Supporting Information

to

PepFormer: End-to-End Transformer based Siamese Network to Predict and Enhance Peptide Detectability based on Sequence Only

Hao Cheng[1,2], Bing Rao[3], Lei Liu[1,2], Lizhen Cui[1,2], Guobao Xiao[4], Ran Su[5*], and Leyi Wei[1,2,4*]

[1]School of Software, Shandong University, Jinan 250101, China

[2]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan 250101, China

[3]School of Mechanical Electronic & Information Engineering, China University of Mining &Technology, Beijing 221008, China;

[4]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, Fuzhou 350000, China；

[5]College of Intelligence and Computing, Tianjin University, Tianjin 300384, China;

*Corresponding author:

L.W: weileyi@sdu.edu.cn

R.S: ran.sdu@tju.edu.cn

Peptide length preference of the proposed model

To investigate whether our model has specific length preference for the peptide detectability task, we compared the length distribution of the samples (true positives and negatives) with the length distribution of the samples predicted by the model, as shown in Figure S1. Here, KL divergence is used to measure the difference between the length distribution of predicted samples and the real samples. As shown in Figure S1, we can see that for both positive and negative samples, the real sample distribution is very similar to the data distribution predicted by the model, and the value of KL divergence is also very small. Specifically, we observed that for both positive (Figure S1A, C) and negative sample (Figure S1, D) distributions, there are more clear gaps between true distribution and predicted distribution around the length interval [10,15], demonstrating that our model is not that good at predicting the peptide detectability in this length interval. Beside the length interval [10,15], our model has no obvious length preference, which shows that it can fully learn the sequence length property in the datasets.

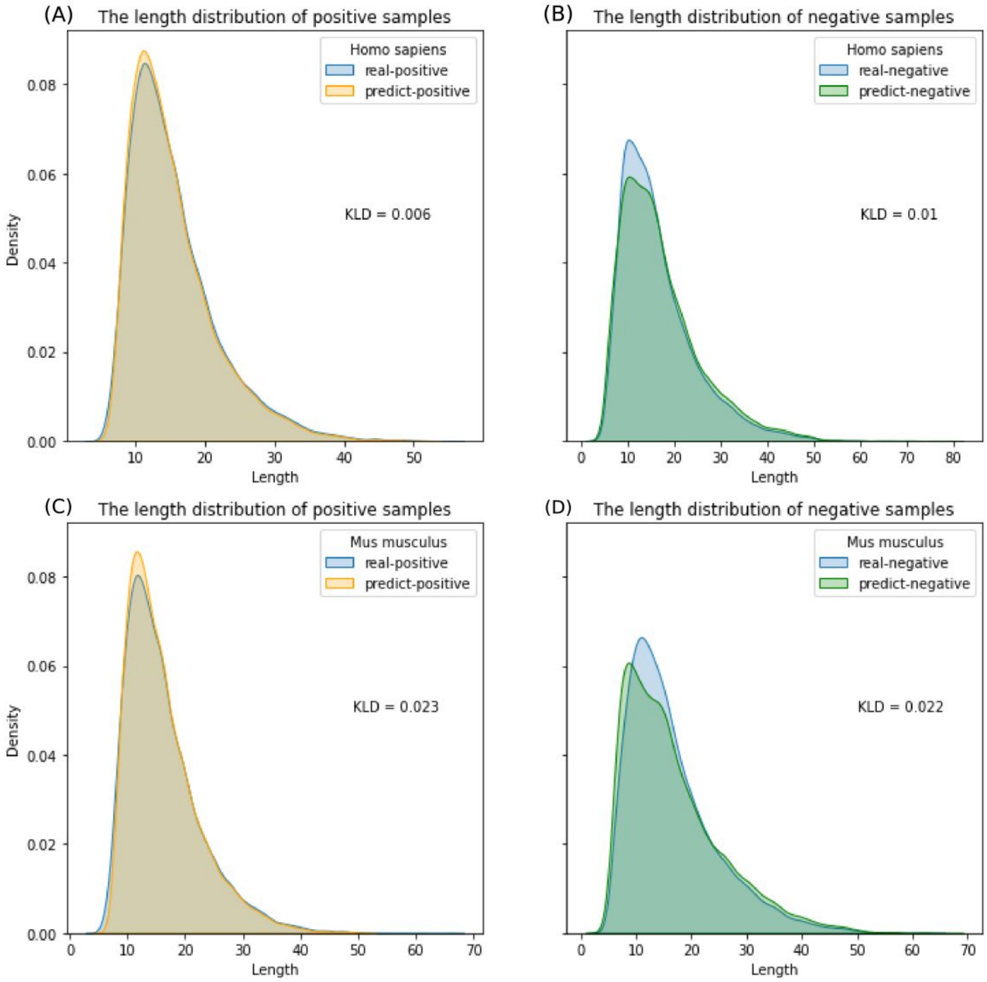

Figure S1. The length distribution of dataset and model prediction results. The "real-positive" and "predict-positive" are the length distribution of positive samples in dataset and model predict results separately. (A) and (B) is the length distribution of positive and negative samples on *Homo sapiens* dataset. (C) and (D) is the length distribution of positive and negative samples on *Mus musculus* dataset.