NEURAL INFORMATION PROCESSING SYSTEMS

MMSite: A Multi-modal Framework for the Identification of Active Sites in Proteins

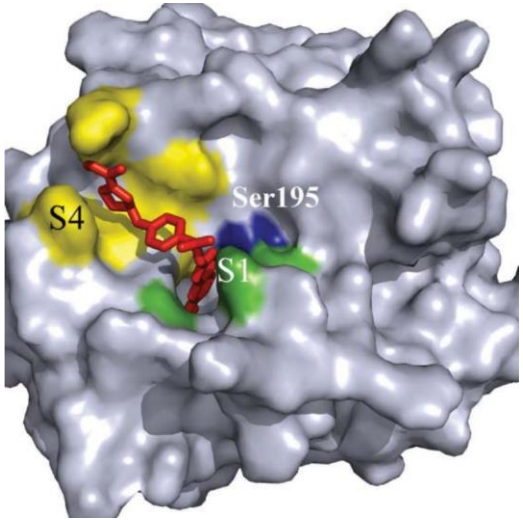
Song Ouyang[1], Huiyu Cai[2,3,4], Yong Luo[1*], Kehua Su[1*], Lefei Zhang[1], Bo Du[1]

[1]Wuhan University [2]BioGeometry [3]Mila – Québec AI Institute [4]Université de Montréal

(*Corresponding authors)

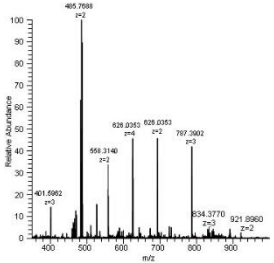
Active sites in Proteins

- **The region where substrate molecules bind and undergo chemical reaction**

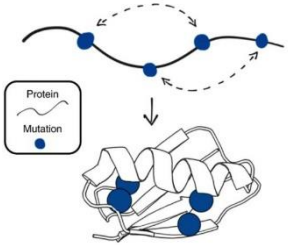


Factor Xa binds **prothrombin**

X-ray Crystallography



Mass Spectrometry



Site-Directed Mutagenesis

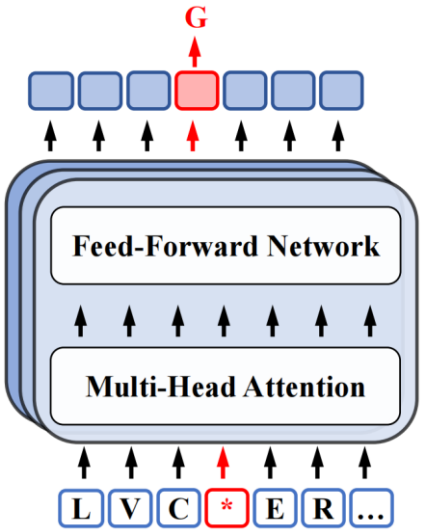
Traditional approaches often:

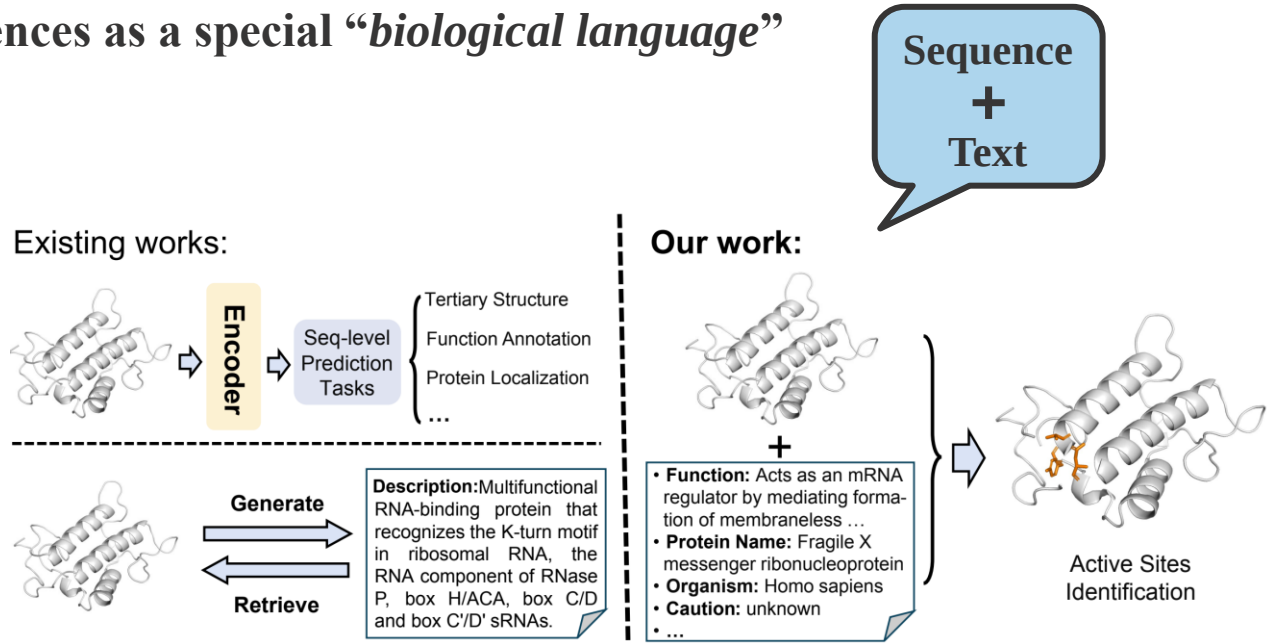
- **Time-consuming**
- **Expensive**
- **Specialized equipment and operations**
- **Extensive experiments**

