BioBlobs: Unsupervised Discovery of Functional Substructures

Xin (Allen) Wang[1,2], Kaiwen Shi[1], Carlos Oliver[1]

[1]Vanderbilt University [2]Yale University

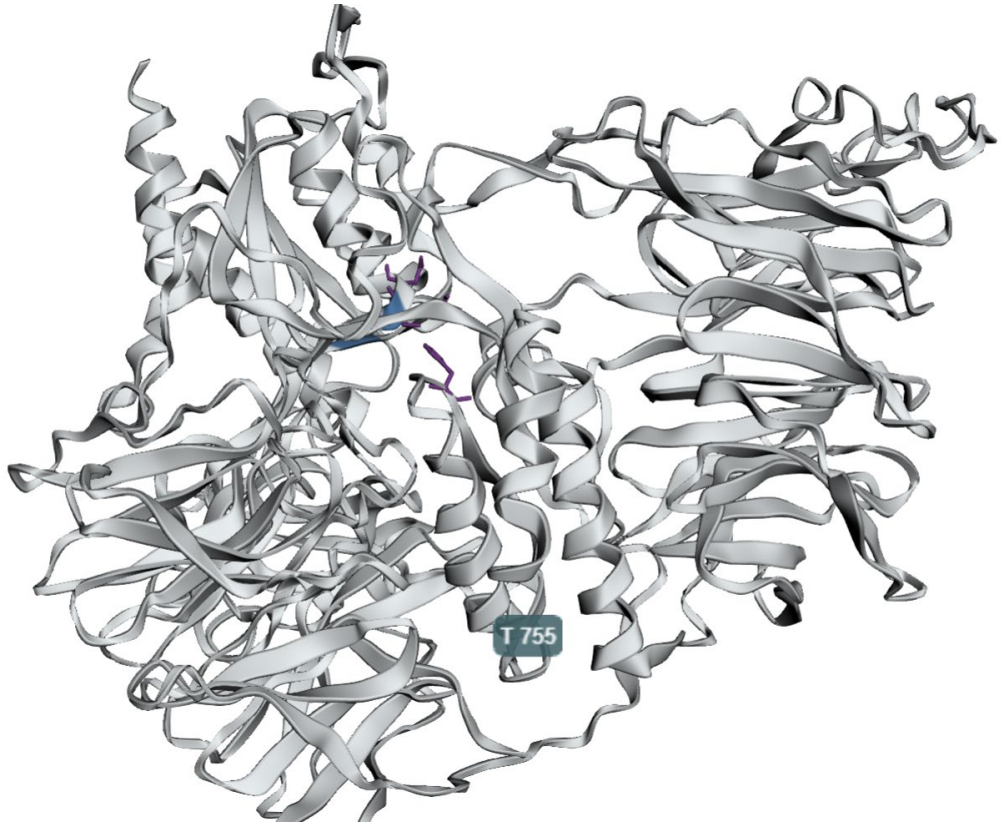
NeurIPS 2026



Protein function is often localized to cohesive substructures

- Functional substructures are cohesive, modular units, such as **catalytic triads**.

Tricorn protease
PDB 1K32, EC 3.4.21.-



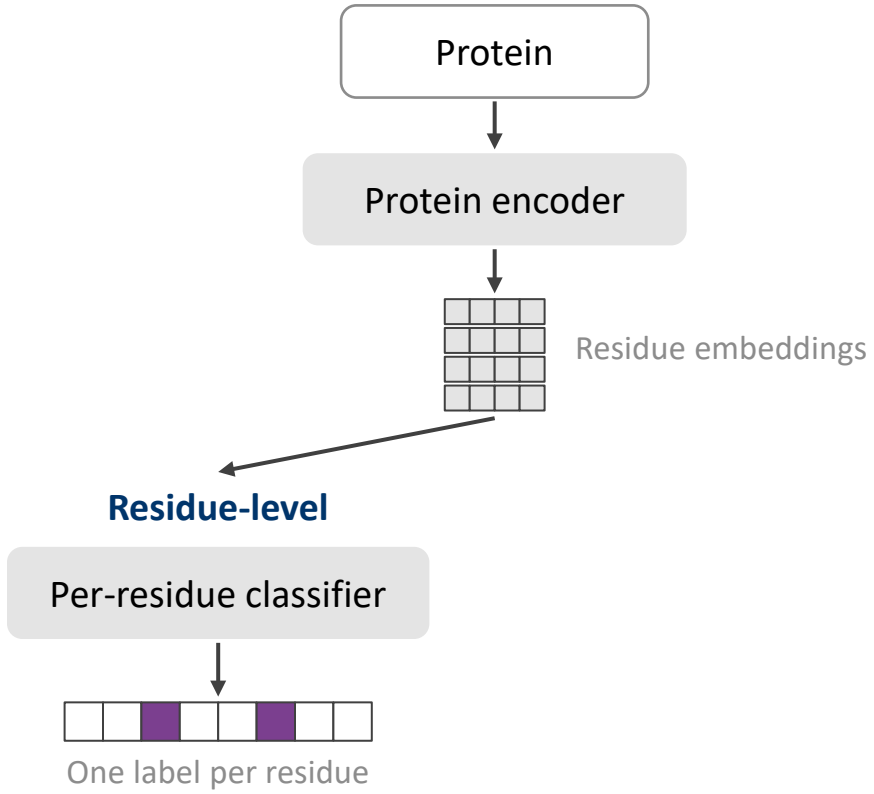
Catalytic site shown by purple sticks

Protein function is often localized to cohesive substructures

- Functional substructures are cohesive, modular units, such as **catalytic triads**.

Residue-level approaches

- Residue annotations are scarce and expensive to obtain.
- A single model predicts only one type of site.



