PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding (*Supplementary Material*)

Minghao Xu[1,2 *] **Zuobai Zhang**[1,2 *] **Jiarui Lu**[1,2] **Zhaocheng Zhu**[1,2]
Yangtian Zhang[3] **Chang Ma**[4] **Runcheng Liu**[5] **Jian Tang**[1,6,7 †]
[*]equal contribution [†]corresponding author
[1]Mila - Québec AI Institute [2]Université de Montréal [3]Shanghai Jiao Tong University
[4]Peking University [5]Tsinghua University [6]HEC Montréal [7]CIFAR AI Research Chair
contacts: <minghao.xu, zuobai.zhang>@mila.quebec, jian.tang@hec.ca

1 Baseline Model Details

Dipeptide Deviation from Expected Mean (DDE) [5]. The DDE protein sequence feature vector is defined by the statistical features of dipeptides, *i.e.*, two consecutive amino acids in the protein sequence. The 400-dimensional feature vector corresponds to the feature of 400 different types of dipeptides. For example, the feature of dipeptide "*st*" is defined by its dipeptide composition (D_c), theoretical mean (T_m) and theoretical variance (T_v) as below:

$$D_c(s,t) = \frac{N_{st}}{N-1}, \quad T_m(s,t) = \frac{C_s C_t}{C_N^2}, \quad T_v(s,t) = \frac{T_m(s,t)\big(1 - T_m(s,t)\big)}{N-1}, \tag{1}$$

$$\mathrm{DDE}(s,t) = \frac{D_c(s,t) - T_m(s,t)}{\sqrt{T_v(s,t)}}, \tag{2}$$

where N_{st} is the number of dipeptide "*st*" occurring in the protein sequence, N denotes the protein sequence length, C_s and C_t are the number of codons for amino acid s and t, and $C_N = 61$ is the total number of possible codons, excluding three stop codons.

Moran correlation [3]. The Moran feature descriptor defines the distribution of amino acid properties along a protein sequence. Following iFeature [2], we retrieve 8 physicochemical properties $\{P^k\}_{k=1}^8$ from AAindex Database [4] to construct the Moran feature vector, and each property is centralized and normalized before calculation. The Moran feature vector is with $8M$ dimensions (M is the parameter of maximum lag, setting as 30 following iFeature). The feature for the k-th property with lag m is defined as below ($1 \leqslant k \leqslant 8$, $1 \leqslant m \leqslant M$):

$$\mathrm{Moran}(k,m) = \frac{\frac{1}{N-m}\sum_{i=1}^{N-m}(P_i^k - \bar{P^k})(P_{i+m}^k - \bar{P^k})}{\frac{1}{N}\sum_{i=1}^{N}(P_i^k - \bar{P^k})^2}, \tag{3}$$

where N denotes the protein sequence length, and $\bar{P^k} = \frac{1}{N}\sum_{i=1}^{N} P_i^k$ is the average of property k over the whole sequence.

2 More Benchmark Results

2.1 Balanced Metrics on Classification Tasks

It should be noted that there are evident class imbalances in two multi-class classification tasks. In particular, on fold classification, the three smallest classes of training, validation and test splits all

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Table 1: Balanced metric (weighted F1) compared with accuracy on multi-class classification tasks. We report *mean (std)* for each experiment. We adopt four color scales of green to denote the first, second, third and fourth best performance among all models.

