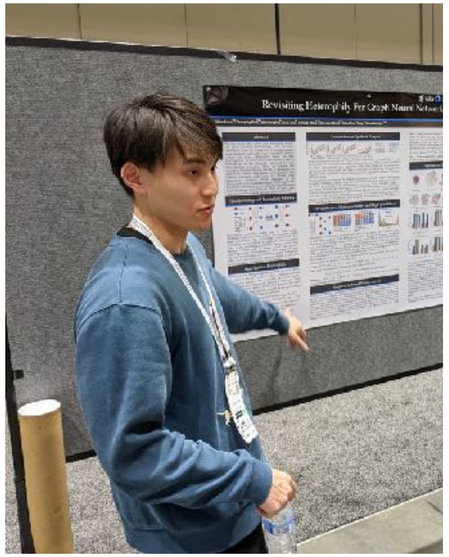
Effective Protein-Protein Interaction Exploration with PPIretrieval

First deep-learning retrieval framework for

https://arxiv.org/pdf/2402.03675.pdf

Chenqing William Hua @ McGill&Mila

- Thanks to all my collaborators :)



McGill&Mila

MIT

UdeM&Mila

DeepMind&McGill&Mila

SJTU&Aureka

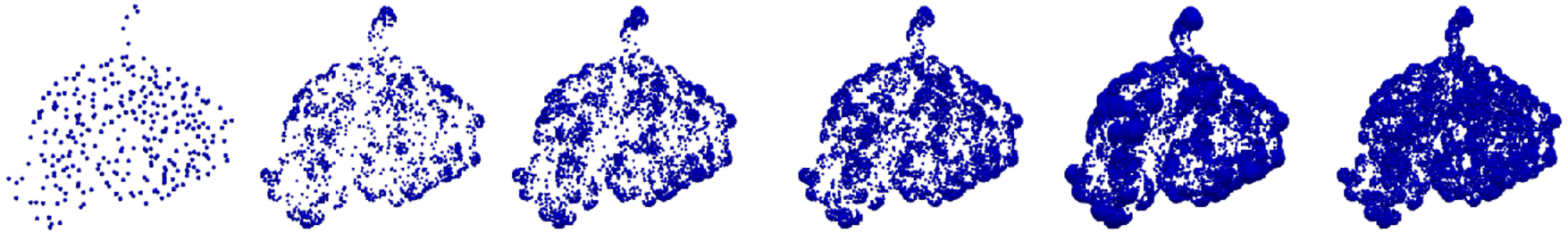
- This work is impossible without their resources and suggestions.

Motivation

- Protein-Protein Interactions (PPIs) are crucial in regulating cellular functions.
- Pressing need for strategies aimed at **designing new protein binders**.
- Help understand underlying mechanisms of protein interactions.
- Help advance therapeutic interventions.

Motivation

- Proteins can be modeled as graphs, meshes, point clouds.

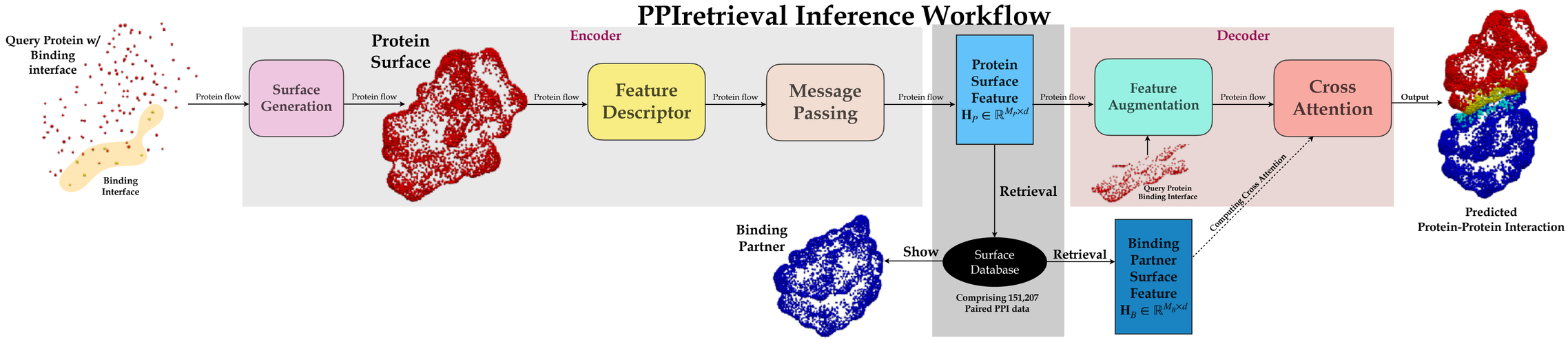


Protein Residue Point Cloud Representation of Protein as different scales

- Geometric deep learning for geometric and chemical mechanisms governing PPIs.