Protein function is often localized to cohesive substructures

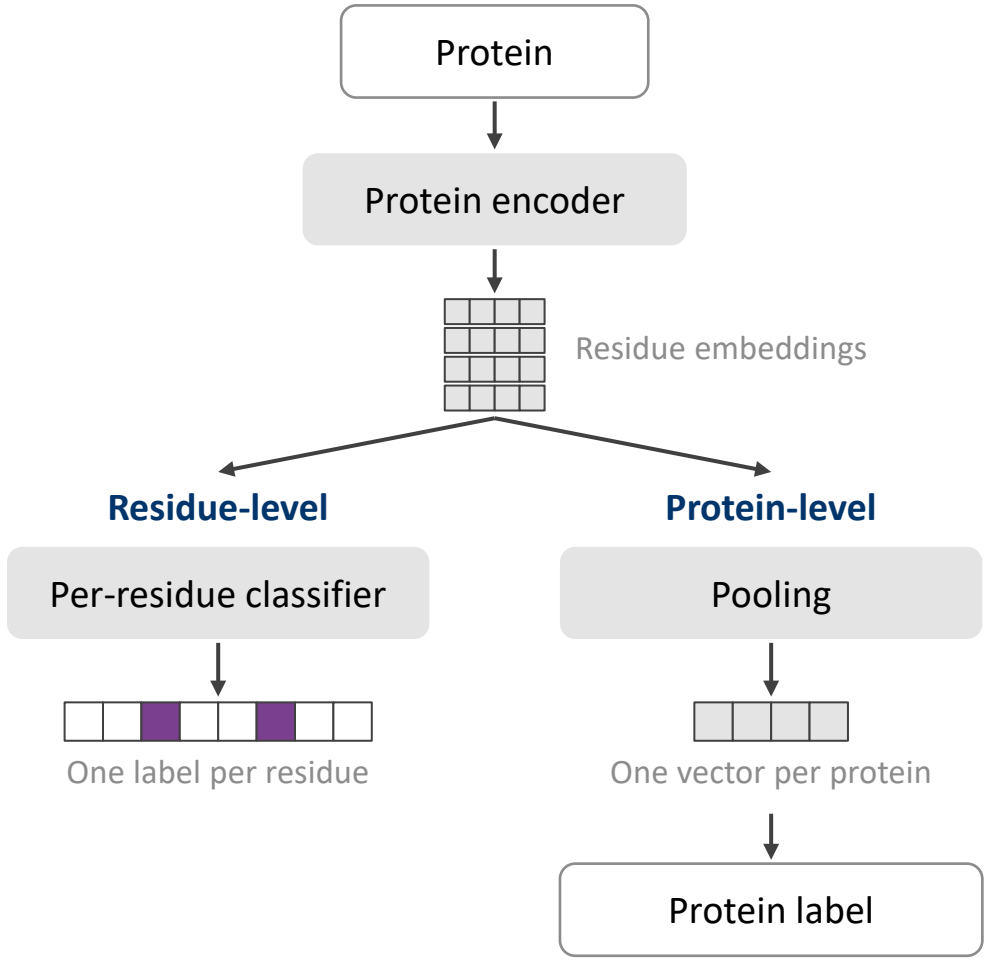
- Functional substructures are cohesive, modular units, such as **catalytic triads**.

Residue-level approaches

- Residue annotations are scarce and expensive to obtain.
- A single model predicts only one type of site.

Protein-level approaches

- Pooling discards local spatial organization.
- Small sites are diluted by non-functional residues.



Protein function is often localized to cohesive substructures

- Functional substructures are cohesive, modular units, such as **catalytic triads**.

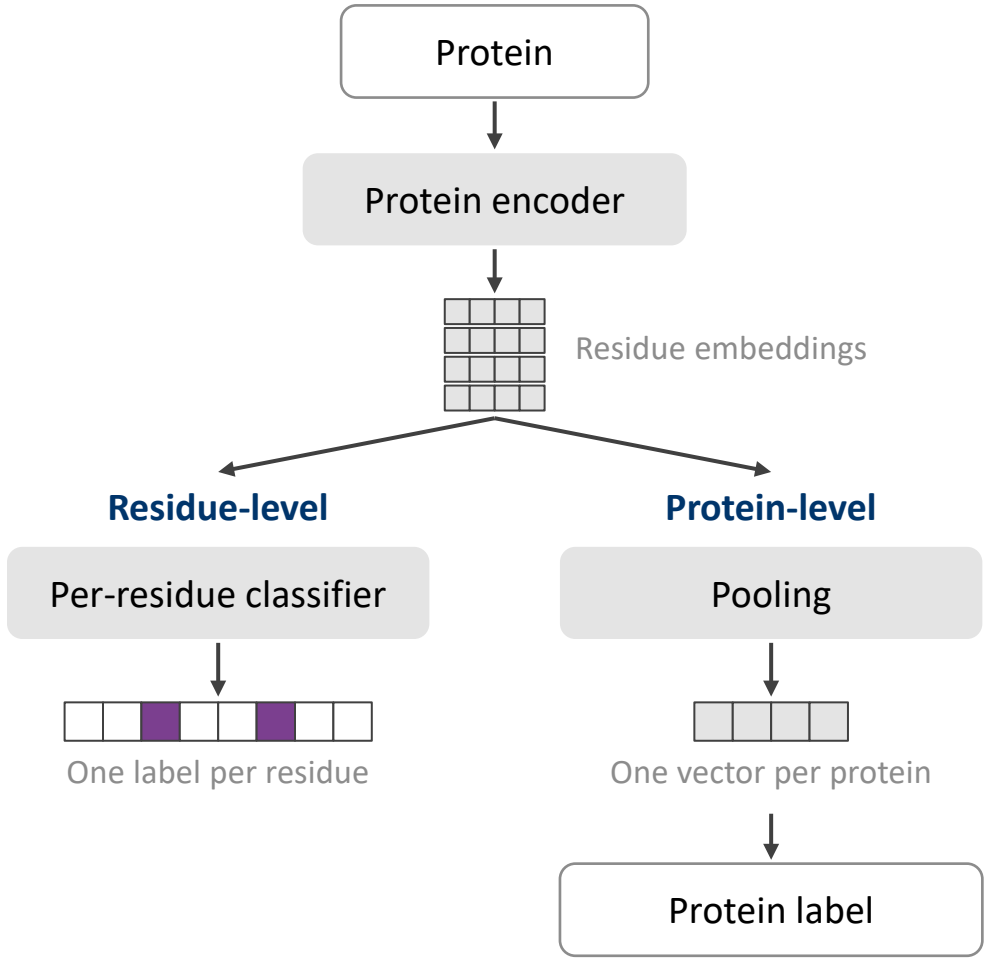
Residue-level approaches

- Residue annotations are scarce and expensive to obtain.
- A single model predicts only one type of site.