Yeh C H, Fredenburgh J C, Weitz J I. Oral direct factor Xa inhibitors[J]. Circulation research, 2012, 111(8): 1069-1078.

Protein Representation Learning

- Treat protein sequences as a special "*biological language*"





Zuobai Zhang. https://drive.google.com/file/d/1RDbaAmajox7vDZN9kqTUHEi6FrxZYzO3/view?usp=sharing. 2024.

ProTAD (ProTein-Attribute text Dataset)

Dataset Preparation

Step 1:
Collect and clean data from Swiss-Prot in UniProt (more than 570,000 pairs of sequence and multi-attribute text description) $T_i = \{(\mathrm{t}^n_{i,j}, \mathrm{t}^c_{i,j})\}_{j=1}^{M}$

Step 2:
Cluster sequences based on the similarity using MMSeqs2

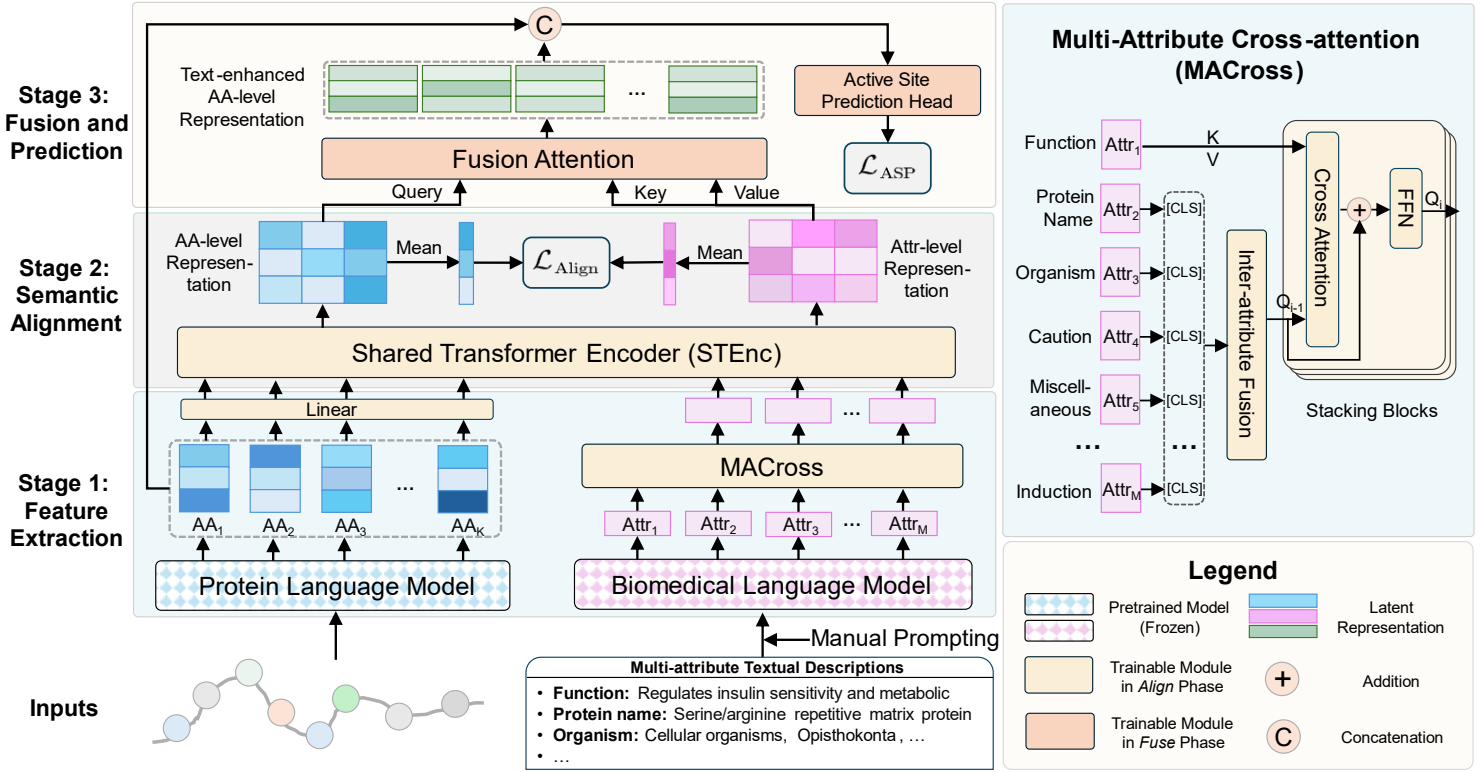
Step 3:
Adopt *k-selected* strategy to build train/validation/ test dataset

