Science Advances
AAAS

Supplementary Materials for

Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning

Jonas B. Nilsson *et al.*

Corresponding author: Morten Nielsen, morni@dtu.dk

Sci. Adv. **9**, eadj6367 (2023)
DOI: 10.1126/sciadv.adj6367

The PDF file includes:

Figs. S1 to S10
Legend for table S1

Other Supplementary Material for this manuscript includes the following:

Table S1

Supplementary Figures

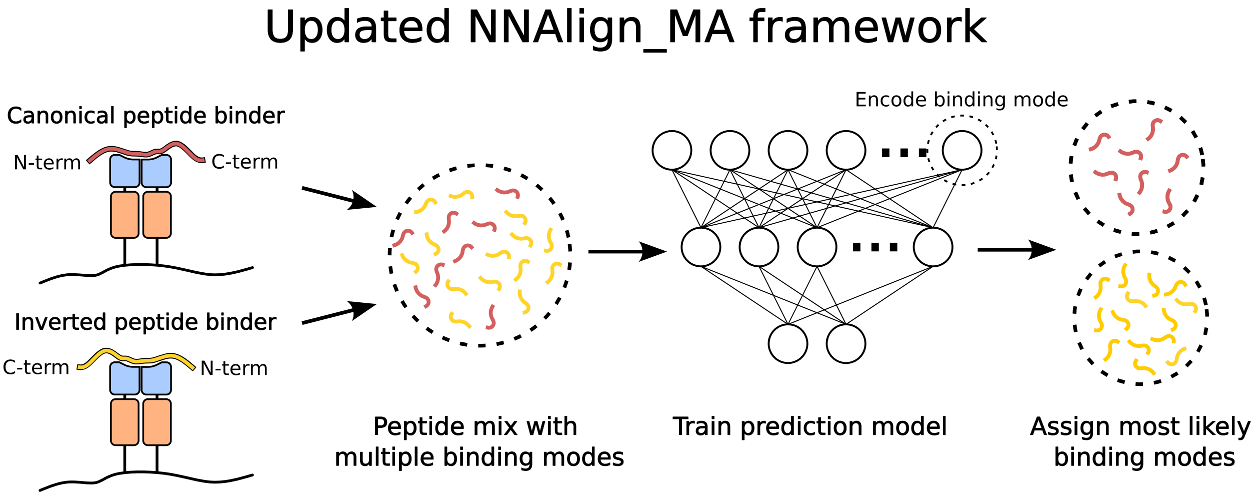

Figure S1: Overview of the updated NNAlign_MA machine learning framework. The method takes into account the presence of multiple peptide binding modes, with some peptides binding in a canonical N-to-C-terminal fashion, while some bind in an inverted C-to-N-terminal mode. The peptide mix is fed to an updated machine learning framework which learns the optimal binding mode for each peptide during training. Our new NNAlign_MA method includes a new input neuron which encodes the peptide binding mode (0 for canonical and 1 for inverted), and the method assigns to each peptide both the optimal binding core offset and the binding mode giving the highest prediction score from the network during training and prediction. This allows for deconvolution of the peptide mix in terms of the two binding modes and thus better interpretation of the binding motifs.

