

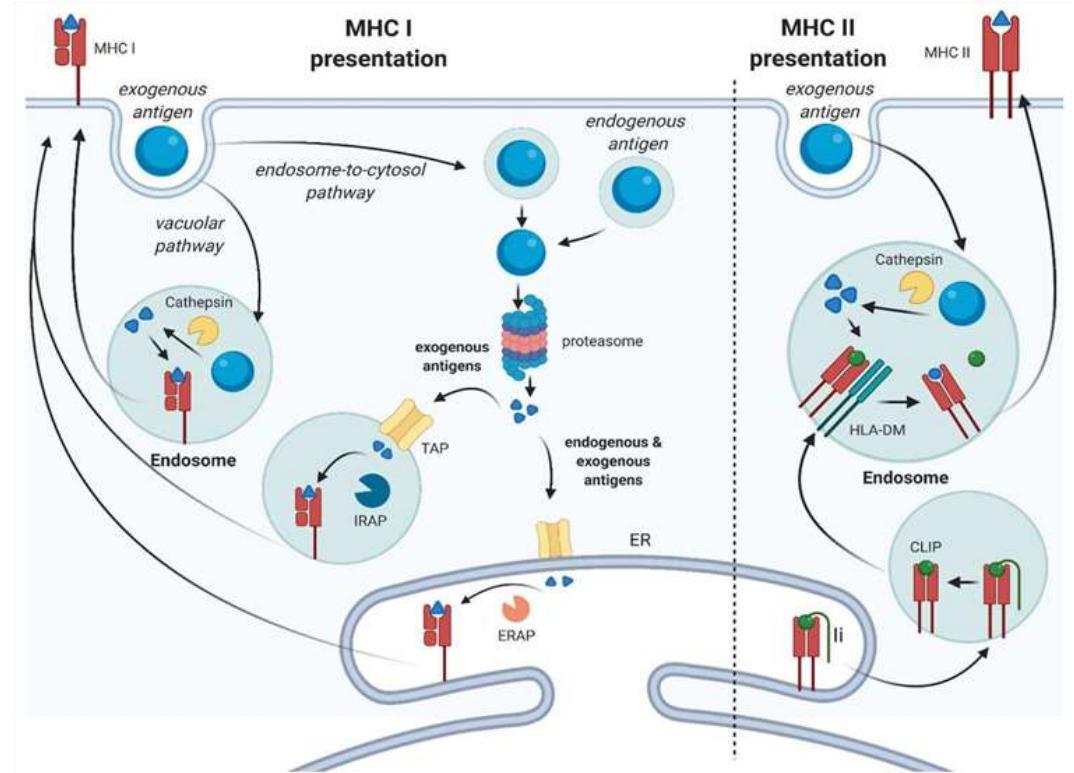
predicting peptide MHC binding and encoding the interaction to generalize over unseen alleles

Amirreza Aleyasin 28.10.2025

Introduction to MHC and peptide-MHC complex (pMHC)

Major Histocompatibility Complex

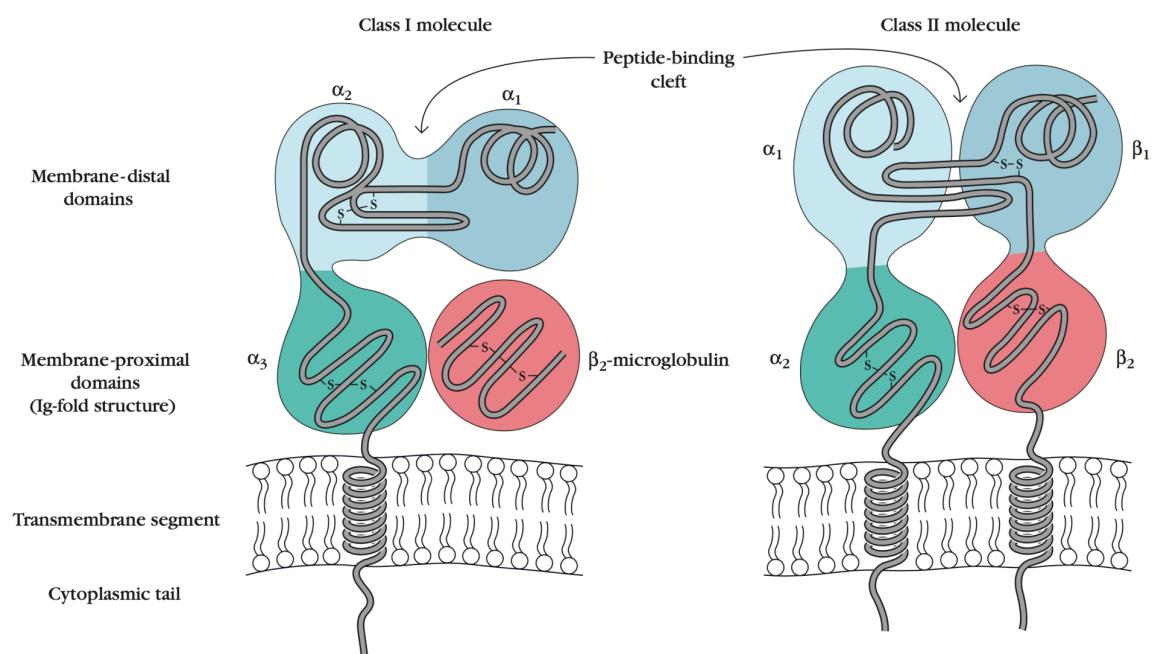
- MHC Class I:
 - present on nearly all nucleated cells,
 - binds to intracellular proteins produced by viruses or mutated in cancer cells
- MHC Class II:
 - primarily expressed on antigen-presenting cells such as dendritic cells, macrophages, and B cells
 - binds to peptides from extracellular pathogens, such as bacteria



Introduction to MHC and peptide-MHC complex (pMHC)

Major Histocompatibility Complex

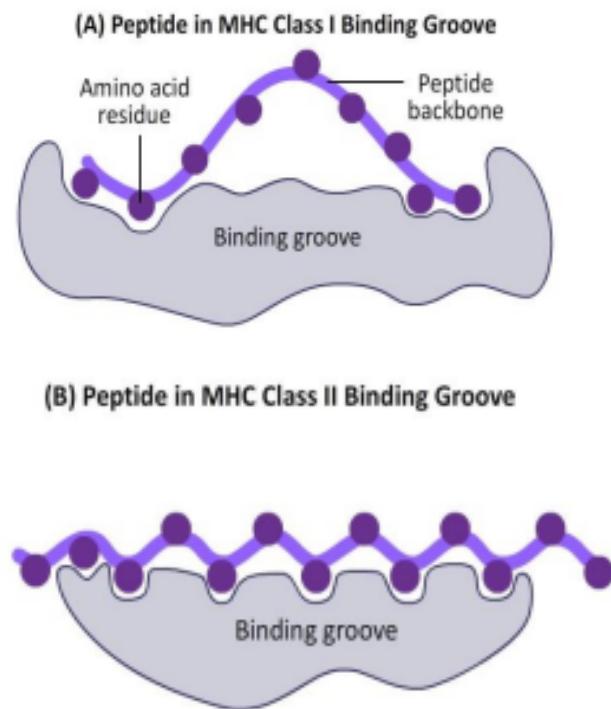
- MHC Class I:
 - heavy chain (alpha chain)
 - alpha1, alpha2, and alpha3. (peptide-binding groove)
 - light chain (beta-2 microglobulin).
- MHC Class II:
 - alpha chain
 - alpha1, alpha2
 - beta chain
 - beta1, beta2



Introduction to MHC and peptide-MHC complex (pMHC)

Major Histocompatibility Complex

- MHC class I:
 - closed groove that typically binds peptides of 8-10 amino acids
- MHC class II:
 - open groove accommodating longer peptides, often 9-30 amino acids



Applications of peptide MHC binding prediction

Accurately predicting the binding of peptides to MHCs help researchers designing tailored therapies against viruses and cancer. (**neoantigen design**)

Why?

