ProteinMAE: masked autoencoder for protein surface self-supervised learning

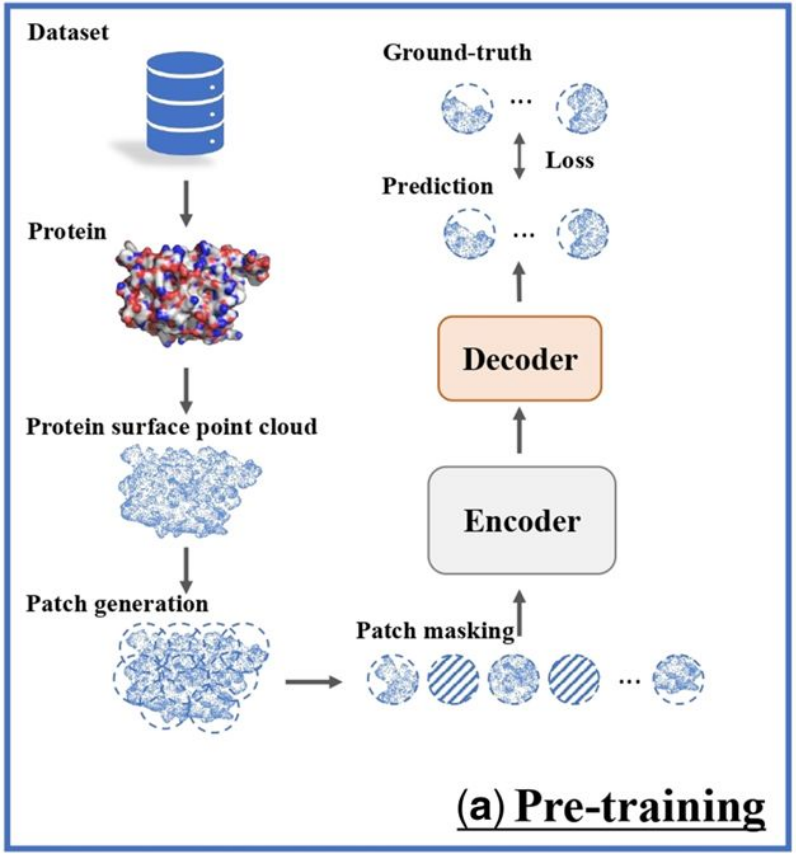
Mingzhi Yuan, Qin Qiao, Ao Shen, Kexue Fu, Manning Wang, Jiaming Guan, Yingfan Ma

