Machine Learning Dataset and Benchmark for Accurate T Cell Receptor-pHLA Binding Prediction

Xinyuan Zhu*[1], Jiadong Lu*[2], Yeqing Lu[2], Yuyan Zhang[3], and Fuli Feng[2]

[1]School of Information Science and Technology, University of Science and Technology of China
[2]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[3]School of Cyber Science and Technology, University of Science and Technology of China

Supplementary Information

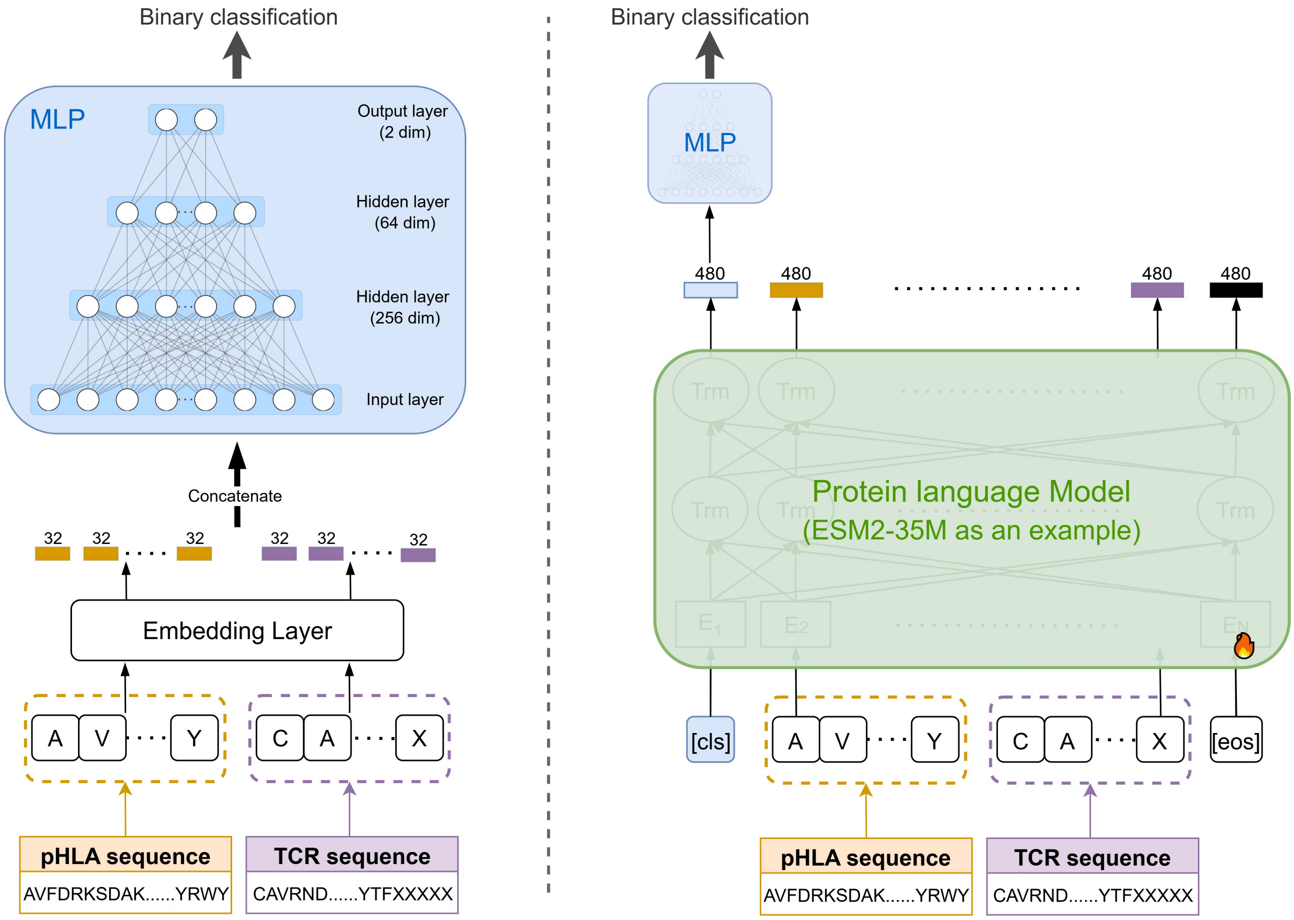
*These authors contributed equally to this work.

Supplementary Table 1: **Statistics of existing datasets for size comparison.** The table details the composition of reference datasets used for a comparative analysis against the Hi-TPH dataset in Figure 2c. For each dataset, the total number of TCR-pHLA pairs, unique peptides, and unique TCRs are enumerated, corresponding to the Hi-TPH level I-IV.

