Pi-SAGE: Permutation-invariant surface-aware graph encoder for binding affinity prediction

Sharmi Banerjee[1], Mostafa Karimi[1], Melih Yilmaz[1], Tommi Jaakkola[2], Bella Dubrov[1], Shang Shang[1], Ron Benson[1]

[1]Grand Challenge Amazon, [2]MIT

Introduction

Protein surface fingerprint encodes chemical and geometric features that govern protein–protein interactions and can be used to predict changes in binding affinity between two protein complexes. Current state-of-the-art models for predicting binding affinity change, such as GearBind, are all-atom based geometric models derived from protein structures.

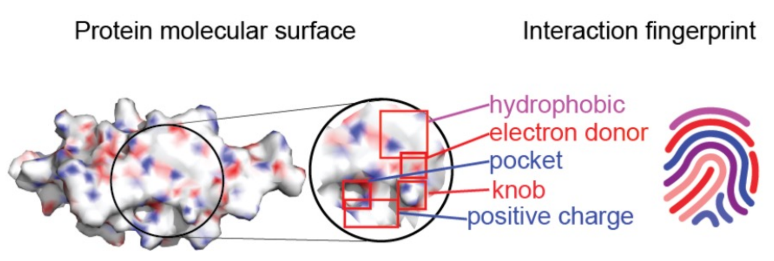
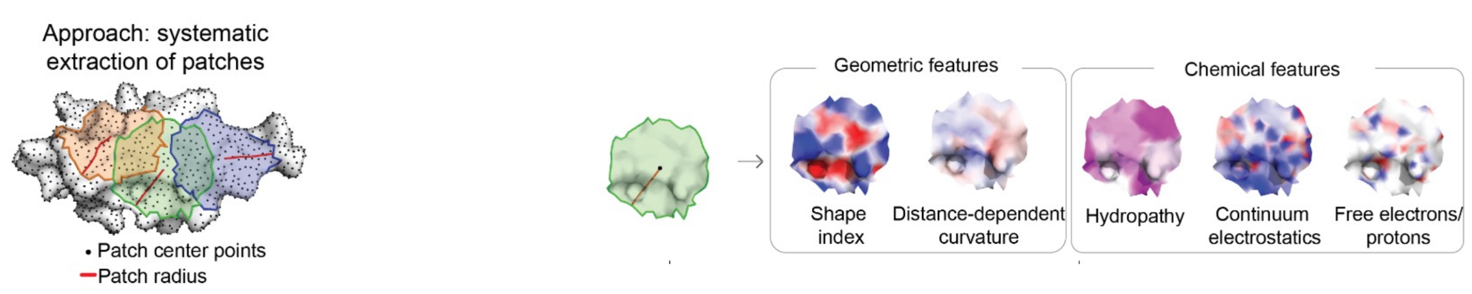


