DePLM: Denoising Protein Language Models for Property Optimization

Zeyuan Wang[1,2], Keyan Ding[2], Ming Qin[1,2], Xiaotong Li[1,2], Xiang Zhuang[1,2], Yu Zhao[4], Jiahua Yao[4], Qiang Zhang[3], Huajun Chen[1,2]

[1]College of Computer Science and Technology, Zhejiang University
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center
[3]The ZJU-UIUC Institute, International Campus, Zhejiang University
[4]Tencent AI Lab, Tencent

Background

Protein Optimization

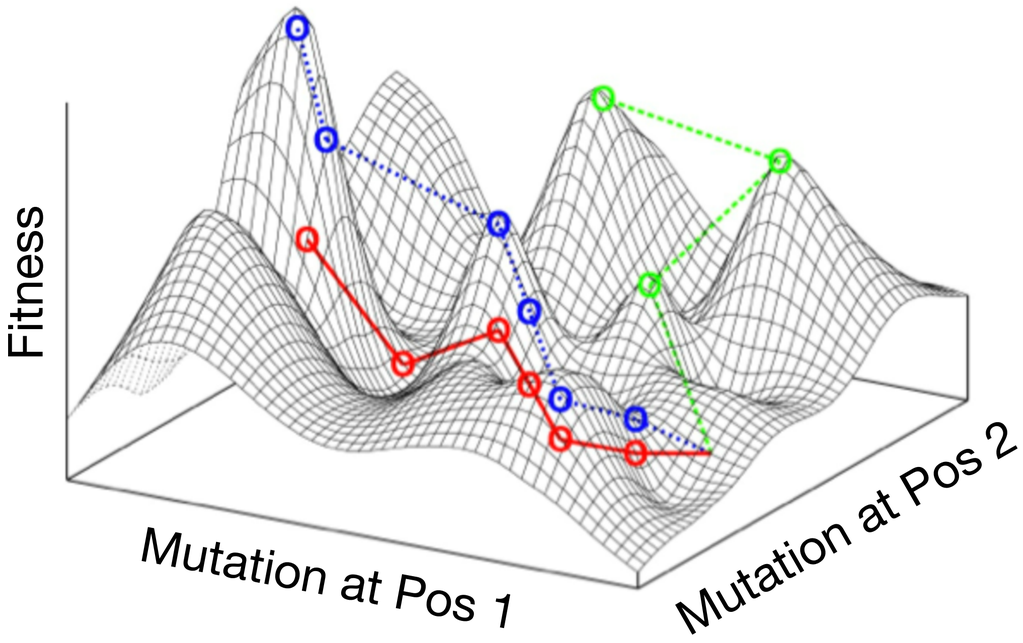
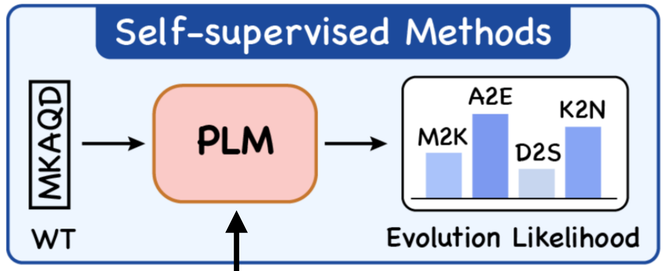
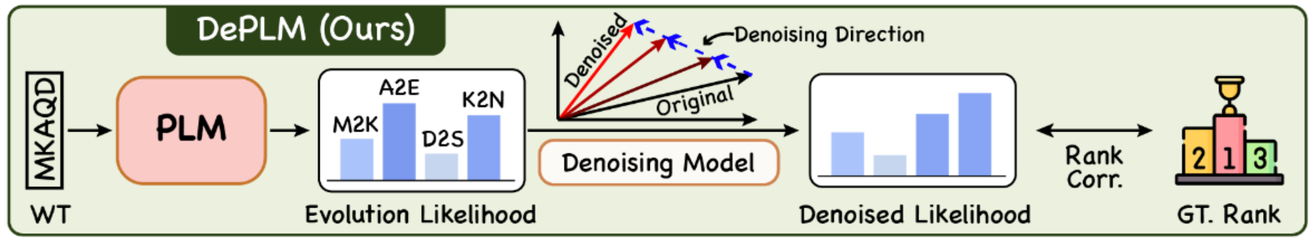


Figure 1: The overview of protein fitness landscape.