Raw Tabular Textual Descriptions of P05117

- `Protein Name:` Polygalacturonase-2 (PG) (PG-2A) (PG-2B) (Pectinase)
- `Organism:` Solanum lycopersicum (Tomato) (Lycopersicon esculentum)
- `Taxonomic Lineage:` cellular organisms (no rank), Eukaryota (superkingdom), Viridiplantae (kingdom), Streptophyta (phylum), Streptophytina (subphylum), Embryophyta (no rank), Tracheophyta (no rank), Euphyllophyta (no rank), Spermatophyta (no rank), Magnoliopsida (class), Mesangiospermae (no rank), eudicotyledons (no rank), Gunneridae (no rank), Pentapetalae (no rank), asterids (no rank), lamiids (no rank), Solanales (order), Solanaceae (family), Solanoideae (subfamily), Solaneae (tribe), Solanum (genus), Solanum subgen. Lycopersicon (subgenus)
- `Function:` Catalytic subunit of the polygalacturonase isozyme 1 and 2 (PG1 and PG2). Acts in concert with the pectinesterase, in the ripening process. Is involved in cell wall metabolism, specifically in polyuronide degradation. The depolymerization and solubilization of cell wall polyuronides mediated by PG2 during ripening seems to be limited by the beta subunit GP1, probably by recruiting PG2 to form PG1.
- `Caution:` **nan**.
- `Miscellaneous:` To avoid liquid rheology of tomato juice, temperature and pressure can be increased to inactivate selectively PG2 during the process.
- `Subunit Structure:` Monomer PG2 (isoenzymes PG2A and PG2B). Also forms heterodimers called polygalacturonase 1 (PG1) with the beta subunit GP1.
- `Induction:` By ethylene.
- `Tissue Specificity:` Expressed only in ripening fruits (at protein level).
- `Developmental Stage:` PG1 appears when fruits start to be coloured. When fruits are orange, both PG2 and PG1 are present. In fully ripe fruit, mostly PG2 is expressed.
- `Allergenic Properties:` **nan**
- `Biotechnological Use:` The effect of PG can be neutralized by introducing an antisense PG gene by genetic manipulation. The Flavr Savr tomato produced by Calgene (Monsanto) in such a manner has a longer shelf life due to delayed ripening.
- `Pharmaceutical Use:` **nan**
- `Involvement in Disease:` **nan**
- `Subcellular Location:` Secreted, extracellular space, apoplast. Secreted, cell wall.
- `Post-translational Modification:` N-glycosylated. PG2B isozyme has a greater degree of glycosylation than PG2A.
- `Sequence Similarities:` Belongs to the glycosyl hydrolase 28 family.