Figure from MaSIF [6] paper showing the chemical and geometric fingerprints

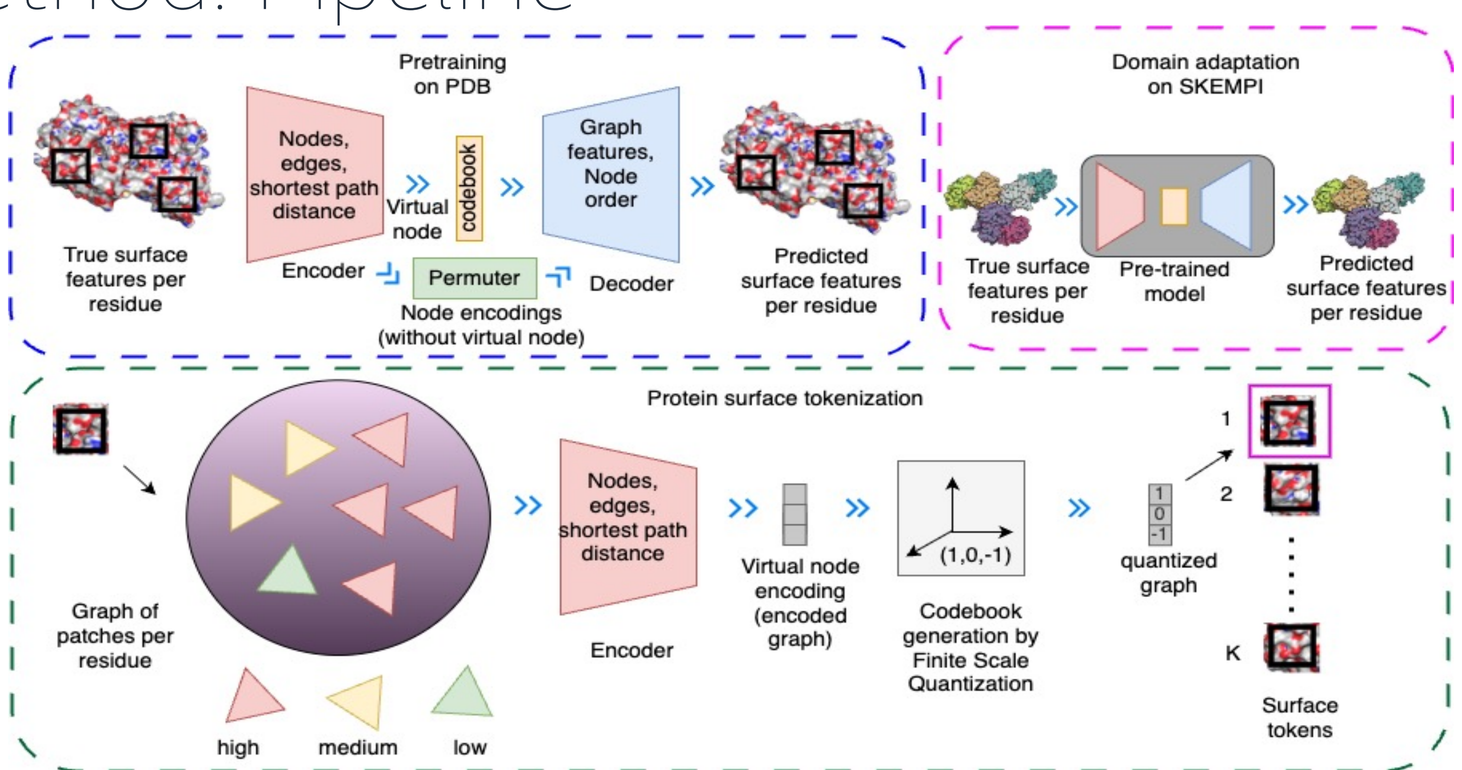
Problem definition

Accurately predicting changes in binding affinity (ΔΔG) is critical for protein design. Sequence-based pLMs learn about structure implicitly while structure aware pLMs learn about surface implicitly. We hypothesize that explicitly encoding protein surface features can improve prediction of ΔΔG. Figures below from MaSIF paper [6].



We propose Pi-SAGE, a Permutation-invariant Surface-Aware Graph Encoder that learns local surface-residue representations via a quantized codebook. We integrate these features into GearBind to improve prediction of binding affinity changes on the SKEMPI dataset.

Method: Pipeline



Permutation Invariance: We represent each surface residue as a graph $G = (V, E)$ where $V = \{n_i\}_{i=1:N}$ correspond to N randomly sampled patches. The edges are defined with a threshold of $3\dot{A}$. We add a virtual node connected to all other nodes use it for tokenized representation of the graph. We used the graph transformer from [1] that uses the vanilla for-product attention and learnable topological relationship between nodes and a learnable edge relationship.

$$a_{(i,j)}^{topology} = q_i \mathcal{P}_{\psi(i,j)}^{query} + k_i \mathcal{P}_{\psi(i,j)}^{key} \quad a_{(i,j)}^{edge} = q_i \mathcal{E}_{\psi(i,j)}^{query} + k_i \mathcal{E}_{\psi(i,j)}^{key}$$

$$a_{(i,j)} = \frac{q_i . k_j + a_{(i,j)}^{topology} + a_{(i,j)}^{edge}}{\sqrt{d_z}}$$

We adopted the Finite Scale Quantization [2] to create surface codebook. We added a Permuter module [3] to infer the node order in the residue graph. The permuter module learns to align the input and output graph through soft alignment. For each node i of input graph the permuter predicts a score s_i corresponding to its probability of having a low node index in the decoder graph. Ther permutation matrix is constructed as [5]

$$p_{ij} = \begin{cases} 1, if\ j = argsort(s)_i \\ 0,\ else \end{cases}, \quad P \approx \hat{P} = softmax(\frac{-d(sort(s)\mathbb{I}^T, \mathbb{I}s^T}{\tau})$$

We used a simple linear layer to project the quantized graph encoding from FSQ to the embedding dimension and defined sinusoidal positional embedding [3] for the nodes.

$$\mathcal{L}_{rec} = \frac{1}{N} \sum_{i=1}^{N} (1 - \frac{m_{node}^T \hat{m}_{node}}{||m_{node}||.||\hat{m}_{node}||}) + ||A_\pi - \sigma(\hat{m}_{edge}.\hat{m}_{edge}^T)||^2$$

