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

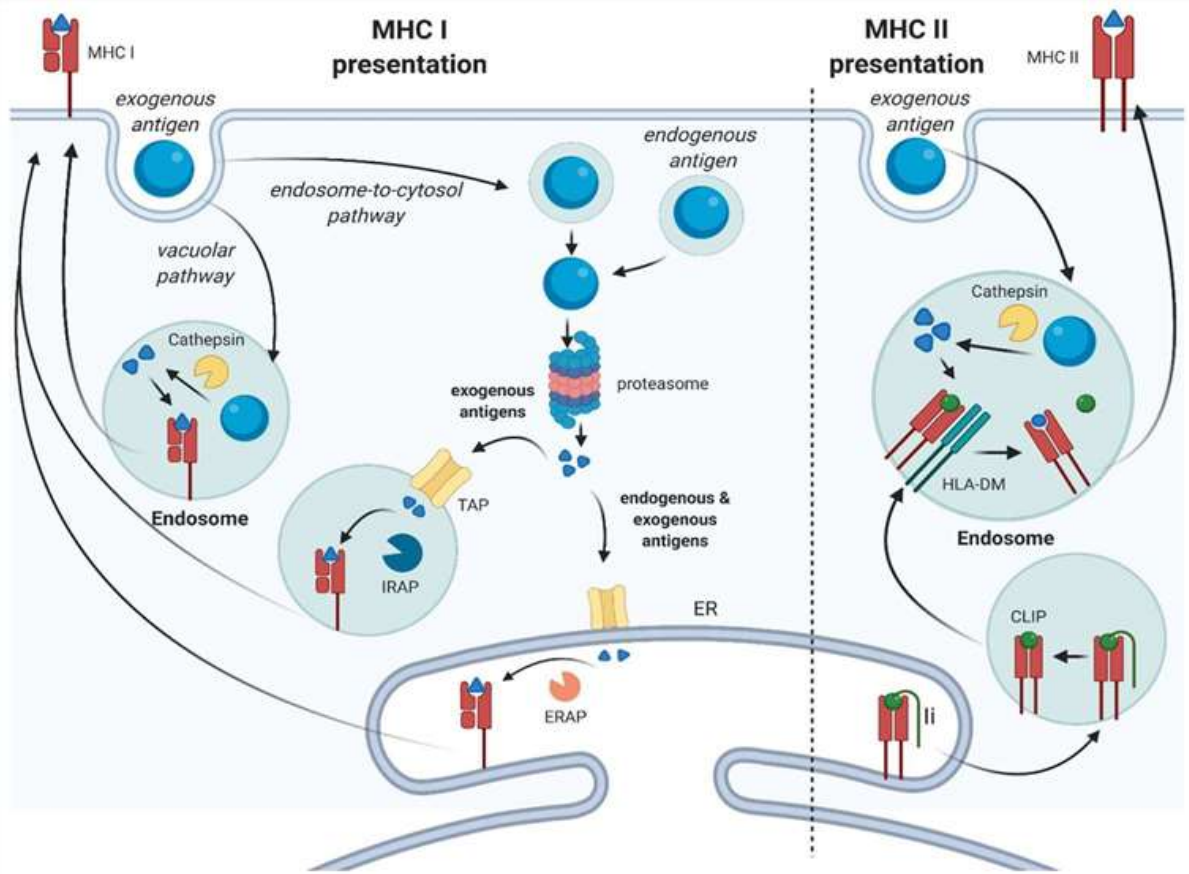
Master thesis report

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

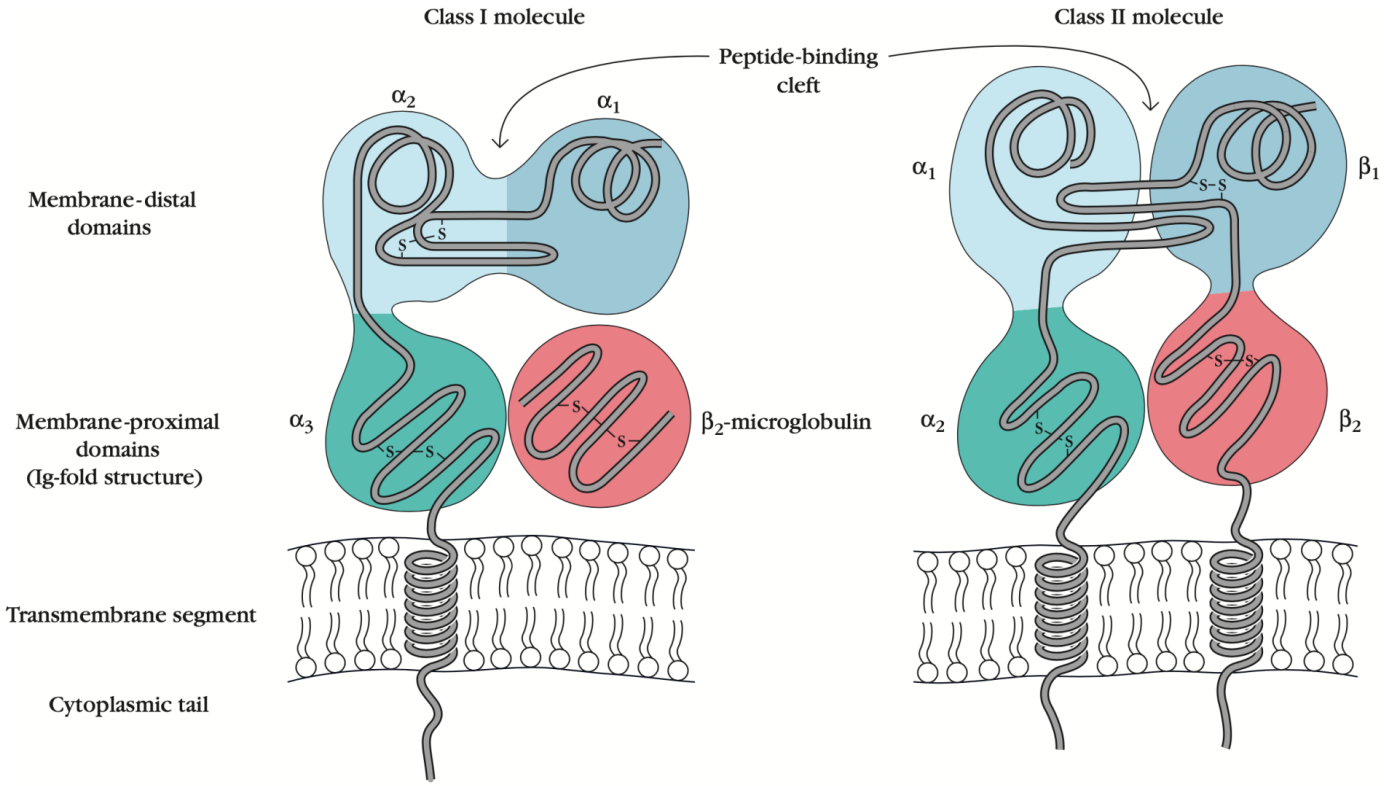


Introduction to MHC and peptide-MHC complex (pMHC)