- Strong binding helps the immune system fight diseases like cancer or viruses better.
- testing every peptide in the lab would take too long and cost too much, slowing down new treatments.

The number of all 9mer peptides = $9^{20} = 12,157,665,459,056,928,801$

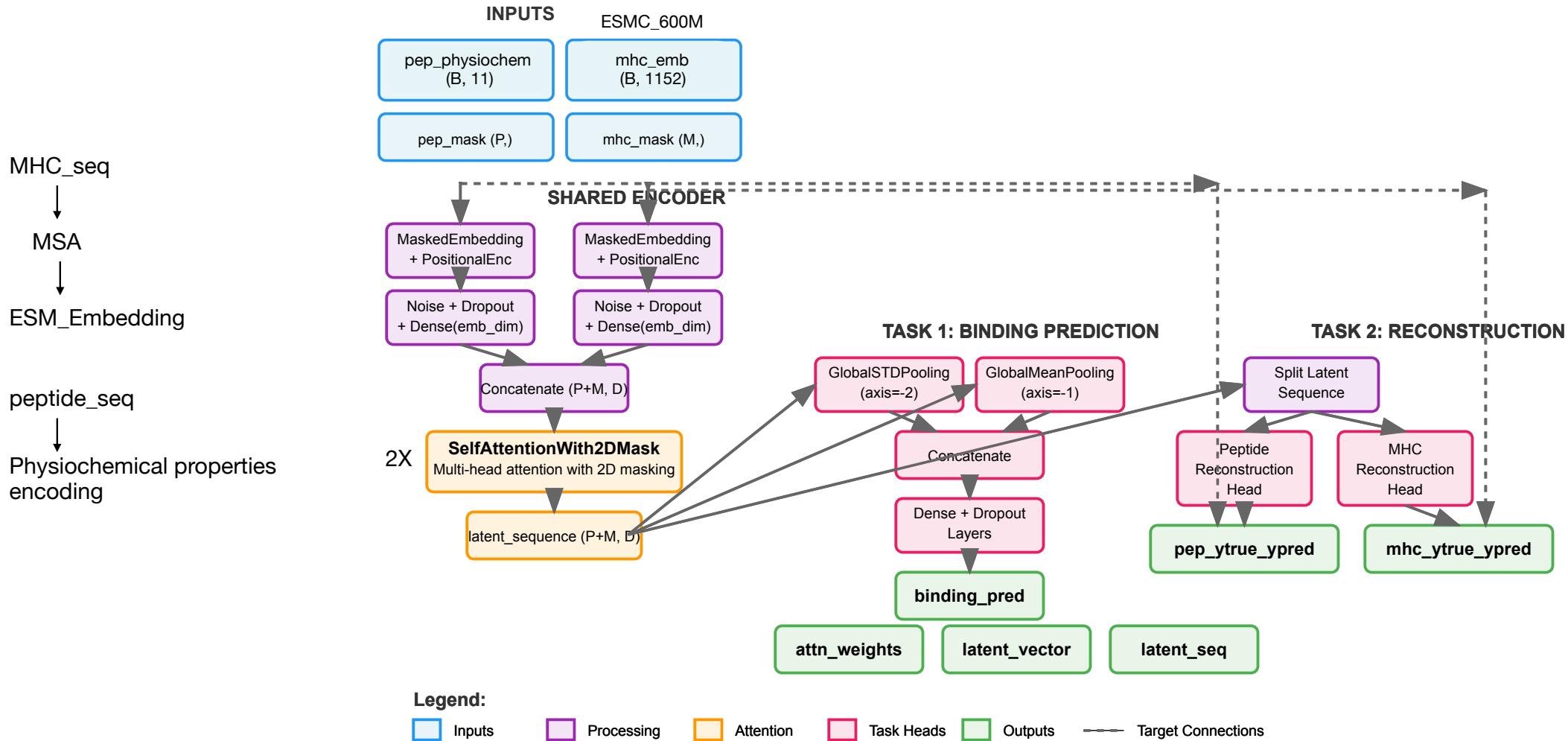
Dataset

Data Source	Positives	Negatives	Alleles	Peptides	Length	Avg.
<i>NetMHCpan Training Data</i>						
NetMHCpan BA 4.1	46,361	—	157	24,777	8–14	9.2
NetMHCpan BA 4.2	1,201	24,342	168	23,180	8–14	10.4
NetMHCpan EL 4.1 (SA)	25,988	—	13	25,746	8–14	9.6
NetMHCpan 4.2 (CEDAR)	926	2,387	89	3,263	8–14	9.5
NetMHCpan 4.2 (IEDB)	5,991	27,768	191	19,340	8–14	9.4
<i>Deep Learning and previous Benchmark Datasets</i>						
CapsNet-MHC-Anthem	163,076	421,546	123	528,290	8–30	9.4
ConvNeXt-MHC	78,354	1,409,981	179	1,438,215	8–14	9.5
HLAB-v1-00	78,588	181,169	125	238,137	8–15	9.5
Pep2Vec	393,838	—	163	251,378	8–12	9.5
RobustMHC-IMGT	188	739	53	906	8–11	9.6
IEDB Weekly 2015–2025	3,170	980	36	3,983	8–11	9.8
<i>Cancer Immunopeptidomics</i>						
TCGA (Xia et al.)	339,183	—	313	303,809	8–11	9.5
<i>Neoantigen-Specific Data</i>						
dbPepNeo (HC)	147	—	27	133	8–13	9.4
dbPepNeo (MC)	51	—	12	51	9–12	9.2
<i>Predicted Negatives</i>						
NetMHCpan pred. negs 4.1	—	40,658,908	202	8,714,986	8–14	10.7
All Sources	1,137,062	42,727,820	477	11,231,758	8–30	10.6

Note: Totals represent unique counts across all datasets.

