probability > 0.95 for positives or < 0.05 for negatives) were selected and assigned corresponding pseudo-labels. These pseudo-labeled samples were then combined with the original DLiP data to train the final AlphaPPIMI-PL model. (3) AlphaPPIMI-CDAN, our proposed model, was trained on the same combination of DLiP and 80% unlabeled target data, but leveraged the unlabeled portion using an end-to-end conditional domain adversarial training mechanism, without relying on hard-threshold pseudo-labels.

S8. Sensitivity Analysis of the Negative Sample Construction Strategy

To rigorously assess the robustness of our model against the negative sampling strategy and to address the potential impact of false negatives, we conducted a sensitivity analysis. The experimental design involved maintaining a constant set of positive samples while systematically varying the down-sampling of putative inactive compounds. This process generated two additional training sets with positive-to-negative sample ratios of 1:2 and 1:5, alongside the original 1:1 balanced set. All models were subsequently retrained and evaluated under the cold-pair split configuration using a five-fold cross-validation scheme. The detailed performance metrics from the sensitivity analysis are presented in **Table S4**. The results demonstrate that while increasing the proportion of negative samples introduces greater class imbalance and predictably tempers absolute performance scores, the superiority of AlphaPPIMI over the baseline model remains stable and significant. This finding confirms that our model's predictive power is not an artifact of a specific sampling ratio but is instead a reflection of its robust learning capabilities, capable of handling challenging data distributions.

Table S4 Model performance comparison under different negative sampling ratios in the cold-pair setting (mean ± SD of 5-fold CV).

Sample Ratio (Positive:Negative)	**Model**	**AUROC**	**AUPRC**	**Specificity**
1:1	MultiPPIMI	0.738 ± 0.185	0.705 ± 0.160	0.350 ± 0.206
	AlphaPPIMI	**0.827 ± 0.033**	**0.781 ± 0.051**	**0.689 ± 0.059**
1:2	MultiPPIMI	0.713 ± 0.142	0.677 ± 0.145	0.272 ± 0.235
	AlphaPPIMI	**0.821 ± 0.045**	**0.772 ± 0.058**	**0.675 ± 0.071**
1:5	MultiPPIMI	0.701 ± 0.131	0.656 ± 0.159	0.249 ± 0.218
	AlphaPPIMI	**0.814 ± 0.041**	**0.758 ± 0.066**	**0.659 ± 0.068**

References

[1] Cortes, C.: Support-vector networks. Machine Learning (1995)

[2] Sun, H., Wang, J., Wu, H., Lin, S., Chen, J., Wei, J., Lv, S., Xiong, Y., Wei, D.-Q.: A multimodal deep learning framework for predicting ppi-modulator interactions. Journal of chemical information and modeling **63**(23), 7363–7372 (2023)

[3] Cankara, F., Senyuz, S., Sayin, A.Z., Gursoy, A., Keskin, O.: Dippi: A curated data set for drug-like molecules in protein–protein interfaces. Journal of Chemical Information and Modeling **64**(13), 5041–5051 (2024)

[4] Labbé, C.M., Laconde, G., Kuenemann, M.A., Villoutreix, B.O., Sperandio, O.: ippi-db: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. Drug discovery today **18**(19-20), 958–968 (2013)

[5] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling **50**(5), 742–754 (2010)

[6] Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of cheminformatics **7**, 1–13 (2015)

[7] ChemDiv Inc.: ChemDiv: Research and Discovery Screening Libraries. Commercial compound library (2024). https://www.chemdiv.com/

[8] Cao, Y., Jiang, T., Girke, T.: A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics **24**(13), 366–374 (2008)

[9] Zhong, S., Guan, X.: Count-based morgan fingerprint: a more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties. Environmental Science & Technology **57**(46), 18193–18202 (2023)