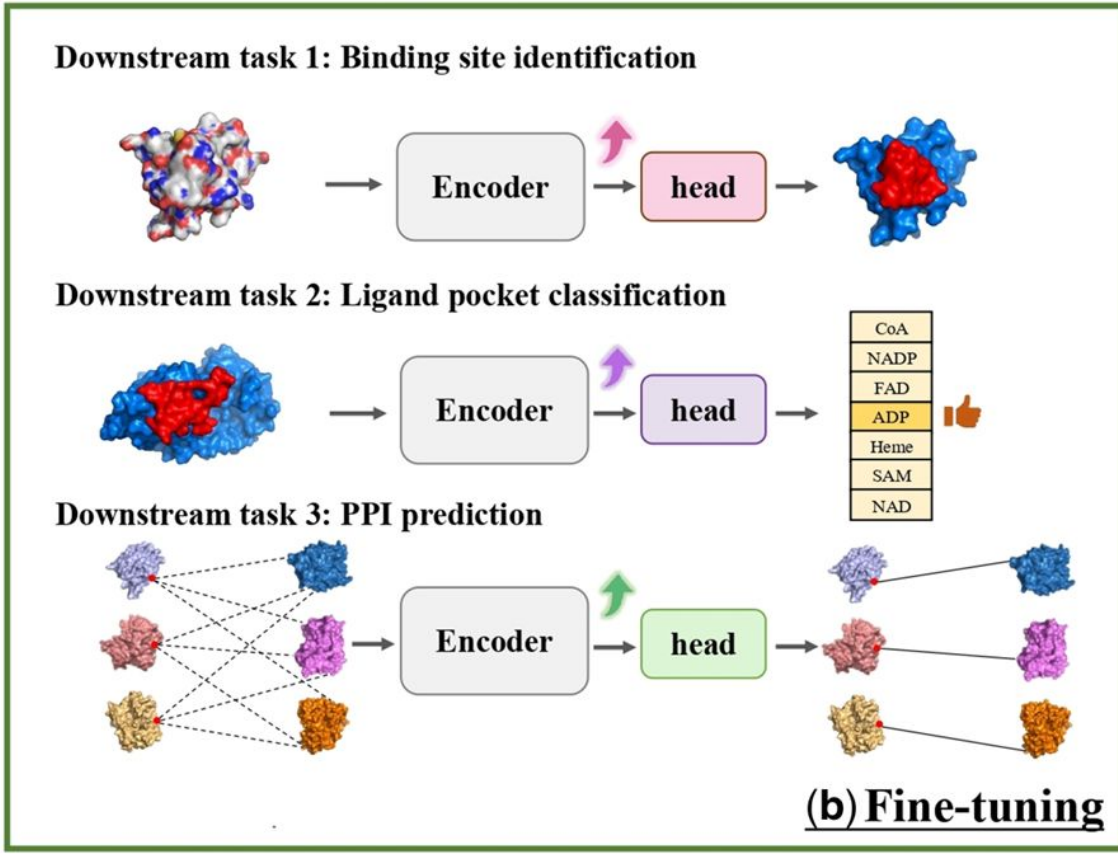
LifeLU reading group

presented by Özdeniz Dolu

17.10.2024

General Outlook





Motivations

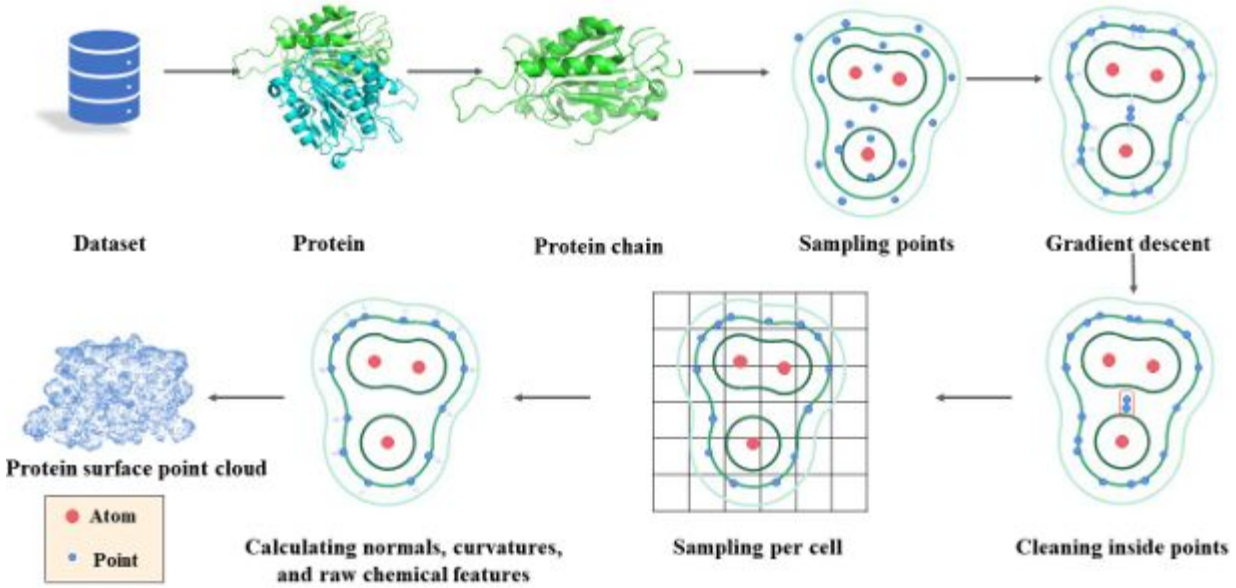
- Labeled data is scarce
- Self-supervised methods from NLP, CV
- Pretrain - Fine-tune paradigm
- Success of protein surface representation learning (MaSIF, dMaSIF etc.)

Related Work

- **MaSIF (Gainza et al. 2020)**: Geometric deep learning approach to protein surface representation. Uses geodesic convolutions on a surface mesh. Has a large computational overhead.
- **dMaSIF (Sverrisson et al. 2021):** Aims to lower MaSIF's computational costs by replacing mesh with point cloud. Applies "quasigeodesic convolution".

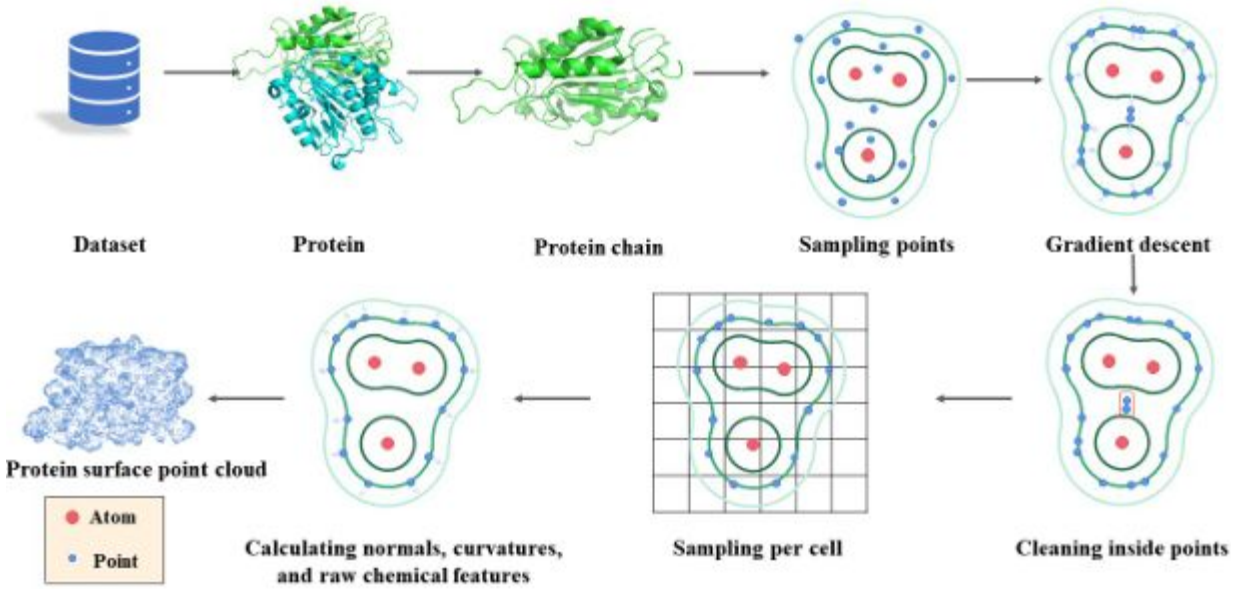
Data Preparation

Starting from 190615 proteins from PDB, data preparation process yields 359255 data points for pretraining. Point sampling process taken from dMaSIF.