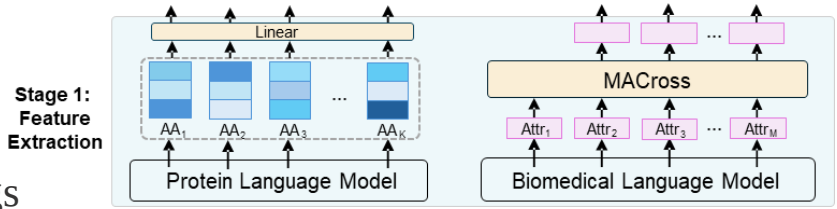
The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1):D523–D531, 11 2022.
Mirdita, Milot, et al. "Fast and sensitive taxonomic assignment to metagenomic contigs." Bioinformatics 37.18 (2021): 3029-3031.

MMSite - First Align, Then Fuse



MMSite - First Align, Then Fuse

- **Stage 1: Feature Extraction**



➢ **Protein Sequence Branch**

Adopt PLM f_ϕ to extract sequence embeddings

$$\boldsymbol{z}^{\mathrm{s}} = f_\phi(S)$$

➢ **Multi-attribute Text Description Branch**

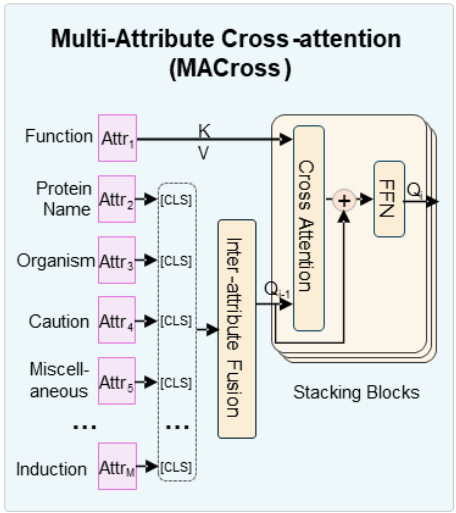
Adopt BLM f_ψ and ***MACross*** to extract text embeddings

$$f_\psi(\tilde{T}) = \{f_\psi(\tilde{\mathrm{t}}_1), f_\psi(\tilde{\mathrm{t}}_2), \cdots, f_\psi(\tilde{\mathrm{t}}_M)\}$$

In MACross:

$$\boldsymbol{x}^{\mathrm{t}}_{-F} = \mathrm{Attention}(\mathrm{Concat}(\{f_\psi(\tilde{\mathrm{t}}_i)_{\texttt{[CLS]}} | 1 \le i \le M, \mathrm{t}^n_i \neq \texttt{Function}\}))$$

$$\mathrm{CrossAttention}(\boldsymbol{x}^{\mathrm{t}}_{-F}, \boldsymbol{x}^{\mathrm{t}}_F, \boldsymbol{x}^{\mathrm{t}}_F) = \mathrm{Softmax}(\frac{\mathbf{Q}_{-F}\mathbf{K}^\top_F}{\sqrt{d}})\mathbf{V}_F$$

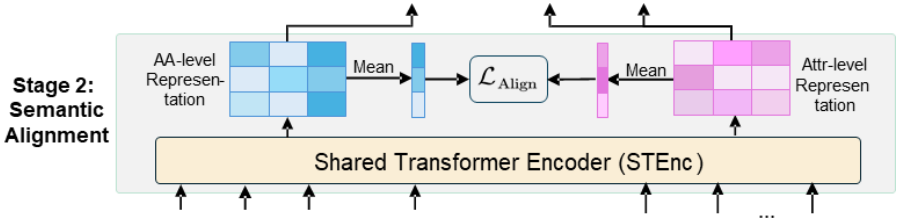


MMSite - First Align, Then Fuse

- **Stage 2: Semantic Alignment**

  - Similar protein sequences give rise to similar structures and functions.

  - Traditional "hard-label" alignment assigns positive pairs a label of 1 and negative pairs a label of 0, pushing them apart in feature space, which is not ideal in our scenario.