pMHC Multitask Transformer Model

Method



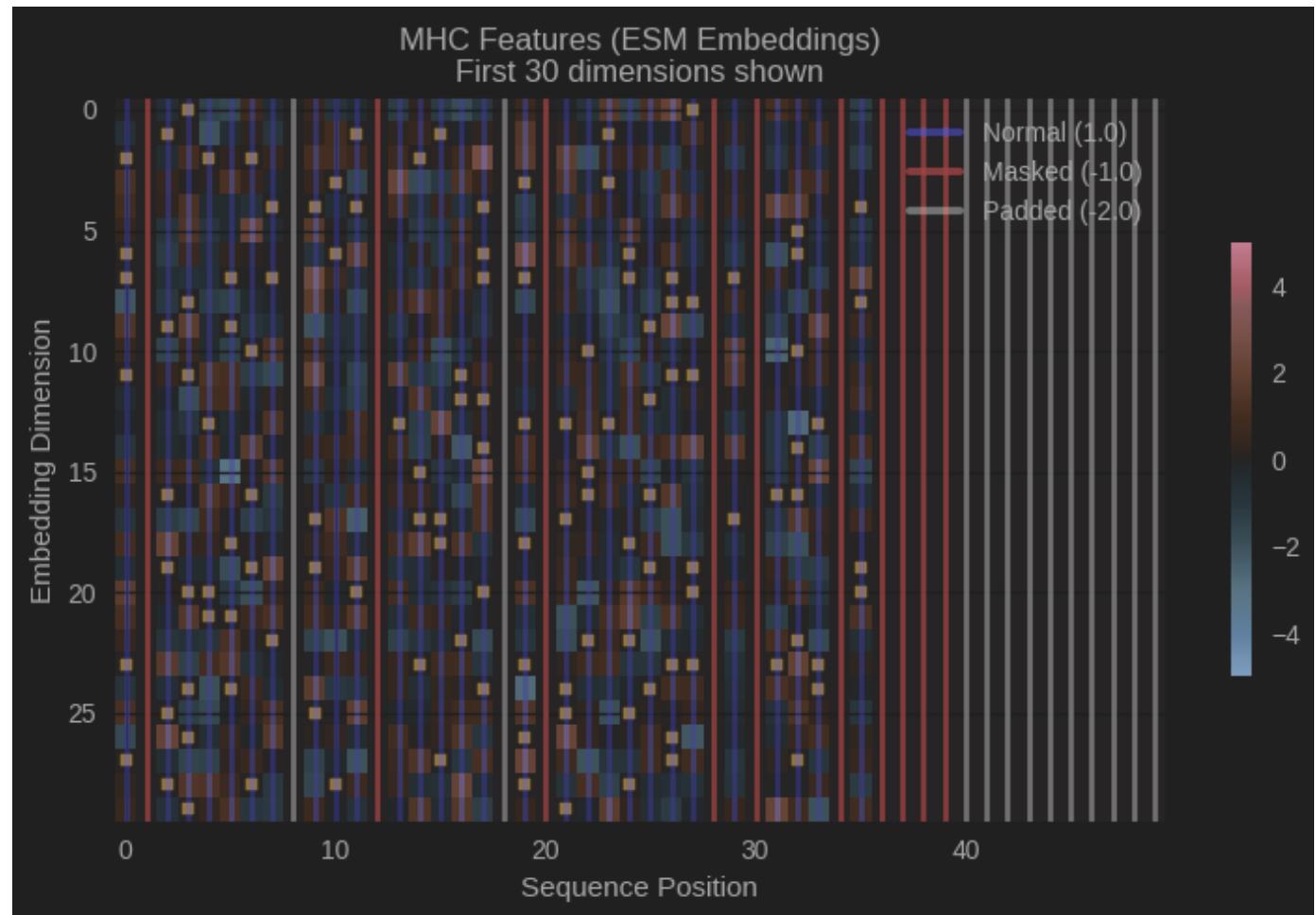
Method

Embedding Masking logic

White lines show padded regions introduced during alignment

Red lines are random 15% of the positions masked for training

Orange squares are random 15% of the embeddings masked for training



Method

Masked Self Attention logic (dummy data)

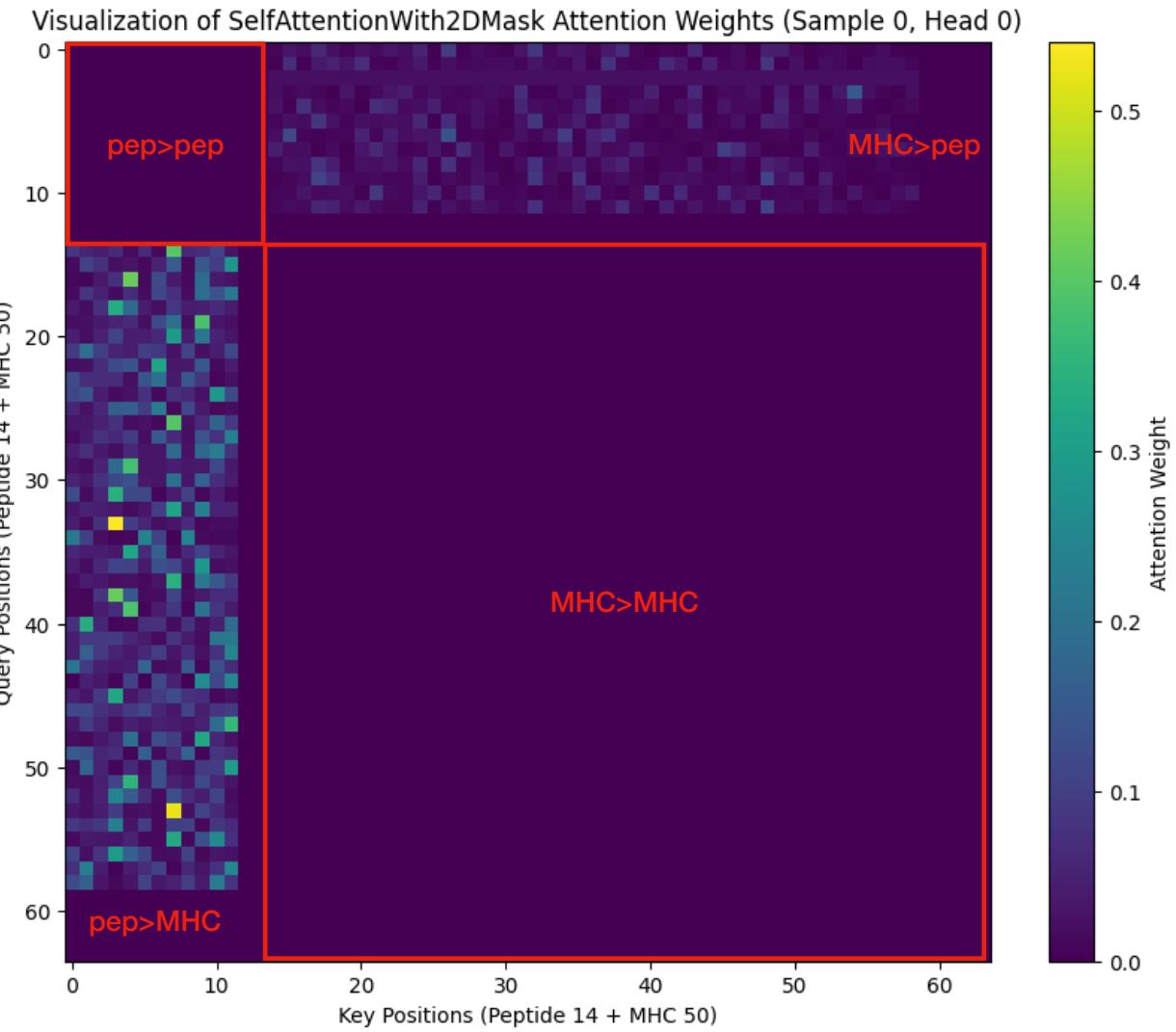
2D masking: Prevents peptide self-attention while enabling full peptide↔MHC cross-attention

Denoted as **M**

RoPE (Rotary Positional Embeddings): Applies position-dependent rotations to Q and K vectors, naturally spreading attention to neighboring residues and encoding relative distances

$$\text{Attn} = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \cdot M \right) V \odot \sigma(G)$$

$$\text{RoPE}(x, pos) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix} \otimes \begin{pmatrix} \cos(pos \cdot \theta) \\ \cos(pos \cdot \theta) \\ \vdots \end{pmatrix} + \begin{pmatrix} -x_2 \\ x_1 \\ \vdots \end{pmatrix} \otimes \begin{pmatrix} \sin(pos \cdot \theta) \\ \sin(pos \cdot \theta) \\ \vdots \end{pmatrix}$$