Data Preparation

Dimension of each point: 122 (16 x 7 + 10) (Types of 16 nearest atoms + 10 geometric features)



Patch Generation

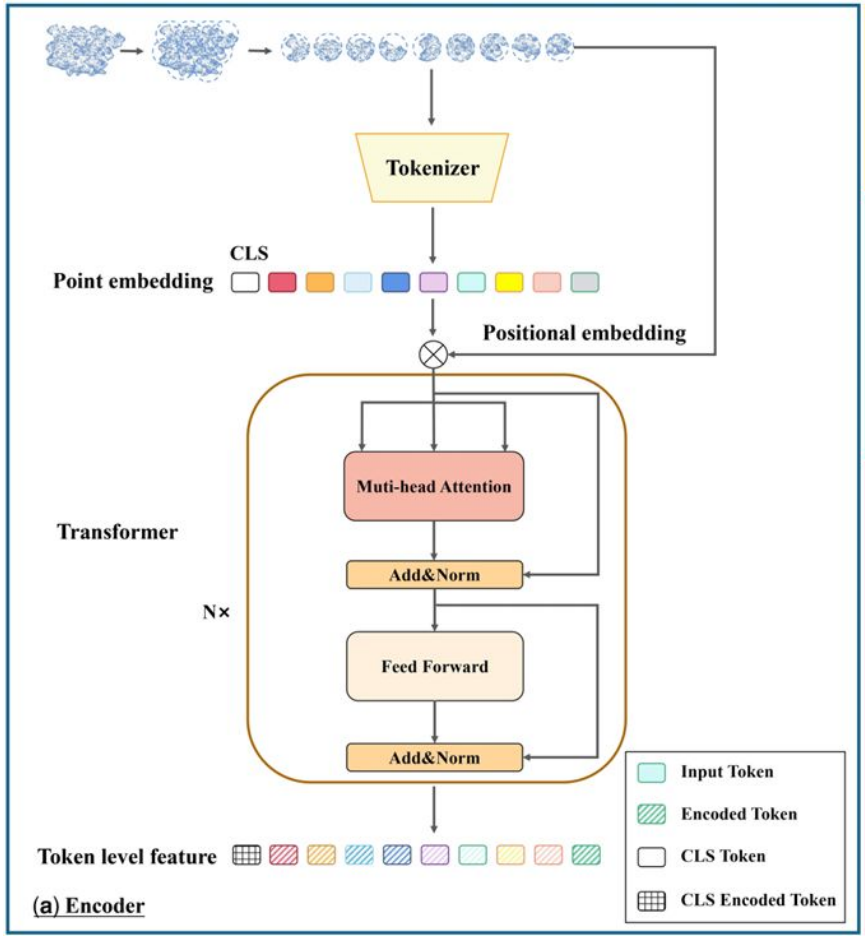
- Given a point cloud,
  a. sample g center points (farthest point sampling).
  b. For each center point sample k’ neighbors using KNN.
  c. We get g patches centered at center points containing k’ points where patches are irregular and possibly overlapping.
- Masking ratio *m*: Percentage of patches to be masked. 60% is used for experiments.

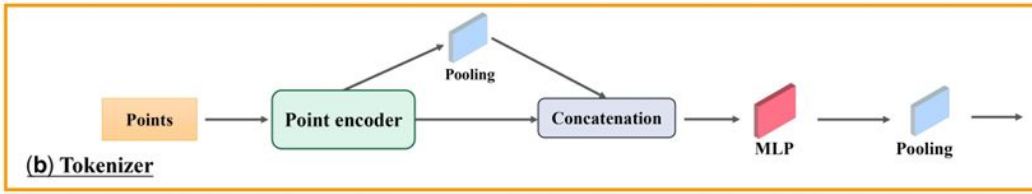
Loss (Pretraining stage)

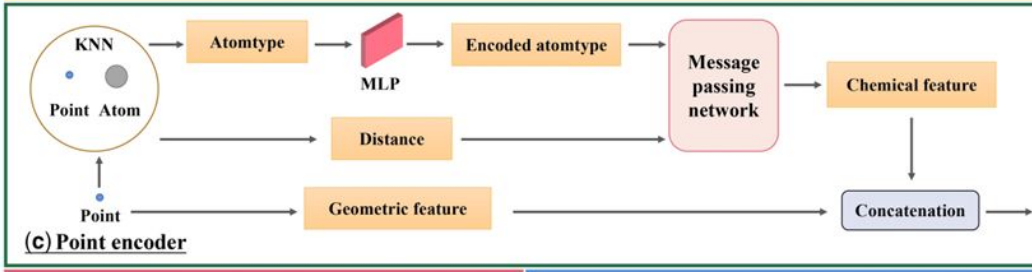
- Encoder takes unmasked patches as input -> Decoder predicts masked patches.
- Chamfer loss.