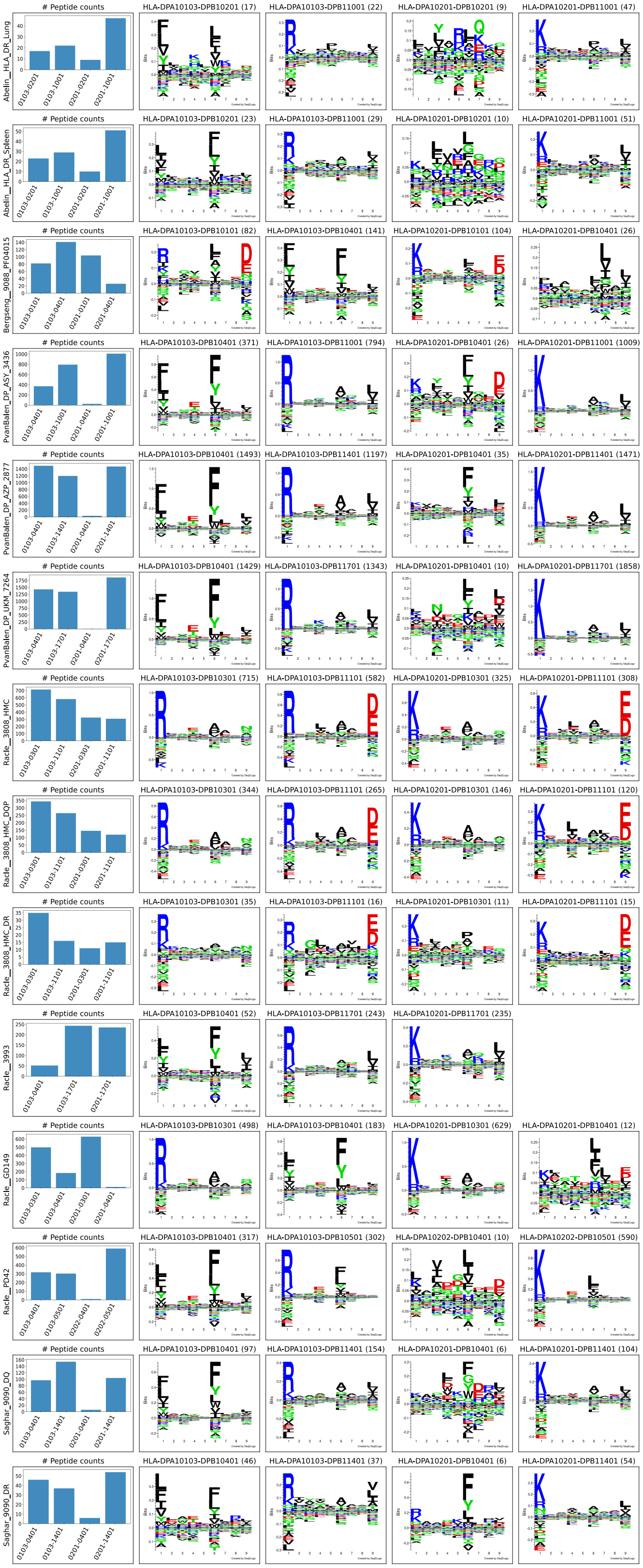

Figure S2: DP motif deconvolution for DP-heterozygous samples. In each sample, trash peptides with percentile rank greater than 20 are not included. For the bar charts in the left-most panel, the molecule names are shortened to include only the HLA type numbers (example: HLA-DPA10103-DPB10401 is displayed as 0103-0401).

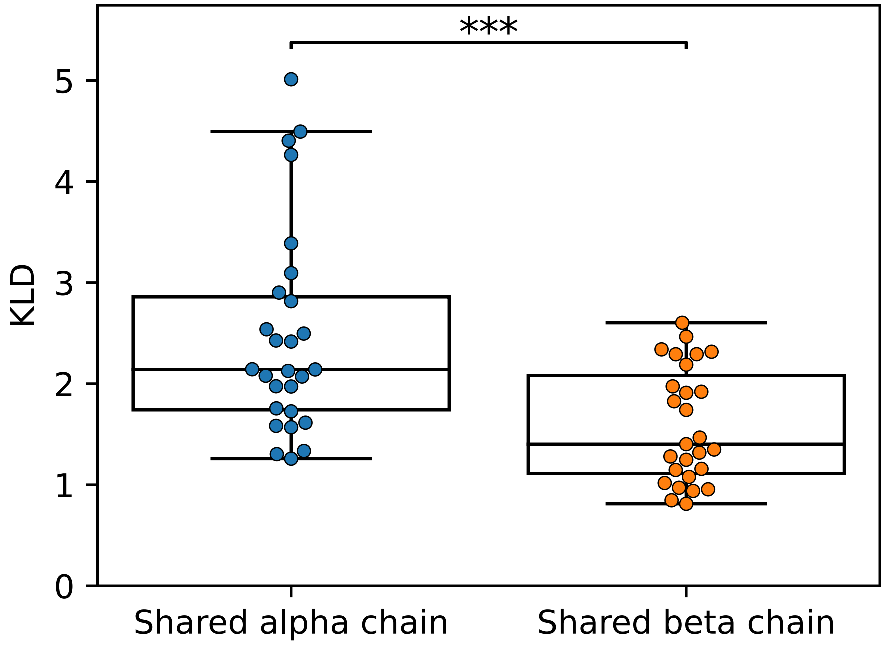

Figure S3: KLDs between motifs of molecules with shared alpha or beta chain in DP-heterozygous sample. Each point is the KLD between the motifs of two molecules within a given heterozygous sample sharing either the same alpha or beta chain. The result of a two-sample unpaired t-test is shown ($N = 27$ in both groups, ***: $p <$ 0.001).