Protein-level approaches

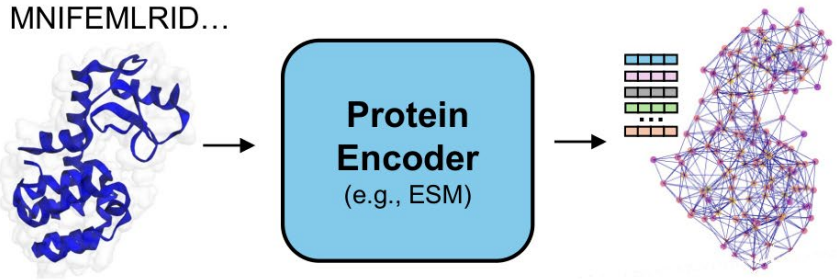
- Pooling discards local spatial organization.
- Small sites are diluted by non-functional residues.

Which substructure of a protein is responsible for its function?



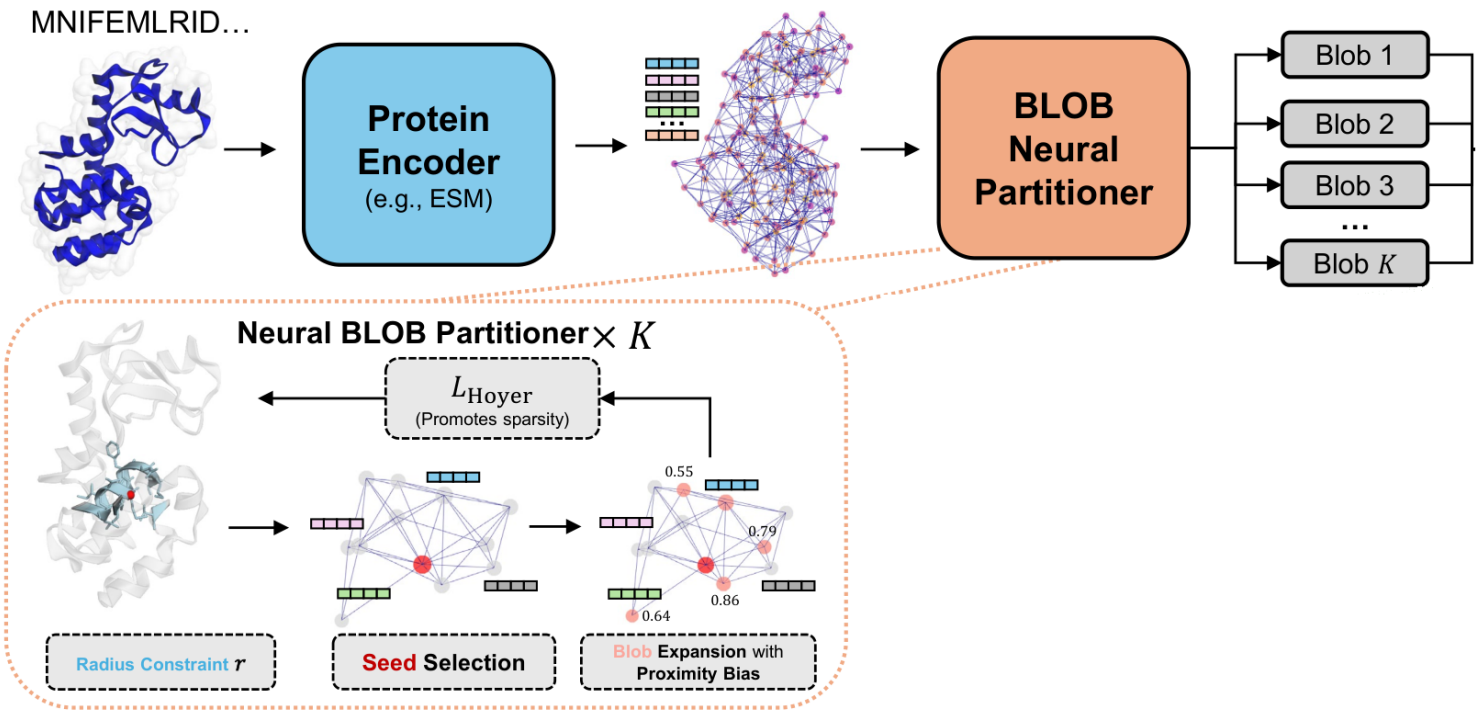
BioBlobs

Contribution: BioBlobs compresses a protein into a small set of cohesive substructures (blobs) and predicts function from these blobs alone.



BioBlobs

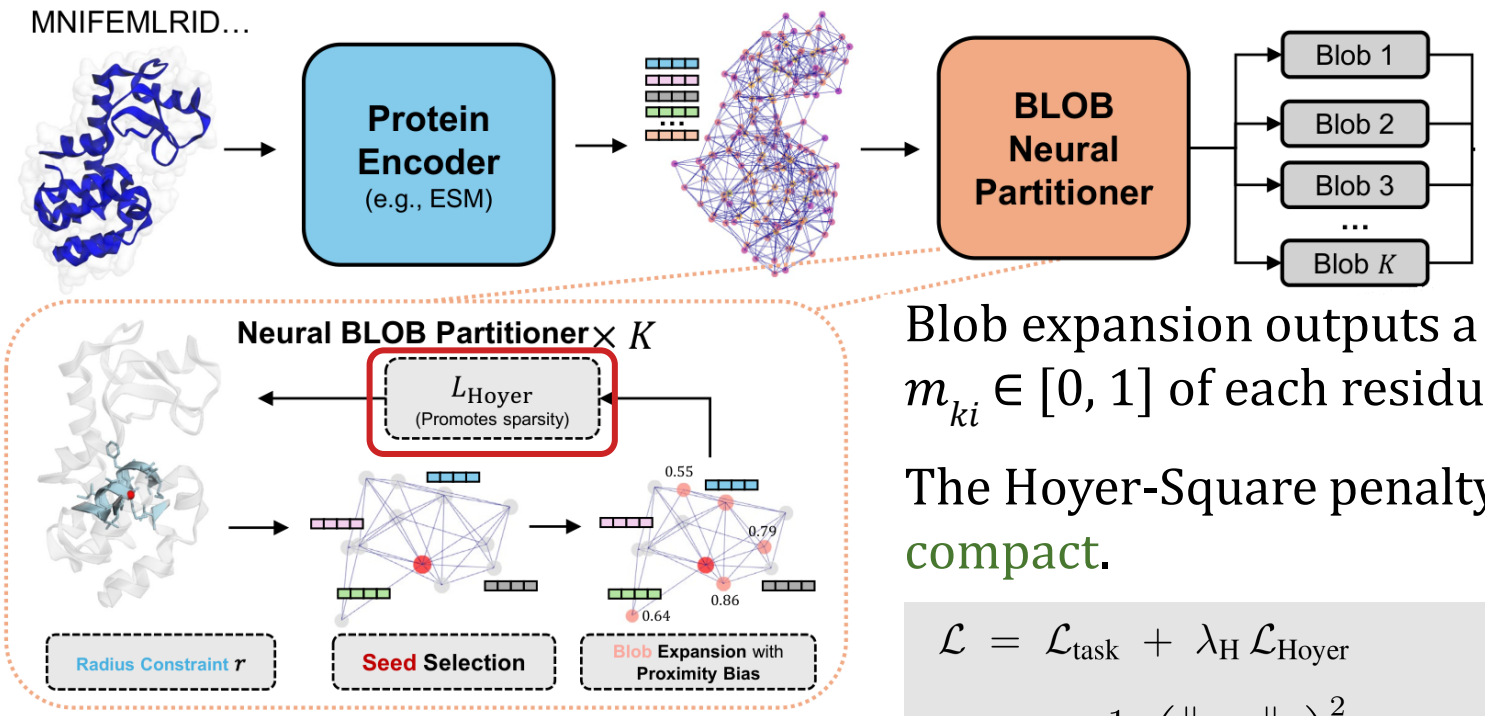
Contribution: BioBlobs compresses a protein into a small set of cohesive substructures (blobs) and predicts function from these blobs alone.