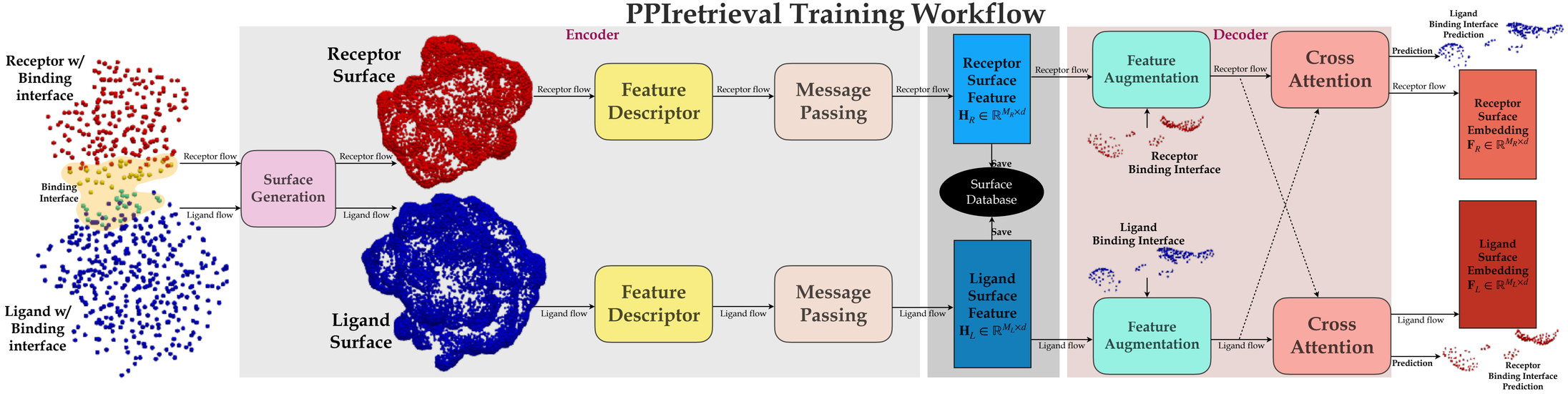
- Geometric deep learning for modeling PPIs.

Retrieval (Inference) Stage



- (1) PPIretrieval takes an unseen protein P with its corresponding binding site.
- (2) Encoder network generates surface representation $\mathbf{H}_P$.
- (3) PPIretrieval identifies a binding partner B with surface $\mathbf{H}_B$ from our database.
- (4) Decoder network takes $\mathbf{H}_P$, $\mathbf{H}_B$ and predictes their binding interface.

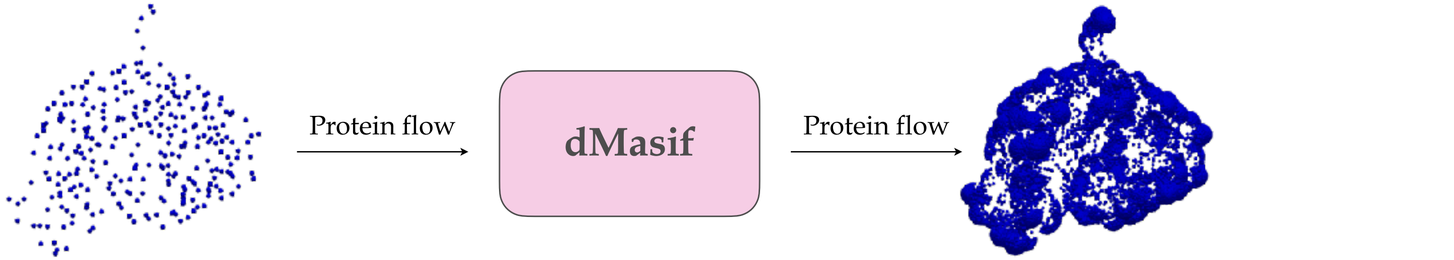
Training Stage



- (1) PPIretrieval takes a paired proteins R, L with their binding interface.
- (2) Encoder network generates surface representations $\mathbf{H}_R, \mathbf{H}_L$.
- (3) These surface representations are stored in our database.
- (4) Decoder network takes $\mathbf{H}_R, \mathbf{H}_L$ and learns their interactions.
- (5) PPIretrieval is optimized to learn the ***'lock-and-key'*** structure between R, L .

