Common Interface Network for Multi-domain Biomolecular Interaction Learning

Jiadong Lu[1,4,†], **Yu Wang**[2,3,†], **Fuming Zeng**[2,3], **Yipin Lei**[8], **Fuli Feng**[1,✉], **Shiwei Sun**[2,3,✉], **and Xin Gao**[5,6,7,✉]

[1]School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei and 230027, China.
[2]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing and 100190, China
[3]University of Chinese Academy of Sciences, Beijing and 100049, China.
[4]Zhongguancun Academy, Beijing and 100094, China
[5]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.
[6]Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.
[7]Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.
[8]Syneron Technology, Guangzhou 510000, China.

Supplementary Note 1: Evaluation of AlphaFold3 for epitope prediction

To establish a comparative baseline for epitope identification within the context of antibody-antigen interactions, we evaluated the predictive performance of AlphaFold3 (1). For the 32 antibody-antigen complexes in epitope prediction, we used the AlphaFold3 server (https://alphafoldserver.com) to predict their complex structures. For each target complex, five candidate structures were generated. To ensure an objective assessment, we selected the top-ranked model for each case based on the 'ranking_score' provided by AlphaFold3.

The accuracy of the predicted binding interface was quantified using the F_{nat} (2, 3), which represents the proportion of native intermolecular contacts successfully recovered in the predicted model compared to the experimental ground truth. The F_{nat} metric serves as a high-fidelity proxy for interface recognition accuracy, with values ranging from 0 to 1. We defined a threshold of $F_{nat} \geq 0.5$ as the criterion for "successful identification" of the antigenic epitope. Models yielding an $F_{nat} < 0.5$ were considered to have failed in accurately localizing the specific interaction site. Our analysis of 32 test samples revealed that 12 samples (37.5%) showed successful identification ($F_{nat} \geq 0.5$), while 20 samples (62.5%) showed suboptimal identification ($F_{nat} < 0.5$).

Reference

1. Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander 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.
2. Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.
3. Keeley W Collins, Matthew M Copeland, Guillaume Brysbaert, Shoshana J Wodak, Alexandre MJJ Bonvin, Petras J Kundrotas, Ilya A Vakser, and Marc F Lensink. Capri-q: The capri resource evaluating the quality of predicted structures of protein complexes. *Journal of molecular biology*, 436(17):168540, 2024.

Supplementary Table 1. Statistics on protein–small molecule interactions for 10 common and 10 unseen small molecules in the training and test sets. Common small molecules refer to those appearing multiple times in both the training and test sets. Unseen small molecules refer to those appearing only a few times in the test sets and not appearing in the training set.

Category	Small molecule	# binding proteins	
		In Train set	In Test_1 and Test_2 sets
Common	ADP	1264	181
	NAD	1,264	95
	HEM	1,007	254
	FAD	1,156	127
	NAP	821	138
	ATP	842	161
	FMN	574	123
	HEC	570	49
	PLP	665	61
	GDP	587	44
Unseen	FLF	0	5
	V3L	0	2
	AZA	0	11
	H6P	0	7
	URC	0	12
	LRY	0	6
	WQT	0	2
	CAP	0	22
	ECH	0	7
	F24	0	4

Supplementary Table 2. Statistics of the Lit-PCBA dataset for small molecule virtual screening evaluation. The table summarizes the count of active and inactive compounds, along with the total number of molecules for each specific protein target.

Target	# actives	# inactives	# actives and inactives
FEN1	369	355,309	355,678
ALDH1	7,166	137,810	144,976
MAPK1	308	62,500	62,808
ADRB2	17	312,401	312,418
TP53	79	4,168	4,247
IDH1	39	361,958	361,997
ESR1_ant	102	4,947	5,049
PPARG	27	5,210	5,237
MTORC1	97	32,972	33,069
ESR1_ago	13	5,582	5,595
KAT2A	194	348,412	348,606
GBA	166	295,871	296,037
PKM2	546	245,457	246,003
VDR	882	355,241	356,123
OPRK1	24	269,745	269,769

Supplementary Table 3. Statistics of HLA alleles, peptides, and total samples within the HLA3DB dataset across the training, validation, and testing partitions.