$$\mathcal{C}^{(k)} = \{i : \|\mathbf{x}_i - \mathbf{x}_{\text{seed}}^{(k)}\|_2 \leq r\}$$

BioBlobs

Contribution: BioBlobs compresses a protein into a small set of cohesive substructures (blobs) and predicts function from these blobs alone.



$$\mathcal{C}^{(k)} = \{i : \|\mathbf{x}_i - \mathbf{x}_{\text{seed}}^{(k)}\|_2 \leq r\}$$

Blob expansion outputs a soft membership $m_{ki} \in [0, 1]$ of each residue i in blob k.

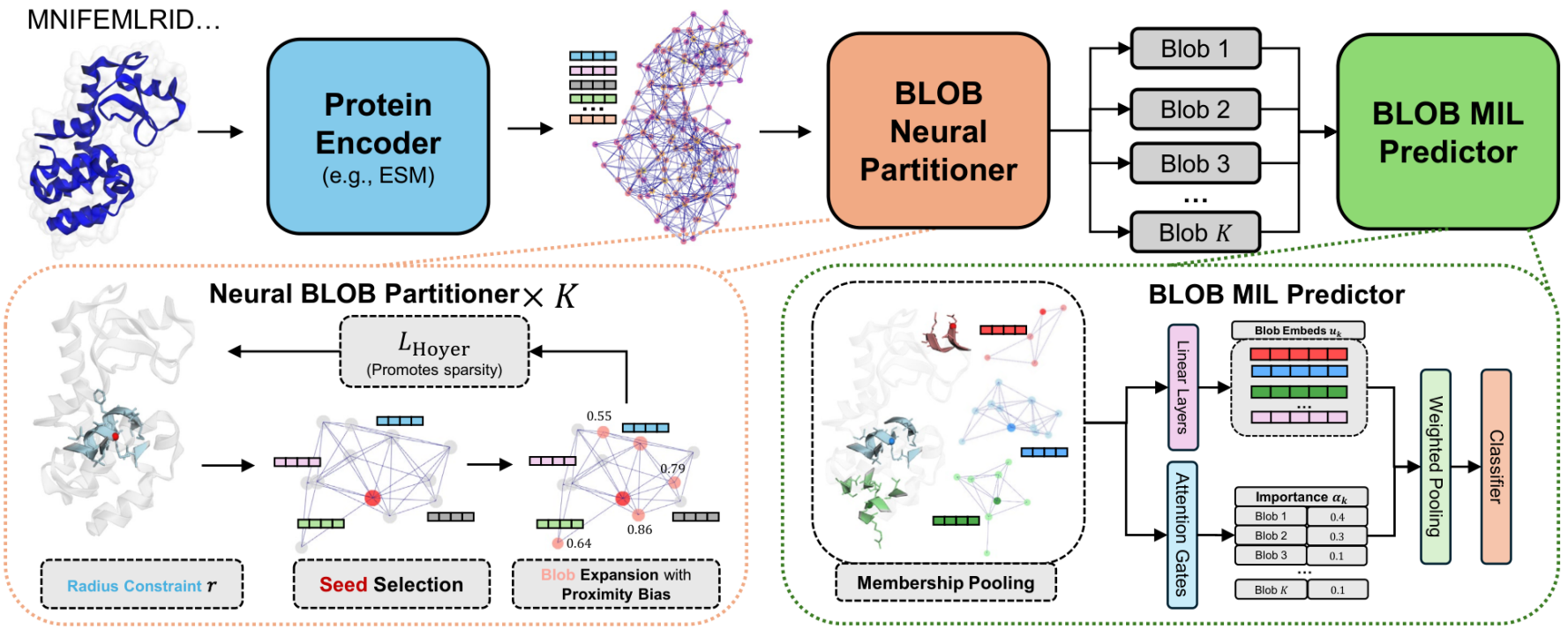
The Hoyer-Square penalty keeps the blobs compact.

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{H}} \mathcal{L}_{\text{Hoyer}}$$

$$\text{HS}_k = \frac{1}{n_k}\left(\frac{\|\mathbf{m}_k\|_1}{\|\mathbf{m}_k\|_2}\right)^2 \in \left[\frac{1}{n_k}, 1\right]$$

BioBlobs

Contribution: BioBlobs compresses a protein into a small set of cohesive substructures (blobs) and predicts function from these blobs alone.



$$\mathcal{C}^{(k)} = \{i : \|\mathbf{x}_i - \mathbf{x}_{\text{seed}}^{(k)}\|_2 \leq r\}$$

The MIL predictor weights the blobs by importance α_k and classifies the protein.