Major Histocompatibility Complex

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

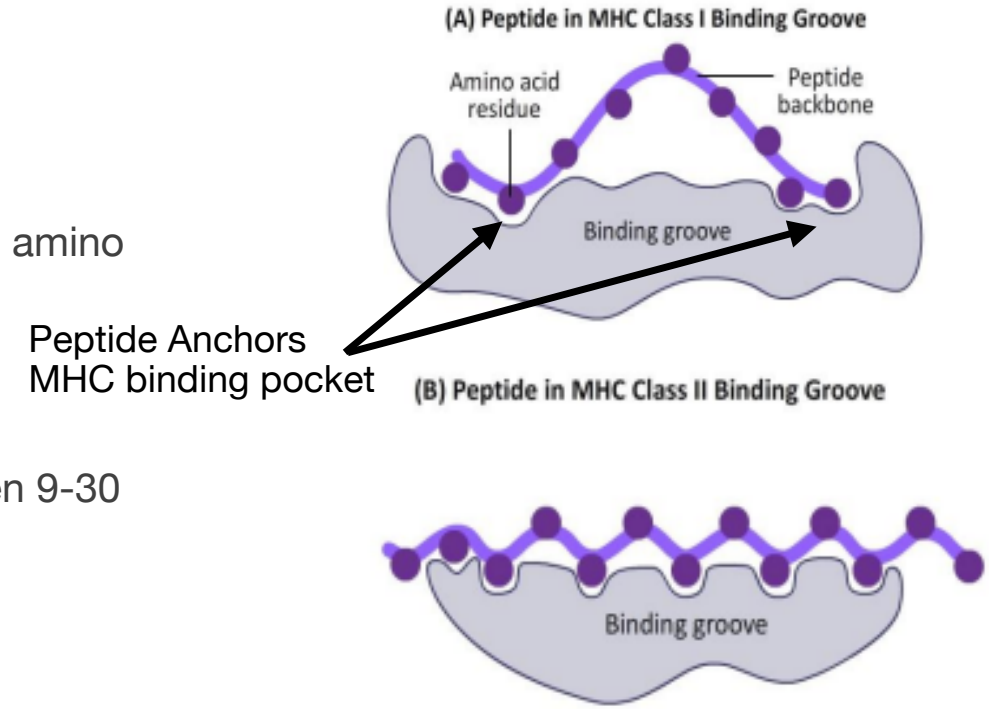
- MHC Class II:
- alpha chain
  - alpha1, alpha2
- beta chain
  - beta1, beta2



Introduction to MHC and peptide-MHC complex (pMHC)

Major Histocompatibility Complex

- MHC class I:
  - closed groove that typically binds peptides of 8-11 amino acids
  - Usually 2 anchor positions (P2 and P9)
- MHC class II:
  - open groove accommodating longer peptides, often 9-30 amino acids
  - Usually 4 anchor positions



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

All recognized MHC-HLA (human) alleles in IMGT database: 29,475 class I
13,521 class II

There are also 96 other species discovered!

https://www.ebi.ac.uk/ipd/mhc/statistics/

Dataset

Data Source	Positives	Negatives	Alleles	Peptides	Length	Avg.
NetMHCpan Training Data						
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
Deep Learning and previous Benchmark Datasets						
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
Cancer Immunopeptidomics						
TCGA (Xia et al.)	339,183	—	313	303,809	8–11	9.5
Neoantigen-Specific Data						
dbPepNeo (HC)	147	—	27	133	8–13	9.4
dbPepNeo (MC)	51	—	12	51	9–12	9.2
Predicted Negatives						
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.

Method

MHC_seq

↓

MSA

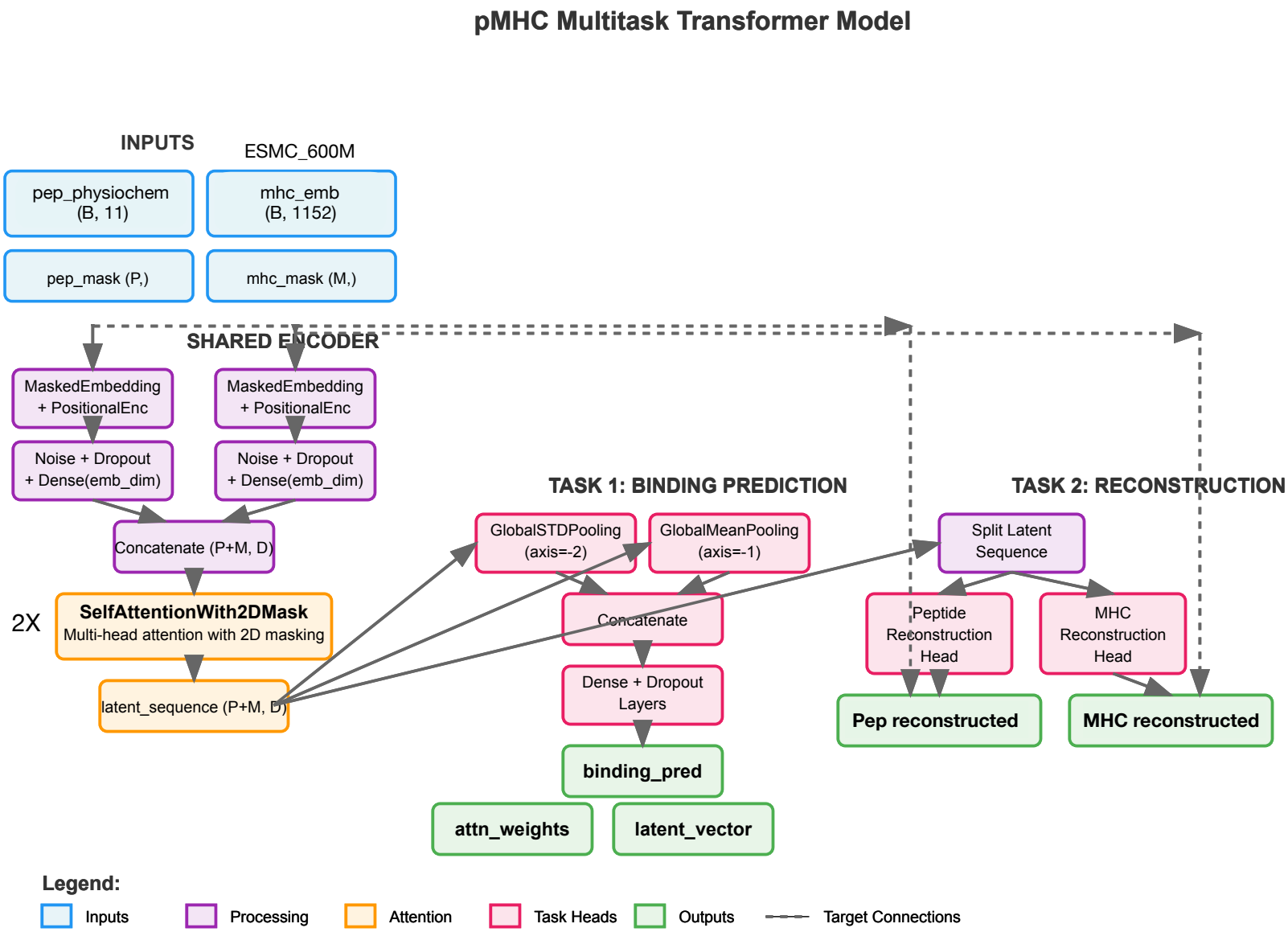
↓

ESM_Embedding

peptide_seq

↓

Physiochemical properties encoding
(from CapsNet repo)