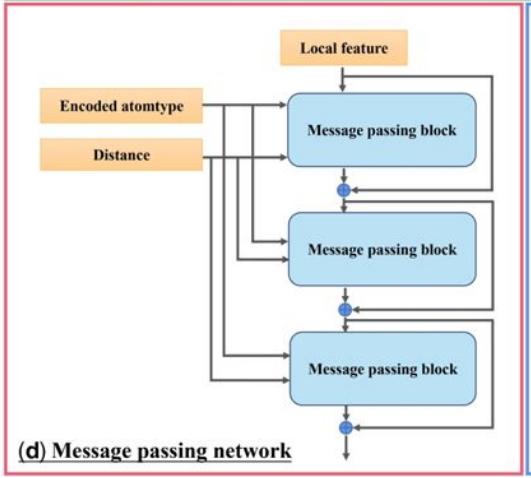
$$\mathcal{L} = \sum_{i=1}^{mg} \left(\frac{1}{|P_i^{\text{mask}}|} \sum_{x \in P_i^{\text{mask}}} \min_{y \in P_i^{\text{pred}}} ||x - y||_2^2 + \frac{1}{|P_i^{\text{pred}}|} \sum_{x \in P_i^{\text{pred}}} \min_{y \in P_i^{\text{mask}}} ||x - y||_2^2 \right),$$