Soft-label Alignment

Inter-modality Similarity: $s_{ij}^{\text{s2t}} = cosine(\tilde{\boldsymbol{z}}_i^{\text{s}}, \tilde{\boldsymbol{z}}_j^{\text{t}})$

Intra-modality Similarity: $r_{ij}^{\text{s2s}} = cosine(\tilde{\boldsymbol{z}}_i^{\text{s}}, \tilde{\boldsymbol{z}}_j^{\text{s}})$ $\quad r_{ij}^{\text{t2t}} = cosine(\tilde{\boldsymbol{z}}_i^{\text{t}}, \tilde{\boldsymbol{z}}_j^{\text{t}})$

Target *Intra*-modality Distribution: $P_{ij}^{\text{s2s}} = \frac{\exp(r_{ij}^{\text{s2s}})}{\sum_{k=1}^{|\mathcal{B}|}\exp(r_{ik}^{\text{s2s}})}$ $\quad P_{ij}^{\text{t2t}} = \frac{\exp(r_{ij}^{\text{t2t}})}{\sum_{k=1}^{|\mathcal{B}|}\exp(r_{ik}^{\text{t2t}})}$

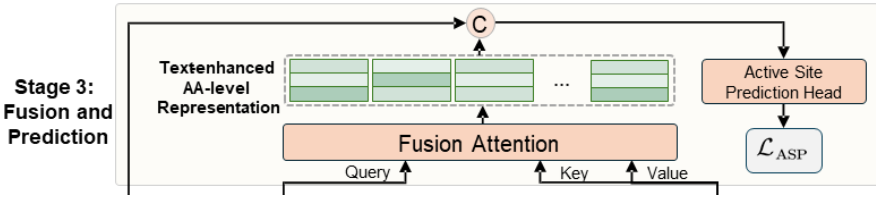
Predicted *Inter*-modality Distribution: $Q_{ij}^{\text{s2t}} = \frac{\exp(s_{ij}^{\text{s2t}}/\tau)}{\sum_{k=1}^{|\mathcal{B}|}\exp(s_{ik}^{\text{s2t}}/\tau)}$ $\quad Q_{ij}^{\text{t2s}} = \frac{\exp(s_{ij}^{\text{t2s}}/\tau)}{\sum_{k=1}^{|\mathcal{B}|}\exp(s_{ik}^{\text{t2s}}/\tau)}$

Loss Function in *Align* Phrase: $\mathcal{L}_{\text{Align}} = \frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}(D_{KL}(P_i^{\text{s2s}} \| Q_i^{\text{s2t}}) + D_{KL}(P_i^{\text{t2t}} \| Q_i^{\text{t2s}}))$

Huang H, Nie Z, Wang Z, et al. Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(16): 18298-18306.

MMSite - First Align, Then Fuse

- **Stage 3: Fusion and Prediction**



  - In *Fuse* phrase, cross-attention mechanism is used to obtain text-empowered sequence representation.

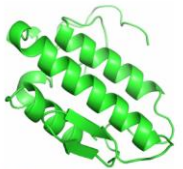
    Then we fuse it with the original sequence representation via skip concatenation

    Loss function: $\mathcal{L}_{\text{ASP}} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left[y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right]$

  - In inference stage, SOTA biomedical text generation model is employed as an agent to complete missing text modality when dealing with newly-discovered proteins, for example:

Sequence: NLYQFGKMINHMVGKSPIFSYGDYGCYCGWGGKGTPVDATDRCCFVHDCCYGRANGCDPKLSTYSYNFQNGNIVCGNKYGCLRHI CECDRVAAICFGENVNTYDKKFLSSSRCRQTSEQC

Agent: Basic phospholipase A2 that inhibits ADP-, thrombin- and arachidonic acid-induced platelet aggregation. Inhibits phospholipase A2 activity in the presence of collagen, and inhibits the binding of monolayers inducing aggregation induced by collagen. Inhibits the formation of superoxide anions in vivo and in vitro. Inhibits the adhesion of endothelial cells to fibronectin and collagen. Inhibits cell migration and invasion in vitro. Inhibits the adhesion of endothelial cells to collagen and human umbilical vein endothelial cells in vitro. Inhibits the adhesion of monocytes to collagen in vitro. Inhibits the binding of human platelets to fibrinogen. Inhibits the adhesion of human platelet receptors to fibrinogen in vitro. Inhibits the binding of human monocytes to dendritic cells in a dose-dependent manner. Inhibits the binding of human monocytes to dendritic cells in a dose-dependent manner. Inhibits the binding of human plasma fibronectin to TNFRSF10 in a dose-dependent manner …