Task	Feature Engineer		Protein Sequence Encoder				Pre-trained Protein Language Model			
	DDE	Moran	LSTM	Transformer	CNN	ResNet	ProtBert	ProtBert*	ESM-1b	ESM-1b*
Fold (weighted F1)	$0.093_{(0.003)}$	$0.030_{(0.003)}$	$0.070_{(0.014)}$	$0.079_{(0.005)}$	$0.120_{(0.011)}$	$0.091_{(0.017)}$	$0.185_{(0.004)}$	$0.069_{(0.005)}$	$0.298_{(0.011)}$	$0.278_{(0.008)}$
Fold (accuracy)	$0.096_{(0.005)}$	$0.071_{(0.006)}$	$0.082_{(0.016)}$	$0.085_{(0.006)}$	$0.109_{(0.004)}$	$0.089_{(0.015)}$	$0.169_{(0.004)}$	$0.107_{(0.009)}$	$0.282_{(0.021)}$	$0.300_{(0.002)}$
Sub (weighted F1)	$0.485_{(0.002)}$	$0.225_{(0.010)}$	$0.624_{(0.009)}$	$0.538_{(0.009)}$	$0.557_{(0.003)}$	$0.515_{(0.005)}$	$0.765_{(0.009)}$	$0.569_{(0.006)}$	$0.778_{(0.005)}$	$0.792_{(0.002)}$
Sub (accuracy)	$0.492_{(0.004)}$	$0.311_{(0.005)}$	$0.630_{(0.004)}$	$0.560_{(0.008)}$	$0.587_{(0.011)}$	$0.523_{(0.035)}$	$0.765_{(0.009)}$	$0.594_{(0.002)}$	$0.781_{(0.005)}$	$0.798_{(0.002)}$

* Used as a feature extractor with pre-trained weights frozen.

contain only one sample, while the largest ones contain tens or hundreds of samples; on subcellular location prediction, the ratio between the number of samples in the largest and smallest classes is greater than 10. Hence, these two tasks are highly imbalanced. We do not observe class imbalances in other classification tasks.

To reflect the ability of models on these imbalanced settings, we report the results of all baselines with a widely used metric to consider data imbalance, weighted F1 [1]. Weighted F1 is defined by first calculating F1 scores on each class and then taking the weighted average according to the number of instances in each class. The results are shown in Tab. 1. The ranking of baselines under weighted F1 is almost unchanged compared to that under accuracy, where shallow CNN is still the best model among models trained from scratch, and ESM-1b remains the SOTA model on these two tasks. Therefore, the conclusions in Section 5.2 of the main paper still hold. Considering the consistency of experimental conclusions and the comparability with previous benchmark results in the literature where accuracy is commonly reported, we still employ accuracy as the metric for these two tasks in the main paper and provide weight F1 performance in the supplement.

3 Ablation Studies

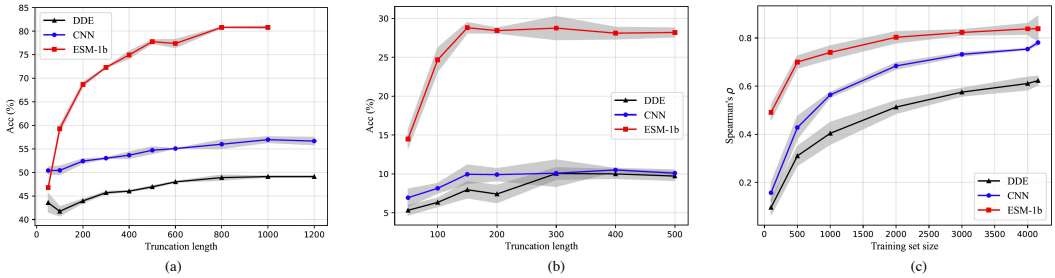

Figure 1: (a) Effect of truncation length on subcellular localization prediction. (b) Effect of truncation length on fold classification. (c) Effect of training set size on β-lactamase activity prediction. All results are averaged over three runs (seeds: 0, 1, 2); the standard deviation is shown by error bar.

3.1 Effect of Truncation Length

In Fig. 1 (a) and (b), we plot the performance of DDE, CNN and ESM-1b on subcellular localization prediction and fold classification under different sequence truncation lengths. The truncation is performed from the start of each protein sequence. It is observed that longer truncation lengths will lead to better performance for all models on both tasks, which matches with the intuition that a longer truncated sequence can contain more information about a protein and thus learn more effective prediction model.