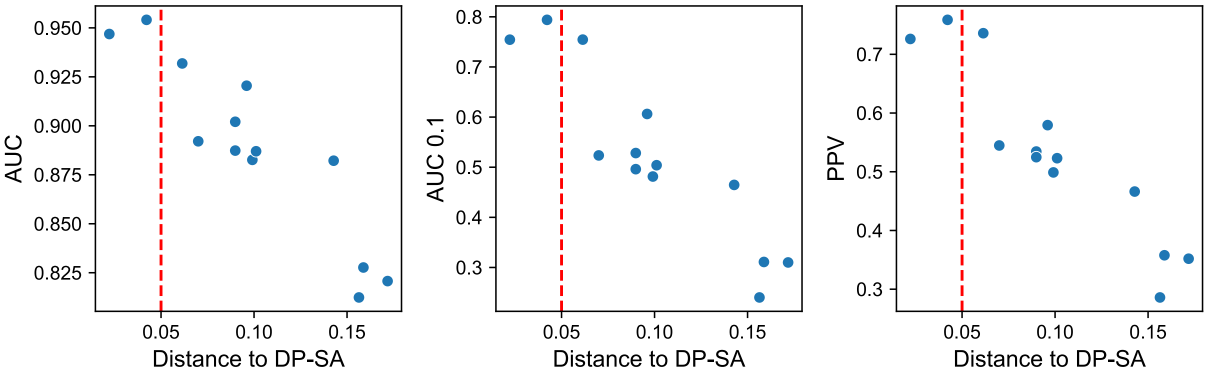

Figure S4: AUC, AUC 0.1 and PPV performance on the DP data from Van Balen *et al.* (*32-34*) by the method trained without this data, as a function of the distance to the DP molecules in the method's SA training data. Each point is a DP molecule in the DP data from Van Balen *et al.* (*32-34*) which is not part of the method's SA training data. The red dashed lines indicate the 0.05 distance threshold used in the coverage analysis.

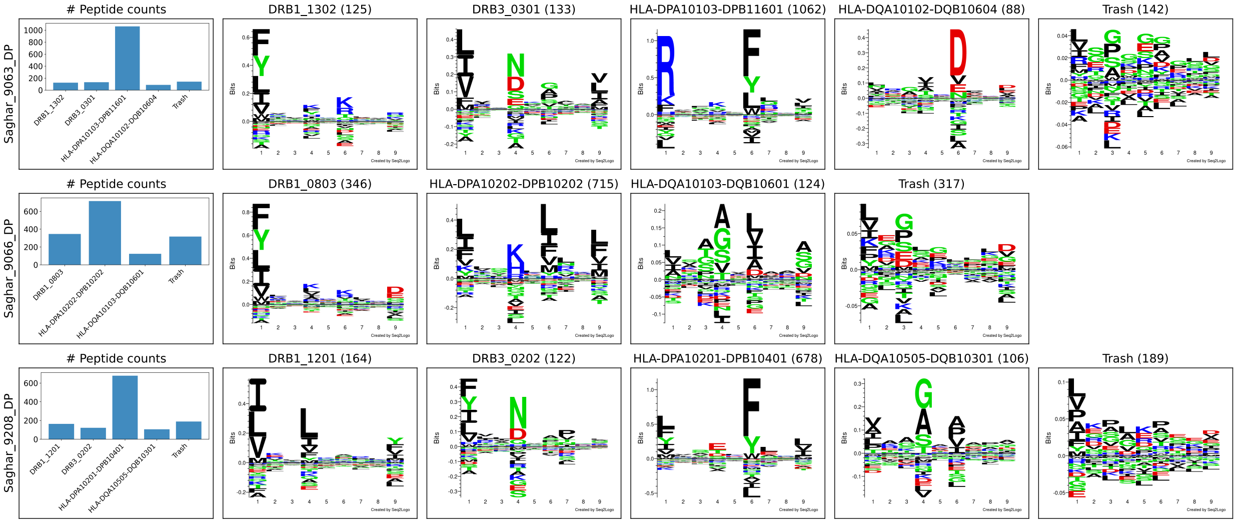

Figure S5: Motif deconvolution of the datasets generated for this study. All peptides with percentile rank greater than 20 are placed in the 'Trash' clusters.