Experiments

- **Main Comparison Results**

Input†	Method	Version	F_{max} ↑	AUPRC ↑	MCC ↑	OS ↑	FPR ↓
Seq.	ESM	1b [44]	0.7052(±0.02)	0.8452(±0.02)	0.7123(±0.02)	0.7211(±0.04)	0.2758(±0.01)
		1v [33]	0.6306(±0.03)	0.7975(±0.02)	0.6382(±0.03)	0.6398(±0.03)	0.3388(±0.02)
		2-650M [29]	0.6517(±0.04)	0.8230(±0.04)	0.6596(±0.04)	0.6652(±0.02)	0.3240(±0.05)
	ProtT5 [11]	BFD	0.4156(±0.05)	0.6773(±0.03)	0.4217(±0.05)	0.4130(±0.05)	0.5509(±0.05)
		UniRef	0.4696(±0.04)	0.7119(±0.02)	0.4767(±0.04)	0.4652(±0.04)	0.4919(±0.04)
	ProtBert [11]	BFD	0.5610(±0.02)	0.7524(±0.02)	0.5715(±0.02)	0.5865(±0.04)	0.4115(±0.02)
		UniRef	0.4817(±0.02)	0.6992(±0.01)	0.4896(±0.02)	0.4915(±0.01)	0.4871(±0.03)
	ProtAlbert [11]		0.6033(±0.03)	0.7519(±0.02)	0.6121(±0.03)	0.6149(±0.03)	0.3636(±0.01)
	ProtXLNet [11]		0.0345(±0.00)	0.0952(±0.02)	0.0409(±0.00)	0.0772(±0.01)	0.9233(±0.00)
	ProtElectra [11]		0.5636(±0.02)	0.7630(±0.02)	0.5732(±0.02)	0.5793(±0.04)	0.4041(±0.01)
	PETA [50]	deep_base	0.6533(±0.02)	0.7994(±0.01)	0.6603(±0.02)	0.6529(±0.02)	0.3134(±0.02)
	S-PLM [56]		0.7262(±0.02)	0.8712(±0.01)	0.7337(±0.02)	0.7322(±0.03)	0.2452(±0.02)
	TAPE [43]		0.3560(±0.02)	0.5413(±0.01)	0.3622(±0.02)	0.3523(±0.02)	0.6096(±0.02)
Seq. & Struct.	MIF [60]	MIF	0.1379(±0.02)	0.3470(±0.02)	0.1393(±0.02)	0.1346(±0.02)	0.8524(±0.03)
		MIF-ST	0.1033(±0.02)	0.2883(±0.03)	0.1034(±0.02)	0.1030(±0.02)	0.8958(±0.02)
	PST [3]	t33	0.6574(±0.01)	0.8139(±0.01)	0.6648(±0.01)	0.6719(±0.01)	0.3219(±0.01)
		t33_so	0.6708(±0.02)	0.8266(±0.03)	0.6793(±0.02)	0.6891(±0.03)	0.3079(±0.01)
Seq. & Text	ProtST [59] w/ retrain	ESM-1b	0.4036(±0.03)	0.6762(±0.02)	0.4144(±0.02)	0.4297(±0.03)	0.5663(±0.02)
		ESM-2	0.1865(±0.01)	0.4220(±0.03)	0.1918(±0.02)	0.1872(±0.05)	0.7897(±0.01)
	ProtST w/o retrain	ESM-1b	0.4632(±0.05)	0.7040(±0.02)	0.4722(±0.05)	0.4779(±0.05)	0.5030(±0.05)
		ESM-2	0.5483(±0.02)	0.7716(±0.01)	0.5562(±0.02)	0.5613(±0.01)	0.4239(±0.02)
	MMSite ‡		**0.8250**(±0.02)	**0.8909**(±0.01)	**0.8319**(±0.02)	**0.8549**(±0.02)	**0.1689**(±0.02)

† This column refers to the modality input in the inference stage.
‡ We report the performance using ESM-1b and PubMedBERT-abs as the PLM and BLM encoders.

MMSite achieves SOTA performance compared with other 21 PRL models

Method	F_{max} ↑	AUPRC ↑	MCC ↑	OS ↑	FPR ↓
ESM-1b	0.7050	0.8443	0.7117	0.7127	0.2705
+MMSite-Abs	↑ 0.120	↑ 0.047	↑ 0.120	↑ 0.142	↓ 0.102
+MMSite-Full	↑ 0.105	↑ 0.044	↑ 0.107	↑ 0.145	↓ 0.076
ESM-1v	0.6267	0.8018	0.6340	0.6351	0.3431
+MMSite-Abs	↑ 0.160	↑ 0.069	↑ 0.159	↑ 0.164	↓ 0.149
+MMSite-Full	↑ 0.172	↑ 0.078	↑ 0.172	↑ 0.184	↓ 0.156
ESM-2-650M	0.6402	0.8068	0.6479	0.6607	0.3434
+MMSite-Abs	↑ 0.156	↑ 0.072	↑ 0.157	↑ 0.175	↓ 0.138
+MMSite-Full	↑ 0.162	↑ 0.075	↑ 0.161	↑ 0.169	↓ 0.152