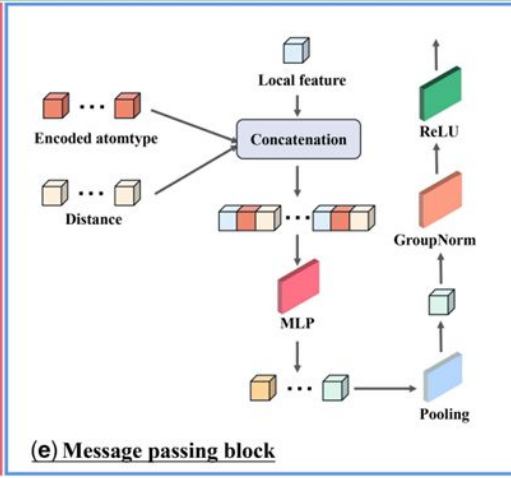
Architecture of the Encoder



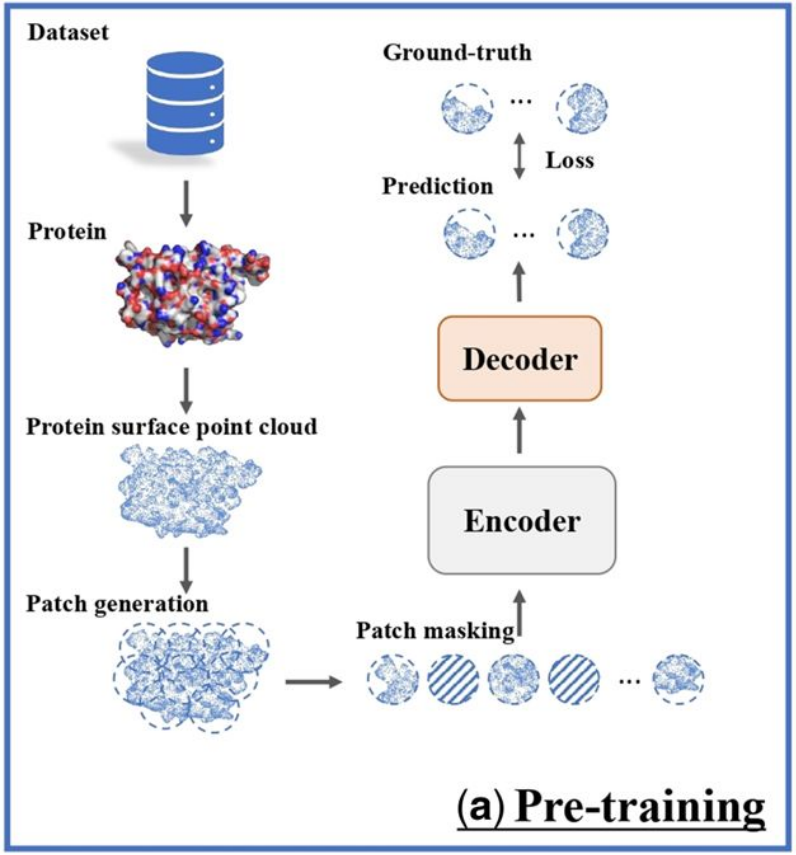


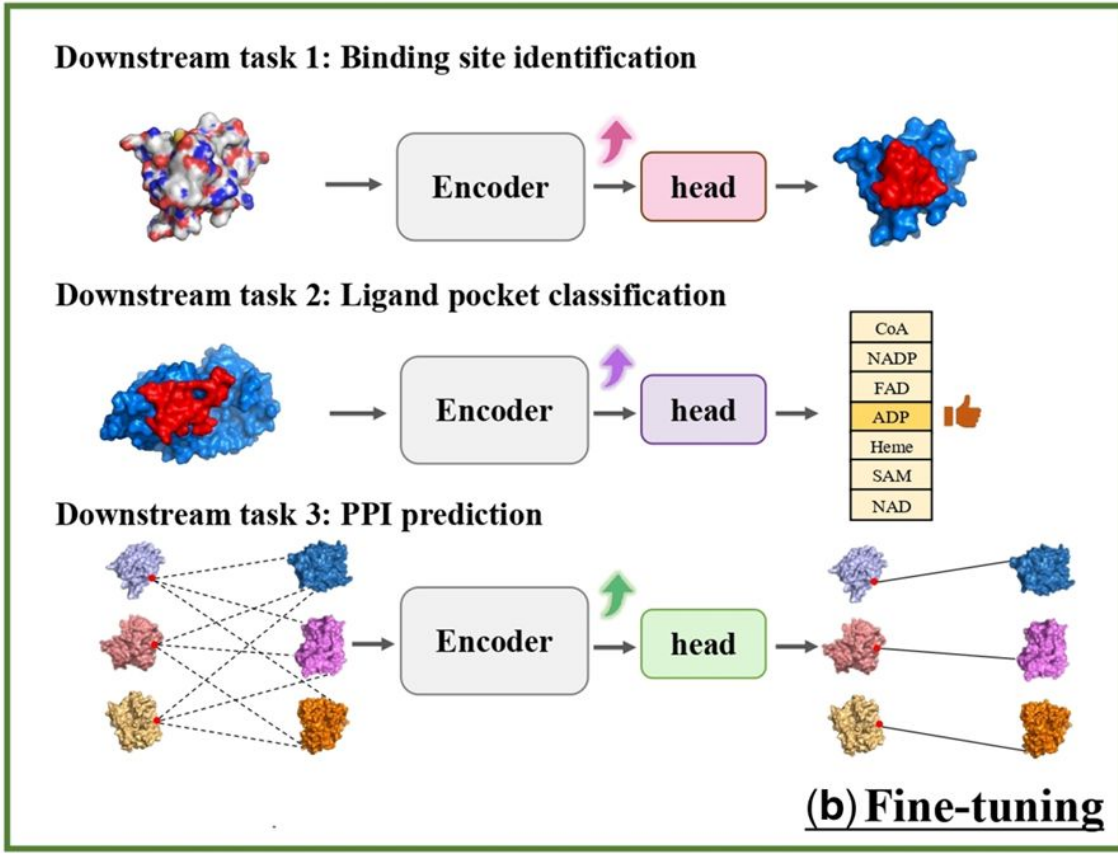






General Outlook

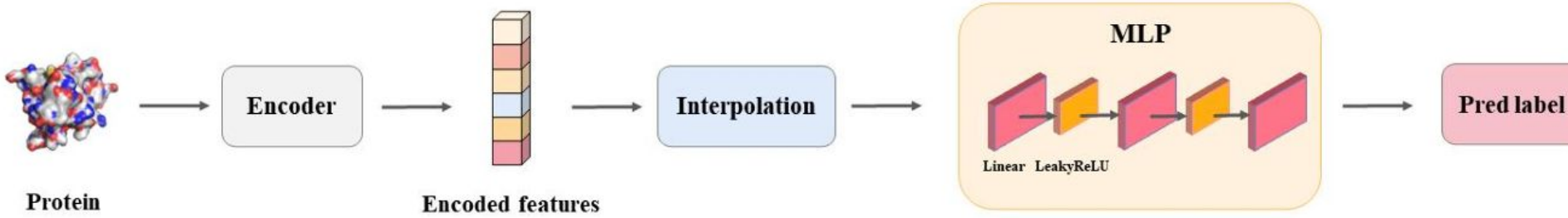




Fine-tuning

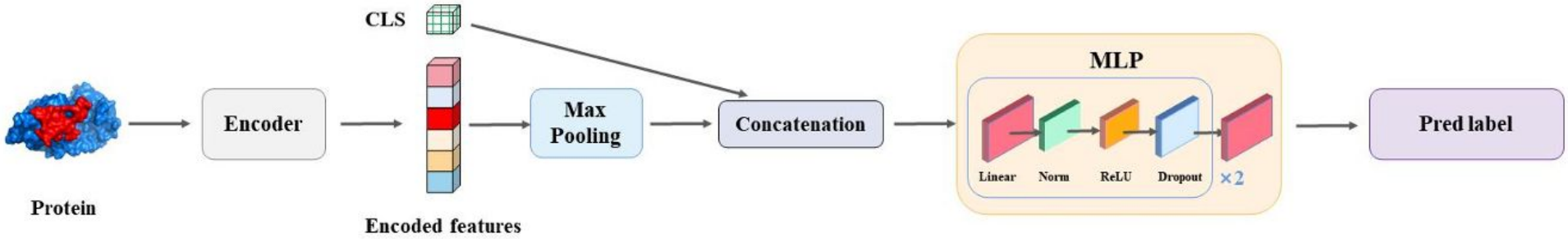
1. Binding site identification: Binary classification task. Classify the points on the surface as “interaction” or “non-interaction” sites. Same balanced cross-entropy loss as dMaSIF.