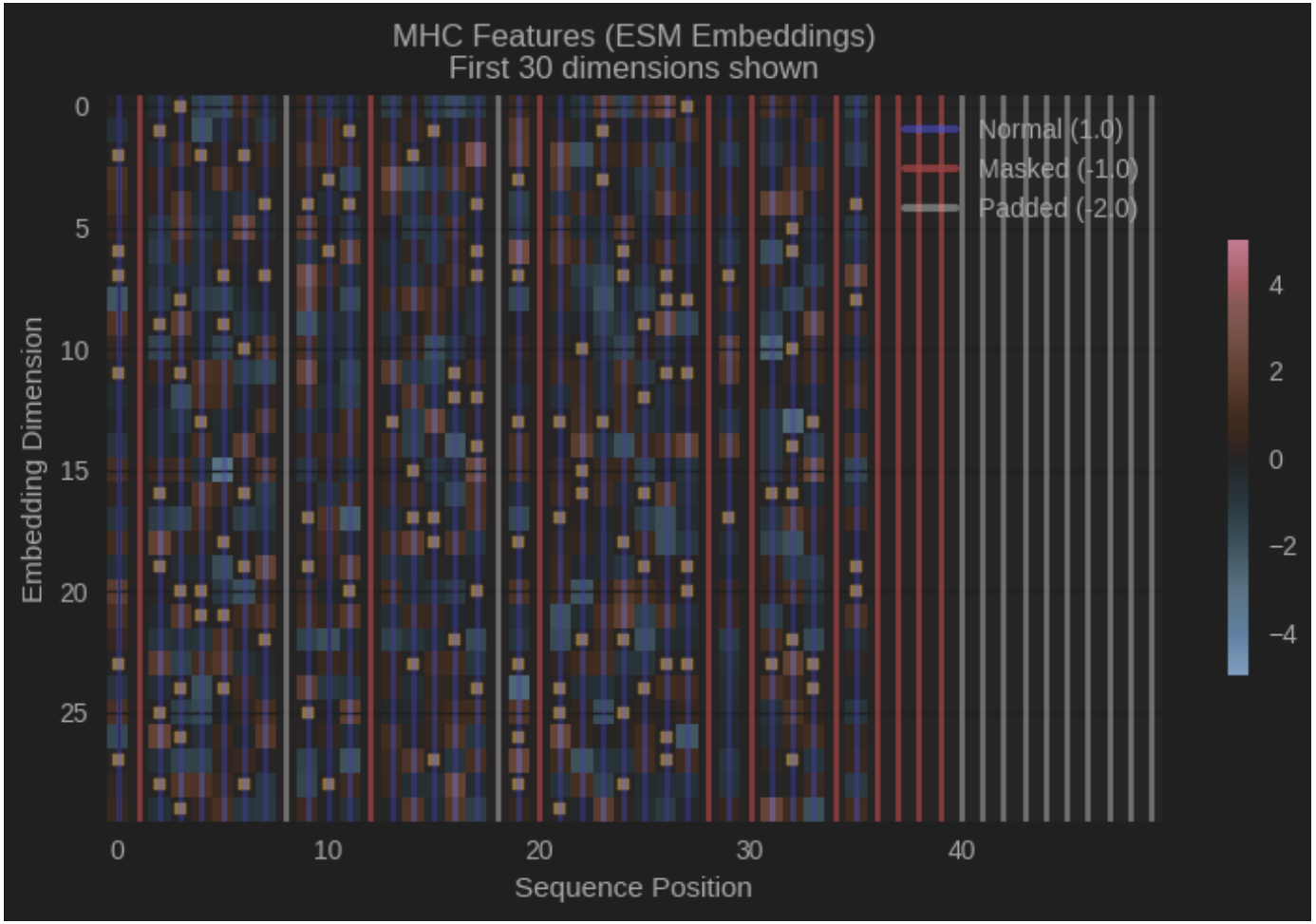
Method

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



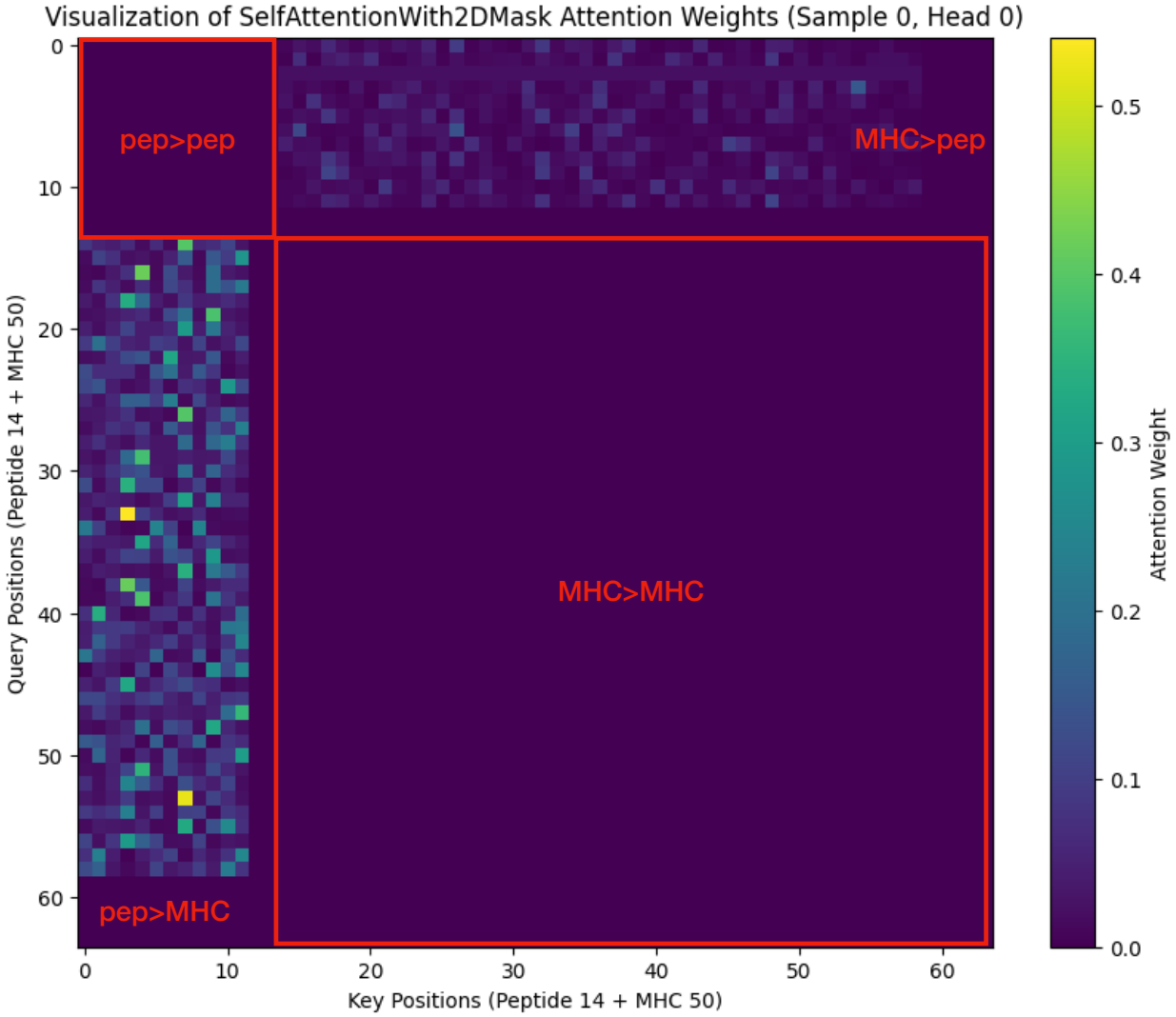
Method

Masked Self Attention logic (dummy data)