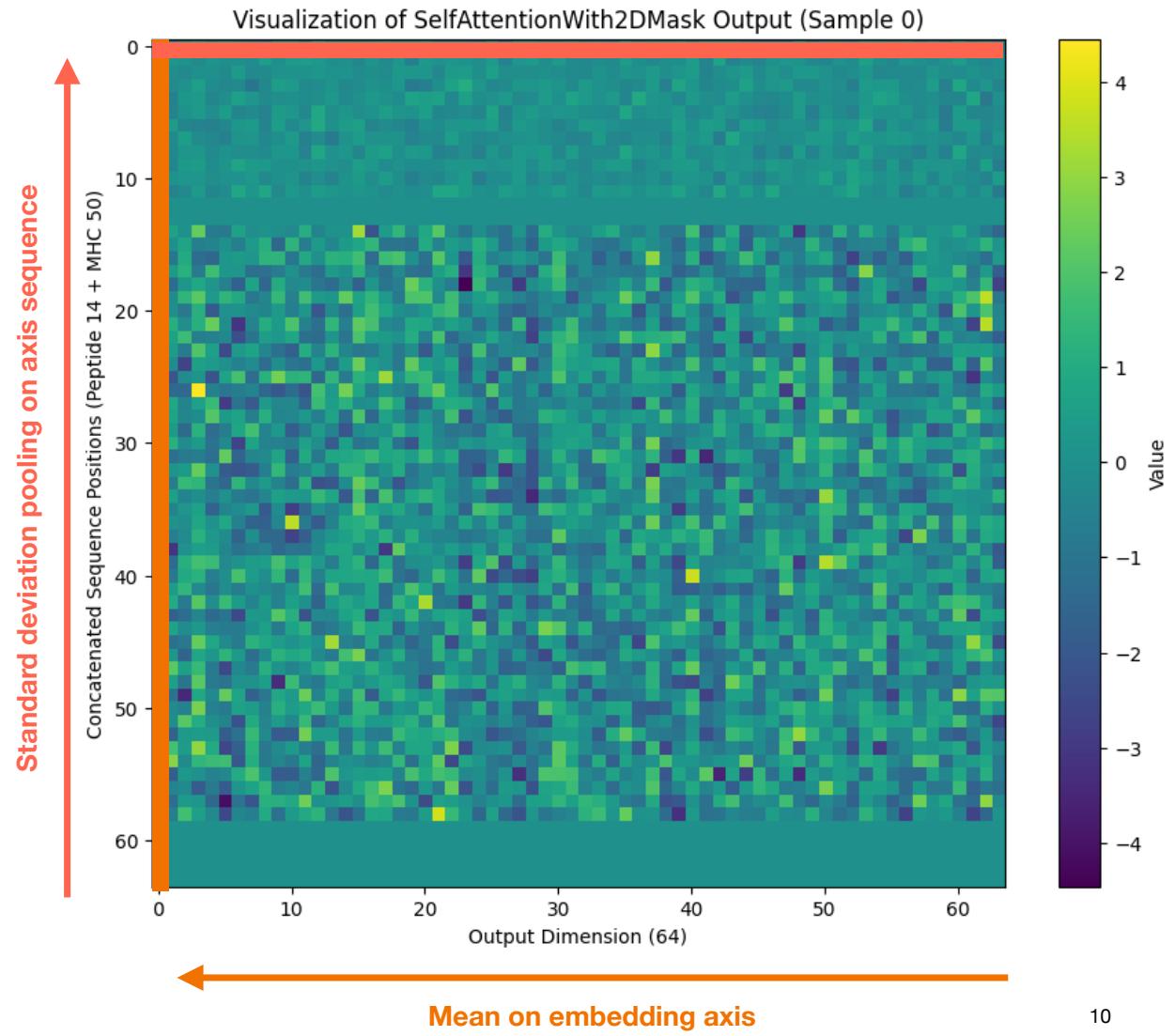


Method

Pooling Trick

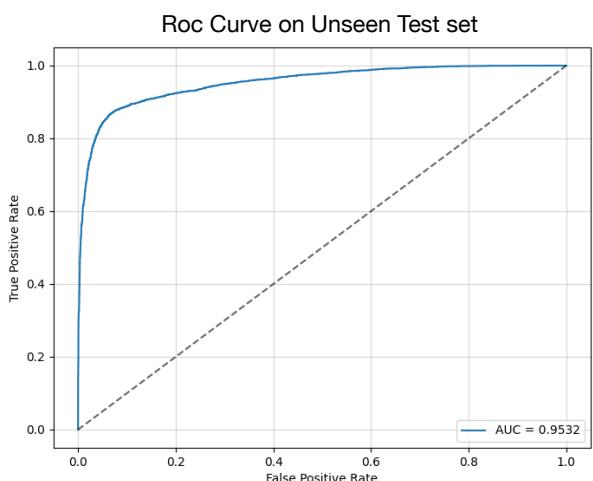
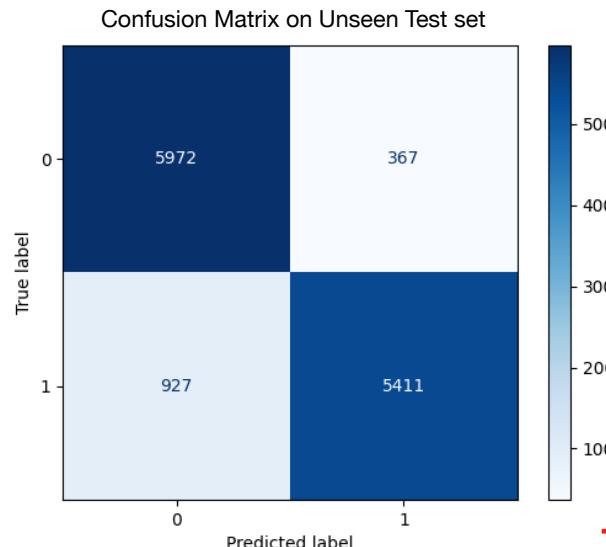
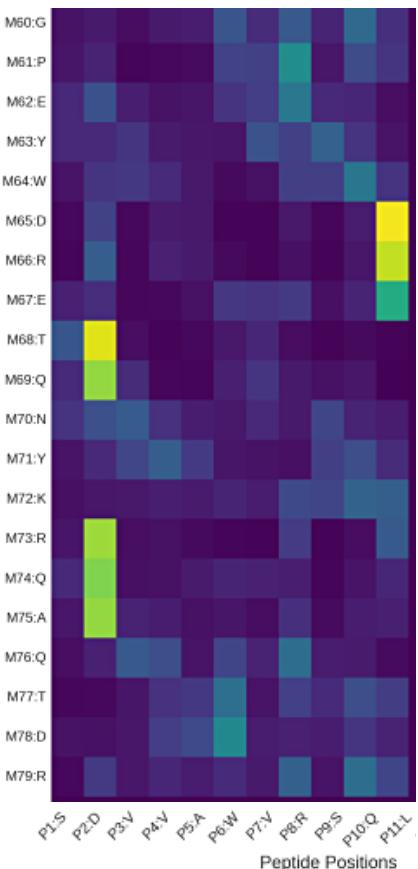
STD pooling Tells how much variation there is in the signal across the sequence