PPIretrieval Encoder

- **Encoding Stage**
  - (1) Surface Generation



Input Protein Residue

Protein Surface Representation

    - Define heat operator Δ_P on P .
    - Calculate first k eigenfunctions of heat operator Φ_P with eigenvalues $\{\lambda_i\}_{i=1}^{k}$.
    - Calculate Moore-Penrose pseudo-inverse Φ_P^+ .
  - (2) Geometric Descriptor
    - Calculate Mean curvature $\mathbf{F}_{\text{Mean}}$ on P .
    - Calculate Gaussian curvature $\mathbf{F}_{\text{Gauss}}$ on P .
    - Calculate Heat Kernel Signatures $\mathbf{F}_{\text{HKS}}$ on P .
    - Transform geometric features $\mathbf{F}_{\text{Geom}} \leftarrow \text{MLP}([\mathbf{F}_{\text{Mean}}, \mathbf{F}_{\text{Gauss}}, \mathbf{F}_{\text{HKS}}])$.

PPIretrieval Encoder

- **Encoding Stage**
  - (3) Chemical Descriptor
    - Compute residue-level chemical features:
      - Compute chemical features $\mathbf{F}_{\text{Res}} \leftarrow \text{MLP}([P, Y_P^{\text{res}}])$.
      - Augment chemical features $\mathbf{F}_{\text{Res}} \leftarrow \text{EGNN}([\mathbf{F}_{\text{Res}}, v_P])$.
    - For every every surface point x_i, compute surface-level chemical features from $\mathbf{F}_{\text{Res}}$:
      - Find k nearest neighboring residues $\{r_1^i, \ldots, r_k^i\}$ with features $\{f_{\text{Res}}^{i,1}, \ldots, f_{\text{Res}}^{i,k}\}$
      - Compute chemical features $f_{\text{Chem}}^i \leftarrow \text{MLP}(\sum_{j=1}^{k} f_{\cos}(\|x_i - r_j^i\|) \cdot \text{MLP}([f_{\text{Res}}^{i,j}, 1/\|x_i - r_j^i\|]))$
    - Compute surface features $\mathbf{F}_{\text{Surf}} \leftarrow \text{MLP}([\mathbf{F}_{\text{Geom}}, \mathbf{F}_{\text{Chem}}])$.
  - (4) Message Passing
    - Compute aggregated surface features $\mathbf{H}_P$ via heat diffusion defined by Δ_P, Φ_P, Φ_P^+ (details see paper).
  - During Training
    - Encoder network generates $\mathbf{H}_R, \mathbf{H}_L$ for input paired receptor and ligand proteins.

PPIretrieval Decoder

- **Decoding Stage**
  - Aiming to predict ligand's binding site conditioned on receptor's binding site
    - (1) Interaction Block
      - Augment receptor surface features $\mathbf{H}_R \leftarrow \text{EGNN}([\text{MLP}(\mathbf{H}_R, \mathbf{Y}_R^{\text{surf}}), x_R])$.
      - Augment ligand surface features $\mathbf{H}_L \leftarrow \text{EGNN}([\text{MLP}(\mathbf{H}_L), x_L])$.
      - Compute cross-attention $\mathbf{F}_L \leftarrow \text{softmax}\left(\frac{(\mathbf{H}_L W_Q)(\mathbf{H}_R W_K)^T}{\sqrt{d}}\right)(\mathbf{H}_R W_V)$.
    - (2) Binding Site Prediction
      - Predict ligand's binding site $\hat{\mathbf{Y}}_L^{\text{surf}} \leftarrow \sigma(\text{MLP}(\mathbf{F}_L))$
    - For every every resiue i, compute residue-level binding site from $\hat{\mathbf{Y}}_L^{\text{surf}}$:
      - Find k nearest neighboring residues $\{\hat{\mathbf{y}}_1^{\text{surf}}, \ldots, \hat{\mathbf{y}}_k^{\text{surf}}\}$ with features $\{\mathbf{f}_1, \ldots, \mathbf{f}_k\}$
      - Compute chemical features $\hat{\mathbf{f}}_i \leftarrow \text{Mean}(\sum_{j=1}^{k} \mathbf{f}_j)$
      - Compute binding site $\hat{\mathbf{y}}_i^{\text{res}} \leftarrow \text{Mean}(\sum_{j=1}^{k} \hat{\mathbf{y}}_j^{\text{surf}})$
  - During Training
    - Each PPI sample is treated as two training instances.