$$\texttt{fitness}(x^{\texttt{mt}}) \simeq \sum_{m \in M} \log p(x_m = x_m^{\texttt{mt}} \,|\, x_{/M}^{\texttt{wt}}) - p(x_m = x_m^{\texttt{wt}} \,|\, x_{/M}^{\texttt{wt}})$$



$$\mathscr{L}_{\texttt{MLM}} = \mathbb{E}_{x \sim X} \left[\mathbb{E}_M \left[\sum_{m \in M} -\log p(x_m \,|\, x_{/M}) \right] \right]$$



Method

Rank-based Denosing Diffusion Process

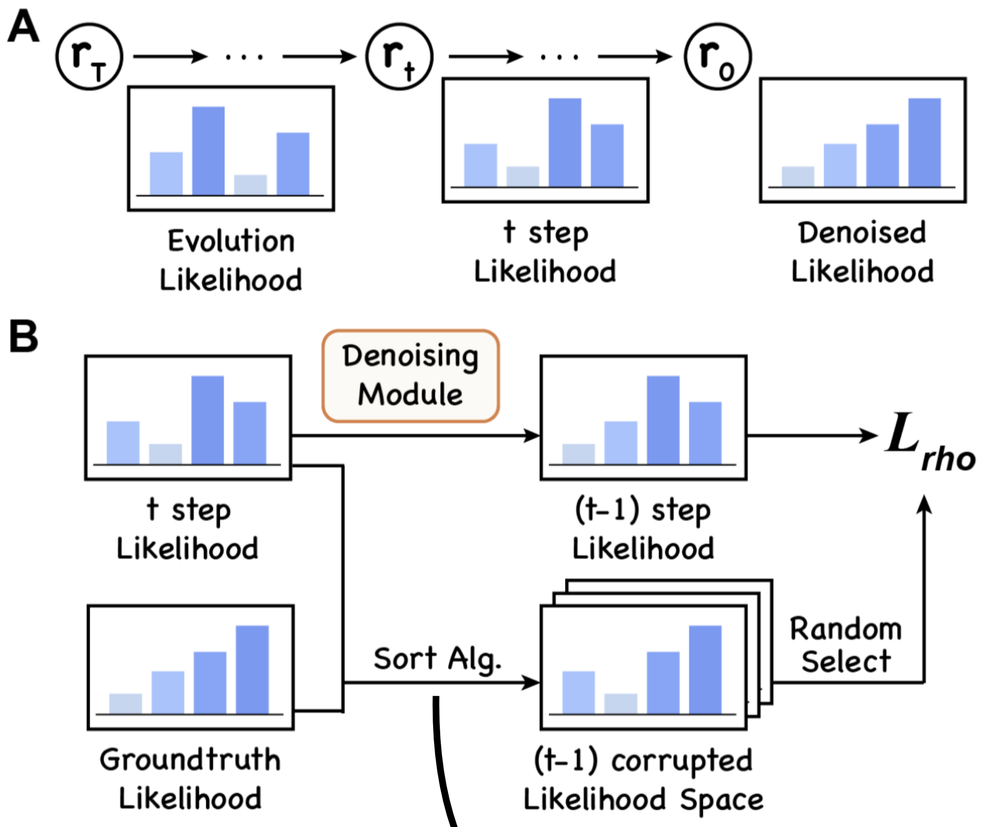


Figure 2: The training process of DePLM

Algorithm 1 Constructing the Space of Rank Variables

Data: The ranks of likelihoods at time steps t_1 and t_2 (where $t_1 < t_2$), represented by $\boldsymbol{r}_{t_1}$ and $\boldsymbol{r}_{t_2}$; the number of sampling trajectories η.

Result: The feasible space of rank variables $\mathbb{S}^{\boldsymbol{r}}_{t_1:t_2}$ between $\boldsymbol{r}_{t_1}$ and $\boldsymbol{r}_{t_2}$.

$\mathbb{S}^{\boldsymbol{r}}_{t_1:t_2} \leftarrow \varnothing, \xi \leftarrow \varnothing, i \leftarrow 0$. // Variable initialization

Compute sorting index $\boldsymbol{I}_{t_1}$ so that $\boldsymbol{r}_{t_1}[\boldsymbol{I}_{t_1}]$ is monotonically increasing and $\boldsymbol{r}_{t_1}[\boldsymbol{I}_{t_1}][\boldsymbol{I}_{t_1}^{-1}] = \boldsymbol{r}_{t_1}$.