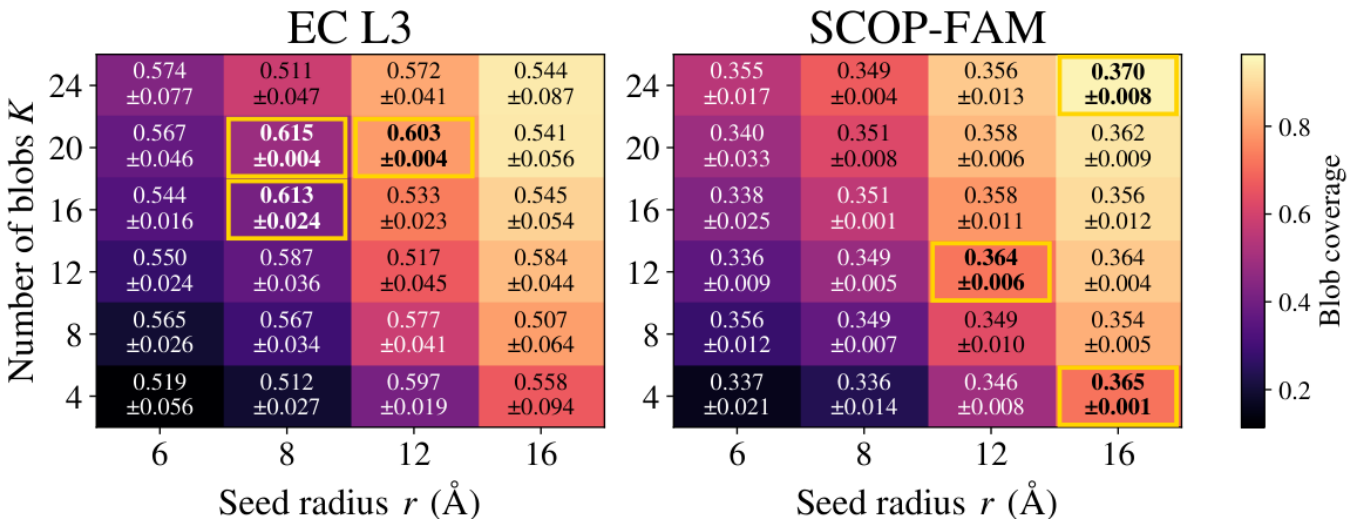
BioBlobs is competitive on protein function prediction

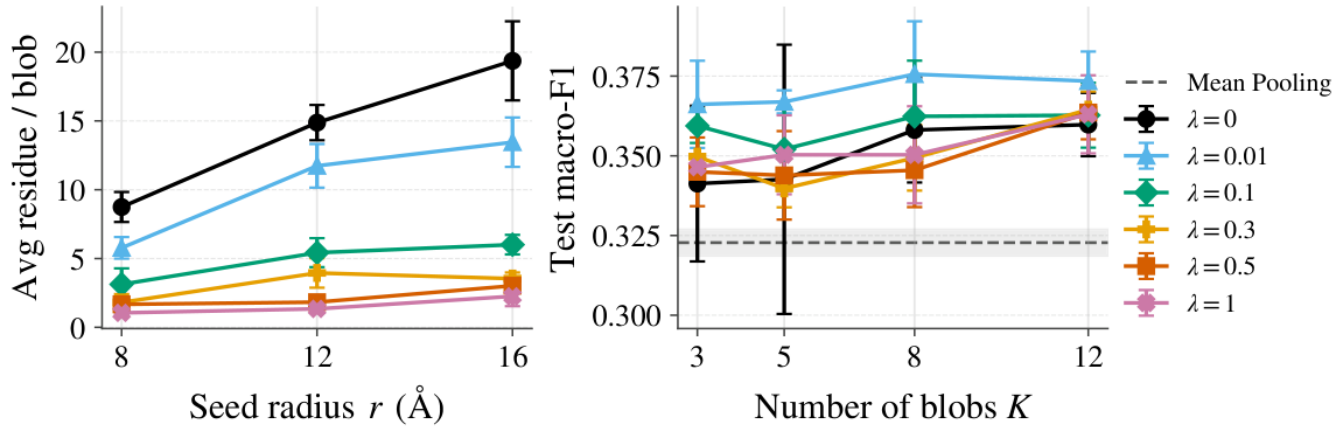
Table 1: **Protein function prediction on ProteinShake.** Bold marks the best mean for each encoder, split and dataset.

Encoder	Method	EC-L3	Pfam	GO-MF	SCOP-FAM	SCOP-SF
Sequence split						
ESM2-T30	Mean pooling	0.562 ± 0.009	0.381 ± 0.004	0.743 ± 0.002	0.340 ± 0.011	0.519 ± 0.012
	Attention pooling	0.594 ± 0.004	0.398 ± 0.015	**0.756 ± 0.005**	0.361 ± 0.005	**0.534 ± 0.012**
	GCN	0.502 ± 0.082	0.366 ± 0.016	0.746 ± 0.010	0.338 ± 0.008	0.495 ± 0.024
	GIN	0.525 ± 0.056	0.390 ± 0.030	0.744 ± 0.010	0.336 ± 0.007	0.509 ± 0.031
	GAT	0.532 ± 0.061	0.391 ± 0.013	0.749 ± 0.011	0.357 ± 0.018	0.524 ± 0.017
	DiffPool	0.572 ± 0.065	0.401 ± 0.006	0.755 ± 0.001	0.373 ± 0.017	0.512 ± 0.018
	TopK	0.563 ± 0.047	0.373 ± 0.014	0.751 ± 0.005	0.343 ± 0.027	0.500 ± 0.013
	SAGPool	0.550 ± 0.045	0.390 ± 0.006	0.754 ± 0.002	0.321 ± 0.005	0.490 ± 0.018
	BioBlobs	**0.615 ± 0.004**	**0.402 ± 0.015**	0.754 ± 0.003	**0.377 ± 0.003**	0.528 ± 0.003
SaProt-650M	Mean pooling	0.586 ± 0.018	0.362 ± 0.006	0.751 ± 0.004	0.353 ± 0.004	0.543 ± 0.022
	Attention pooling	**0.688 ± 0.050**	0.400 ± 0.013	**0.763 ± 0.004**	0.360 ± 0.008	0.565 ± 0.015
	GCN	0.587 ± 0.053	0.358 ± 0.001	0.747 ± 0.005	0.353 ± 0.012	0.543 ± 0.008
	GIN	0.544 ± 0.119	0.363 ± 0.034	0.746 ± 0.010	0.338 ± 0.031	0.518 ± 0.008
	GAT	0.531 ± 0.123	0.392 ± 0.014	0.759 ± 0.006	**0.373 ± 0.016**	**0.570 ± 0.001**
	DiffPool	0.546 ± 0.061	0.384 ± 0.009	0.762 ± 0.003	0.369 ± 0.009	0.565 ± 0.003
	TopK	0.509 ± 0.042	0.397 ± 0.003	0.751 ± 0.005	0.332 ± 0.005	0.542 ± 0.018
	SAGPool	0.579 ± 0.050	0.377 ± 0.023	0.748 ± 0.005	0.323 ± 0.021	0.506 ± 0.004
	BioBlobs	0.647 ± 0.047	**0.403 ± 0.012**	0.745 ± 0.003	0.371 ± 0.009	0.565 ± 0.010

- BioBlobs stays competitive across different tasks using only a subset of residues.
- Two splits: sequence and structure splits
- Two encoders: ESM2 and SaProt
- The blobs hold about **a quarter** of the residues or fewer.