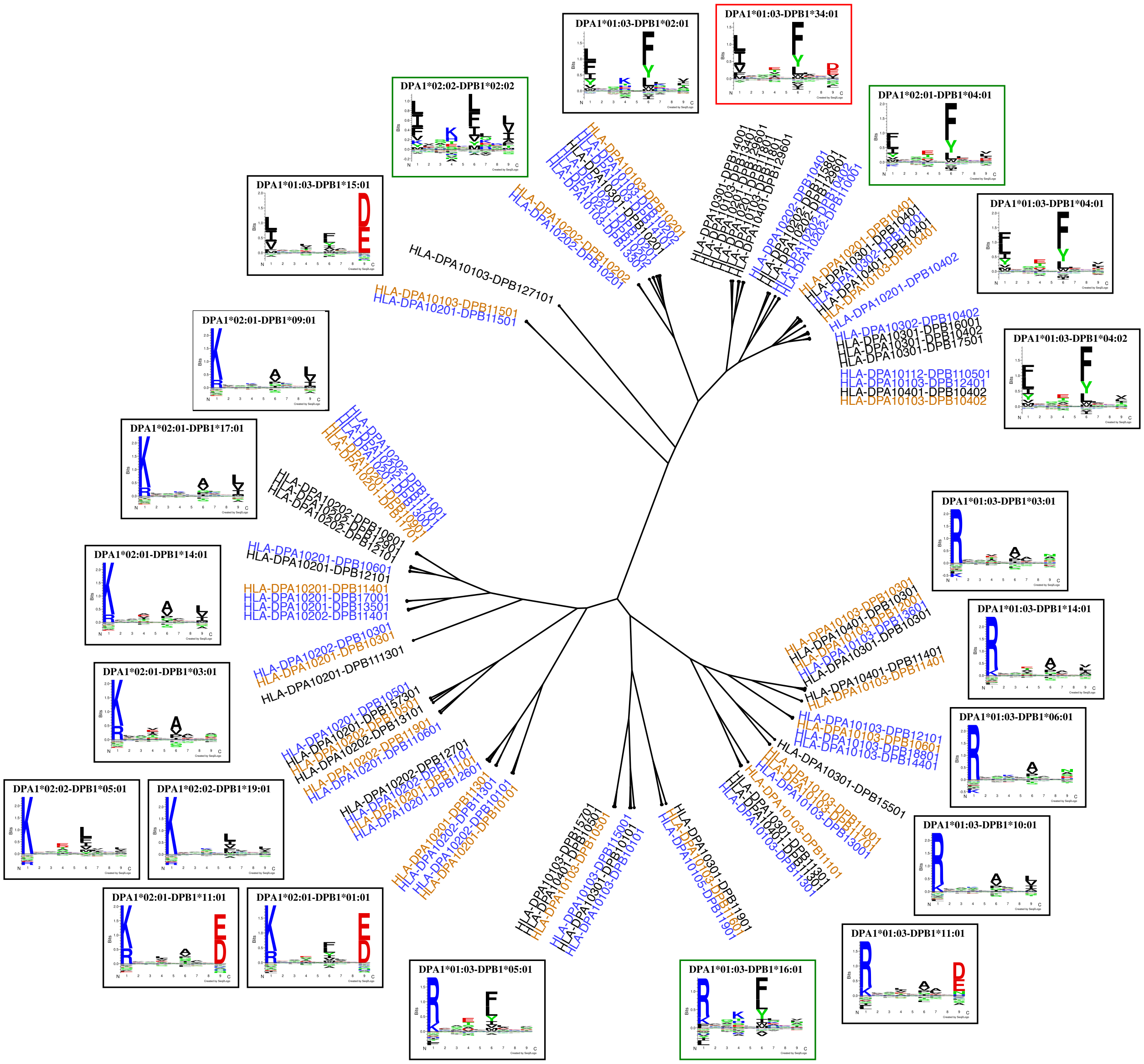

Figure S6: DP specificity tree for the retrained method. Orange molecules have peptide coverage corresponding to at least 50 high-confidence ligands, and blue molecules have a pseudo-sequence distance of at most 0.05 to an orange molecule. Logos in green frames correspond to molecules in the DP data generated for our study.