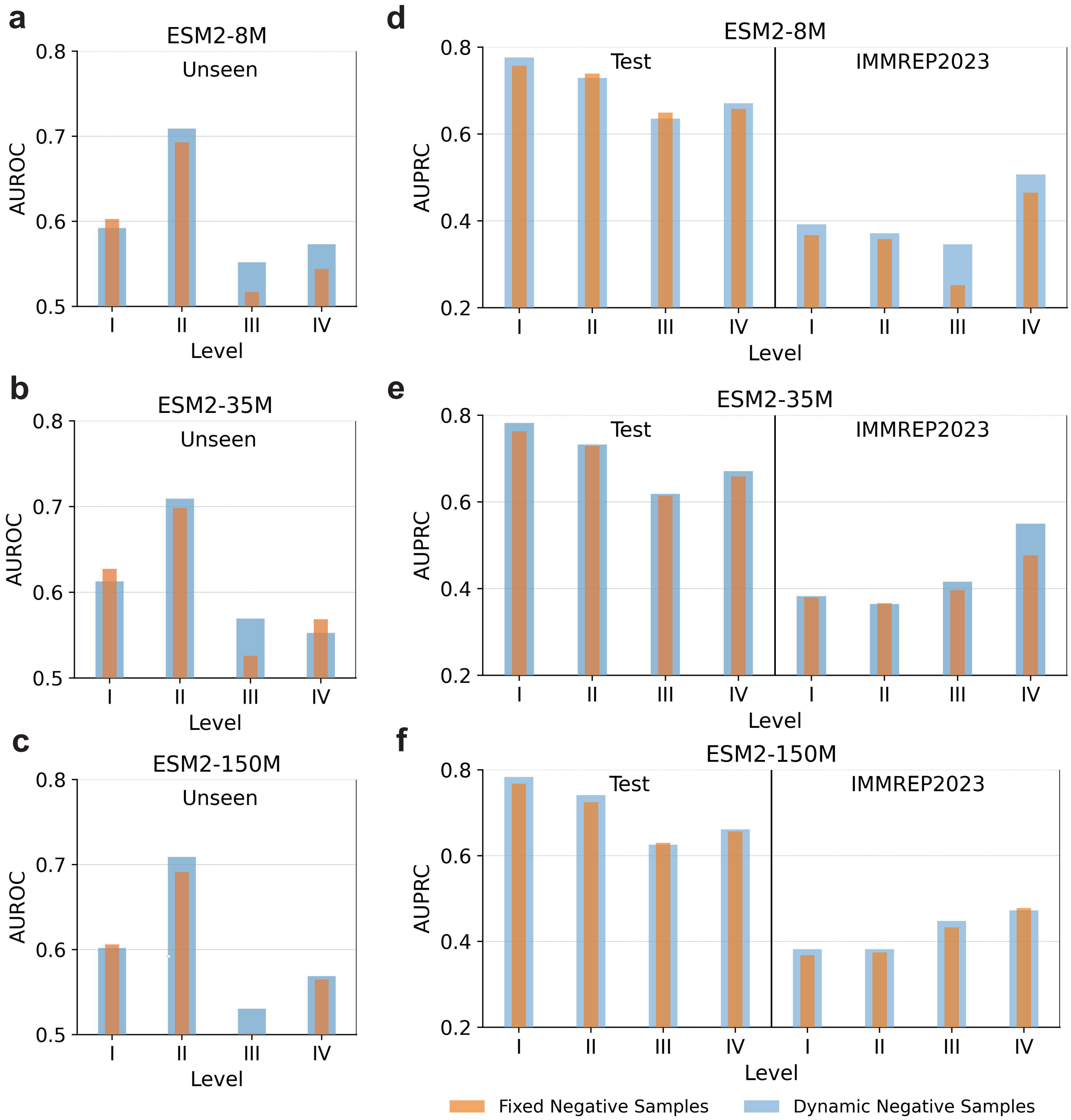
	Pairs	Peptides	TCRs
Level I	UnifyImmun (137,400)	UnifyImmun (1,488)	UnifyImmun (128,169)
Level II	epiTCR (66,471)	epiTCR (1,391)	epiTCR (61,159)
Level III	TCRAI (8,130)	STAPLER (604)	TCRAI (8,101)
Level IV	STAPLER (4,457)	STAPLER (604)	STAPLER (4,253)

Supplementary Table 2: **AUROC performance of Hi-TPH-trained models on internal evaluation sets.** This table highlights the performance difference between the in-distribution **Test set** and the **Unseen set** (containing strictly novel peptides). Previously published tools are excluded from this analysis to prevent confounding results from data leakage, as their original training data may overlap with our evaluation splits.