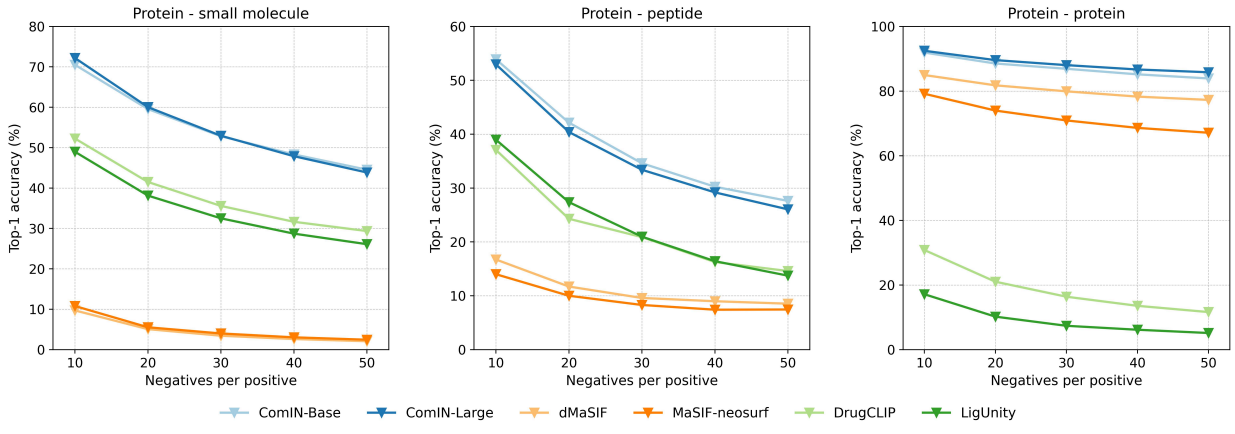
Dataset partition	# HLA alleles	# peptides	# samples
Train set	43	173	200
Valid set	14	36	37
Test set	42	167	200

Supplementary Table 4. PDB IDs of 32 antibody–antigen complex structures used in the epitope prediction task.

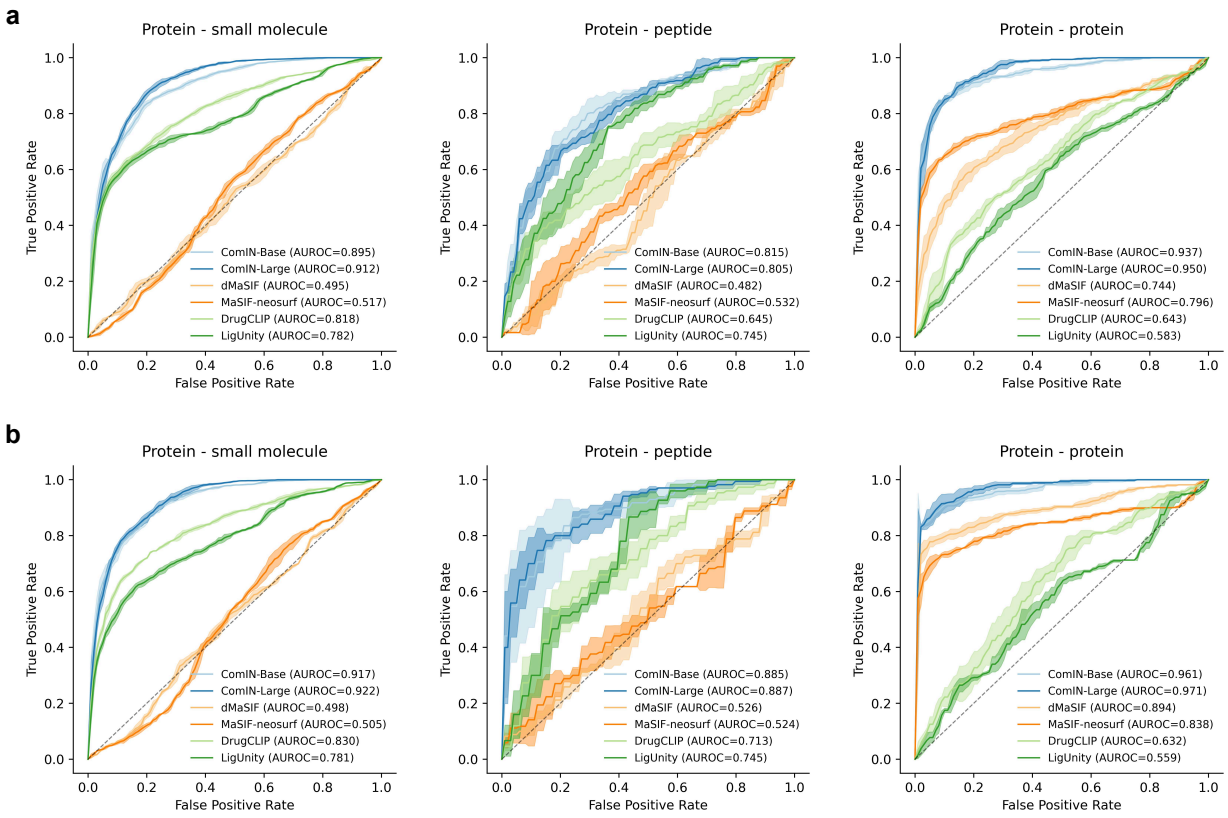
PDB IDs			
8EZL	8SIR	4FP8	8EEE
6MTN	8V2E	8HRD	6Q18
6OOR	4YE4	5UEM	4J4P
8VDL	4YFL	4YDJ	4XMP
6NNJ	6NM6	4JB9	3LD8
5VLP	6SVL	6C08	4QTI
6NNF	4RX4	7LFB	5F96
1NFD	6BCK	4J6R	7S13