PPIretrieval Optimization

- Utilize the ***'lock-and-key'*** structure between the receptor and ligand.

- Achieve by three training objectives:
  - (1) Lock-and-Key Optimization
    - Assemble the Jigsaw puzzles of receptor and ligand.

  - (2) Contrastive Optimization
    - Bring residues of the binding interface closer.
    - Push residue that do not belong to the binding interface farther apart.

  - (3) Binding Interface Optimization
    - Optimize the predictions of binding interfaces.

Lock-and-Key Optimization

- Ground-truth matching matrix $\mathbf{X} \in \{0,1\}^{N_R \times N_L}$, $\mathbf{x}_{ij} = 1$ if $d_{ij} \leq d_{\text{cut}}$.
- Construct a soft-matching matrix
  - $\hat{\mathbf{X}} = \text{sinkhorn}(\exp(\hat{\mathbf{F}}_R^T W \hat{\mathbf{F}}_L / \tau)) \in [0,1]^{N_R \times N_L}$.
  - $\hat{\mathbf{X}}$ is a doublystochastic matrix, $\hat{\mathbf{X}}_{ij}$ measures soft-matching score between residues i, j in receptor and ligand.

- $\mathscr{L}_{\text{match}} = \text{cross-entropy}(\mathbf{X}, \hat{\mathbf{X}})$.
- $\mathscr{L}_{\text{match}}$ serves the dual purpose:
  - Encourage alignment between soft-matching scores and the ground truth.
  - Ensure residues are matched with its complementary part in the opposite protein.

Contrastive Optimization

- Bring residues of the binding interface closer.
- Push residue that do not belong to the binding interface farther apart.

- $$\mathscr{L}_{\text{contra}} = -\sum_{i \in \mathbf{Y}_R^{\text{res}}} \sum_{j \in \mathbf{Y}_L^{\text{res}}} \log \frac{\exp(\hat{\mathbf{f}}_R^i, \hat{\mathbf{f}}_L^j / \tau)}{\sum_{k \in \mathbf{Y}_L^{\text{res}}} \exp(\hat{\mathbf{f}}_R^i, \hat{\mathbf{f}}_L^k / \tau)}$$

- $\mathscr{L}_{\text{contra}}$ minimizes the distance between corresponding residues and maximizes the distance between non-corresponding residues.