Mean pooling Tells how strong the overall signal is across the features



Results - MHC Class I

~12K samples
6 Unseen alleles



Benchmark on Liepe lab dataset

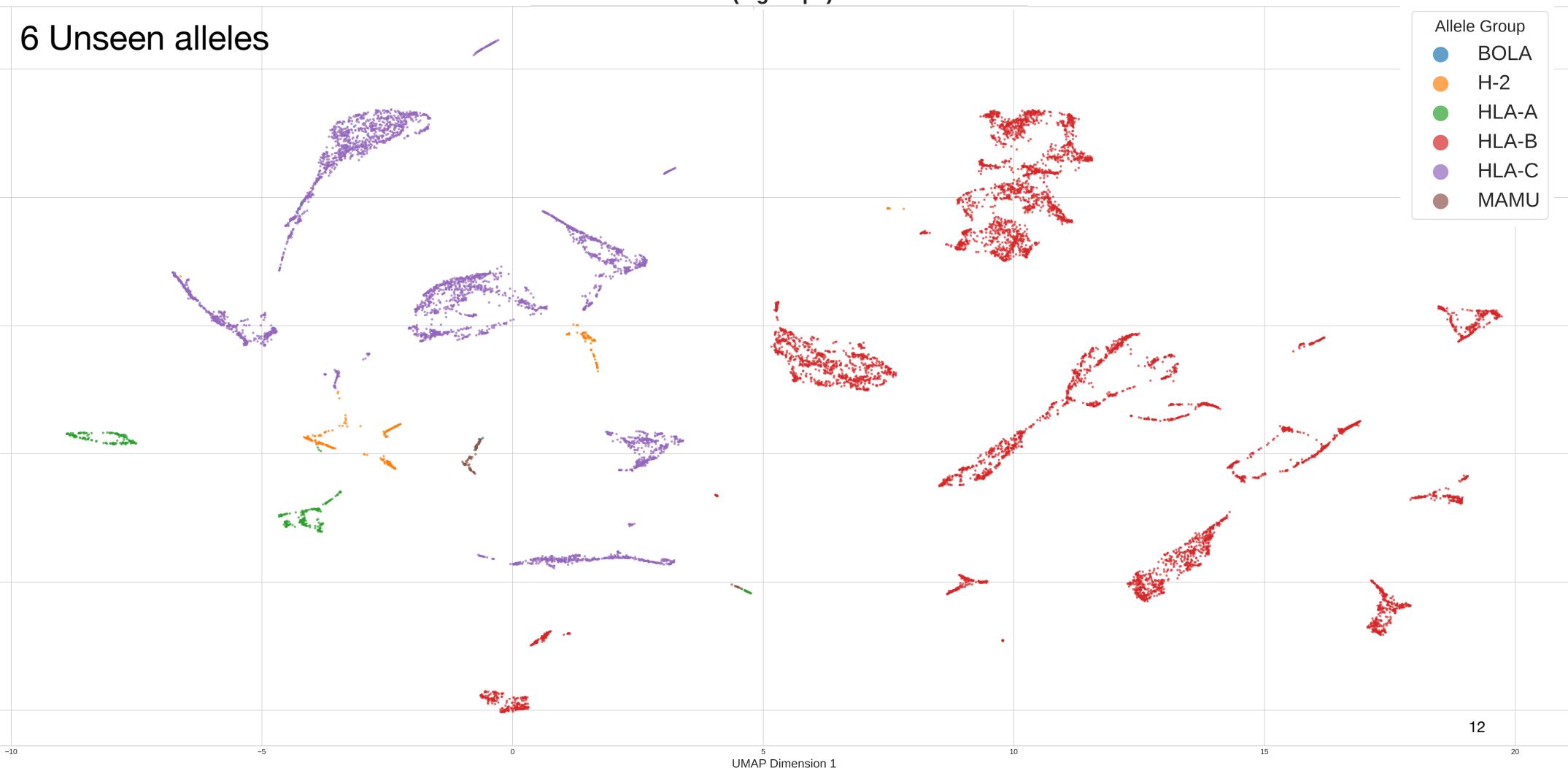
	MHCflurry	NetMHC	NetMHCons	NetMHCpan	PMBind	SMM	PickPocket	SMPMBEC
H-2K δ	0.893	1.000	1.000	0.875	0.885	0.786	0.917	1.000
HLA-A*01:01	0.873	0.750	1.000	0.875	0.908	0.786	0.833	1.000
HLA-A*02:01	0.881	0.750	1.000	1.000	0.911	0.714	0.875	1.000
HLA-A*26:01	0.877	0.875	1.000	0.875	0.934	0.643	0.917	1.000
HLA-B*07:02	0.846	1.000	1.000	1.000	0.894	0.500	0.958	1.000
HLA-B*08:01	0.845	0.750	1.000	1.000	0.863	0.500	0.667	1.000
HLA-B*15:01	0.868	0.875	1.000	1.000	0.884	0.857	0.875	1.000
HLA-B*18:01	0.871	0.750	1.000	1.000	0.892	0.714	0.917	1.000
HLA-B*40:01								