Supplementary Table 5. Hyperparameter configurations for the ViSNet encoder within ComIN.

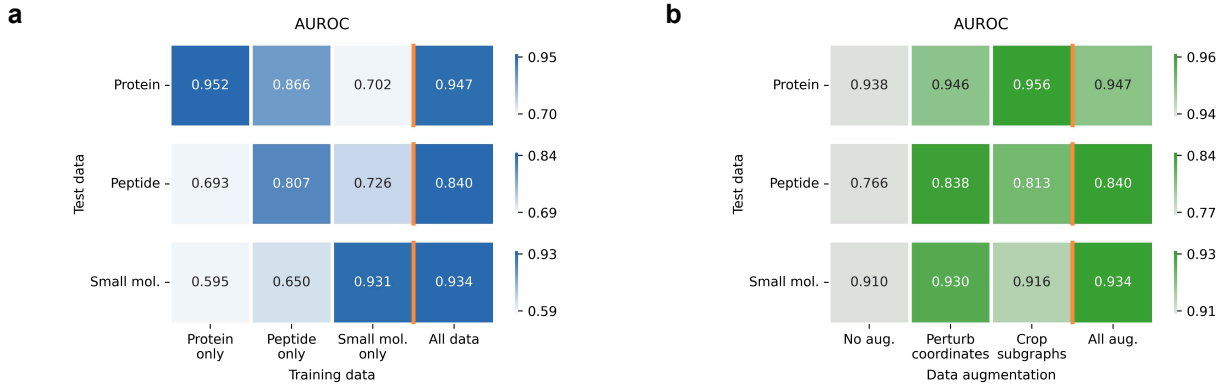
Hyperparameter	Value
Number of layers (`num_layers`)	6 for ComIN-Base, 8 for ComIN-Large
Number of hidden channels in the node embeddings (`hidden_channels`)	160
Number of output channels in the node embeddings (`output_channels`)	160
Maximum degree of spherical harmonics (`lmax`)	1
Number of attention heads (`num_heads`)	8
Number of radial basis functions (`num_rbf`)	32
Radius cutoff distance (`cutoff`)	4.5 (Å)
Maximum number of neighbors (`max_num_neighbors`)	16
Maximum atomic numbers (`max_z`)	34
Aggregation operation	mean