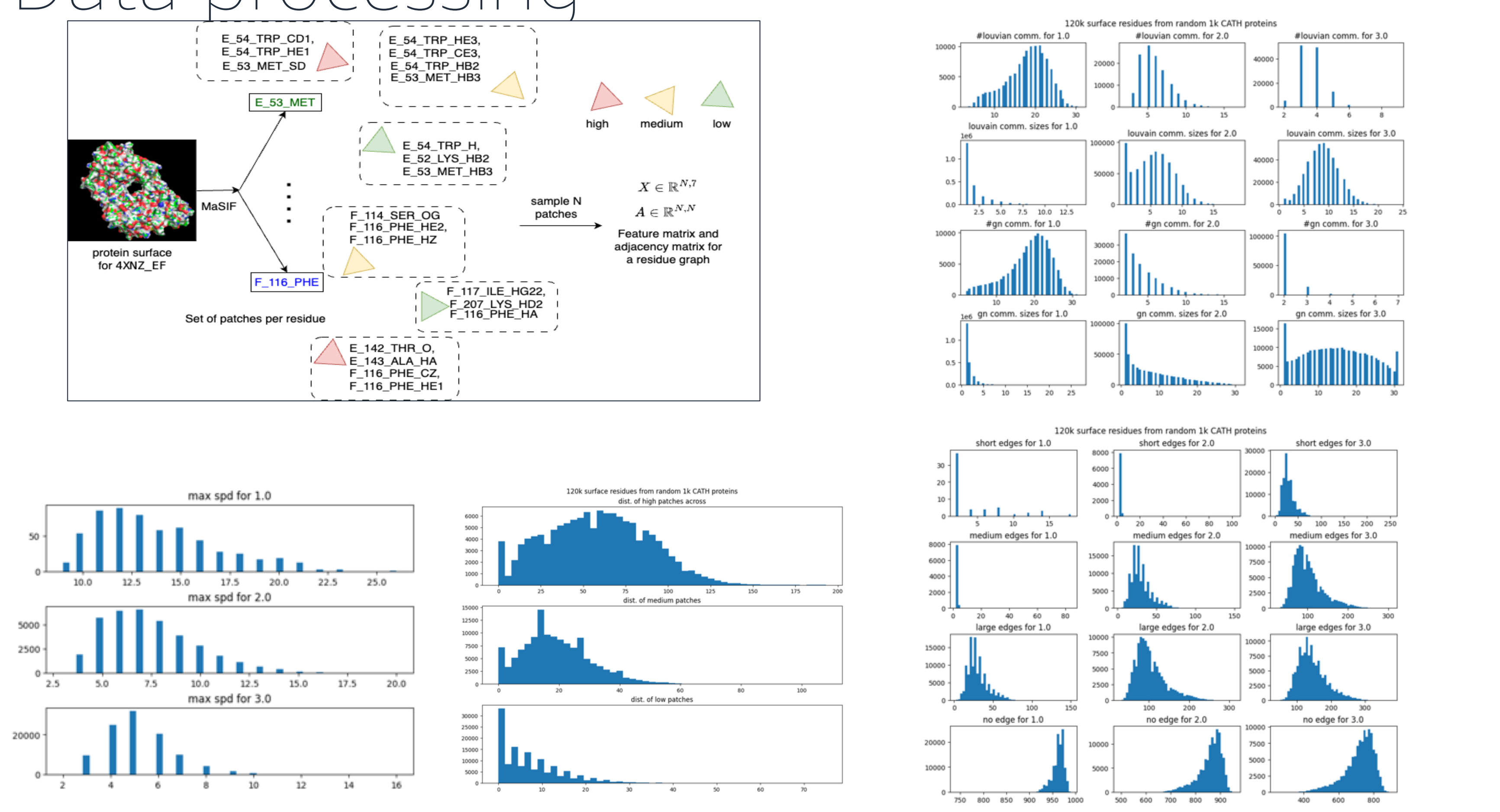
where $\hat{m}_{node}$ is used to reconstruct initial node features, $m_{node}$and $\hat{m}_{edge}$ is used to reconstruct the undirected adjacency matrix A_π