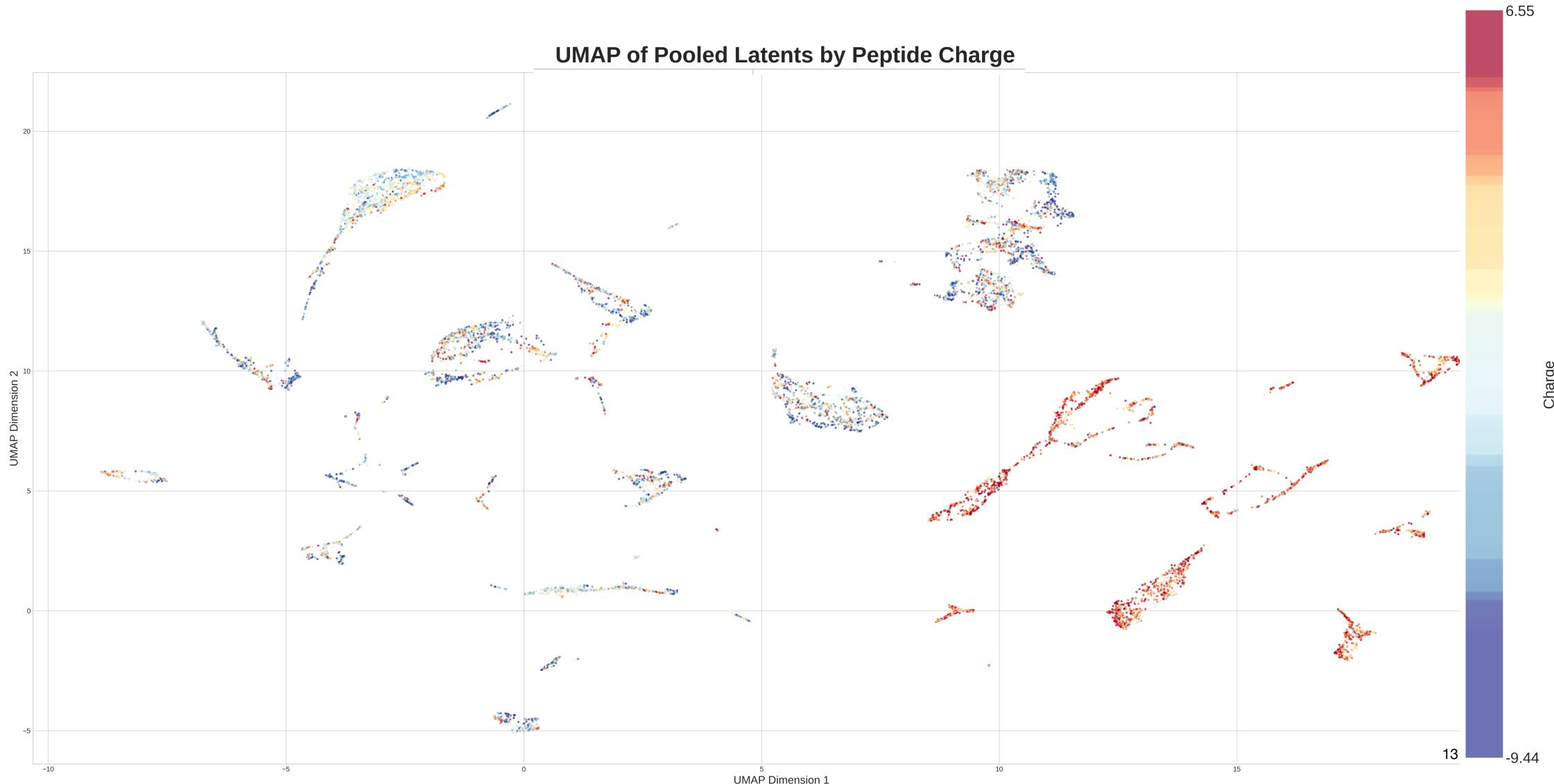
Results - MHC Class I

~12K samples

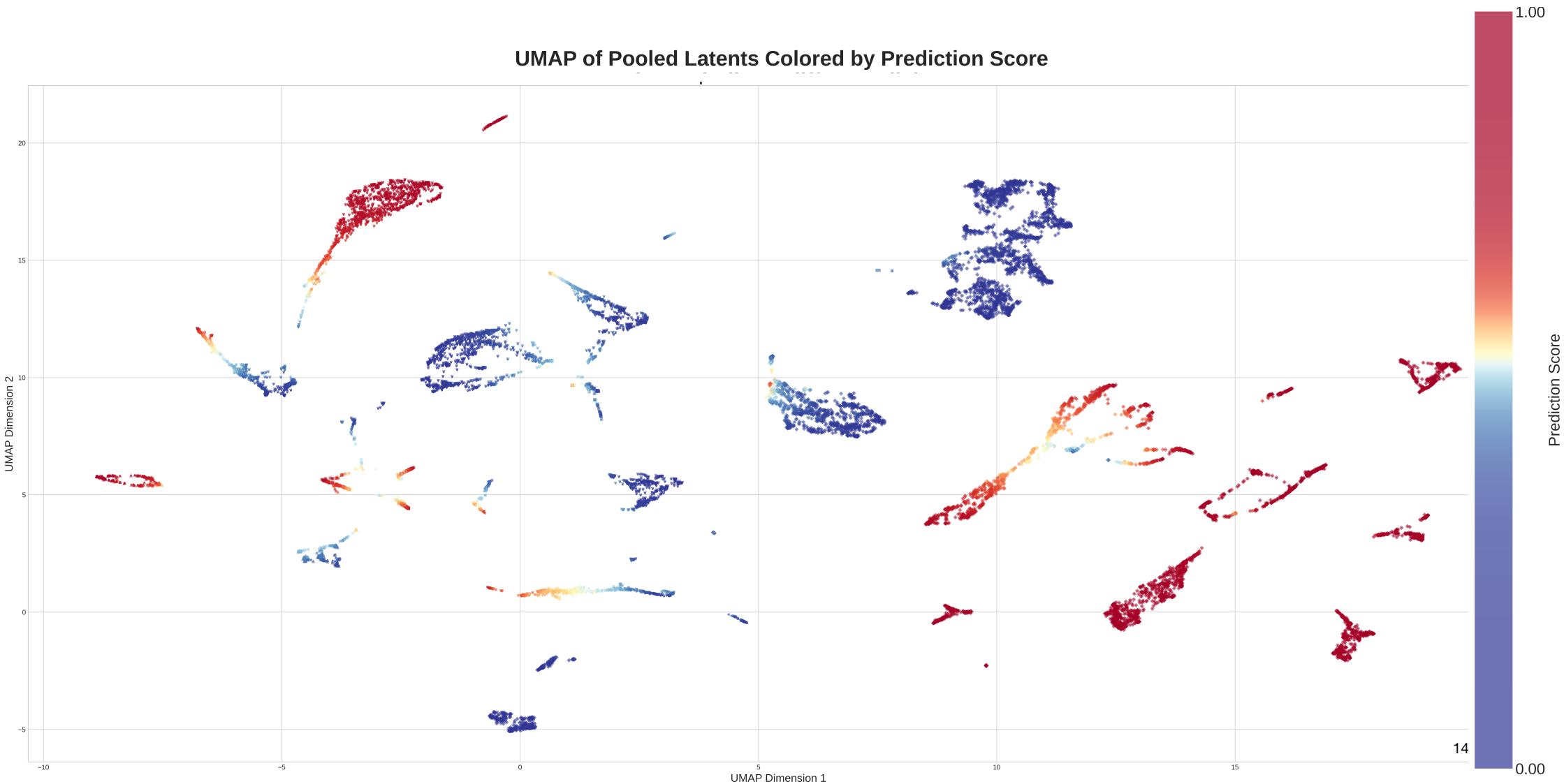
UMAP of Pooled Latents by Major Allele Groups
(6 groups)