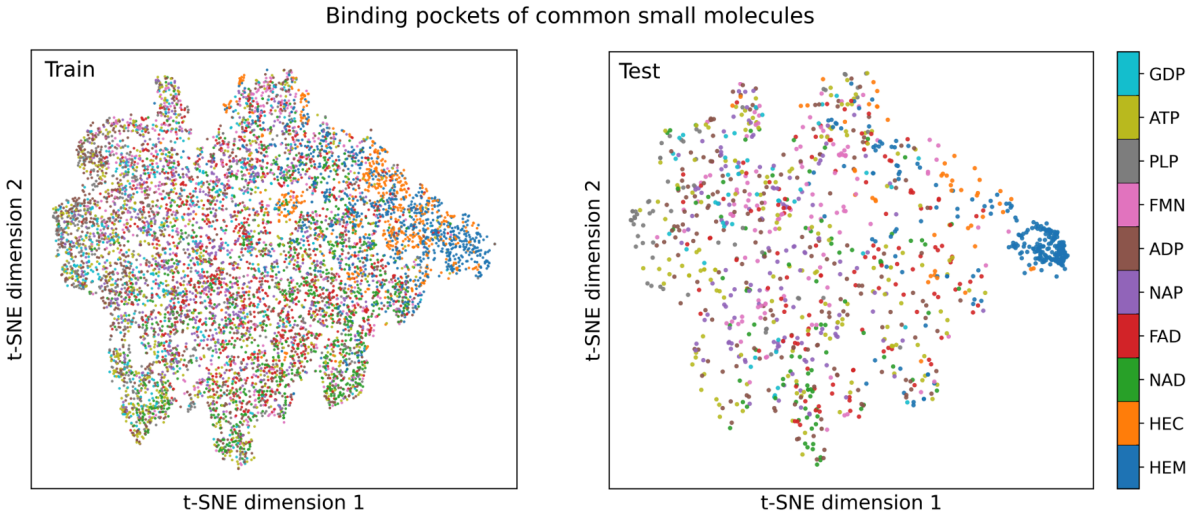
Supplementary Fig. 1. Ranking performance in predicting protein–small molecule, protein–peptide, and protein–protein interactions on the Test_1 set.



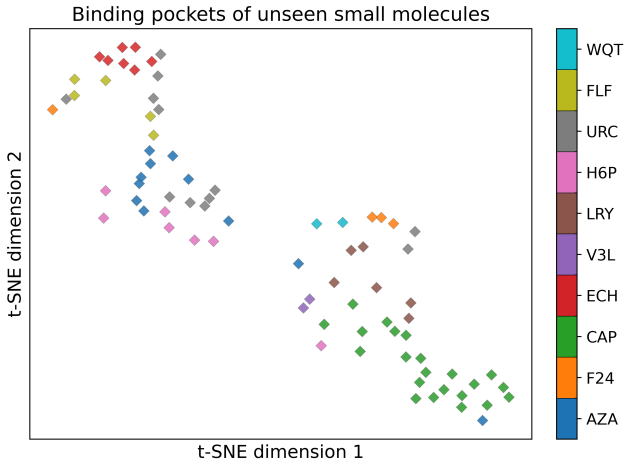
Supplementary Fig. 2. Robustness analysis of ComIN on the Test_2 set. a, ROC curves and AUC values achieved by ComIN and baseline methods for relaxed proteins. **b,** ROC curves and AUC values for ESMfold-predicted proteins.