Patch level features > Point level features > Prediction



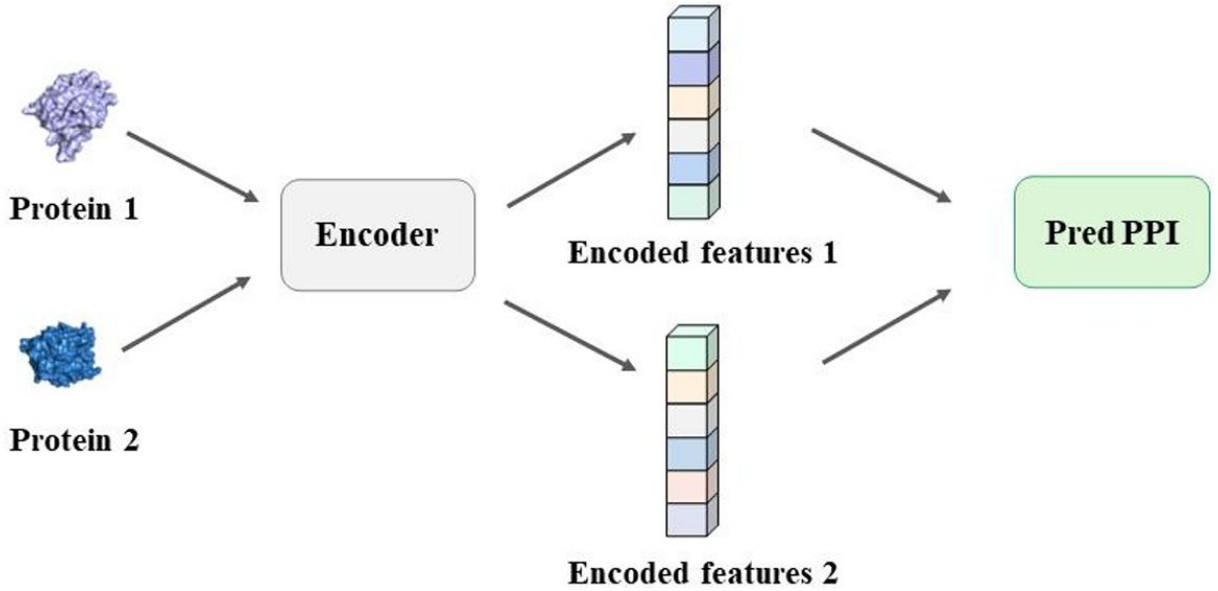
Fine-tuning

2. Ligand-binding protein pocket classification: Classification task at protein level with multiple classes. Task of estimating the binding preferences of the protein to 7 metabolites. Cross-entropy loss was used.



Fine-tuning

3. Protein–protein interaction prediction: Given two proteins, estimate the probability of their binding. Binary classification task. Balanced metric loss was used.



Results

1. Binding site identification in protein surface
Dataset: 2958 training 356 test

Table 1. Performance on binding site identification.

Method	Accuracy↑	Recall↑	F1 score↑	ROC-AUC↑
MaSIF	0.741	**0.864**	0.760	0.847
dMaSIF	0.774	0.781	0.763	0.865
Ours (from scratch)	0.765	0.785	0.756	0.843
Ours (contrastive)	0.788	0.772	0.769	0.866
Ours	**0.793**	0.799	**0.782**	**0.871**

Results

2. Ligand-binding protein pocket classification
Dataset: 1459 structures (72%, 8%, 20%) (train, val, test)

Table 2. Performance on ligand-binding pocket classification.

Method	Balanced accuracy↑
MaSIF	**0.74**
dMaSIF	0.623
Ours (from scratch)	0.666
Ours (contrastive)	0.667
Ours	0.707

Results

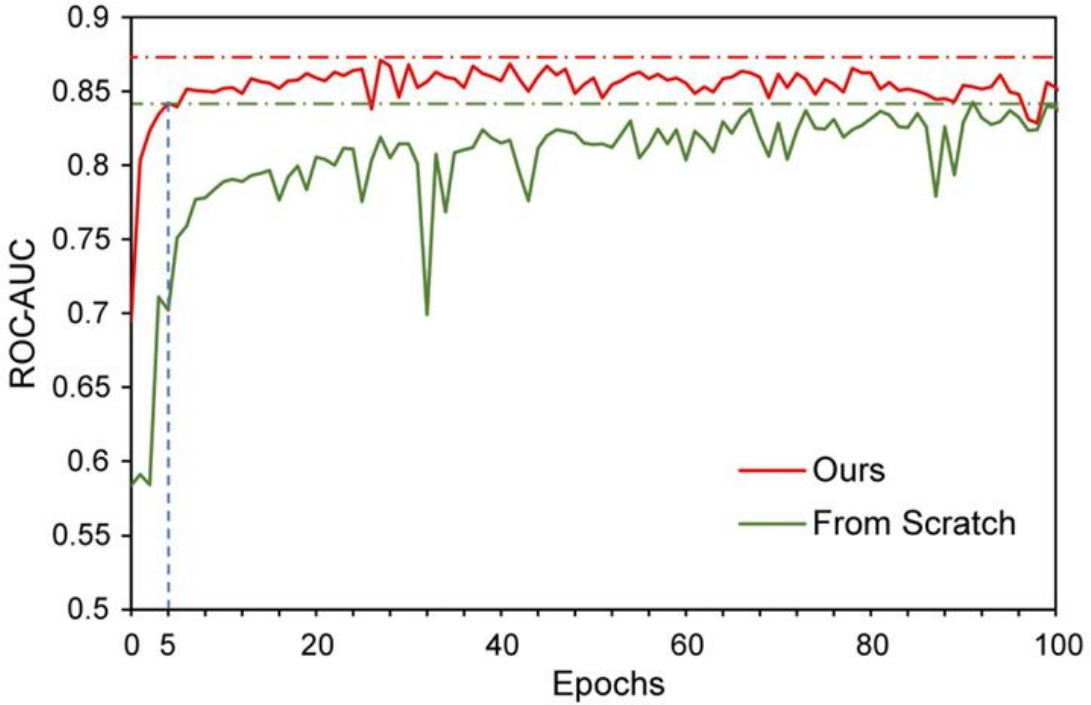
3. Protein–protein interaction prediction
Dataset: 4614 training 912 test

Table 3. Performance on protein–protein interaction prediction.