Experiments

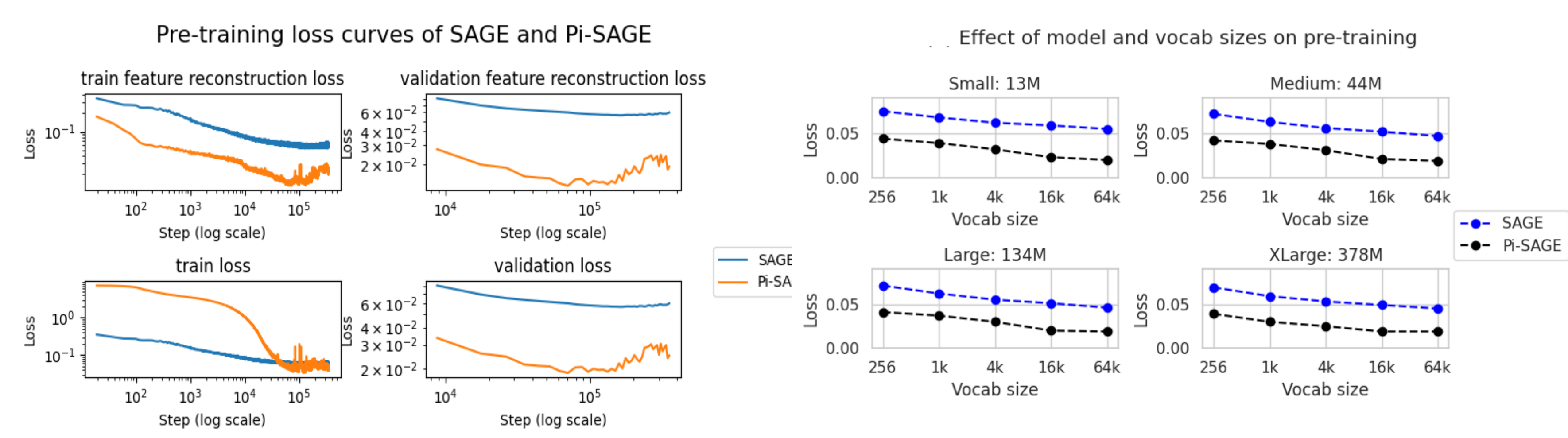
We evaluate Pi-SAGE on the SKEMPI v2.0 dataset on two-stages. ***1) Pretraining***: Pi-SAGE is trained on ~200K protein structures from the RCSB PDB to learn a surface codebook. ***2) Finetuning***: The surface tokenizer is applied to SKEMPI complexes, and surface tokens are integrated into the GearBind model for ΔΔG prediction.

Model	#layers	#heads	hdim	#params
Small	2	2	512	13M
Medium	4	4	768	44M
Large	8	8	1024	134M
XLarge	16	16	1280	378M