2D masking: Prevents peptide self-attention while enabling full peptide↔MHC **<u>cross-attention</u>**

Denoted as **M**

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



Method

Masked Self Attention logic (dummy data)

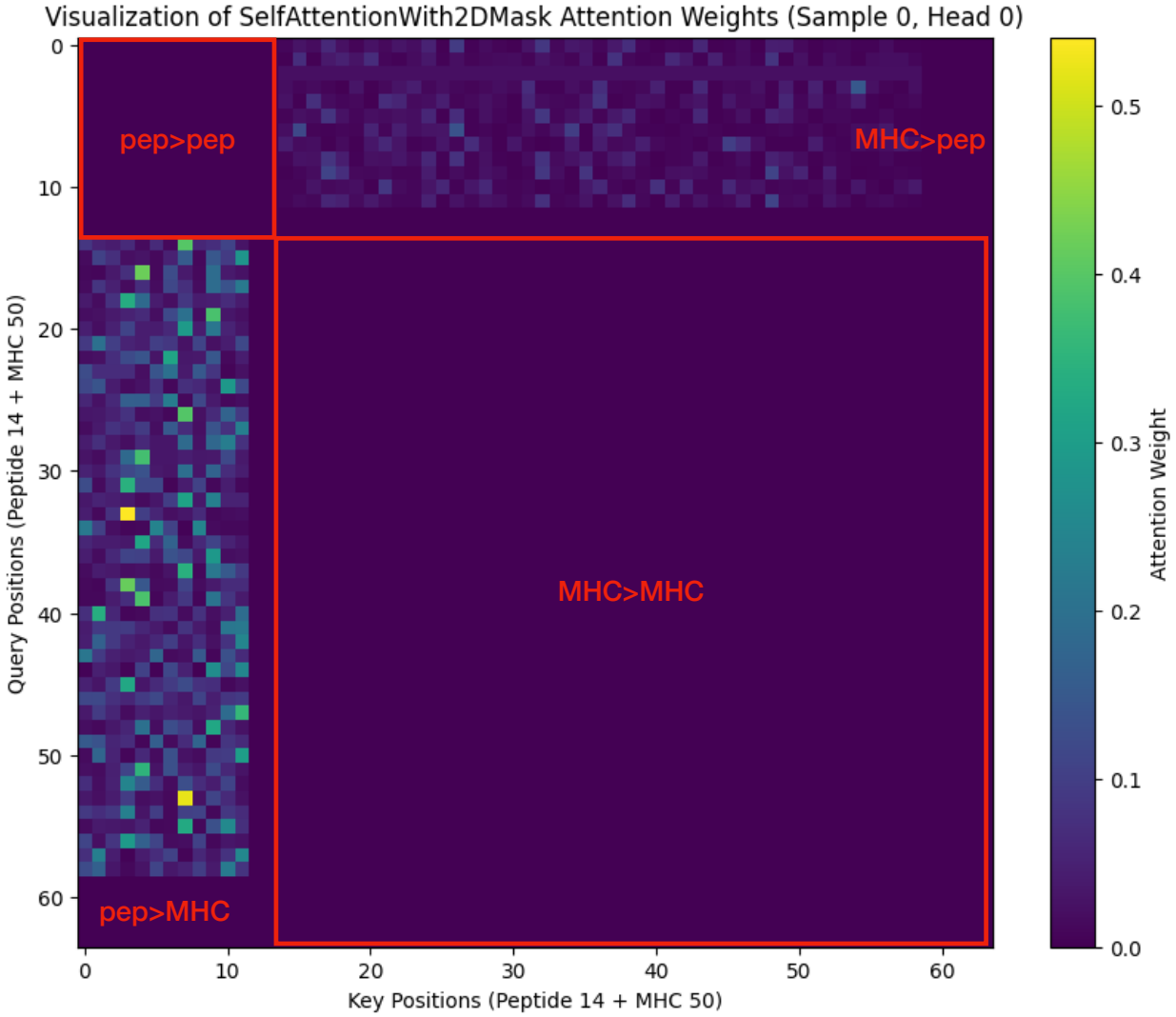
2D masking: Prevents peptide self-attention while enabling full peptide↔MHC **cross-attention**

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

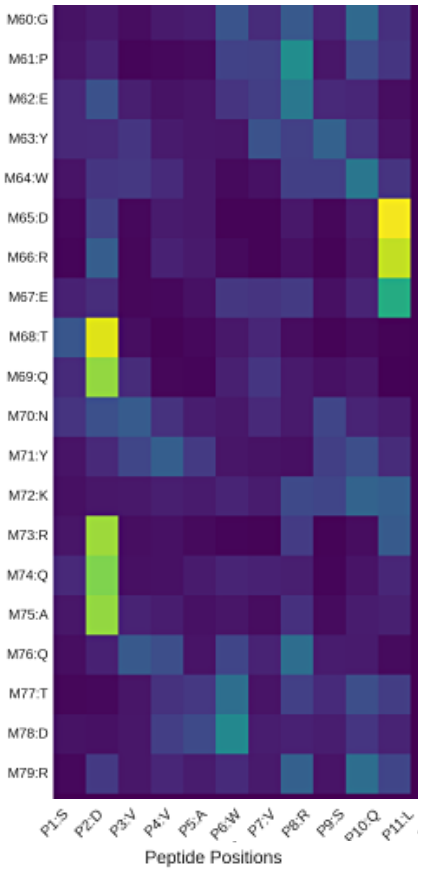
$$\text{Attn} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} . M\right) V$$

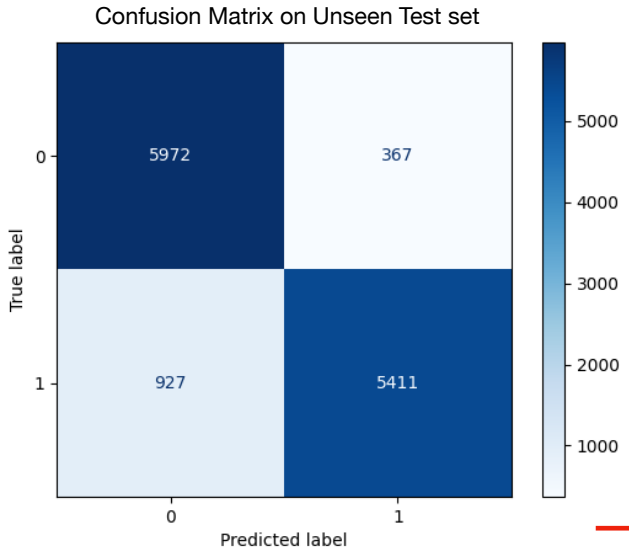
$$\text{RoPE}(x, \text{pos}) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix} \otimes \begin{pmatrix} \cos(\text{pos} \cdot \theta) \\ \cos(\text{pos} \cdot \theta) \\ \vdots \end{pmatrix} + \begin{pmatrix} -x_2 \\ x_1 \\ \vdots \end{pmatrix} \otimes \begin{pmatrix} \sin(\text{pos} \cdot \theta)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