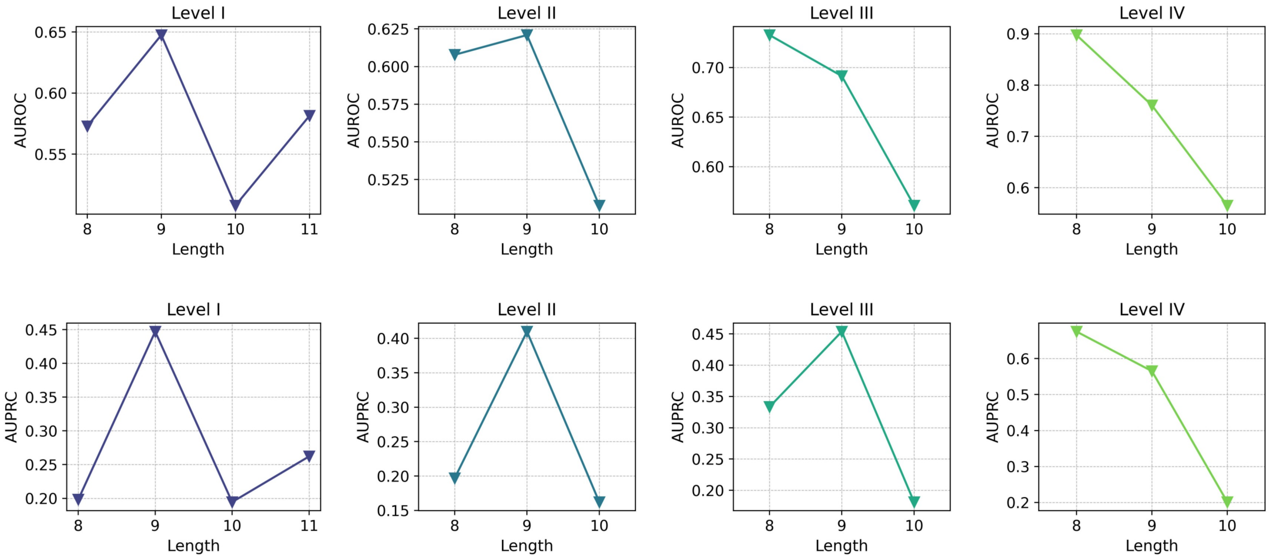
Model	Level I		Level II		Level III		Level IV	
	Test	Unseen	Test	Unseen	Test	Unseen	Test	Unseen
MLP	0.7592	*0.6053*	0.7404	0.6878	0.6370	0.5317	0.6509	0.5687
ESM2-8M	0.7854	0.5919	0.7498	0.7089	*0.6391*	0.5517	0.6601	*0.5729*
ESM2-35M	**0.7892**	**0.6125**	0.7503	*0.7093*	0.6194	**0.5692**	*0.6655*	0.5526
ESM2-150M	*0.7889*	0.6018	**0.7567**	0.7089	0.6251	0.5303	0.6589	0.5686
TAPE-BERT	0.7839	0.5919	0.7494	0.7015	**0.6497**	0.5521	**0.6671**	0.5583
AMPLIFY-120M-base	0.7688	0.5965	*0.7505*	**0.7130**	0.6362	0.5605	0.6441	0.5499
AMPLIFY-120M	0.7635	0.6040	0.7492	0.7071	0.6155	*0.5612*	0.6411	**0.5801**



Supplementary Figure 1: **Architectures of the MLP (left) and Hi-TPH-PLMs (right) evaluated in our benchmark.** The MLP model first converts the pHLA and TCR amino acid sequences into fixed-dimensional vectors using an embedding layer. These vectors are then concatenated and fed into an MLP with two hidden layers (with 256 and 64 dimensions, respectively). Finally, a 2-dimensional output layer performs the binary classification Hi-TPH-PLMs, which are based on PLMs, concatenate the pHLA and TCR sequences, adding a special classification token [cls] at the beginning and a separator/end token [eos] at the end. The entire sequence is input into a pre-trained PLM (e.g., ESM2-35M). The output representation corresponding to the [cls] token is then extracted and passed to a simple MLP head for binary classification.



Supplementary Figure 2: **Additional performance comparison of models trained with fixed versus dynamic negative sampling across different datasets and model sizes. (a-c)** The plots show the AUROC on the Unseen set for the ESM2-8M (a), ESM2-35M (b), and ESM2-150M (c) models. **(d-f)** The plots show the AUPRC on the Test set and the external IMREP2023 set for the ESM2-8M (d), ESM2-35M (e), and ESM2-150M (f) models.



Supplementary Figure 3: **Evaluation on the IMMREP2023 benchmark reveals that model performance can be influenced by peptide length.** The figure illustrates the performance of the ESM2-35M model, evaluated using the AUROC (top row) and the AUPRC (bottom row) across four levels of the IMMREP2023 benchmark. Performance is stratified by the length of the test peptides. Both metrics demonstrate a performance peak for peptides of length 9 (9-mers). Notably, a marked decrease in performance is observed for peptides of length 10. This trend suggests a potential model bias stemming from the preponderance of 9-mer peptides in the training data, which may limit the model's ability to generalize effectively to peptides of other lengths.