BioBlobs adapts to the natural granularity of functional units

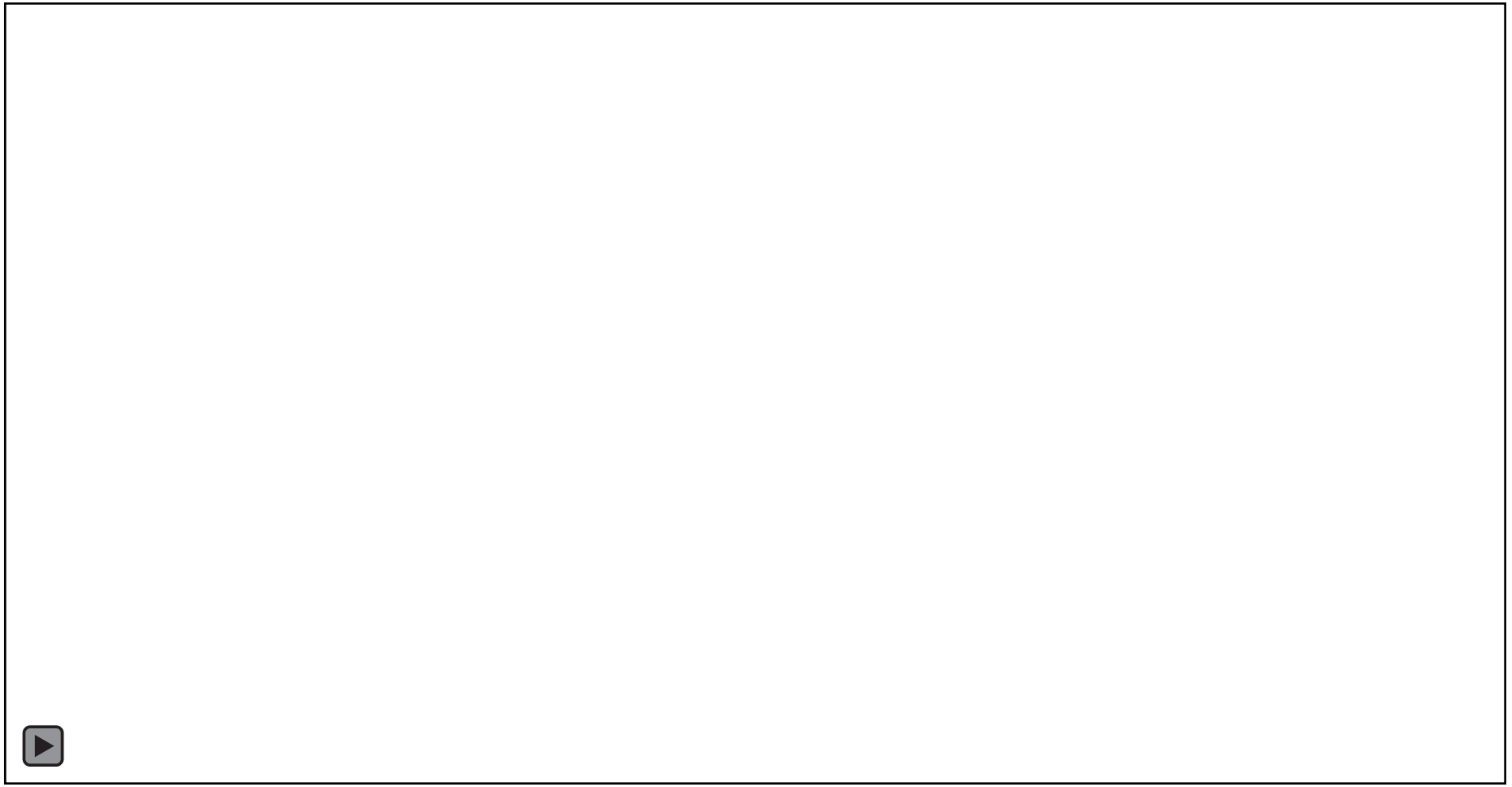




- Enzyme classes depend on local catalytic sites.
- Structural domains span most of a chain.

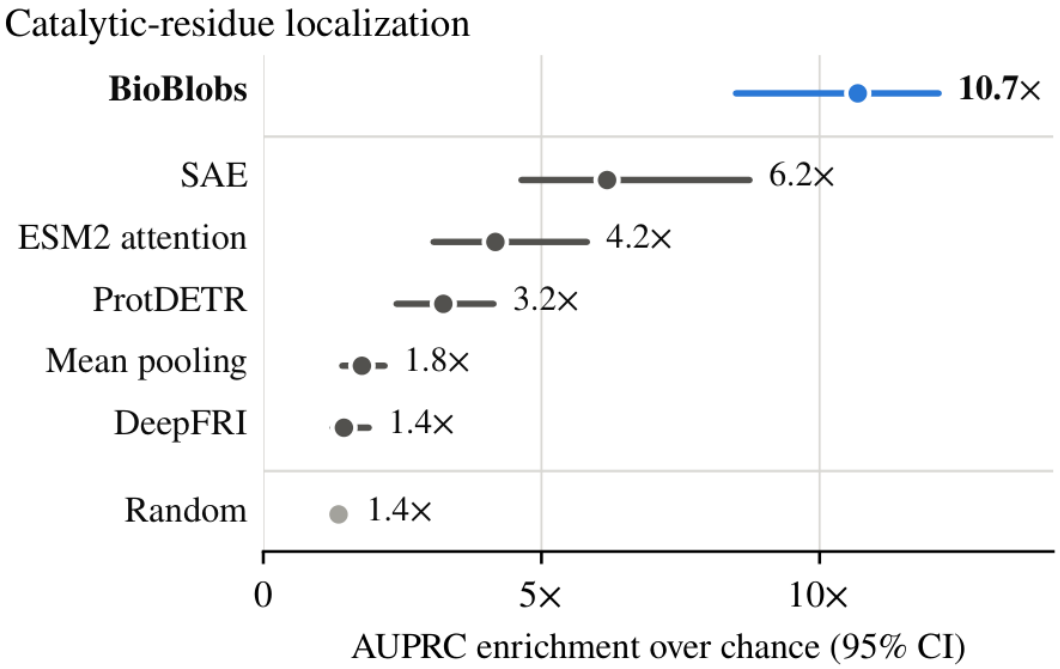
- Increasing λ_H compacts blobs by **an order of magnitude** at nearly no cost in accuracy.
- The prediction therefore needs **only a few residue embeddings** per blob.