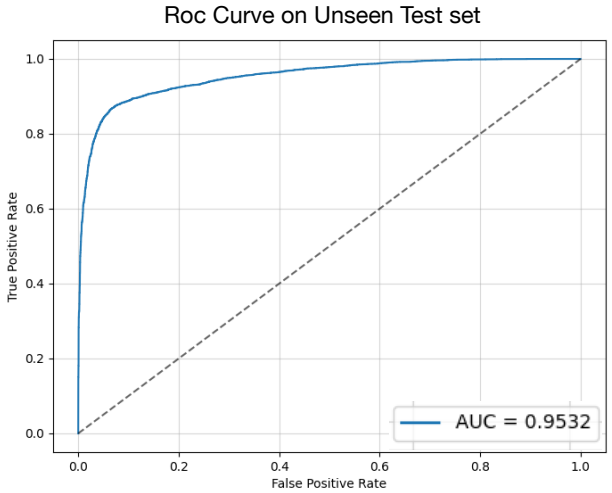


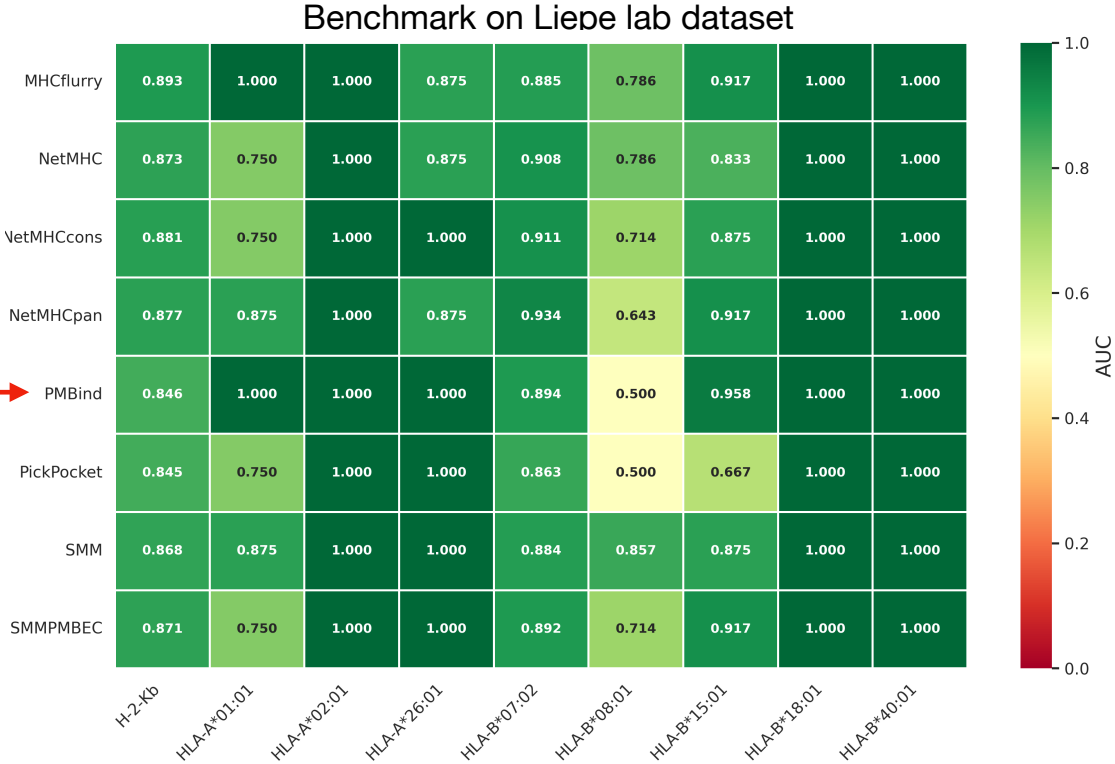
Results - MHC Class I

~12K samples
6 Unseen alleles





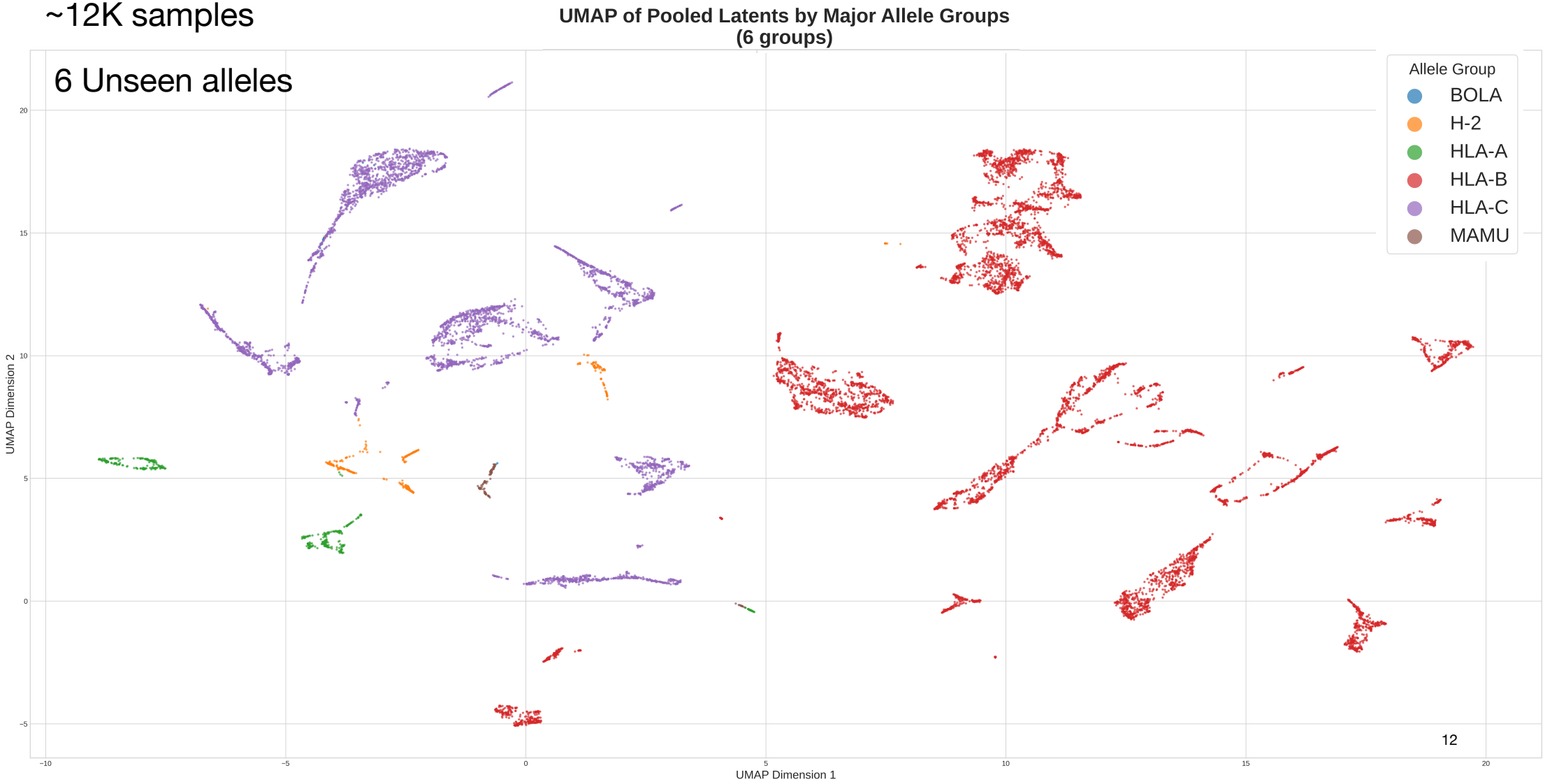




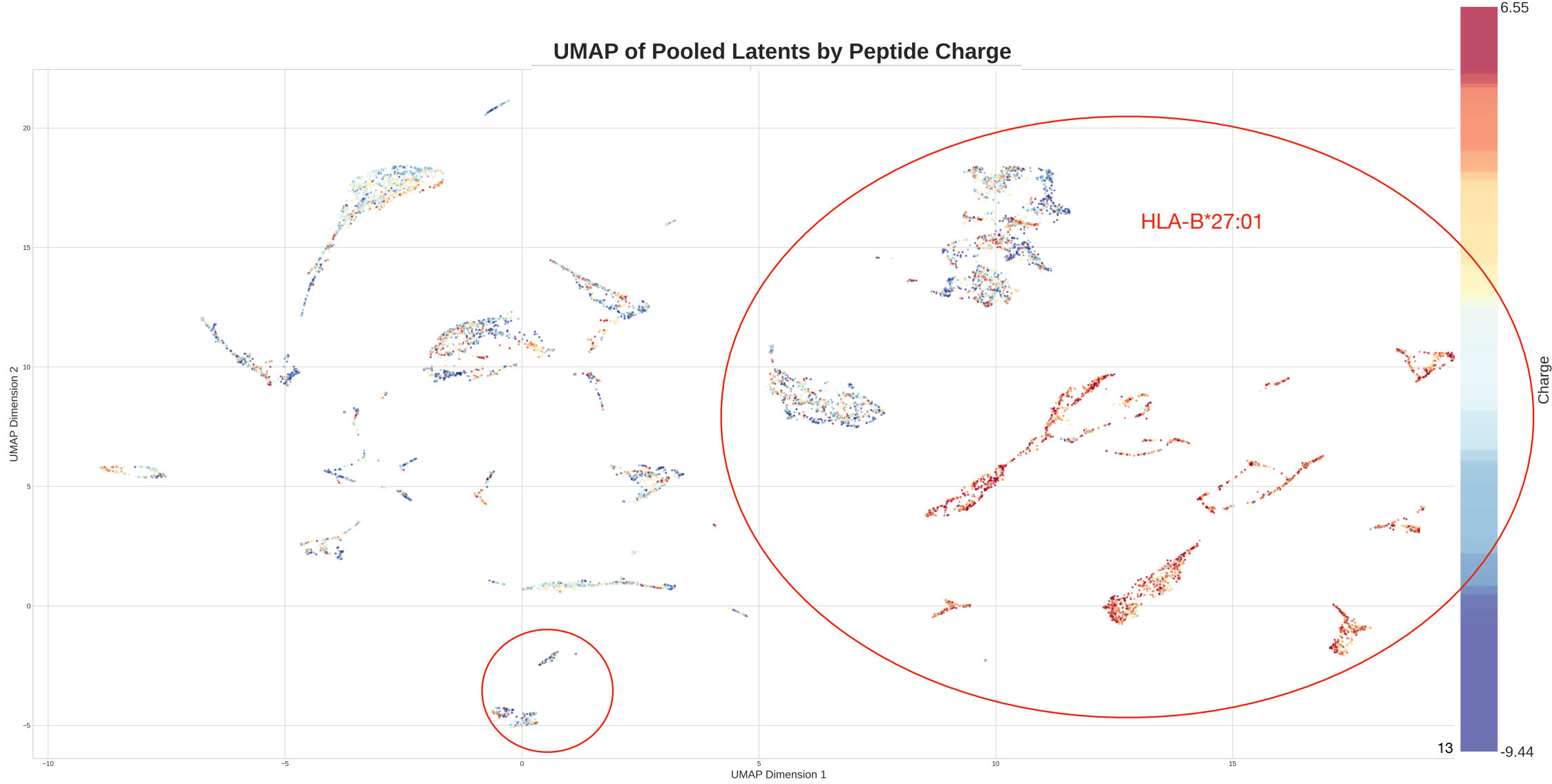