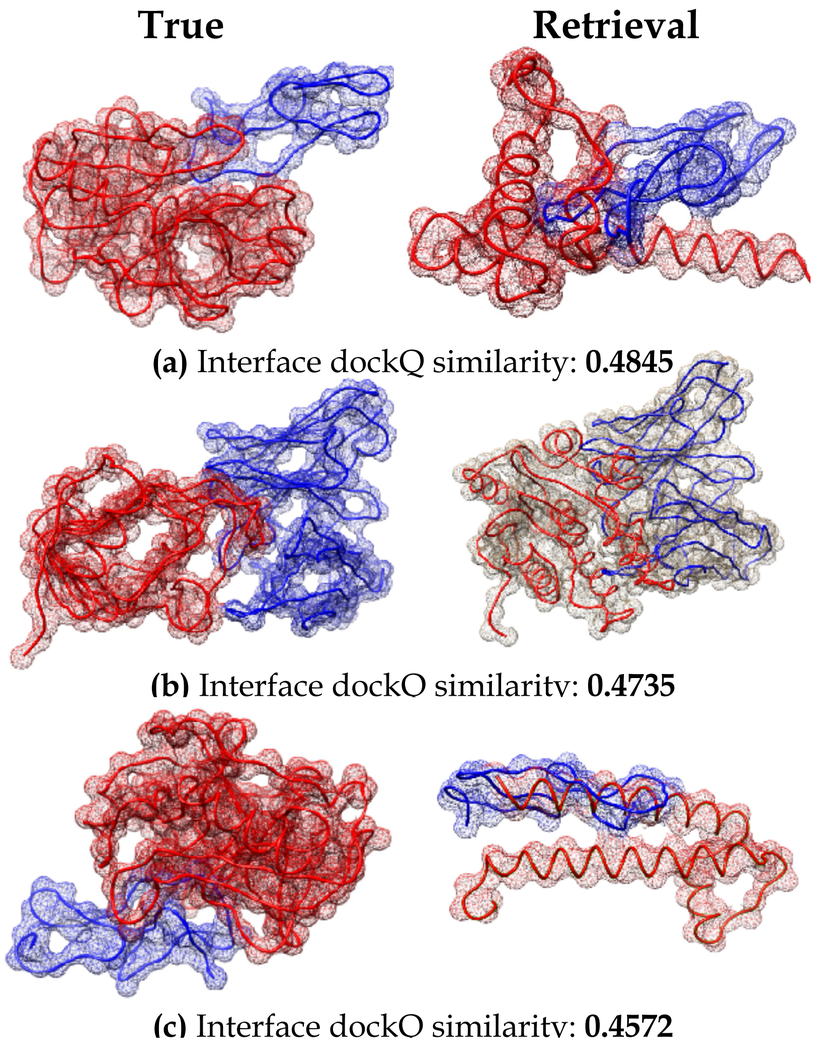
Binding Interface Optimization

- Directly optimize the predictions of binding interfaces.
- $\mathscr{L}_{\text{bind}} = \text{cross-entropy}(\mathbf{Y}_R^{\text{res}}, \hat{\mathbf{Y}}_R^{\text{res}}) + \text{cross-entropy}(\mathbf{Y}_L^{\text{res}}, \hat{\mathbf{Y}}_L^{\text{res}})$.

- $\mathscr{L} = \mathscr{L}_{\text{match}} + \mathscr{L}_{\text{contra}} + \mathscr{L}_{\text{bind}}$

- Overall optimization strategy is to leverage the ***'lock-and-key'*** structure inherent in PPI complexes.