BioBlobs recovers known functional sites without residue-level supervision



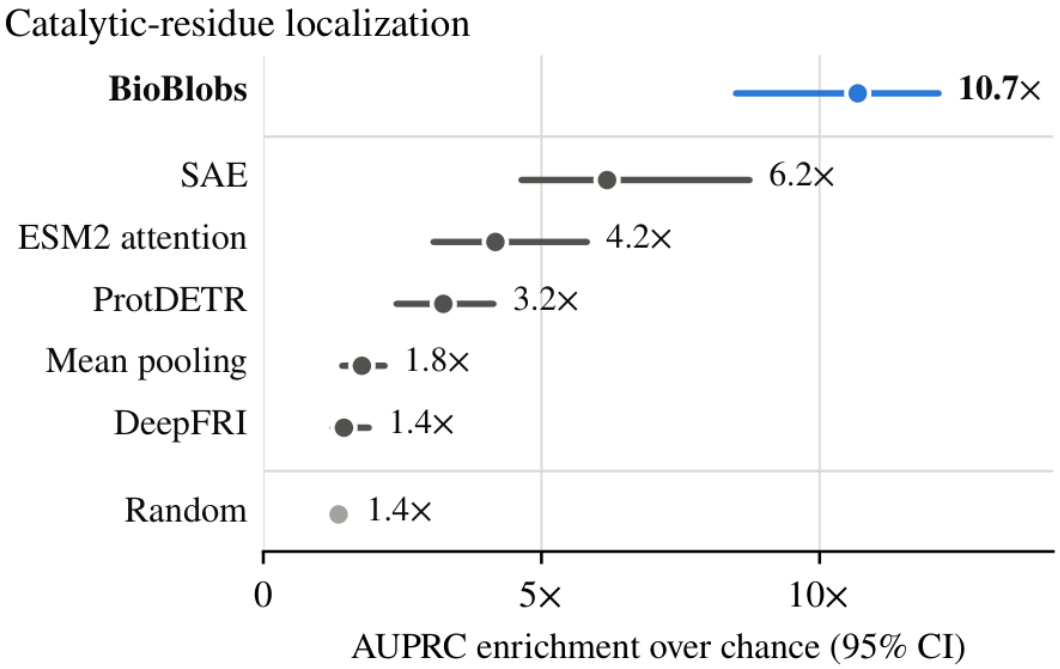
Anthrax lethal factor (PDB 1PWV): the 8 blobs hold 10 of 733 residues and cover 4 of the 5 M-CSA catalytic residues.

BioBlobs recovers known functional sites without residue-level supervision



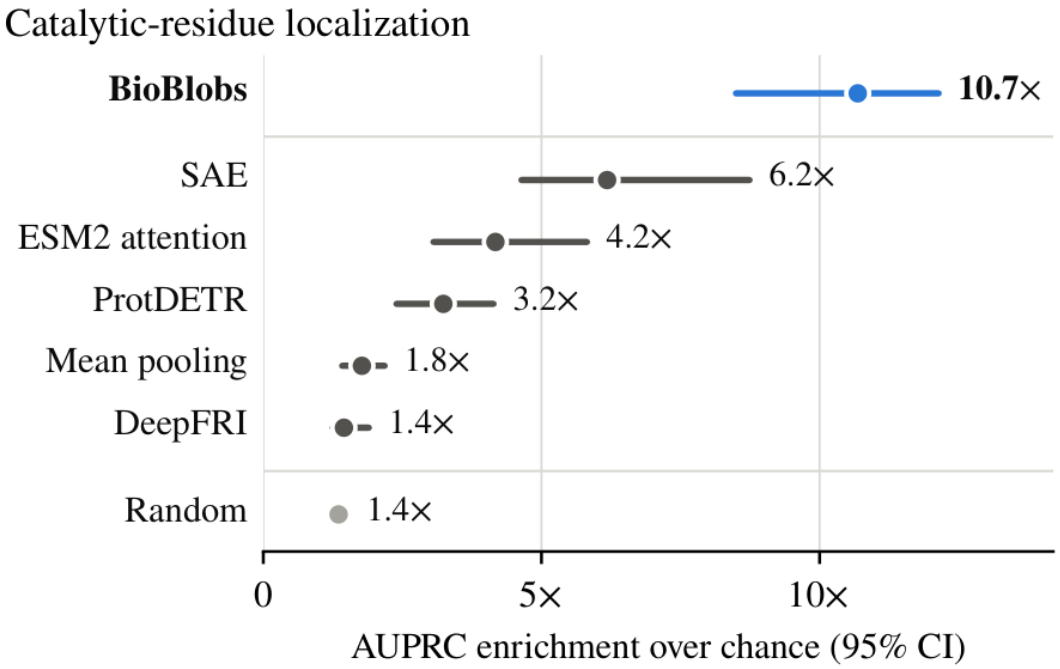
- The blobs alone localize catalytic residues better than **every baseline**.