$\boldsymbol{r}_{t_1} \leftarrow \boldsymbol{r}_{t_1}[\boldsymbol{I}_{t_1}], \boldsymbol{r}_{t_2} \leftarrow \boldsymbol{r}_{t_2}[\boldsymbol{I}_{t_1}]$.

$\xi \leftarrow \xi \cup \{[0, \text{len}(\boldsymbol{r}_{t_1}) - 1]\}$ // Set left index ϕ to 0 and right index ψ to $\text{len}(\boldsymbol{r}_{t_1}) - 1$

while $i < \eta$ **do**

  while Stack $\neq \varnothing$ **do**

    $\tau \leftarrow \varnothing$.

    for $[\phi, \psi] \in \xi$ **do**

      $\boldsymbol{r}_{t_2}, \varphi = \text{Sort}(\boldsymbol{r}_{t_2}, \phi, \psi)$ // No element in $[\phi, \varphi]$ is greater than any element in $[\varphi, \psi]$.

      $\tau \leftarrow \tau \cup \{[\phi, \varphi - 1]\}$ if $\phi < \varphi - 1$.

      $\tau \leftarrow \tau \cup \{[\varphi + 1, \psi]\}$ if $\psi > \varphi + 1$.

    end for

    $\xi \leftarrow \tau$

    $\mathbb{S}^{\boldsymbol{r}}_{t_1:t_2} \leftarrow \mathbb{S}^{\boldsymbol{r}}_{t_1:t_2} \cup \boldsymbol{r}_{t_2}[\boldsymbol{I}_{t_1}^{-1}]$