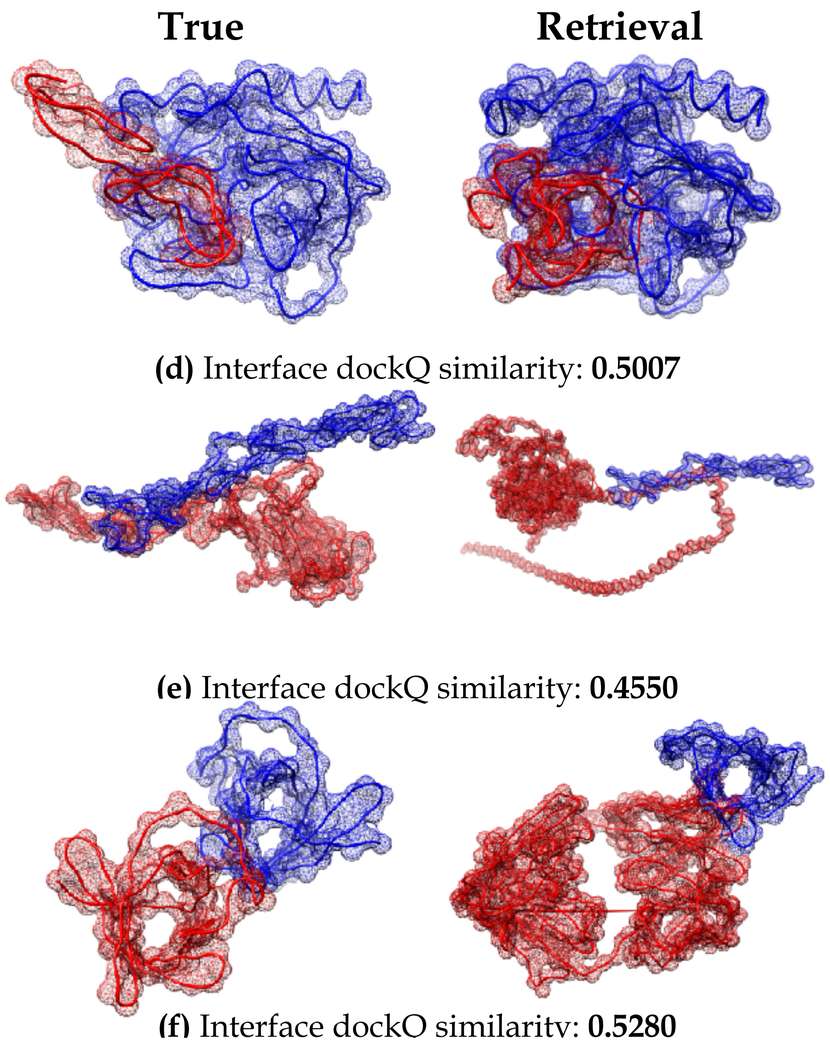
Retrieval Visualization



(a) Interface dockQ similarity: **0.4845**

(b) Interface dockQ similaritv: **0.4735**

(c) Interface dockQ similaritv: **0.4572**



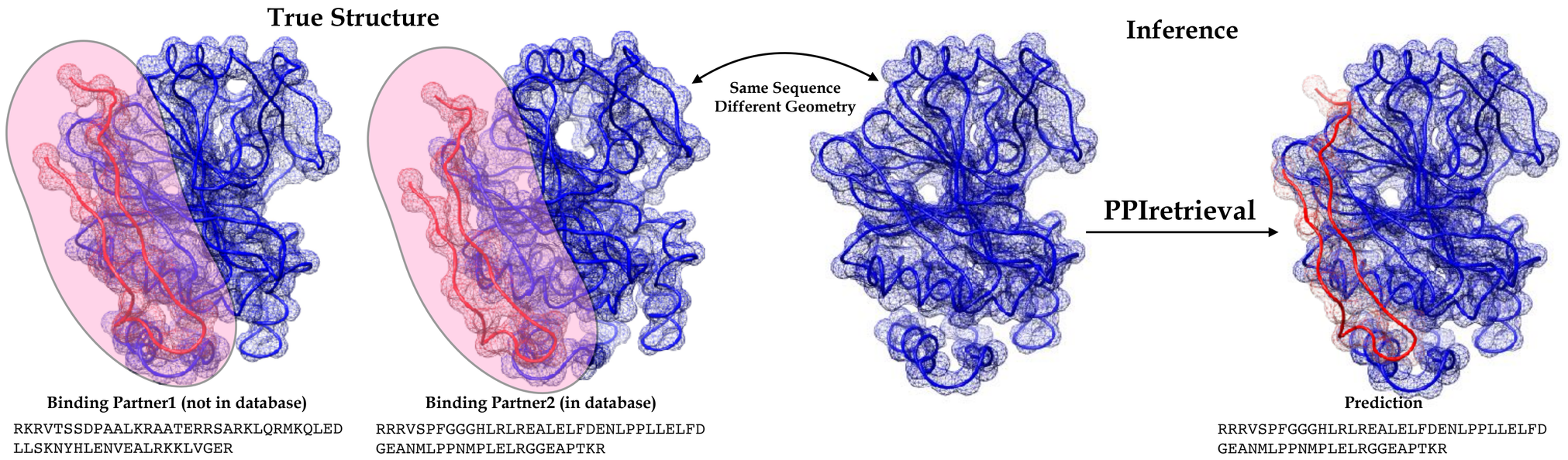
(d) Interface dockQ similarity: **0.5007**

(e) Interface dockQ similarity: **0.4550**

(f) Interface dockQ similaritv: **0.5280**

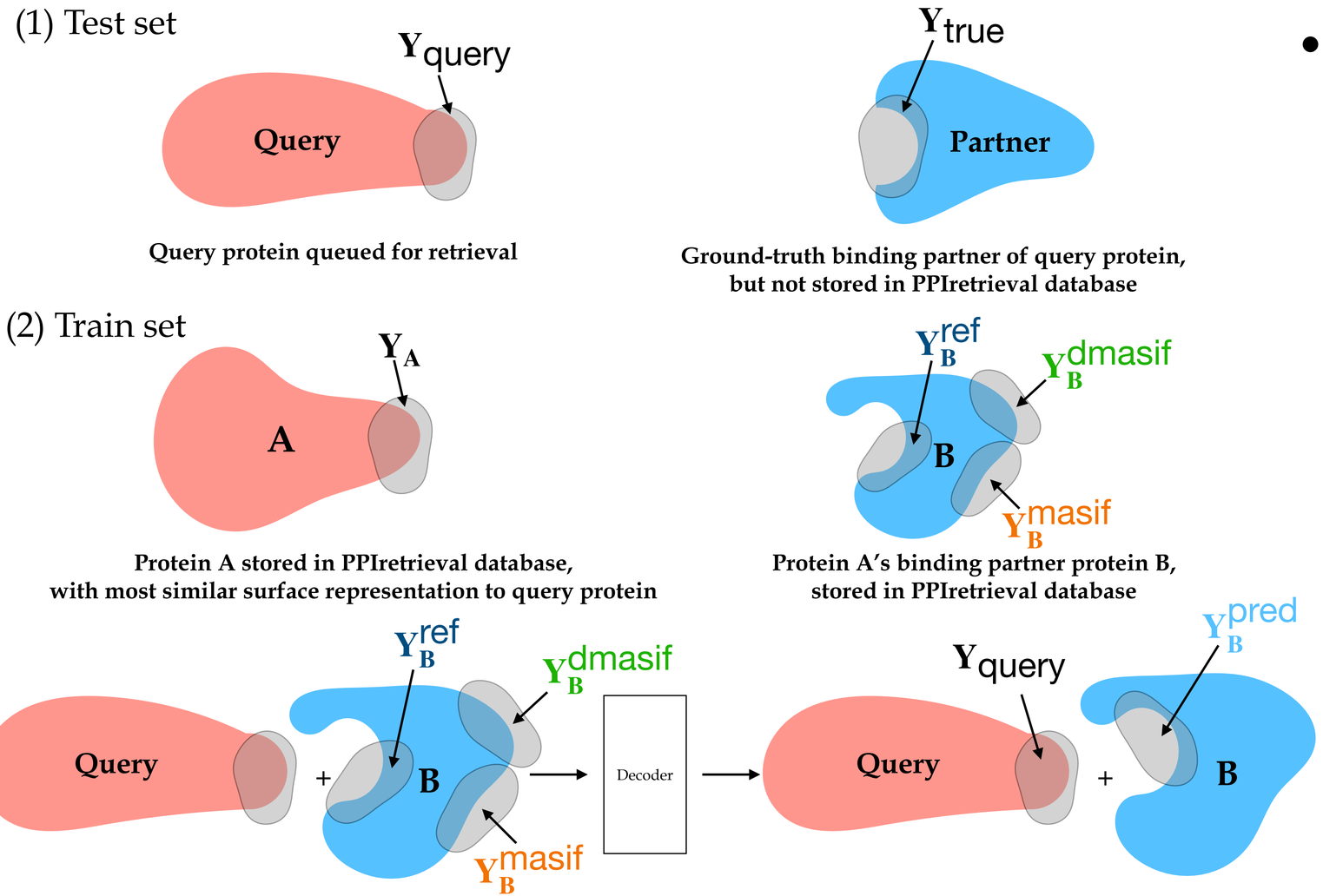
- Visualization of PPIretrieval results for proteins in the PDB test set, evaluated by dockQ. Proteins colored in blue are input query proteins; proteins colored in red are binding partners. Left column displays the ground-truth structures; right column shows the structures predicted by PPIretrieval.