Method	Accuracy↑	Recall↑	F1 score↑	ROC-AUC↑
MaSIF	–	–	–	0.813
dMaSIF	0.795	0.823	0.793	0.862
Ours (from scratch)	0.922	0.990	0.930	0.944
Ours (contrastive)	0.926	**0.994**	0.933	0.945
Ours	**0.927**	**0.994**	**0.934**	**0.948**

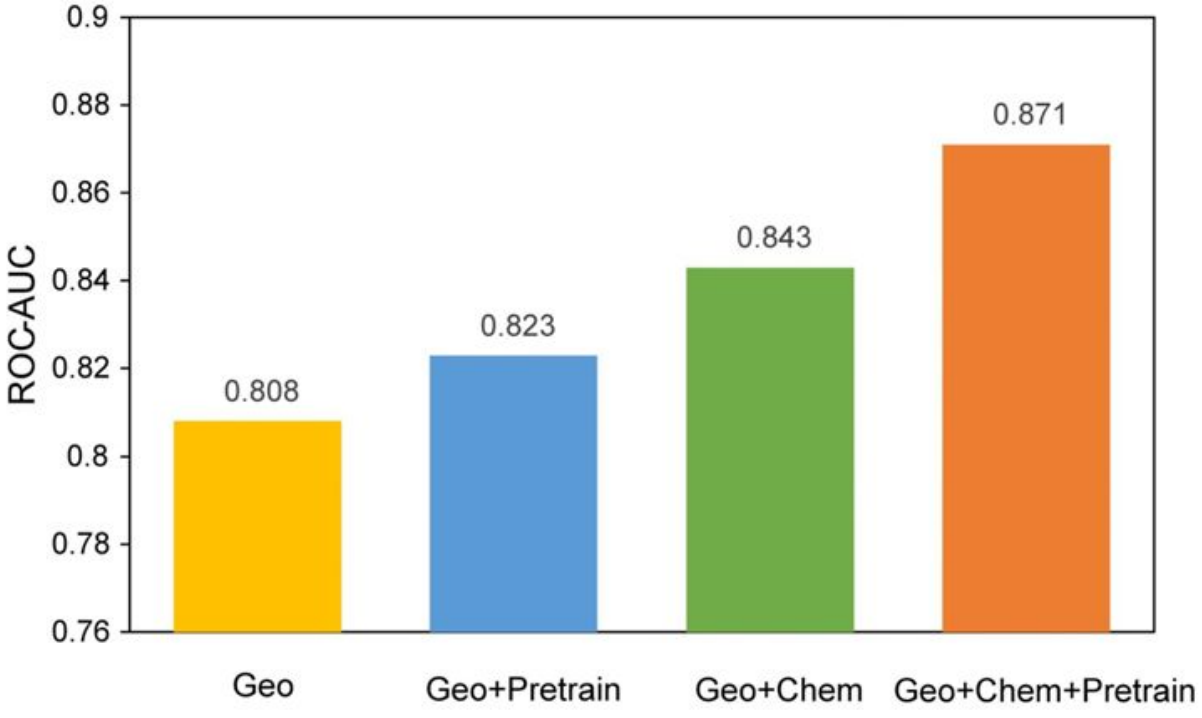
Results

Pretraining leads to faster convergence on downstream task.



Results

Ablation study on the first task (binding site identification).



Results

Performance by mask ratio on the first task (binding site identification).

Mask ratio (%)	ROC-AUC ↑
10	0.861
20	0.857
30	0.861
40	0.860
50	0.866
60	**0.871**
70	0.868
80	0.860
90	0.852
From scratch	0.843

Results

Computational efficiency (binding site identification).

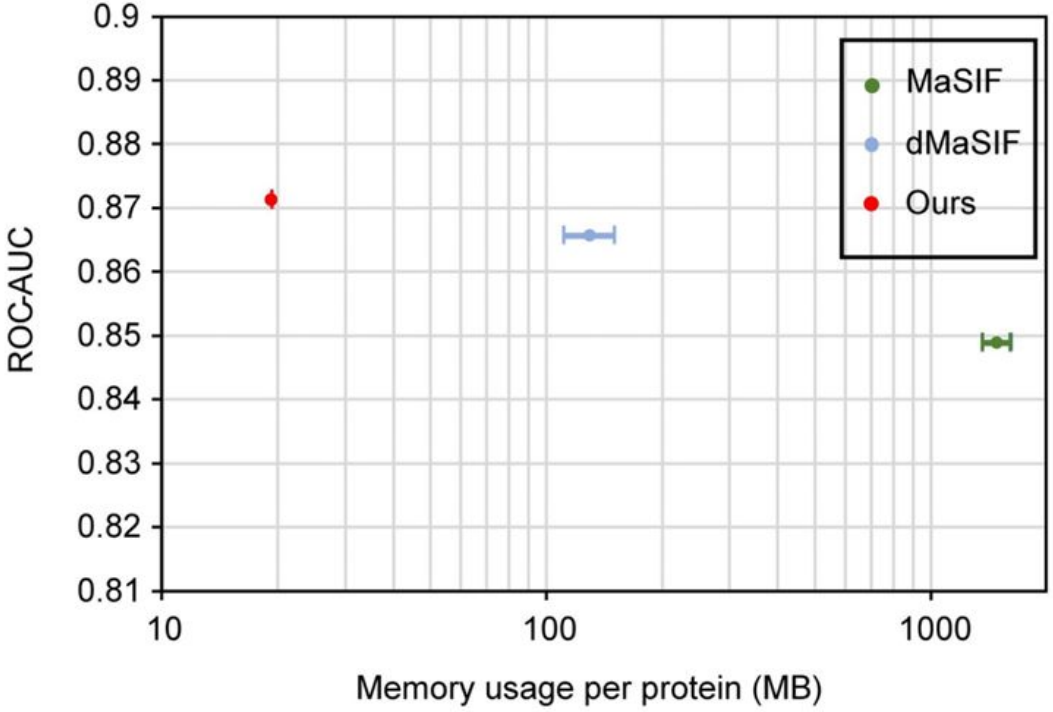
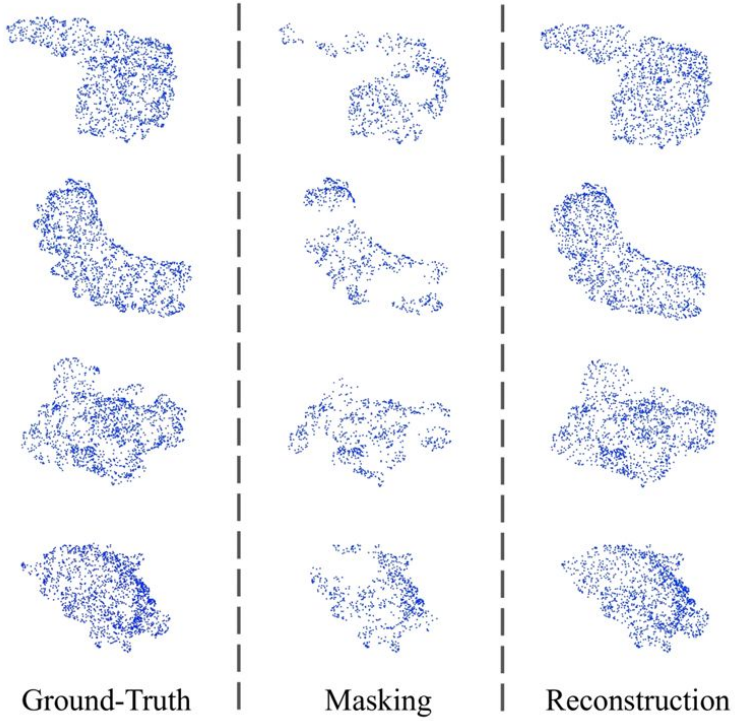