BLM enhances PLM's performance

Ablation Study

- **Effectiveness of each components in MMSite**

Seq-M	Text-M	Align	MACross	SENC	F_{max} ↑	AUPRC ↑	MCC ↑	OS ↑	FPR ↓
✓	✓	✓	✓	✓	**0.8250**	**0.8909**	**0.8319**	**0.8549**	**0.1689**
✓	✓	✓	✓		0.8021	0.8819	0.8071	0.8027	0.1738
✓	✓	✓		✓	0.8152	0.8908	0.8214	0.8379	0.1757
✓	✓	✓			0.8037	0.8850	0.8105	0.8241	0.1847
✓	✓				0.7911	0.8710	0.7980	0.8150	0.1978
✓					0.7052	0.8452	0.7123	0.7211	0.2758

- **"First Align, Then Fuse" *vs* "Align While Fusing"**

PLM	Strategy	PubMedBERT-abs					PubMedBERT-full				
		F_{max} ↑	AUPRC ↑	MCC ↑	OS ↑	FPR ↓	F_{max} ↑	AUPRC ↑	MCC ↑	OS ↑	FPR ↓
ESM-1b	Single	0.8086	0.8772	0.8158	0.8329	0.1798	0.8055	0.8766	0.8121	0.8253	**0.1818**
	Two	**0.8250**	**0.8909**	**0.8319**	**0.8549**	**0.1689**	**0.8099**	**0.8882**	**0.8183**	**0.8574**	0.1950
ESM-1v	Single	0.7369	0.8525	0.7440	0.7576	0.2480	0.7713	0.8589	0.7780	0.7924	0.2155
	Two	**0.7864**	**0.8705**	**0.7933**	**0.7987**	**0.1942**	**0.7988**	**0.8795**	**0.8058**	**0.8194**	**0.1871**
ESM-2 -650M	Single	0.7522	0.8572	0.7591	0.7664	0.2296	0.7603	0.8638	0.7669	0.7677	0.2171
	Two	**0.7965**	**0.8789**	**0.8046**	**0.8358**	**0.2052**	**0.8018**	**0.8814**	**0.8091**	**0.8294**	**0.1916**

Ablation Study

- **Ablations on training strategies**

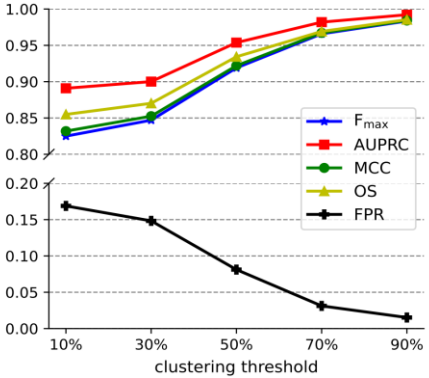
Method	Avg. (token)	Avg. (region)
MMSite	**0.8493**	**0.8430**
Func. as Q	0.8394	0.8279
Hard align	0.8334	0.8221

- **Temporal-based evaluation**

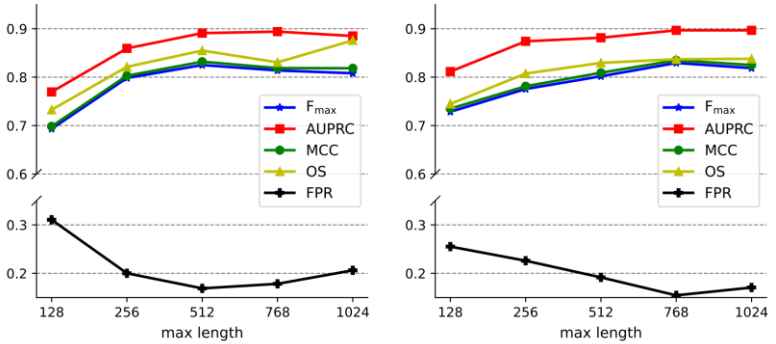
(Taking 115 newly discovered samples as test data)