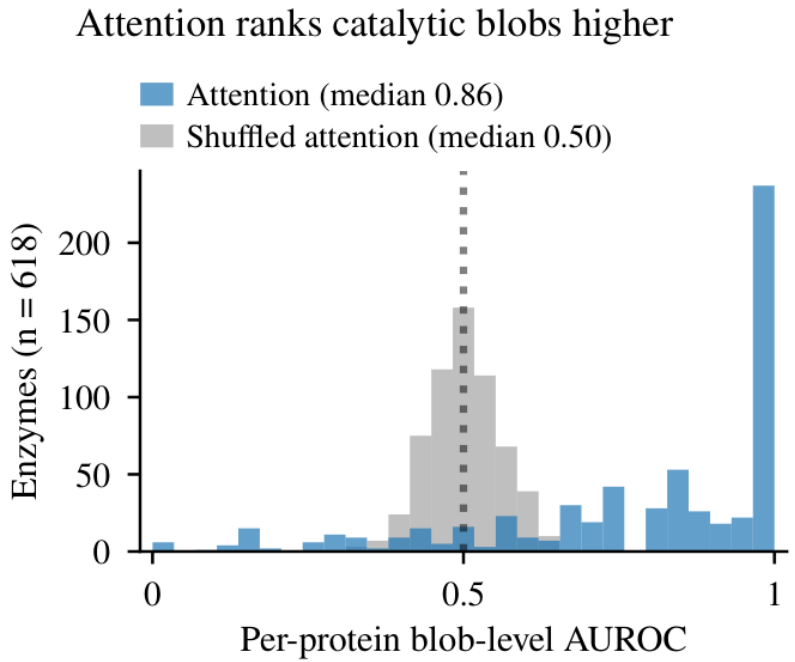
BioBlobs recovers known functional sites without residue-level supervision



- The blobs alone localize catalytic residues better than **every baseline**.
- They hold **4%** of the residues but **48%** of the catalytic residues.

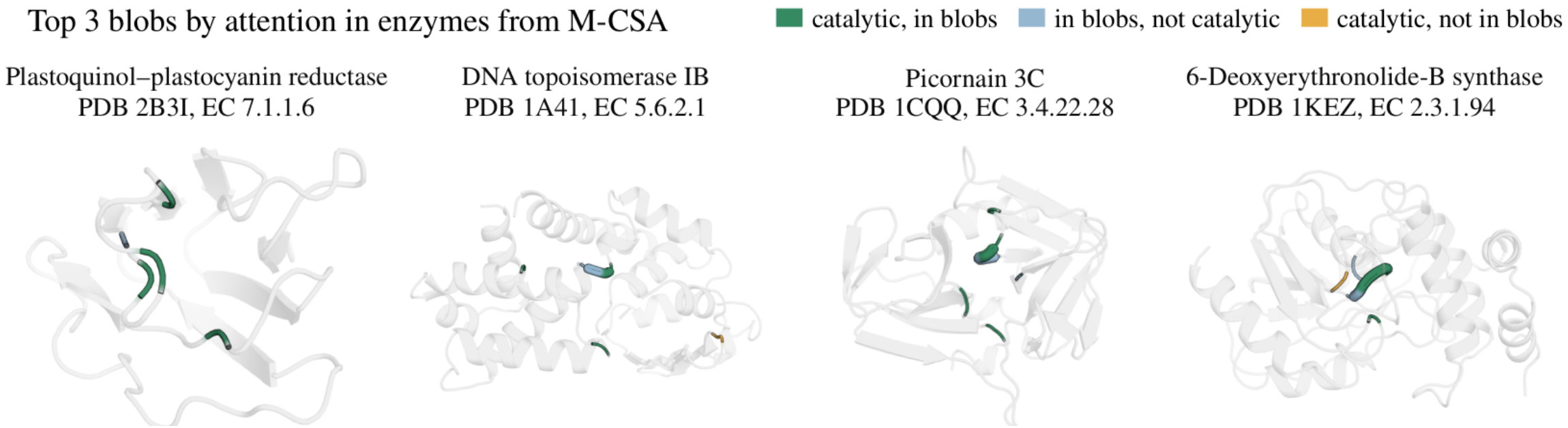
BioBlobs recovers known functional sites without residue-level supervision





- The blobs alone localize catalytic residues better than **every baseline**.
- They hold **4%** of the residues but **48%** of the catalytic residues.
- MIL attention picks out the blobs that contain **catalytic residues**.

BioBlobs recovers known functional sites without residue-level supervision



– In four held-out enzymes, the three highest-attention blobs cover most **catalytic residues**.

BioBlobs recovers known functional sites without residue-level supervision

Table 2: **Functional site localization on VenusX.** Mean over three seeds. Bold marks the best mean without seed alignment within each encoder.

	AUROC ↑				AUPRC ↑				F1@L ↑			
Method	Act	BindI	Evo	Motif	Act	BindI	Evo	Motif	Act	BindI	Evo	Motif
ESM2-T30												
Uniform random	0.500	0.499	0.498	0.500	0.050	0.057	0.078	0.109	0.000	0.049	0.068	0.109
Residue burial	0.795	0.676	0.670	0.577	0.121	0.103	0.120	0.127	0.121	0.105	0.111	0.082
Integrated gradients	0.656	0.806	0.659	**0.703**	0.083	0.168	0.166	**0.302**	0.086	0.174	0.173	**0.311**
Attention pooling	0.851	0.793	0.751	0.673	0.187	**0.246**	0.221	0.215	0.197	**0.257**	**0.236**	0.236
BioBlobs	**0.886**	**0.831**	**0.803**	0.662	**0.235**	0.207	**0.233**	0.228	**0.245**	0.197	0.235	0.197
BioBlobs (+ seed align)	0.873	0.886	0.824	0.902	0.276	0.368	0.290	0.658	0.301	0.376	0.299	0.608
SaProt-650M												
Uniform random	0.500	0.499	0.498	0.500	0.050	0.057	0.078	0.109	0.000	0.049	0.068	0.109
Residue burial	0.795	0.676	0.670	0.577	0.121	0.103	0.120	0.127	0.121	0.105	0.111	0.082
Integrated gradients	0.591	0.676	0.642	0.693	0.060	0.118	0.126	0.228	0.000	0.110	0.120	0.251
Attention pooling	0.767	0.741	0.728	0.678	0.128	0.165	0.196	0.250	0.121	0.188	0.206	0.257
BioBlobs	**0.814**	**0.780**	**0.776**	**0.747**	**0.173**	**0.211**	**0.230**	**0.397**	**0.177**	**0.233**	**0.239**	**0.395**
BioBlobs (+ seed align)	0.845	0.845	0.809	0.910	0.199	0.301	0.282	0.692	0.208	0.311	0.290	0.642

- With SaProt, BioBlobs leads on **every dataset** in all three metrics.
- With residue labels, seed alignment more than **doubles** the best baseline's Motif AUPRC.

Thank you!

Paper arxiv.org/abs/2510.01632

Code github.com/OliverLaboratory/BioBlobs

Viewer oliverlaboratory.com/resources/bioblobs