Enhancing predictive performance of pMHC class I presentation models through multi-allelic data integration techniques and comparative analysis of architectural choices

T. Borck[2], M. Perier-Dulhoste[2], D. Schaller[1], I. Kizhvatov[1], R. Eloff[2], Y. Fu[2], A. Tovchigrechko[1], M. Rooney[1], M. Lang[1], N. Lopez Carranza[2], K. Beguir[2], U. Sahin[1]

[1]BioNTech SE, Mainz, Germany, [2]InstaDeep, London, United Kingdom. Presenting author: n.lopezcarranza@instadeep.com

Introduction & Aims

Background

Accurately predicting the presentation of peptides by MHC-I molecules is a critical challenge, particularly in designing individualized neoantigen-based cancer vaccines. To train machine learning models that target pMHC-I binding or presentation, various types of data can be utilized:

- Binding affinity (BA) data
- Single-allelic (SA) eluted ligand data (mass spec. experiments) e.g., from engineered cells expressing just one MHC-I allele
- Multi-allelic (MA) eluted ligand data (mass spec. experiment) from natural cells expressing up to six different MHC-I alleles

Incorporating the large amount of publicly available MA data has been shown to improve performance [1, 2, 3]. To that end, various approaches to deconvolution (= assignment of the most-likely presenting allele to the peptides in MA samples) have been proposed.

Objectives

We develop a new Transformer-based architecture for MHC-I presentation prediction adapted to handle BA, SA, and MA data. We aim to disambiguate the contribution of architectural choices and different strategies to include MA data from the gain provided by additional training data to, ultimately, improve machine learning predictors for pMHC presentation.

Materials & Methods

Model Architecture

Our baseline for benchmarking and the base for our deconvolution approaches is the Transformer-based architecture in Fig. 1. It takes as input the peptide sequence and the allele-specific MHC receptor pseudo-sequence. We distinguish between

- *Offline deconvolution* (Fig. 2A,B) = the model input consists of a peptide and a single allele at a time, and
- *Online deconvolution* (Fig. 2C,D) = the model input consists of a peptide and a bag of alleles using techniques from the multiple instance learning (MIL) paradigm.

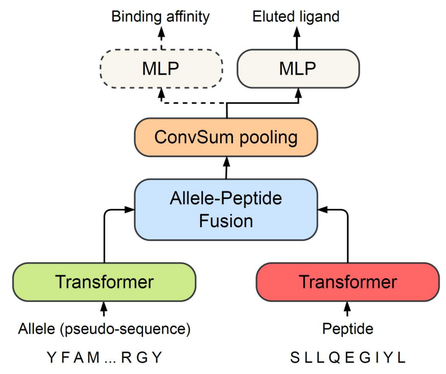


Figure 1. We use a Transformer architecture as base for all approaches

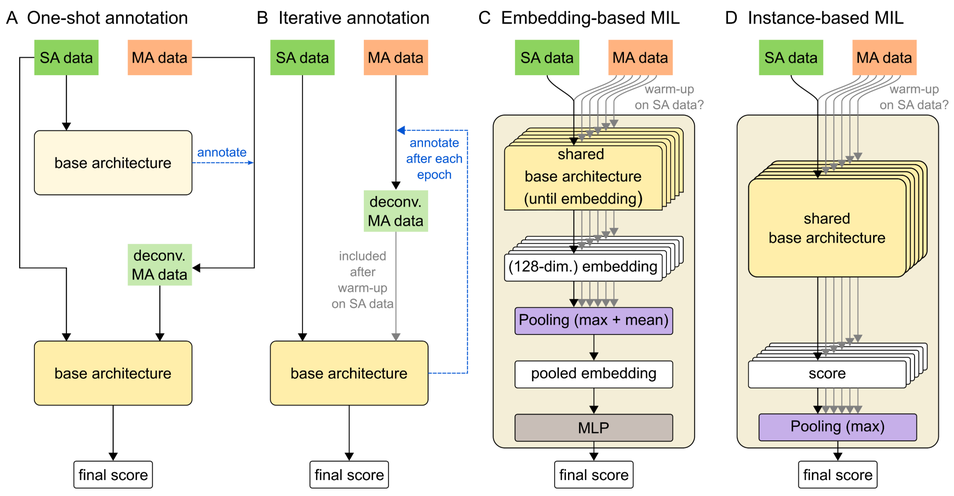


Figure 2. We compare two offline and two online deconvolution approaches

Training Data

NetMHCpan4.1 training data [1], which is split into 5 folds (with 8mer overlap minimization) used for cross validation, filtered to include only HLA alleles:

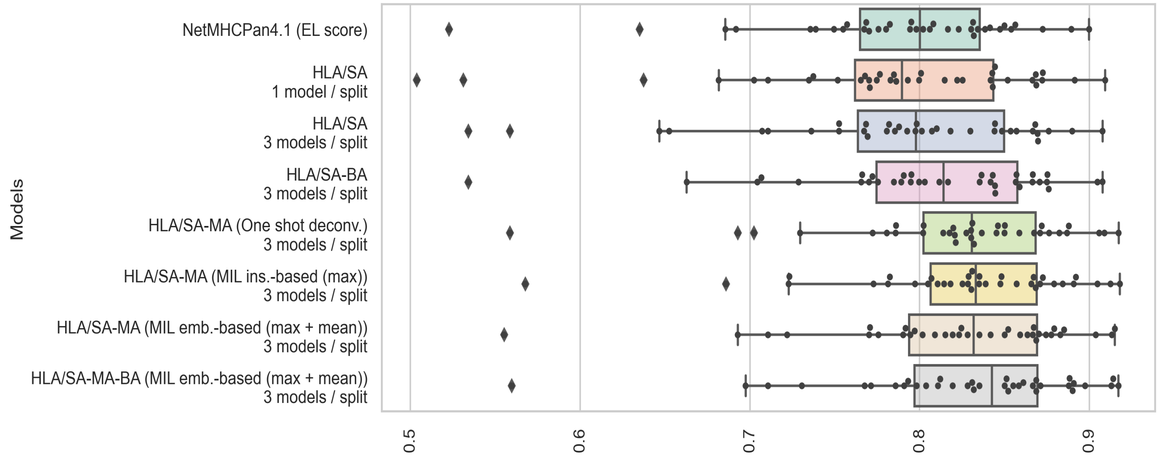
- SA eluted ligand data (≈ 3.7M data points, 5.67% positives, unequal across 130 HLA alleles)
- MA eluted ligand data (≈ 7M data points, 5.27% positives, unequal across 102 cell lines, covering 91 HLA alleles)
- BA data (≈ 170K data points, scaled IC50 values, unequal across 111 HLA alleles)

Results

Evaluation Data

- SA eluted ligands evaluation set from [1] (≈ 946K data points, 5.04% positives, unequal across 36 HLA alleles)
- CD8 epitopes evaluation set from [1] containing 1,660 experimentally validated epitopes from IEDB (combined from ELISpot and multimer assays)
- Schuster & Löffler MA evaluation set
  - Ligandome of ovarian [4] and colorectal [5] cancer cells
  - Raw mass spectrometry data have been re-processed with Spectrum Mill
  - Negatives of uniform length have been generated from proteins known to be expressed with roughly 1:5000 balance for each genotype yielding ≈ 12M data points

Figure 3. Per-allele PPV on the HLA MS ligands



Evaluation Metrics

- **PPV** = (# true positives among the k highest-scoring peptides) / k where k is the total number of positives → **the higher the better**
- **FRANK score** = (# negatives ranked higher than the true epitope) / # negatives where the negatives are all other peptides from the epitope's source protein → **the lower the better**

Both BA and MA boost mean PPV but not in a complementary manner.

Embedding-based MIL performs best, but performance is highly impacted (57% on scores' mean) without inclusion of BA data.

Trade-off for embedding-based MIL: high performance in online scoring vs. higher interpretability (i.e., most likely presenting allele) in offline scoring.