Test Data	F_{max}	AUPRC	MCC	OS	FPR
Newly discovered proteins	0.8432	0.8865	0.8460	0.8465	0.1420

- **Impact of different clustering thresholds**

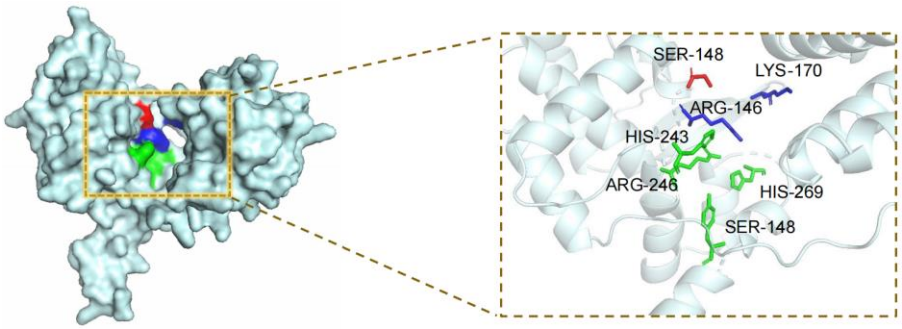


- **Impact of the length of protein sequence**

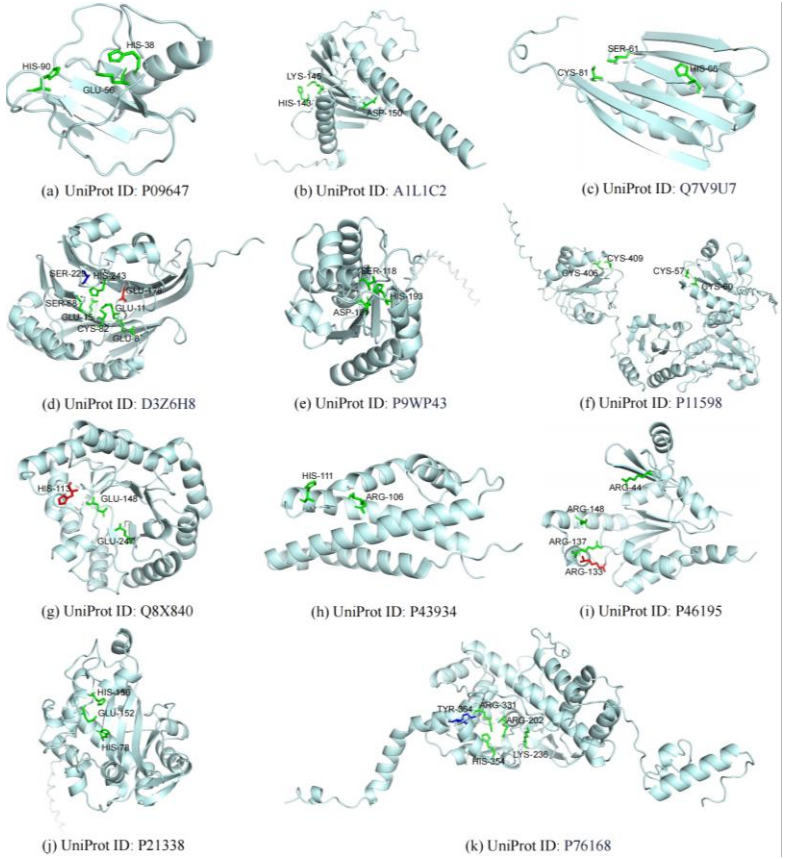


(a) ESM-1b & PubMedBERT-abs (b) ESM-2-650M & PubMedBERT-full

Visualization Results



Justification: The **palecyan** surface/sticks (residues) represent the background, while the **green**, **blue**, and **red** surface/sticks (residues) indicate the **correctly predicted** sites, **unpredicted** sites, and **incorrectly predicted** sites, respectively.



Conclusion

- ✓ We propose a new and meaningful task in biological science: identifying active sites in proteins using both sequence and textual descriptions, and construct the ProTAD dataset.
- ✓ We introduce a framework that integrates both modalities for predicting protein active sites, utilizing a “First Align, Then Fuse” strategy.
- ✓ Our comprehensive experimental validations confirm the effectiveness of our approach, demonstrating that our framework can be effectively applied to different PLMs and BLMs.

Thanks for watching!

Project

Also feel free to contact me at *ouyangsong@whu.edu.cn*