Data processing

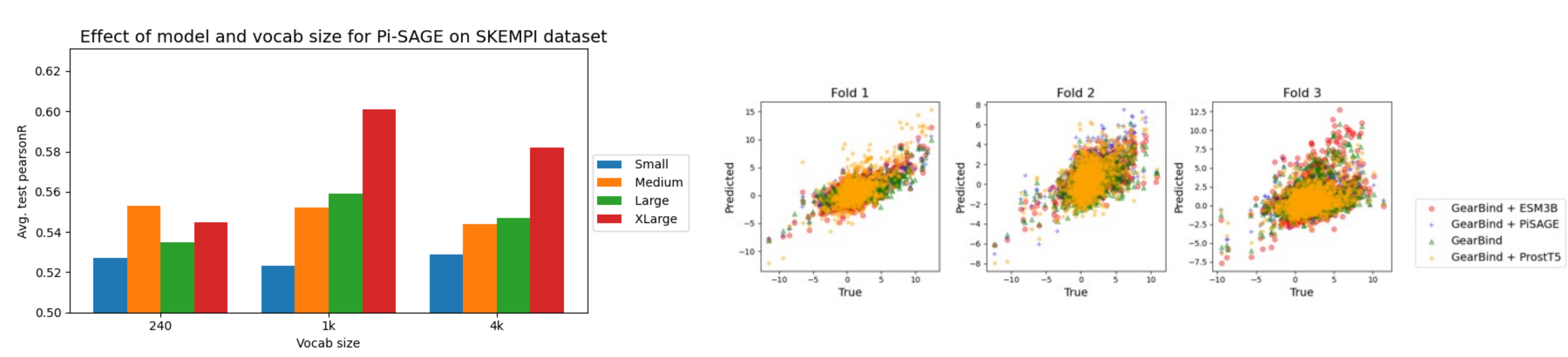


Results: Pre-training



Results: Performance on SKEMPI

Model	Per structure		Overall				
	Pearson ↑	Spear. ↑	Pearson ↑	Spear. ↑	RMSE ↓	MAE ↓	AUROC ↑
Gearbind	0.365 +/- 0.082	0.299 +/- 0.053	0.525 +/- 0.106	0.372 +/- 0.035	1.921 +/- 0.277	1.403 +/- 0.208	0.650 +/- 0.006
+ ESM150M	0.378 +/- 0.050	0.326 +/- 0.047	0.563 +/- 0.088	0.400 +/- 0.014	1.866 +/- 0.259	1.359 +/- 0.209	0.655 +/- 0.028
+ ESM650M	0.381 +/- 0.063	0.316 +/- 0.052	0.539 +/- 0.096	0.377 +/- 0.047	1.852 +/- 0.226	1.349 +/- 0.170	0.652 +/- 0.032
+ ESM3B	0.418 +/- 0.088	0.338 +/- 0.067	0.567 +/- 0.057	0.425 +/- 0.039	1.834 +/- 0.144	1.331 +/- 0.114	0.671 +/- 0.026
+ ProtT5	0.376 +/-0.112	0.325 +/- 0.080	0.551 +/- 0.088	0.400 +/- 0.056	1.873 +/- 0.179	1.375 +/- 0.135	0.665 +/- 0.019
+ ProstT5 (seq)	0.372 +/- 0.094	0.316 +/- 0.087	0.540 +/- 0.085	0.390 +/- 0.070	1.90 +/- 0.173	1.401 +/- 0.146	0.660 +/- 0.046
+ ProstT5 (struct)	0.400 +/- 0.076	**0.347 +/- 0.049**	0.545 +/- 0.092	0.408 +/- 0.032	1.953 +/- 0.190	1.436 +/- 0.137	0.662 +/- 0.020
+ SaProt	0.332 +/- 0.092	0.268 +/- 0.071	0.527 +/- 0.065	0.362 +/- 0.014	1.948 +/- 0.234	1.439 +/- 0.183	0.659 +/- 0.009
+ SAGE	0.386 +/- 0.082	0.314 +/- 0.068	0.546 +/- 0.114	0.383 +/- 0.039	1.864 +/- 0.246	1.350 +/- 0.176	0.660 +/- 0.013
+ Pi-SAGE	**0.423 +/- 0.091**	0.345 +/- 0.077	**0.600 +/- 0.084**	**0.428 +/- 0.038**	**1.817 +/- 0.241**	**1.306 +/- 0.200**	**0.691 +/- 0.026**



Results: Ablation studies

Pi-SAGE	Per structure		Overall				
	Pearson ↑	Spear. ↑	Pearson ↑	Spear. ↑	RMSE ↓	MAE ↓	AUCROC ↑
- Finetune	0.386 +/- 0.071	0.321 +/- 0.052	0.549 +/- 0.101	0.400 +/- 0.048	1.883 +/- 0.191	1.355 +/- 0.134	0.67 +/- 0.025
+ Finetune	0.423 +/- 0.091	0.345 +/- 0.077	0.600 +/- 0.084	0.428 +/- 0.038	1.817 +/- 0.241	1.306 +/- 0.200	0.691 +/- 0.026
+ VQ	0.359 +/- 0.078	0.281 +/- 0.053	0.512 +/- 0.105	0.353 +/- 0.013	1.998 +/- 0.277	1.477 +/- 0.232	0.634 +/- 0.007

References

- [1] Park, W., Chang, W., Lee, D., Kim, J., and Hwang, S.w. Grpe: Relative positional encoding for graph transformer. arXiv preprint arXiv:2201.12787, 2022
- [2] Mentzer, F., Minnen, D., Agustsson, E., and Tschannen, M.Finite scalar quantization: Vq-vae made simple. arXiv preprint arXiv:2309.15505, 2023.
- [3] Winter, R., No´ e, F., and Clevert, D.-A. Permutation-invariant variational autoencoder for graph-level representation learning. Advances in Neural Information Processing Systems, 34:9559–9573, 2021.
- [4] Yang, L., Tian, Y., Xu, M., Liu, Z., Hong, S., Qu, W.,Zhang, W., Bin, C., Zhang, M., and Leskovec, J. Vqgraph: Rethinking graph representation space for bridging gnns and mlps. In The Twelfth International Conference on Learning Representations, 2024.
- [5] Prillo, S. and Eisenschlos, J. Softsort: A continuous relaxation for the argsort operator. In International Conference on Machine Learning, pp. 7793–7802. PMLR, 2020.
- [6] Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods, 17(2): 184–192, 2020.

 *Corresponding author: sharmiba@amazon.com