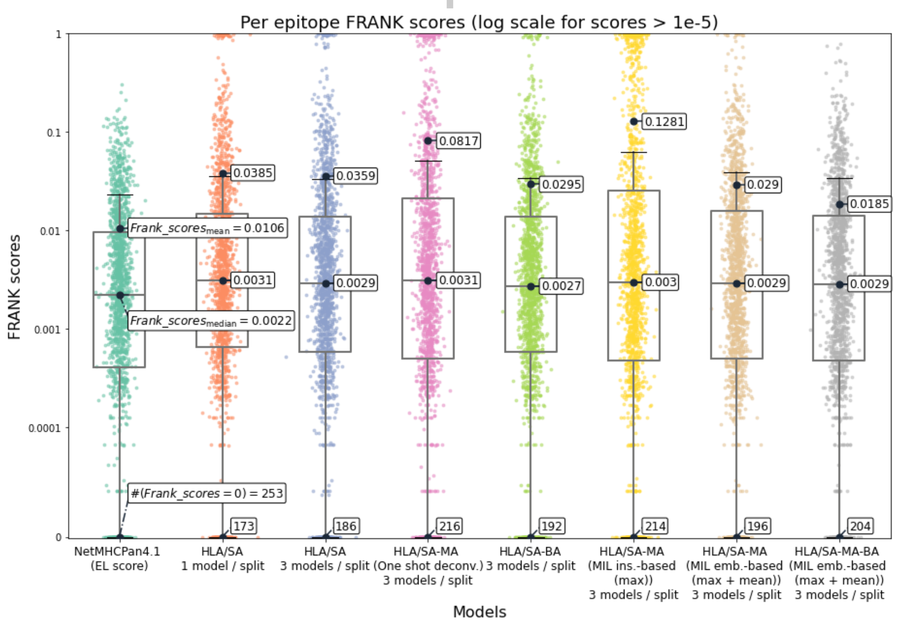


Figure 4. Per-epitope FRANK scores on CD8 epitopes

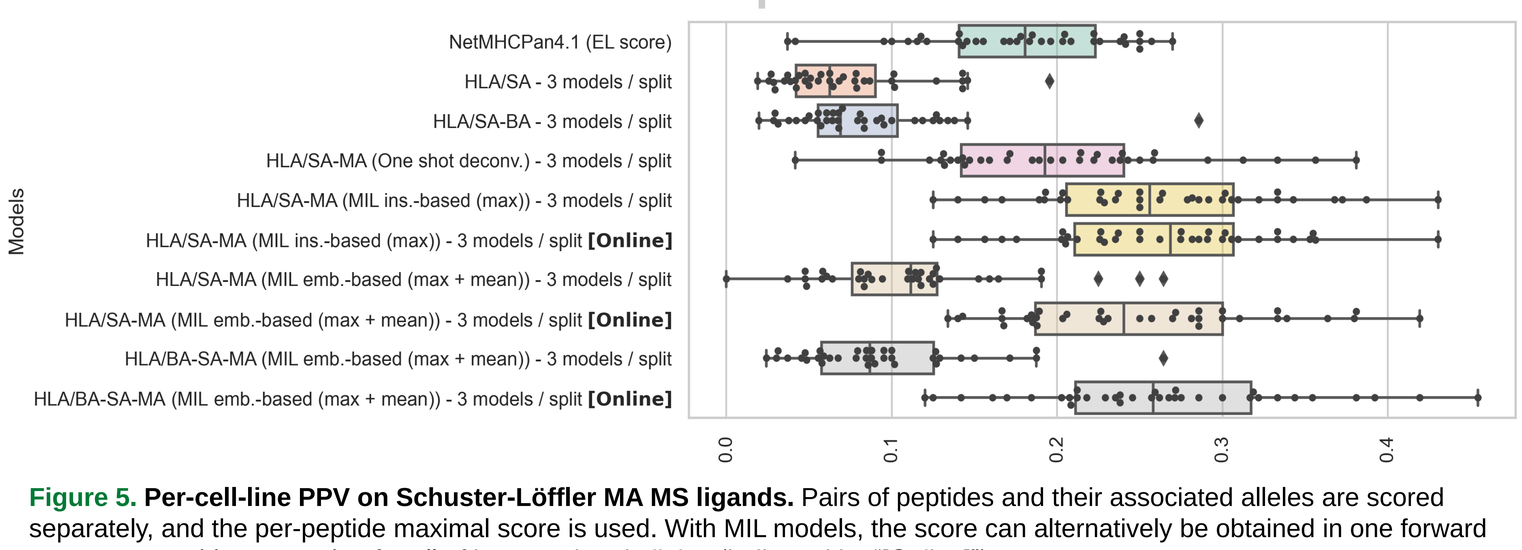


Figure 5. Per-cell-line PPV on Schuster-Löffler MA MS ligands. Pairs of peptides and their associated alleles are scored separately, and the per-peptide maximal score is used. With MIL models, the score can alternatively be obtained in one forward pass on a peptide accounting for all of its associated alleles (indicated by "[Online]").

Conclusions

- **Integrating MA data through the use of MIL architectures improves performance on ligand data (+4.7% on per-allele mean PPV on the SA eluted ligands set, +3.5-fold improvement on per-cell-line mean PPV on the Schuster & Löffler MA set).**
- **On the CD8 evaluation set, both BA and MA data drive performance with complementary respective added values, while NetMHCPan4.1 still outperforms the best MIL approaches.**
- **Instance-based MIL with max pooling has higher interpretability than embedding-based MIL architectures.**
- **Next steps include investigating iterative annotation, one-shot annotation with warm-up training and BA data integration, as well as completing the benchmark with additional evaluation data.**

References

1. Reynisson et al., 2020, Nucleic Acids Research, DOI: 10.1093/nar/gkaa379
2. Pyke et al., 2021, Molecular Cellular Proteomics, DOI: 10.1016/j.mcpro.2021.100111
3. Thrift et al., 2022, bioRxiv, DOI: 10.1101/2022.12.08.519673
4. Schuster et al., 2017, Proceedings of the National Academy of Sciences DOI: 10.1073/pnas.1707658114
5. Löffler et al., 2018, Cancer Research, DOI: 10.1158/0008-5472.CAN-17-1745

Presented at CIMT 2023