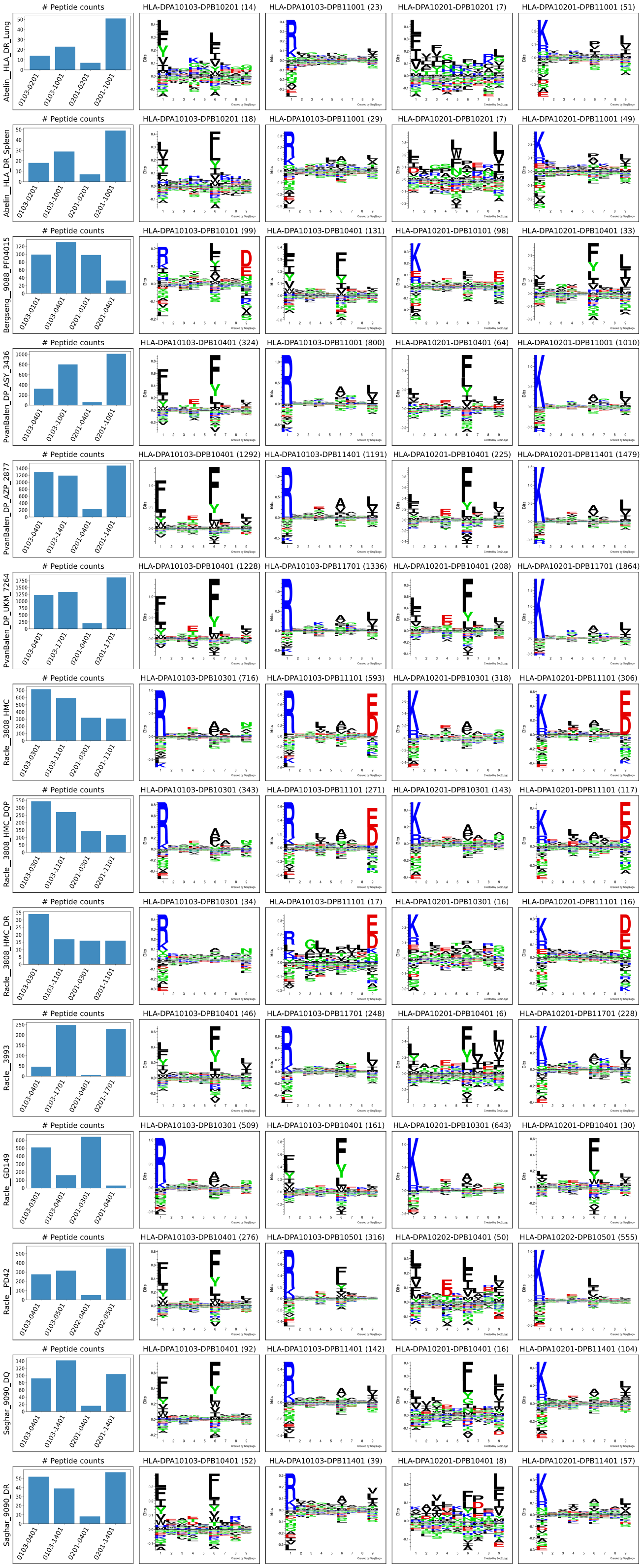

Figure S7: DP motif deconvolution for DP-heterozygous samples in the final method trained with the DP-eluted datasets generated in this study. In each sample, trash peptides with percentile rank greater than 20 are not included. For the bar charts in the left-most panel, the molecule names are shortened to include only the HLA type numbers (example: DPA10103-DPB10401 is displayed as 0103-0401).