Case Study



- The query protein has two binding partners: one stored in our surface database (pdb id: 5J28), while the other not (pdb id: 1DGC).
- Although the query protein in the two ground-truth structures shares the same sequence representation, there are slightlt different in their geometry.
- PPIretrieval identifies the protein in the database that most closely matches in both sequential and geometric representation

Evaluation



- For a PPI in the test set, a query protein with a known binding site, we compare between

$\text{dockQ}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}})$, $\text{dockQ}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}})$, $\text{dockQ}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}})$, $\text{dockQ}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}})$, $\text{TM}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}})$, $\text{TM}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}})$, $\text{TM}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}})$, $\text{TM}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}})$, $\text{rmsd}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}})$, $\text{rmsd}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}})$, $\text{rmsd}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}})$, $\text{rmsd}(\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}})$.

Emperical Results

Dataset	Metrics		PDB	DIPS	PPBS
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4073	0.4177	0.5535
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4220	0.4304	0.5946
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.1334	0.1021	0.1228
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.1155	0.0837	0.1036
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.2134	0.6617	0.4622
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.2366	0.6649	0.4735
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0773	0.0981	0.0911
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0665	0.0831	0.0871
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	11.40	11.33	8.20
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	10.44	6.02	9.77
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	15.73	19.66	17.32
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	17.87	23.55	19.65

Table 1. dockQ, TM, and *rmsd* for evaluation of **Top1 hit** binding sites predicted by PPIretrieval in comparison with other binding sites over three runs. The database for each test set comprises surface representations from the training and validation sets of each respective dataset.

Dataset	Metrics		PDB	DIPS	PPBS
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4250	0.4367	0.6014
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4678	0.4410	0.6045
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.1345	0.1026	0.1249
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.1235	0.0998	0.1261
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.3222	0.6914	0.6014
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.3300	0.6944	0.6045
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0833	0.1002	0.1004
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0823	0.0911	0.1144
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	9.30	9.65	9.96
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	10.70	5.67	6.52
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	15.56	19.05	16.82
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	16.81	21.22	16.08

Table 2. dockQ, TM, and *rmsd* for evaluation of **Top1 hit** binding sites predicted by PPIretrieval in comparison with other binding sites over three runs. The database comprises all surface representations from the training and validation sets of PDB, DIPS, and PPBS datasets.

Cross-Dataset Validation

- We take the model trained on PDB training set only to encode the PPIs in DIPS and PPBS training and validation sets, respectively.