Results - Latent vector clusterings

~12K samples

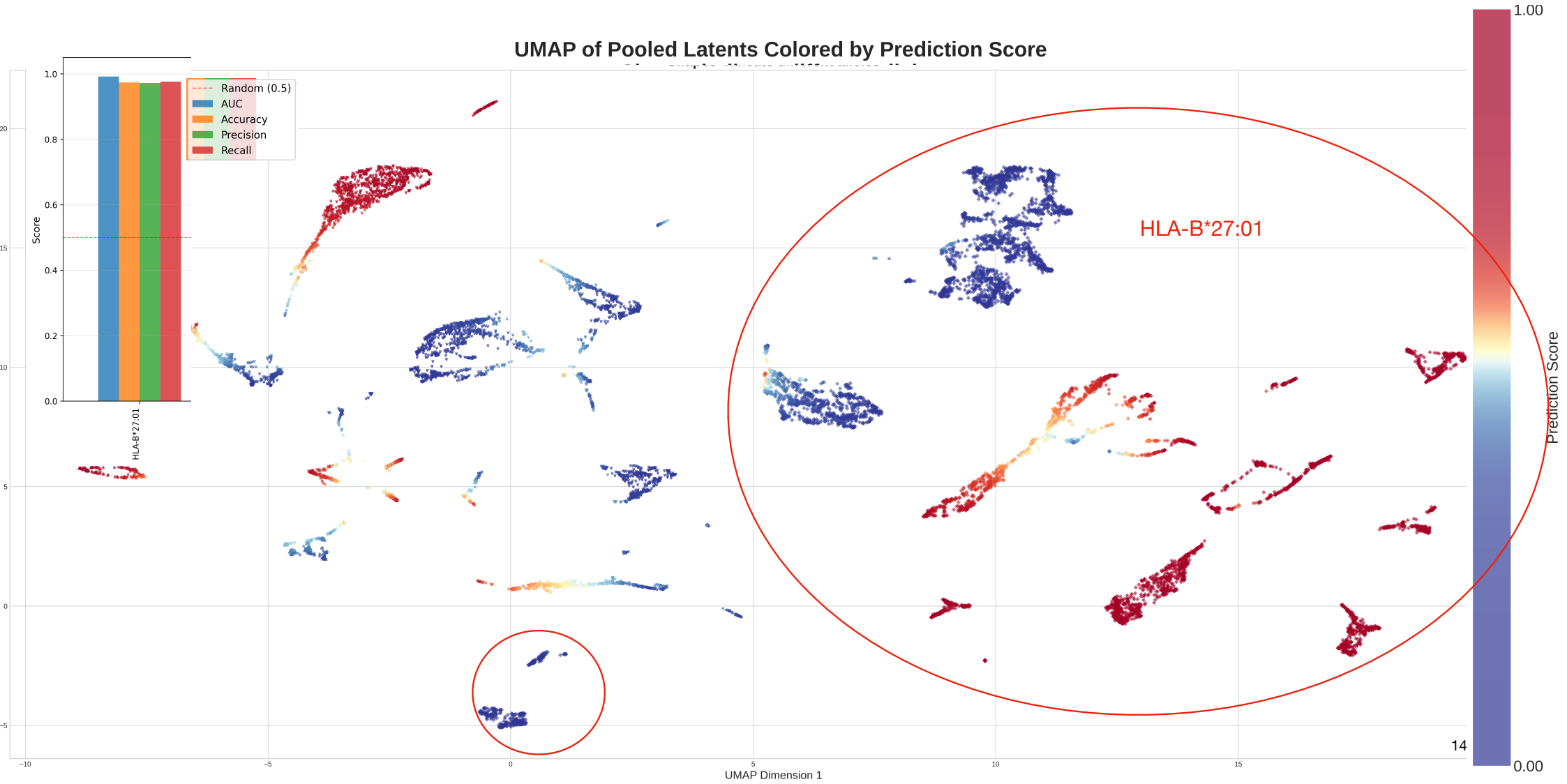
6 Unseen alleles



Results - Peptide charge values

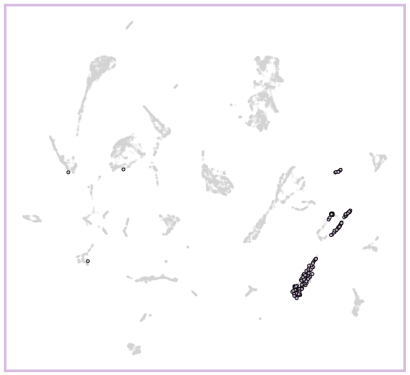


Results - Predicted probabilities