  end while

  $i \leftarrow i + 1$

end while

Method

Denoising Module

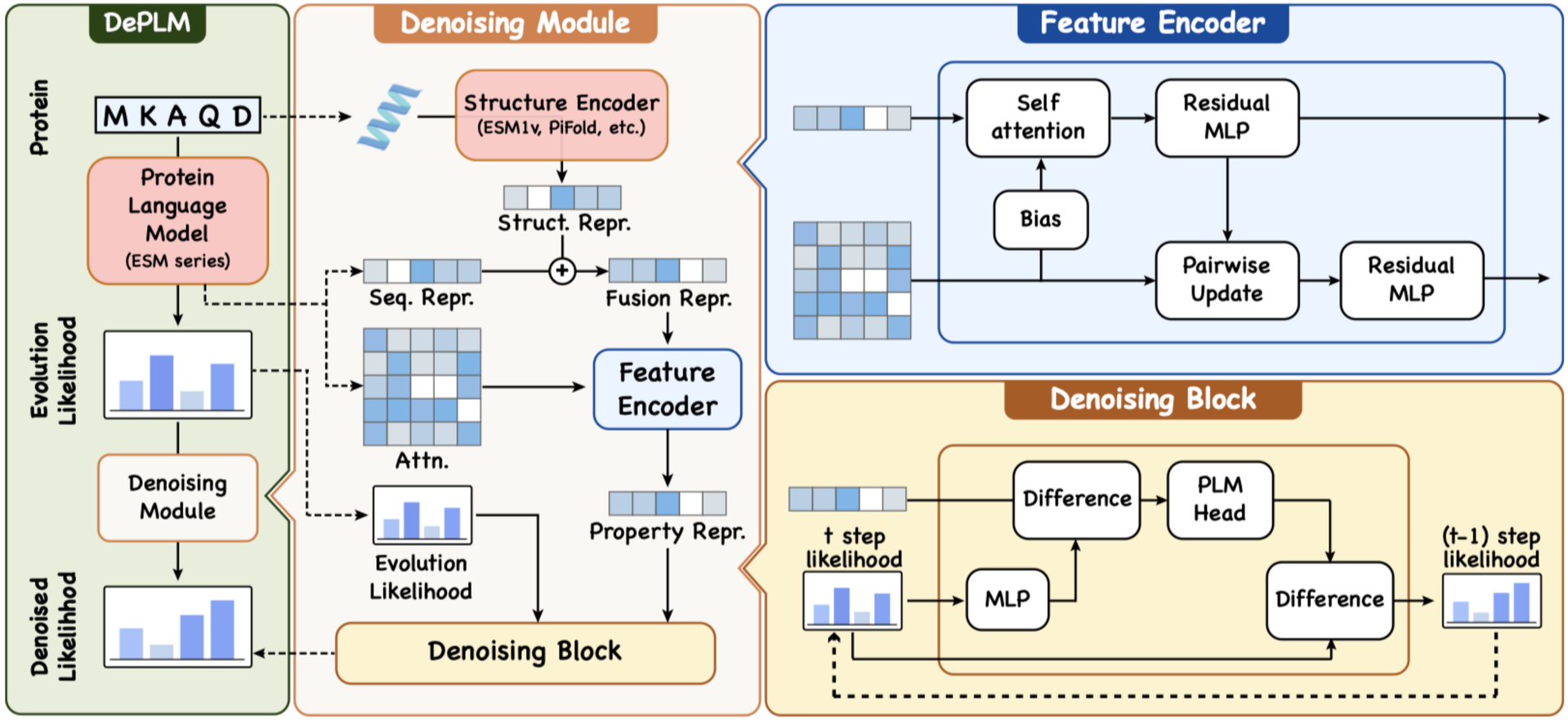


Figure 3: The architecture overview of DePLM.

Results

Performance comparison

Table 1: Model performance on protein engineering tasks. The **best** and suboptimal results are labeled with bold and underline, respectively. ProteinGym results of OHE, ESM-MSA, Tranception, and ProteinNPT are borrowed from Notin et al. [46]. Other results are obtained by our own experiments.

Model	ProteinGym					β-lact.	GB1	Fluo.
	Stability	Fitness	Expression	Binding	Activity			
CNN	0.788	0.588	0.627	0.599	0.573	0.781	0.502	**0.682**
ResNet	0.734	0.489	0.521	0.525	0.481	0.152	0.133	0.636
LSTM	0.745	0.413	0.477	0.496	0.408	0.139	-0.002	0.494
Transformer	0.560	0.149	0.156	0.172	0.155	0.261	0.271	0.643
OHE	0.718	0.545	0.573	0.562	0.555	0.823	0.533	0.657
ESM-1v	0.880	0.566	0.642	0.596	0.572	0.536	0.394	0.438
ESM-2	0.882	0.573	0.645	0.587	0.576	-	-	-
ESM-MSA	0.885	0.568	0.632	0.565	0.600	-	-	-
ProtSSN	0.877	0.692	0.718	0.757	0.678	-	-	-
SaProt	0.882	0.686	0.716	0.749	0.677	-	-	-
Tranception	0.871	0.632	0.704	0.671	0.623	-	-	-
ProteinNPT	**0.904**	0.668	0.736	0.706	0.680	-	-	-
DePLM (ESM1v)	0.887	0.704	0.738	**0.773**	0.688	0.900	**0.676**	0.662
DePLM (ESM2)	0.897	**0.707**	**0.742**	0.764	**0.693**	**0.904**	0.665	0.662

Results

Generalization ability

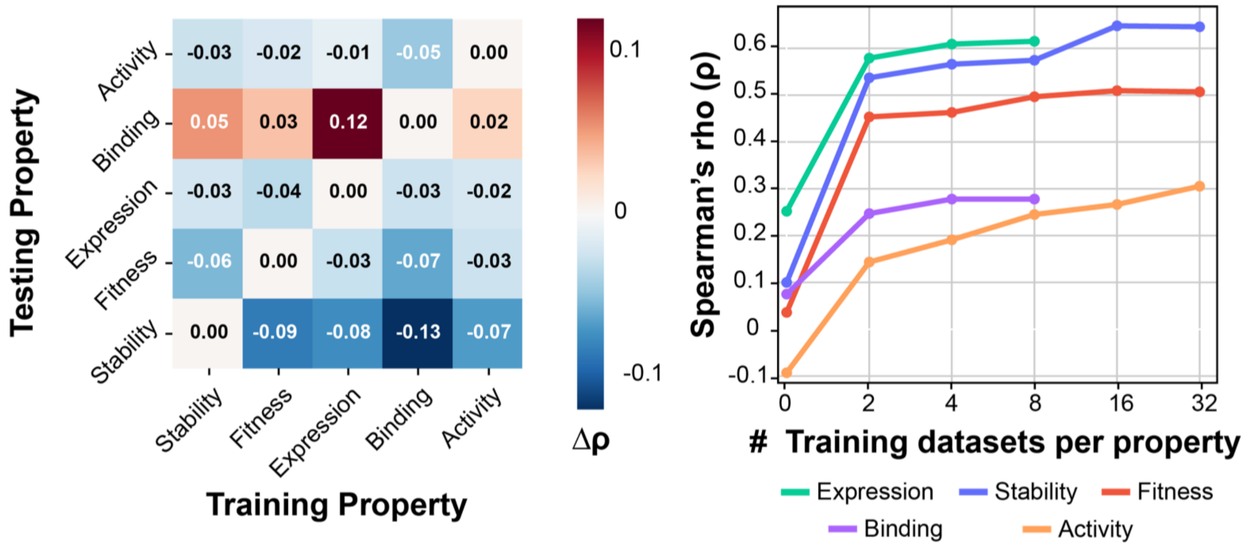
Table 2: Generalization ability evaluation. The **best** and suboptimal results are labeled with bold and underline, respectively. The information (evolutionary, structural or experimental) involved in each model is provided. Results of unsupervised methods are borrowed from Notin et al.[43]. Other results are obtained by our own experiments. (FT=Fine-tuned version)