3.2 Effect of Training Set Size

In Fig. 1 (c), we plot the performance of DDE, CNN and ESM-1b on β-lactamase activity prediction under the training set sizes from 100 to 4,158 (the full training set). Training samples are randomly sampled from the full set in required cases. As expected, the performance of all three models

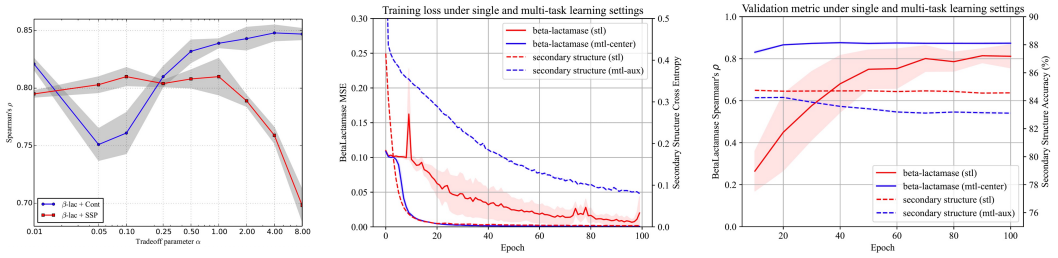

Figure 2: (a) Effect of tradeoff parameter α on the MTL of CNN. (b,c) Training loss curves (b) and validation metric curves (c) on β-lacatamase and secondary structure prediction under single and multi-task settings. All results are averaged over three runs (seeds: 0, 1, 2); the standard deviation is shown by error bar.

monotonously increases as the increase of training set size. These results illustrate the benefit of collecting more labeled data when predicting the fitness landscape of proteins.

3.3 Effect of Tradeoff Parameter

In Fig. 2 (a), we study how the tradeoff parameter α affects the MTL of CNN. When contact prediction serves as the auxiliary task, the center task, β-lactamase activity prediction, is well enhanced by using a larger tradeoff parameter (*i.e.*, $0.5 \leqslant \alpha \leqslant 8.0$). By comparison, when using secondary structure prediction as the auxiliary task, the center task suffers from severe performance decay under large tradeoff parameters, and the peak performance is achieved when α is between 0.1 and 1.0. Both cases suggest $\alpha = 1.0$ as a configuration that achieves stable performance gain, which can be used as a good candidate configuration for MTL.

3.4 Training Center and Auxiliary Tasks under Single- and Multi-Task Learning

In our benchmark experiments, the tradeoff parameter α is generally set to 1. Therefore, it is interesting to see how the optimization of center and auxiliary tasks under the multi-task setting differ from that under the single-task setting. In Fig. 2 (b) and (c), we draw the training loss curves and validation metric curves of β-lacatamase and secondary structure prediction under single- and multi-task learning settings. For multi-task learning, we choose the β-lacatamase as the center task. It can be observed that the model converges faster and generalizes better on the center task under the multi-task setting, which shows the optimization of the center task benefits from the auxiliary task. However, the training of the auxiliary task is worse under multi-task learning. This can be understood since the number of training iterations is determined by the center task, which leads to the insufficient sampling of the auxiliary dataset. Interestingly, we find that the variance of β-lacatamase is largely decreased (almost stable under different seeds) when including secondary structure prediction as the auxiliary task. One potential reason is that the low-variance auxiliary task makes the training of the center task more stable.

4 Broader Societal Impacts

This work focuses on building a comprehensive and multi-task benchmark for protein sequence understanding. In this benchmark, five types of protein understanding tasks are leveraged to evaluate the general effectiveness of protein sequence encoding methods. By evaluating on the proposed benchmark, we can comprehensively assess whether a protein sequence encoding approach could be promising in various real-world applications. Therefore, this benchmark lays a solid foundation for the application of machine learning techniques on pharmaceutical research.

However, it cannot be denied that some harmful activities could be boosted by the powerful models validated by our benchmark, *e.g.*, designing harmful drugs. Therefore, our future works will seek to mitigate these issues by formulating guidelines for the responsible usage of our benchmark.

References

[1] sklearn.metrics.f1 score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

[2] Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14):2499–2502, 2018.

[3] Zhi-Ping Feng and Chun-Ting Zhang. Prediction of membrane protein types based on the hydrophobic index of amino acids. *Journal of Protein Chemistry*, 19(4):269–275, 2000.

[4] Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic Acids Research*, 28(1):374–374, 2000.

[5] Vijayakumar Saravanan and Namasivayam Gautham. Harnessing computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor. *Omics: A Journal of Integrative Biology*, 19(10):648–658, 2015.