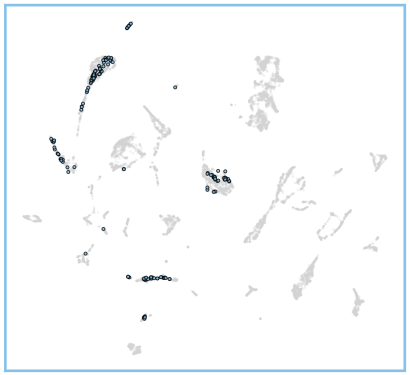


Results - Anchor combinations

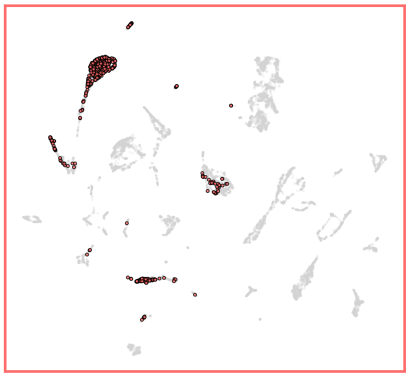
R-M
(n=209)



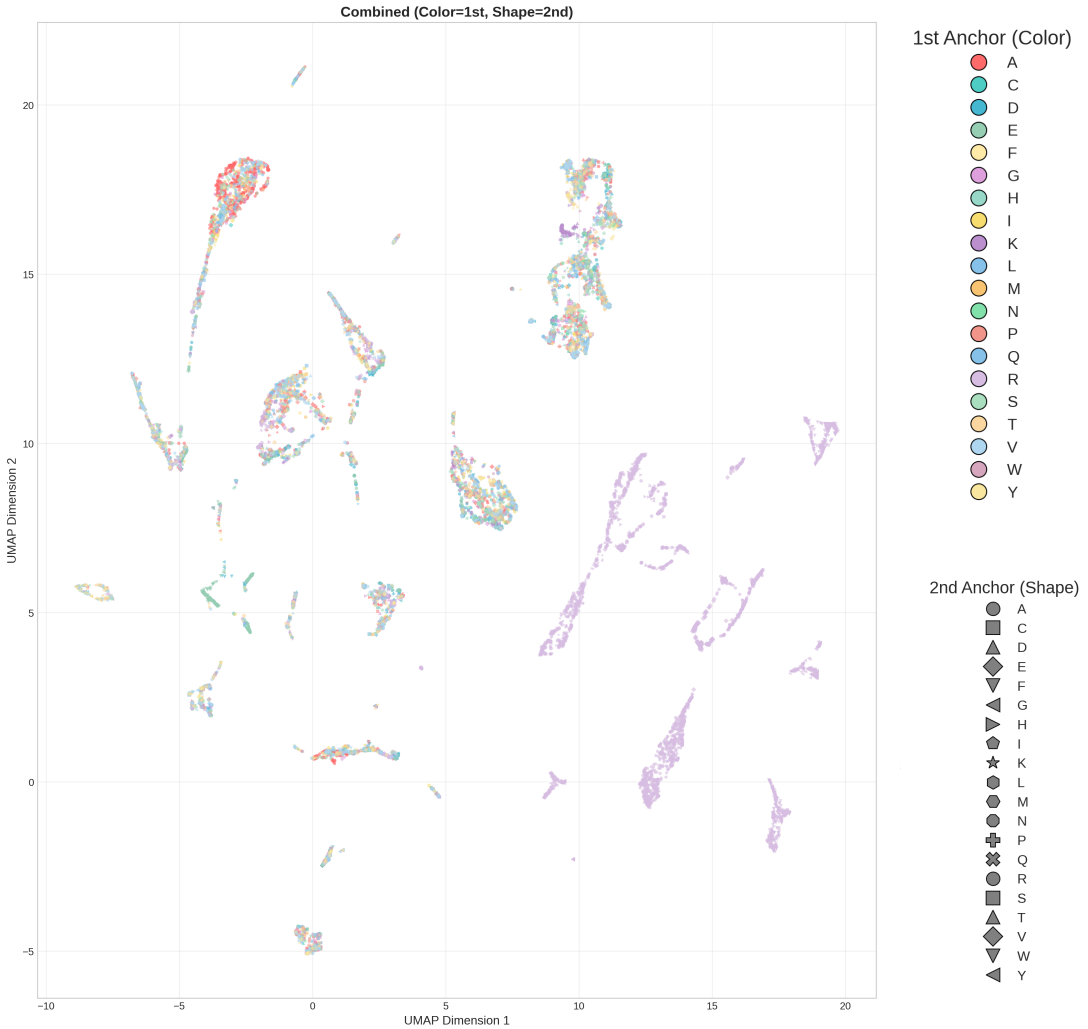
L-L
(n=190)



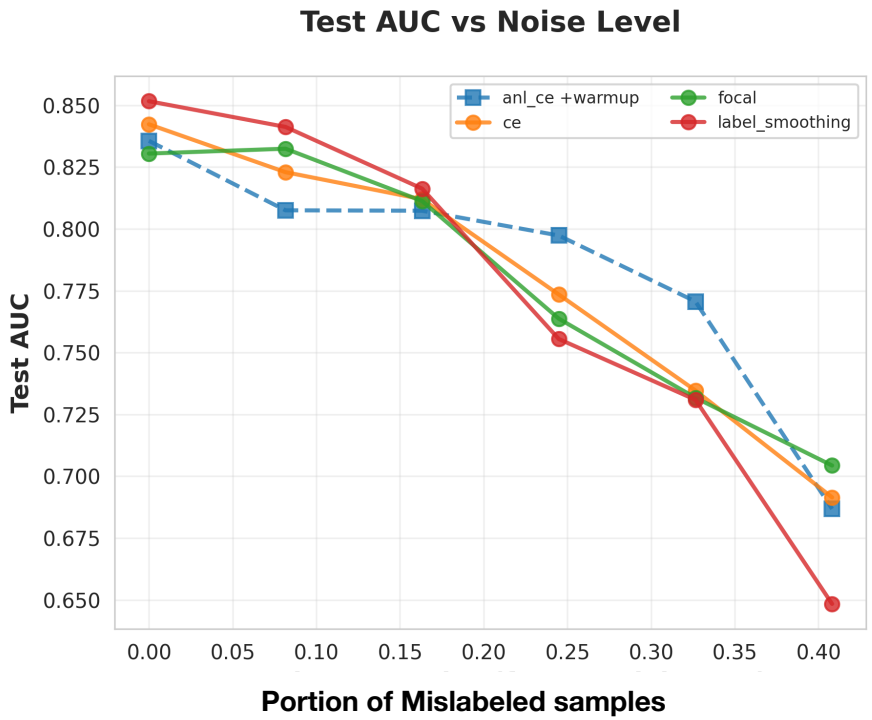
A-L
(n=747)