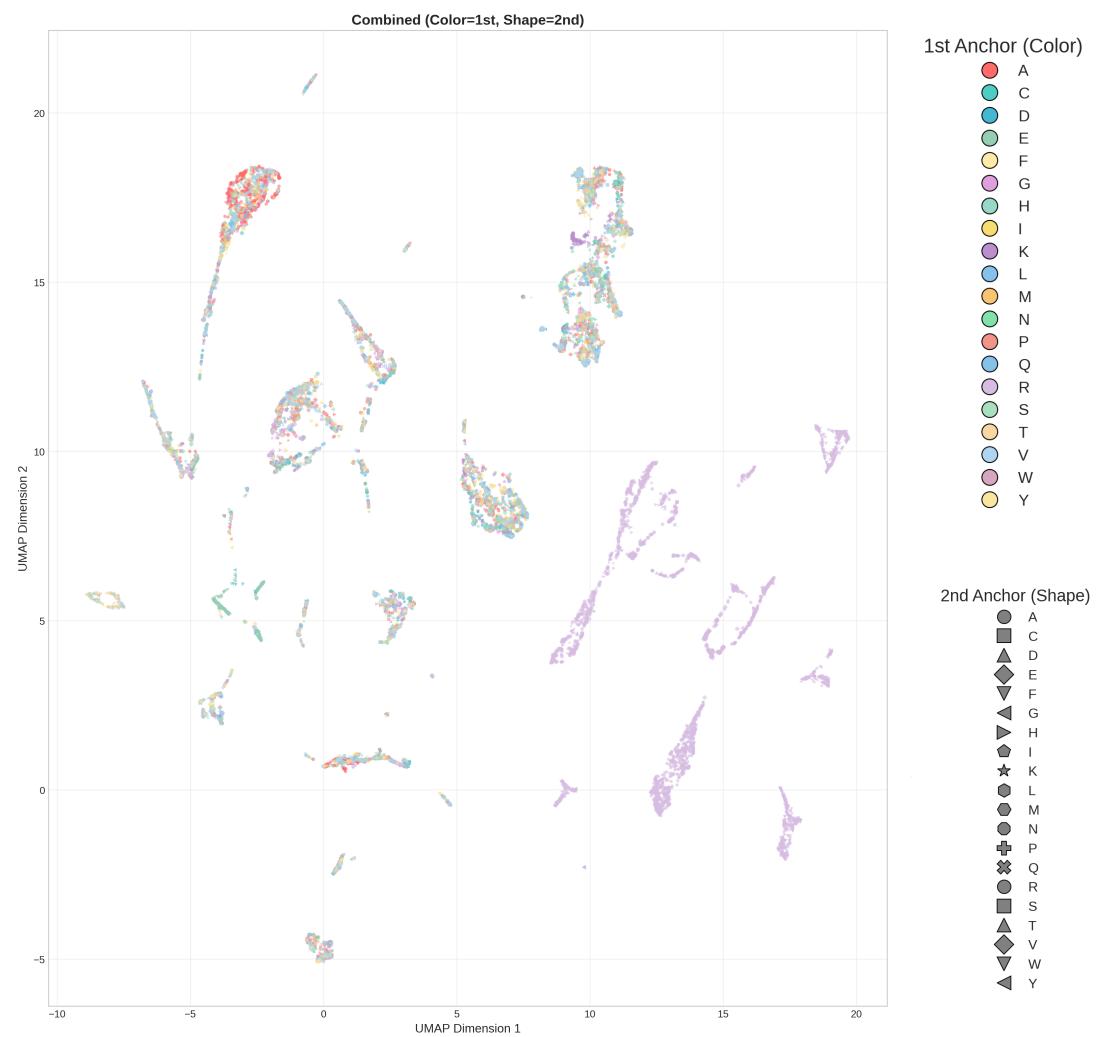
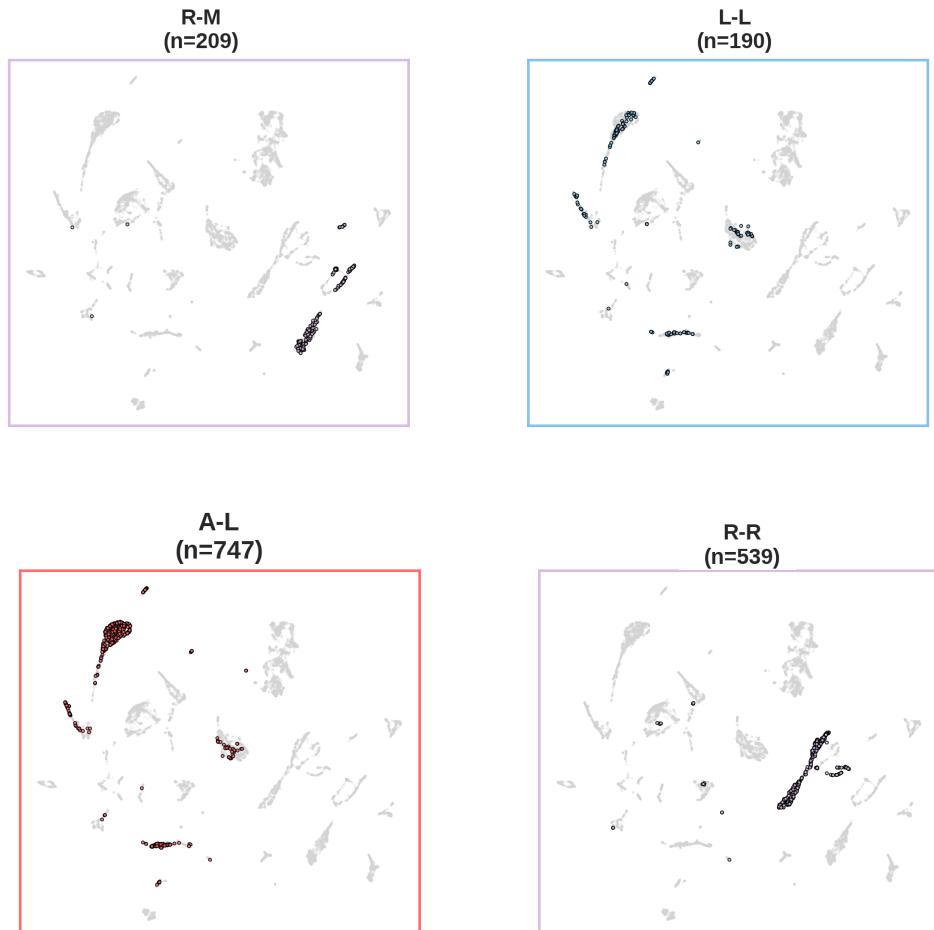
Results - Peptide charge values



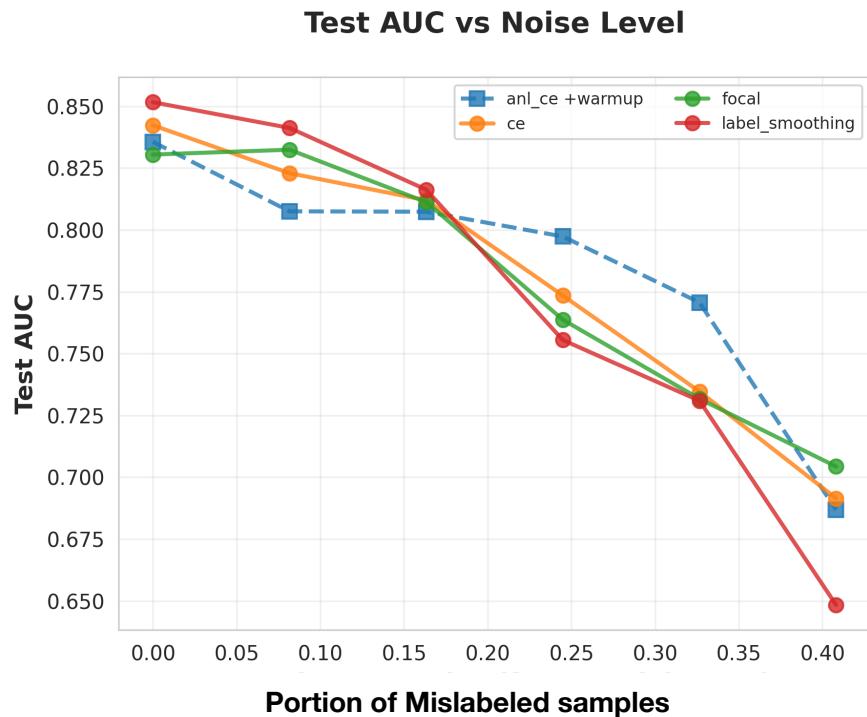
Results - Predicted probabilities



Results - Anchor combinations



Noise robust loss functions experiment



$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

p_t : predicted probability for true class

α_t : class-balancing weight

γ : focusing parameter ($\gamma = 2$)

$$\mathcal{L}_{\text{ANL-CE}} = \alpha \cdot \mathcal{L}_{\text{NCE}} + \beta \cdot \mathcal{L}_{\text{NNCE}}$$

