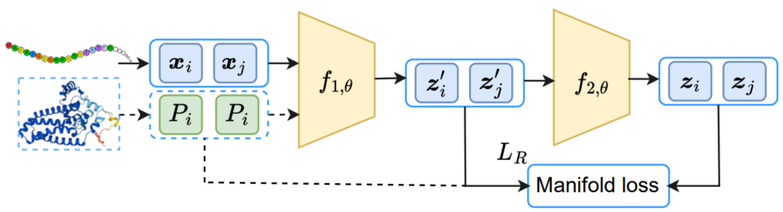
Deep Manifold Transformation For Protein Representation Learning

Bozhen Hu[1,2], Zelin Zang[2,1], Cheng Tan[2,1], Stan Z. Li[2*]
[1] Zhejiang University, [2] Westlake University, ⋆ denotes the corresponding author

Backgrounds

Protein representation learning is critical in various tasks in biology, such as drug design and protein structure or function prediction, which has primarily benefited from the protein language models and graph neural networks. These models can capture intrinsic patterns from protein sequences and structures by masking and task-related losses. However, the learned protein representations are usually not constrained, leading to performance degradation due to data scarcity, task adaptation, etc. We propose a novel Deep Manifold Transformation method for universal Protein Representation Learning (DMTPRL), which adopts manifold learning strategies to improve the quality and adaptation of learned embeddings. A specifically designed manifold learning loss is applied during training based on the graph node-to-node similarity. The proposed method surpasses state-of-the-art baseline algorithms by a significant margin on different downstream tasks across popular datasets, which validates our solutions.



The pipeline of the proposed model.

Methods

Protein Graph Construction

$$(P_i, \boldsymbol{x}_i) \leftarrow (P_i, \boldsymbol{x}_i) + \Theta, \Theta \sim (\mu_k, \sigma_k^2)$$

$$\boldsymbol{Q}_i = [\boldsymbol{b_i} \quad \boldsymbol{n_i} \quad \boldsymbol{b_i} \times \boldsymbol{n_i}]$$

$$\mathcal{F}(G)_{ij} = (d_{ij}, \boldsymbol{Q}_i^T \cdot \frac{P_i - P_j}{d_{ij}}, \boldsymbol{Q}_i^T \cdot \boldsymbol{Q}_j)$$

$$\boldsymbol{e}_{ij} = \mathrm{Cat}(\mathcal{F}(G)_{ij}, |i-j|)$$

Deep Manifold Transformation [1]

$$d_{ij}^{G_{Z',\mathcal{P}}} = \begin{cases} \|P_i - P_j\| & \text{if } \varepsilon_{ij} \in E \\ \Lambda & \text{otherwise} \end{cases}$$

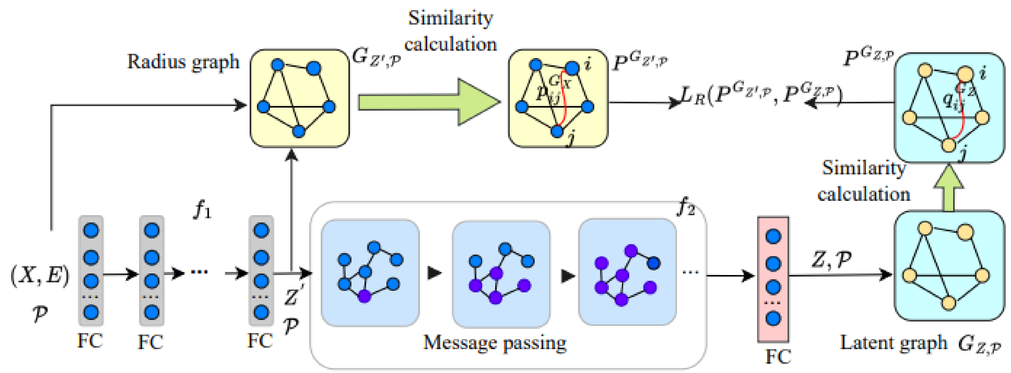
$$p_{ij}^G(\sigma_i, \nu) = g(d_{ij}^G, \sigma_i, \nu) = C_\nu (1 + \frac{d_{ij}^G}{\sigma_i \nu})^{-\frac{(\nu+1)}{2}}$$

$$C_\nu = \sqrt{2\pi} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}$$

$$L_R(a,b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$$

Final Loss

$$L = L_{task} + \beta L_R(P^{G_{Z',\mathcal{P}}}, P^{G_Z})$$



Model architecture

Experiments

Protein-Protein Interaction Identification

Methods	SHS27K		SHS148K	
	BFS	DFS	BFS	DFS
DNN-PPI [18]*	48.90(7.24)	54.34(1.30)	57.40(9.10)	58.42(2.05)
DPPI [18]*	41.43(0.56)	46.12(3.02)	52.12(8.70)	52.03(1.18)
PIPR [19]*	44.48(0.75)	57.80(3.24)	61.83(10.23)	63.98(0.76)
GNN-PPI [20]*	63.81(1.79)	74.72(5.26)	71.37(5.33)	82.67(0.85)
SemiGNN-PPI [17]	72.15(2.87)	78.32(3.15)	71.78(3.56)	**85.45**(1.17)
KeAP [16]†	75.74(5.14)	79.39(3.52)	73.19(7.13)	82.54(2.10)
DMTPRL (Ours)	**77.68**(2.19)	**79.44**(2.52)	**76.81**(4.83)	83.34(2.13)

Reference

[1] Zang Z, Li S, Wu D, et al. DLME: Deep Local-flatness Manifold Embedding. In: *ECCV*, 2022

Email: hubozhen@westlake.edu.cn