R-R
(n=539)



Noise robust loss functions experiment



params:
anl_ce: alpha = 0.5 beta =1
fl: gamma = 2
label_smoothing: 0.1

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

p_t: predicted probability for true class
α_t: class-balancing weight
γ: focusing parameter (γ = 2)

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

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

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

A = -log(p_min) where p_min = 10^-7. This is a normalization constant that prevents numerical instability.
K = number of classes

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
- Testing MHC class II predictions

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://microscopiaiwm.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
8. Ye, X., Wu, Y., Xu, Y., Li, X., Zhang, W., & Chen, Y. (2024, December 3). *Active Negative Loss: A Robust Framework for Learning with Noisy Labels*. arXiv.org. https://arxiv.org/abs/2412.02373

Noise robust loss functions experiment

Standard cross-entropy
Easy negative example: p_t = 0.95 → CE = -log(0.95) = 0.05
Hard negative example: p_t = 0.60 → CE = -log(0.60) = 0.51

Total loss from easy examples: 6,300 × 0.05 = 315
Total loss from 100 hard examples: 100 × 0.51 = 51

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

p_t: predicted probability for true class
α_t: class-balancing weight
γ: focusing parameter (γ = 2)

Focal Loss:
For easy example (p_t = 0.95):

- Modulating factor: (1 - 0.95)^2 = 0.0025
- FL = 0.0025 × 0.05 = 0.000125

For hard example (p_t = 0.60):

- Modulating factor: (1 - 0.60)^2 = 0.16
- FL = 0.16 × 0.51 = 0.0816

Now the ratio is reversed:

- 6,300 easy examples: 6,300 × 0.000125 = 0.79
- 100 hard examples: 100 × 0.0816 = 8.16

Noise robust loss functions experiment

Standard cross-entropy:

p(binder|x) = 0.7
p(non-binder|x) = 0.3 # Mislabeled

The gradient pushes the model to INCREASE p(non-binder) and DECREASE p(binder):

After training, the model learns:

- p(binder|x) → 0.1
- p(non-binder|x) → 0.9

The model memorized the WRONG label!

NCE:

- Numerator: -log p(non-binder|x) = -log(0.3) = 1.20
- Denominator: -log(0.7) + (-log(0.3)) = 0.36 + 1.20 = 1.56

L_NCE = 1.20 / 1.56 = 0.77

Key property - Symmetry:

If we sum NCE over all possible labels:

- L_NCE(y=0) = 1.20/1.56 = 0.77
- L_NCE(y=1) = 0.36/1.56 = 0.23
- Sum = 0.77 + 0.23 = **1.0** (constant!)

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

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

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

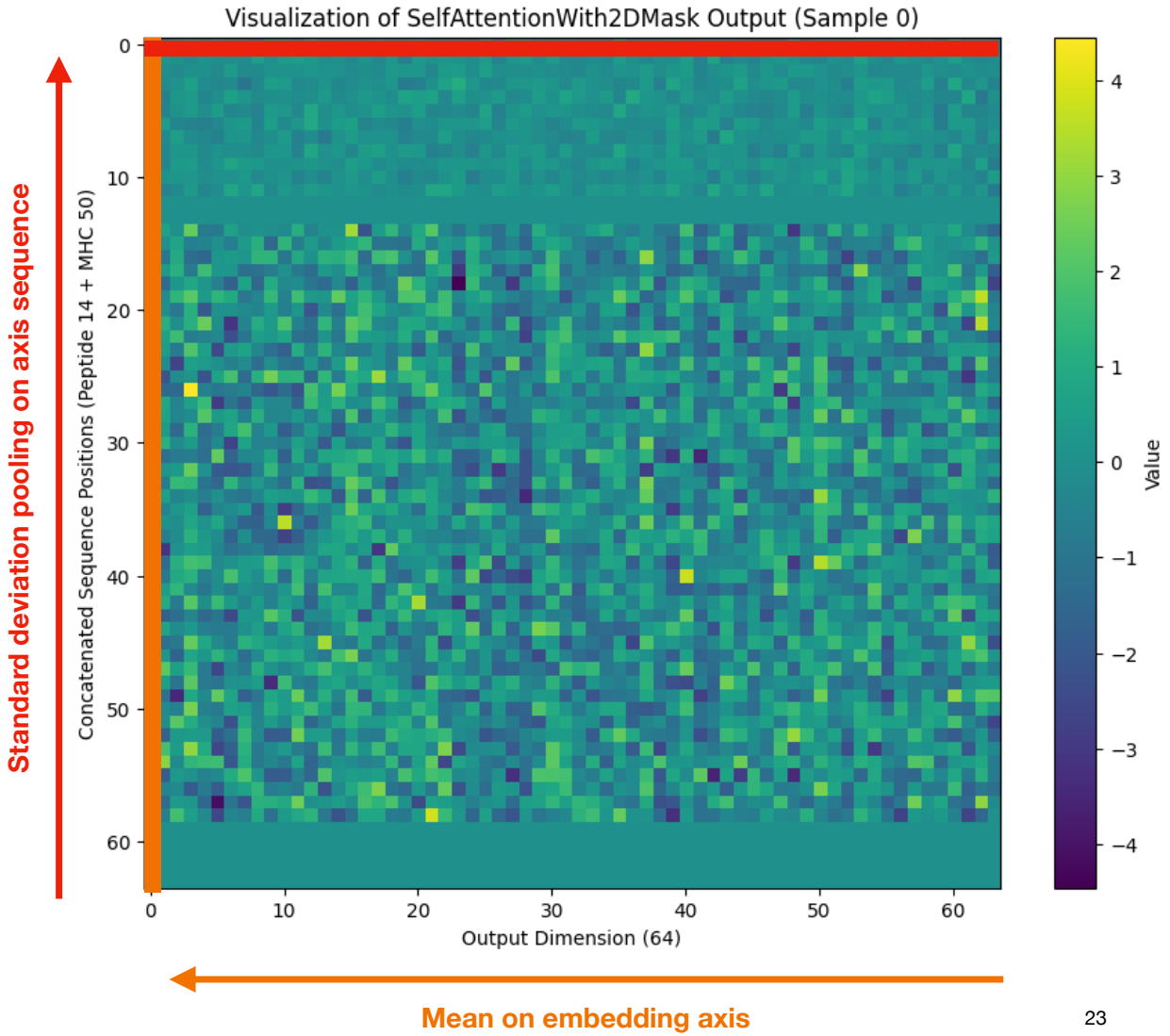
A = -log(p_min) where p_min = 10^-7. This is a normalization constant that prevents numerical instability.
K = number of classes

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



Noise in the pMHC data