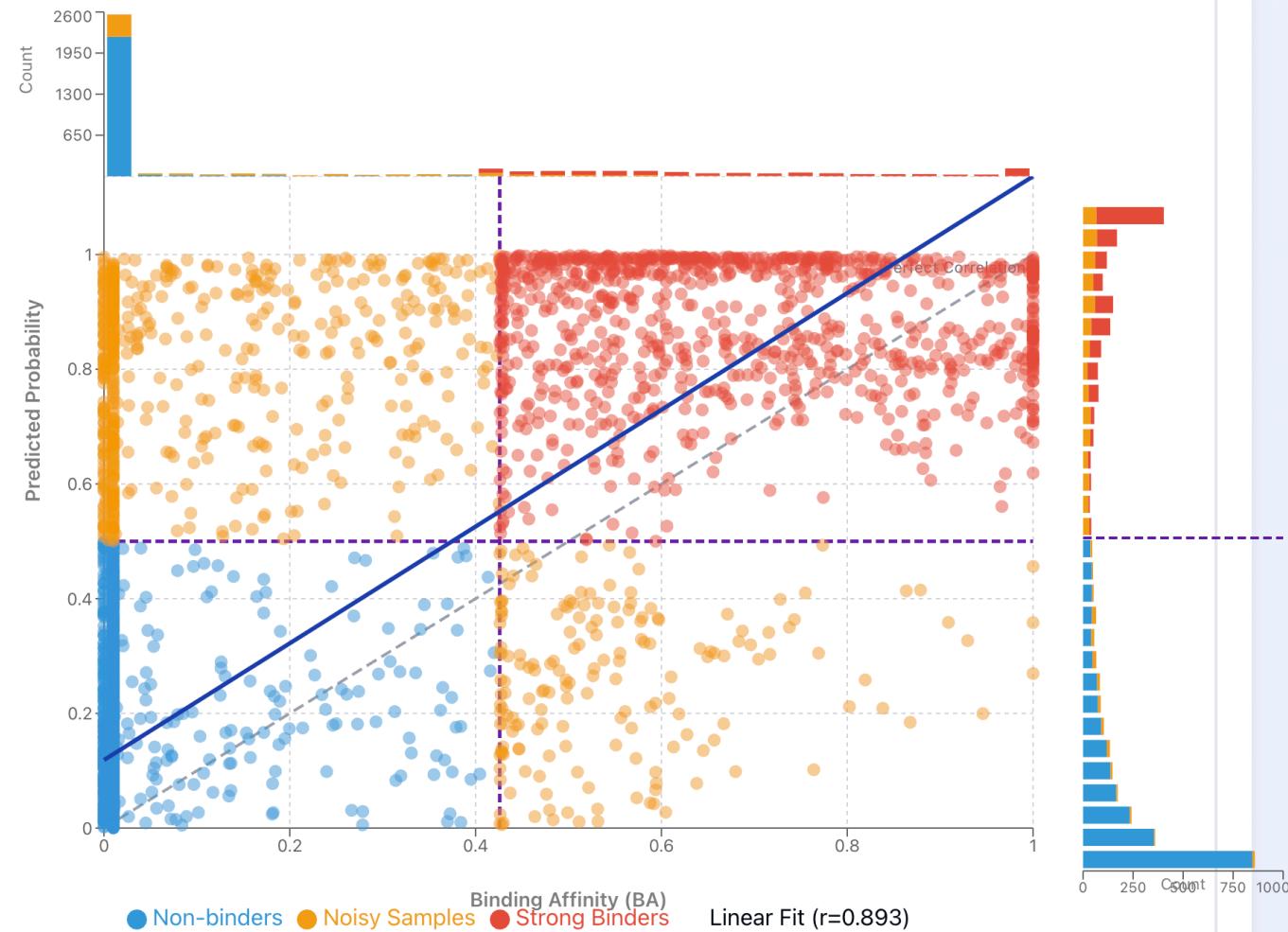
$$\mathcal{L}_{\text{NCE}} = \frac{-\log p(y|x)}{\sum_{k=1}^K [-\log p(k|x)]}$$

$$\mathcal{L}_{\text{NNCE}} = 1 - \frac{A + \log p(y|x)}{\sum_{k=1}^K [A + \log p(k|x)]}$$

$A = -\log(p_{\min})$ where $p_{\min} = 10^{-7}$. This is a normalization constant that prevents numerical instability.
 K = number of classes in your classification problem

Noise in the pMHC data

Binding Affinity vs Predicted Probability



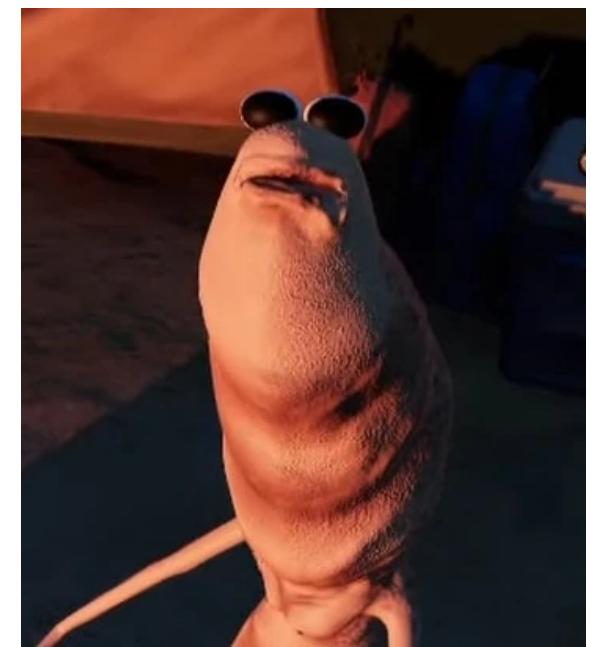
Conclusion

- Collected a huge pMHC dataset
- We identified noise in the pMHC binding affinity datasets
- Trained a model with novel cross attention layer to capture interaction
- Achieved competitive performance compared to SOTA methods

Future direction

- Incorporating Structural information using PMGen/AlphaFold - ESM3-open failed to provide us structural information due to numerical instability
- Using PMbind and other state-of-the-art methods to get high confident samples as a cleaner dataset for binding prediction with structures

Questions



References

1. Evaluation of antigen presentation pathways - Creative proteomics. (n.d.). <https://www.creative-proteomics.com/nucleic-acid-evaluation-of-antigen-presentation-pathways.html>
2. Iwm, M. (2020, October 16). MHC molecules. <https://microscopaiwm.wordpress.com/2020/10/16/mhc-molecules/>
3. Poiret, Thomas. (2018). DIVERSITY AND FOCUS OF CMV SPECIFIC T-CELL RESPONSES IN PATIENTS POST-HSCT AND WITH SOLID TUMOR.
4. Gao, Y., Gao, Y., Fan, Y., Zhu, C., Wei, Z., Zhou, C., Chuai, G., Chen, Q., Zhang, H., & Liu, Q. (2023). Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, 5(3), 236–249. <https://doi.org/10.1038/s42256-023-00619-3>
5. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017, February 6). Outrageously large neural networks: the Sparsely-Gated Mixture-of-Experts layer. OpenReview. <https://openreview.net/forum?id=B1ckMDqlg>
6. Van Den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017, November 2). Neural Discrete representation learning. arXiv.org. <https://arxiv.org/abs/1711.00937v2>
7. Srivastava, A., Ghorai, S. M., Department of Zoology, University of Delhi, Singh, S., Banaras Hindu University, & Singh, D. K. (n.d.). Structure and function of Major Histocompatibility Complex. In N. Sehgal, *ZOOLOGY Immunology* (pp. 1–4). https://epgp.inflibnet.ac.in/epgpdata/uploads/epgp_content/S000035ZO/P001308/M020587/ET/1498561836StructureandfunctionofMajorHistocompatibilityComplexQuad1%281.pdf