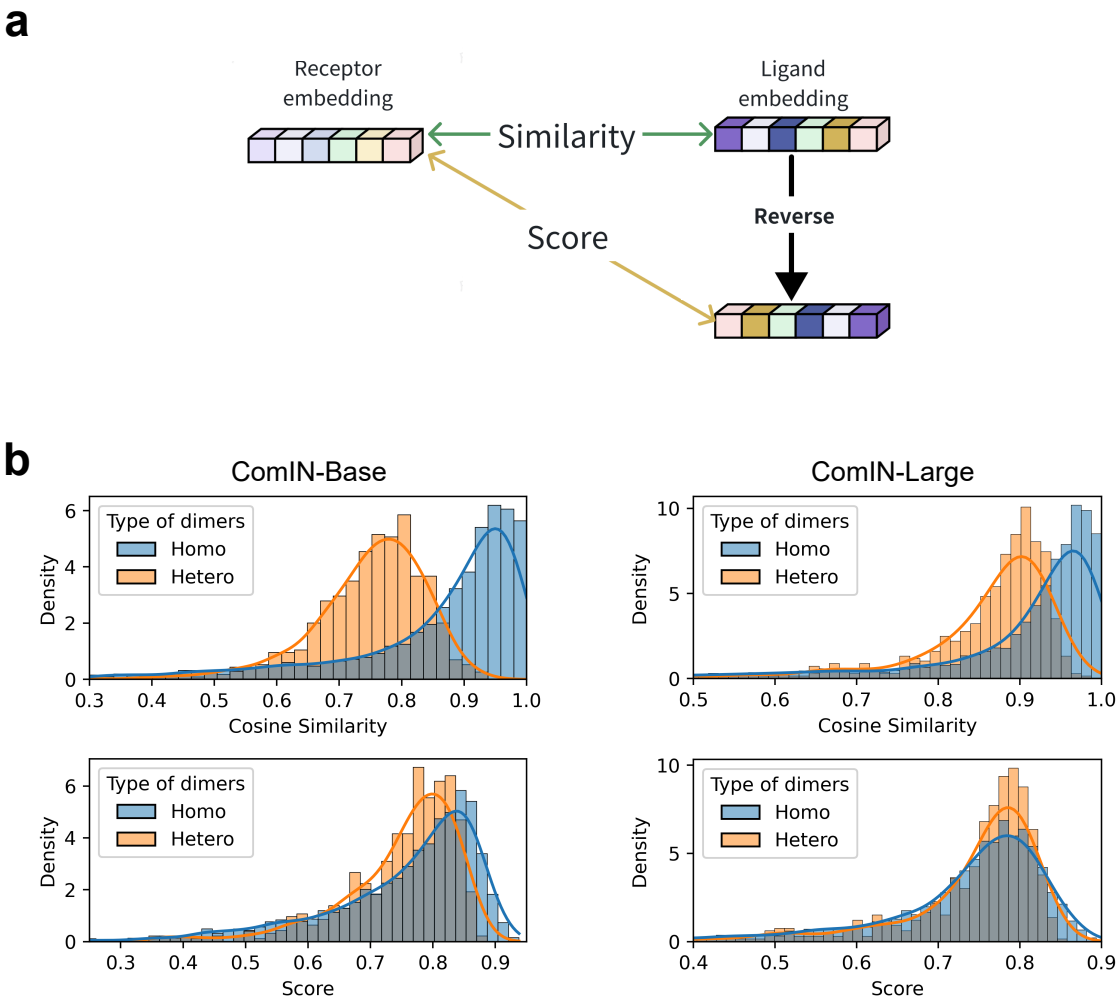
Supplementary Fig. 3. Ablation analysis of ComIN on the Test_2 set. a, Ablation study comparing ComIN with models trained on a single interaction type. **b,** Effect of data augmentation strategies on model performance.



Supplementary Fig. 4. t-SNE visualization of pocket embeddings learned by the trained ComIN encoder for 10 common small molecules. Points of the same color are not clustered together, and there are no noticeable gaps between points of different colors.



Supplementary Fig. 5. t-SNE visualization of pocket embeddings for 10 unseen small molecules. No obvious clustering was observed.



Supplementary Fig. 6. Effect of reverse operation in interaction score calculation. **a,** Schematic diagram of cosine similarity and interaction score calculation. **b,** Difference of cosine similarity and interaction score distributions for protein-protein interactions. Results are shown for homodimers (identical chains) and heterodimers (40%–70% sequence similarity). While a notable shift is observed in the cosine similarity distributions between the two dimer types (top), the score distributions (bottom) remain highly consistent, demonstrating the stability of the scoring mechanism across varying sequence similarities.