Model	Information			ProteinGym				
	Evo.	Struct.	Exp.	Stability	Fitness	Expression	Binding	Activity
ESM1v	✓			0.437	0.395	0.427	0.287	0.415
ESM2	✓			0.523	0.396	0.439	0.356	0.433
ProtSSN	✓	✓		0.560	0.408	0.435	0.362	0.458
TranceptEVE L	✓			0.500	0.477	0.457	0.360	0.487
ESM-IF		✓		0.624	0.346	0.436	0.380	0.412
ProteinMPNN		✓		0.564	0.166	0.209	0.159	0.203
CNN			✓	0.141	0.053	0.043	0.056	0.095
ESM1v (FT)	✓		✓	0.497	0.318	0.301	0.216	0.385
ESM2 (FT)	✓		✓	0.454	0.359	0.338	0.276	0.391
ProtSSN (FT.)	✓	✓	✓	0.689	0.448	0.478	0.421	0.488
SaProt (FT.)	✓	✓	✓	0.703	0.442	0.496	0.391	0.495
DePLM (ESM1v)	✓	✓	✓	0.763	0.467	0.506	0.409	0.499
DePLM (ESM2)	✓	✓	✓	**0.773**	**0.480**	**0.510**	**0.441**	**0.518**

Results

Ablation study

Figure4. Visualization of the impact of optimization targets and size of training data on performance



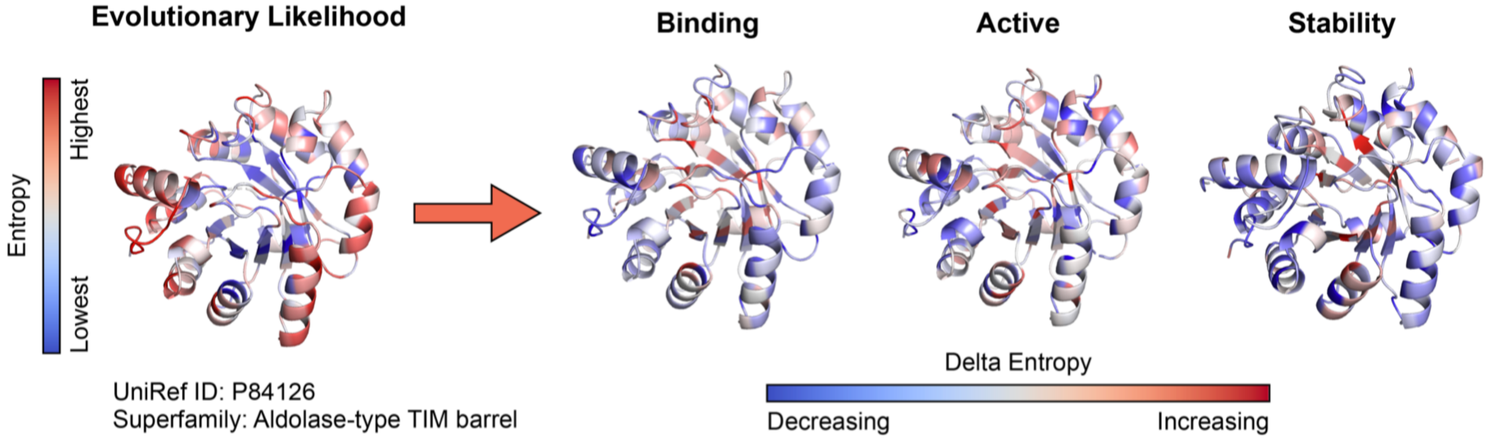


Figure 5.Visualization of the impact of denoising process on the evolutionary likelihood

Thanks!

Codes are available at https://github.com/HICAI-ZJU/DePLM

Email: yuanzew@zju.edu.cn