Dataset	Metrics		PDB	DIPS	PPBS
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4073	0.4177	0.5535
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4220	0.4304	0.5946
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.1334	0.1021	0.1228
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.1155	0.0837	0.1036
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.2134	0.6617	0.4622
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.2366	0.6649	0.4735
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0773	0.0981	0.0911
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0665	0.0831	0.0871
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	11.40	11.33	8.20
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	10.44	6.02	9.77
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	15.73	19.66	17.32
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	17.87	23.55	19.65

Table 1. dockQ, TM, and *rmsd* for evaluation of **Top1 hit** binding sites predicted by PPIretrieval in comparison with other binding sites over three runs. The database for each test set comprises surface representations from the training and validation sets of each respective dataset.

Dataset	Metrics		DIPS-Top1	DIPS-Top10	PPBS-Top1	PPBS-Top10
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4030	0.4156	0.5231	0.5611
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4207	0.4435	0.5579	0.5857
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0515	0.0523	0.0621	0.0633
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0434	0.0515	0.0601	0.0633
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.5330	0.6714	0.4202	0.3725
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.5419	0.6792	0.4421	0.3889
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0499	0.0511	0.0611	0.0620
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0433	0.0491	0.0519	0.0602
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	11.12	7.35	8.92	8.91
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	5.84	10.50	11.76	10.49
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	20.73	19.21	19.81	19.55
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	23.89	22.05	20.66.	20.04

Table 3. dockQ, TM, and *rmsd* for evaluation of **Top1, Top10 hit** binding sites predicted by PPIretrieval in comparison with other binding sites on cross-datasets over three runs. The database for each test set comprises surface representations from the training and validation sets of each respective dataset.

More Emperical Results

PDB Dataset	Metrics		Top1	Top10	Top20	Top50	Top100
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4073	0.4362	0.4379	0.4411	0.3507
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4220	0.4375	0.4459	0.4569	0.4688
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.1334	0.1338	0.1355	0.1401	0.1405
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.1155	0.1194	0.1212	0.1247	0.1301
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.2134	0.2078	0.2059	0.2059	0.2108
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.2366	0.2266	0.2241	0.2231	0.2265
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0773	0.0775	0.0774	0.0758	0.0702
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0665	0.0668	0.0679	0.0698	0.0701
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	11.40	9.74	9.59	9.50	9.34
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	10.44	9.33	8.94	8.52	8.16
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	15.98	15.88	15.84	15.76	15.53
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	17.87	17.31	17.03	16.55	16.02
Cost	*PPIretrieval runtime*(↓)	second/protein	0.29	0.91	1.97	4.64	9.44

Table 4. dockQ, TM, and *rmsd* for evaluation of **Top1, Top10, Top20, Top50, Top100 hit** binding sites predicted by PPIretrieval in comparison with other binding sites in the PDB test set over three runs. The database comprises surface features from training and validation sets from PDB dataset only.

PDB Dataset	Metrics		Top1	Top10	Top20	Top50	Top100
Site Quality	*dockQ*(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.4126	0.4331	0.4480	0.4491	0.4490
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4235	0.4402	0.4531	0.4649	0.4708
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.1433	0.1436	0.1455	0.1458	0.1478
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.1225	0.1266	0.1301	0.1398	0.1405
	TM(↑)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	0.3944	0.3877	0.3833	0.3554	0.3422
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	0.4041	0.3969	0.3863	0.3625	0.3528
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	0.0787	0.0766	0.0750	0.0721	0.0709
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	0.0536	0.0548	0.0588	0.0582	0.0601
	rmsd(↓)	$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{ref}}$	10.41	9.73	9.70	9.49	9.32
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{pred}}$	10.04	8.97	8.66	8.20	7.35
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{masif}}$	15.73	15.71	15.54	15.26	15.19
		$\mathbf{Y}_{\text{true}}, \mathbf{Y}_B^{\text{dmasif}}$	17.75	17.22	17.02	16.40	16.11

Table 5. dockQ, TM, and *rmsd* for evaluation of **Top1, Top10, Top20, Top50, Top100 hit** binding sites predicted by PPIretrieval in comparison with other binding sites in the PDB test set over three runs. The database comprises surface features from training and validation sets from PDB, DIPS, and PPBS dataset.

Future Direction

- Improving model size and training.
- Including more high-qiality PPI data.
- Designing interpretable embedding space for visualization, like foldseek.
- Establishing webserve for PPIretrieval, like foldseek.

Discussion

Will Hua 华辰清
Montreal, Canada

Personal Website

Paper

Code (private)

微信

Scan the QR code to add me as a friend.