Table 5. Average running time per protein on binding site identification task of different networks.

	MaSIF	dMaSIF	Ours
Time (s/protein)	187.79	0.21	**0.17**

Results

Reconstruction of masked patches



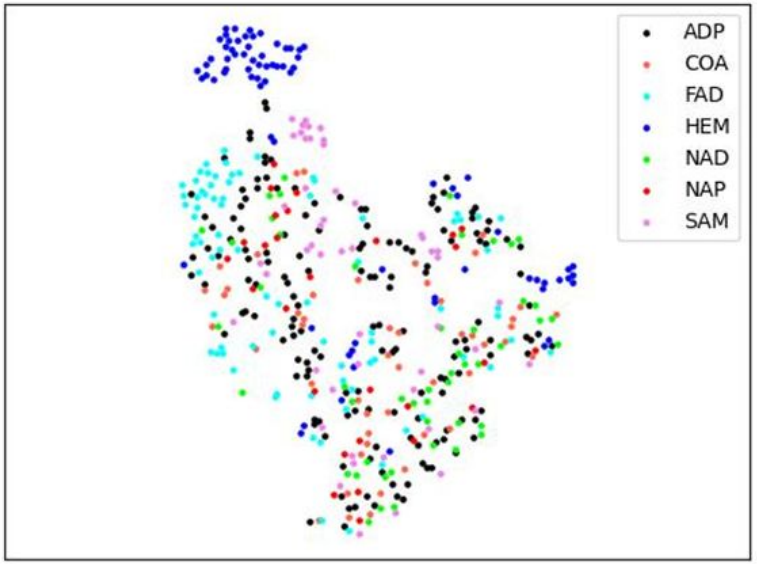
Results

Given some amount of labeled data, the effect of pretraining.

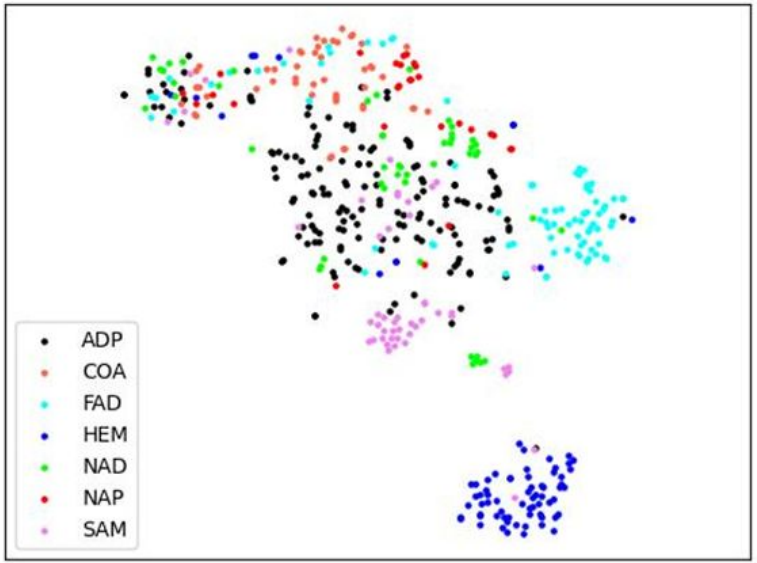
Percentage of labeled data (%)	From scratch (ROC-AUC↑)	Ours (ROC-AUC↑)
1	0.592	0.713 (+0.121)
2	0.603	0.747 (+0.144)
5	0.717	0.811 (+0.094)
10	0.780	0.829 (+0.049)
20	0.799	0.844 (+0.045)
30	0.816	0.847 (+0.031)
50	0.830	0.863 (+0.033)
100	0.852	0.871 (+0.019)

Results

Feature distribution in ligand-binding protein classification task



(a) Features trained from scratch



(b) Features fine-tuned

Conclusion

- ProteinMAE provides a framework for self supervised protein surface representation learning.
- Compared to its competitors, it’s computationally lighter while having competitive results.
- It does not require complex preprocessing or labeled data.

Thanks for listening