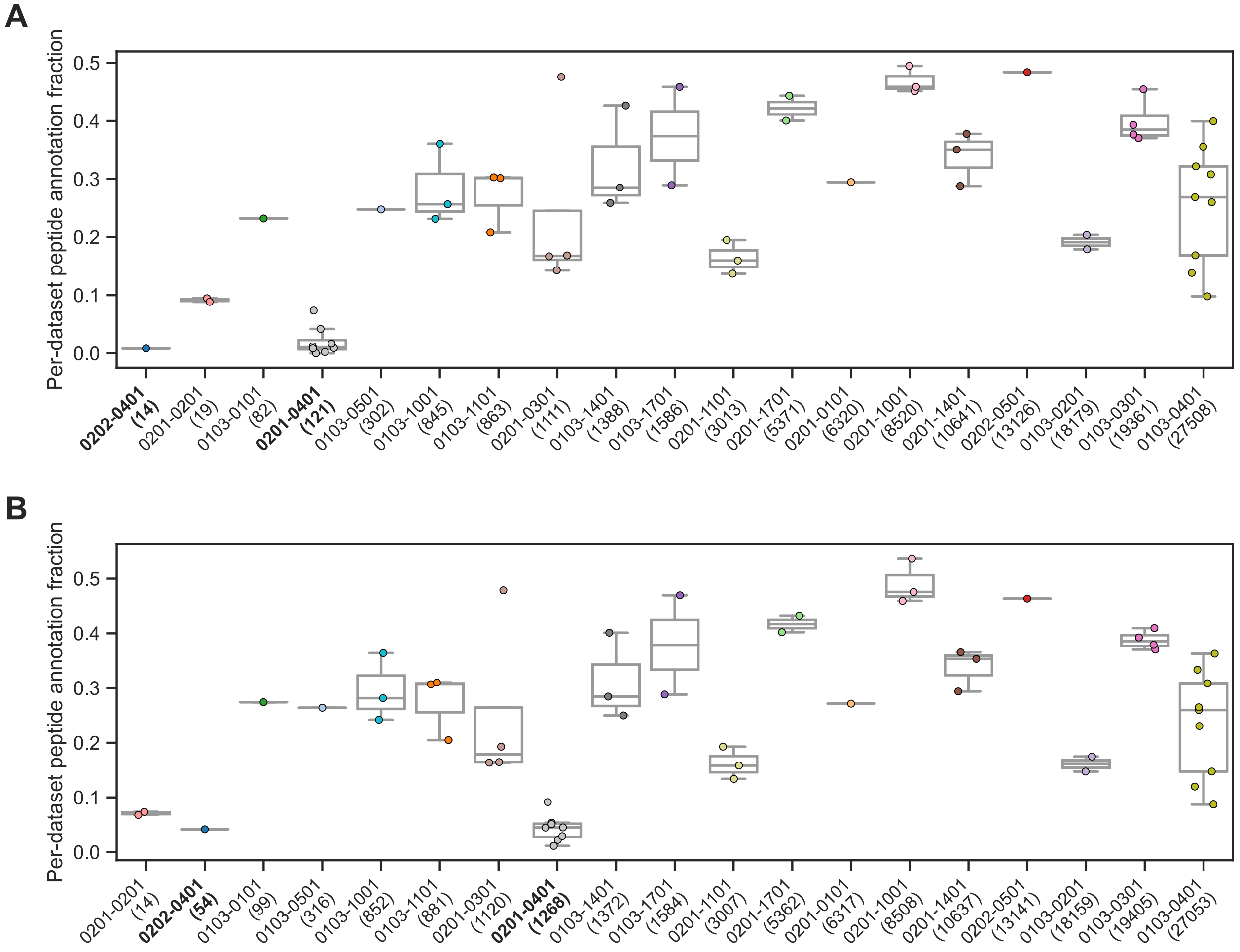

Figure S8: Contribution of HLA-DP molecules in DP-heterozygous datasets for the methods trained without (A) and with (B) the data generated in this study covering DPA1*02:01-DPB1*04:01, DPA1*02:02-DPB1*02:02 and DPA1*01:03-DPB1*16:01. Each boxplot shows the DP peptide annotation fraction for the given molecule across DP-heterozygous datasets. The molecules are labelled by the HLA type numbers, such that e.g. DPA1*02:01-DPB1*04:01 is shown as 0201-0401. Molecules are sorted by their total peptide annotation count across datasets (note that the order is not the same for each panel). The number in parenthesis below each molecule label indicates the total number of peptide annotations for the given molecule across all datasets. Labels for DPA1*02:01-DPB1*04:01 and DPA1*02:02-DPB1*04:01 are highlighted in bold in both panels.

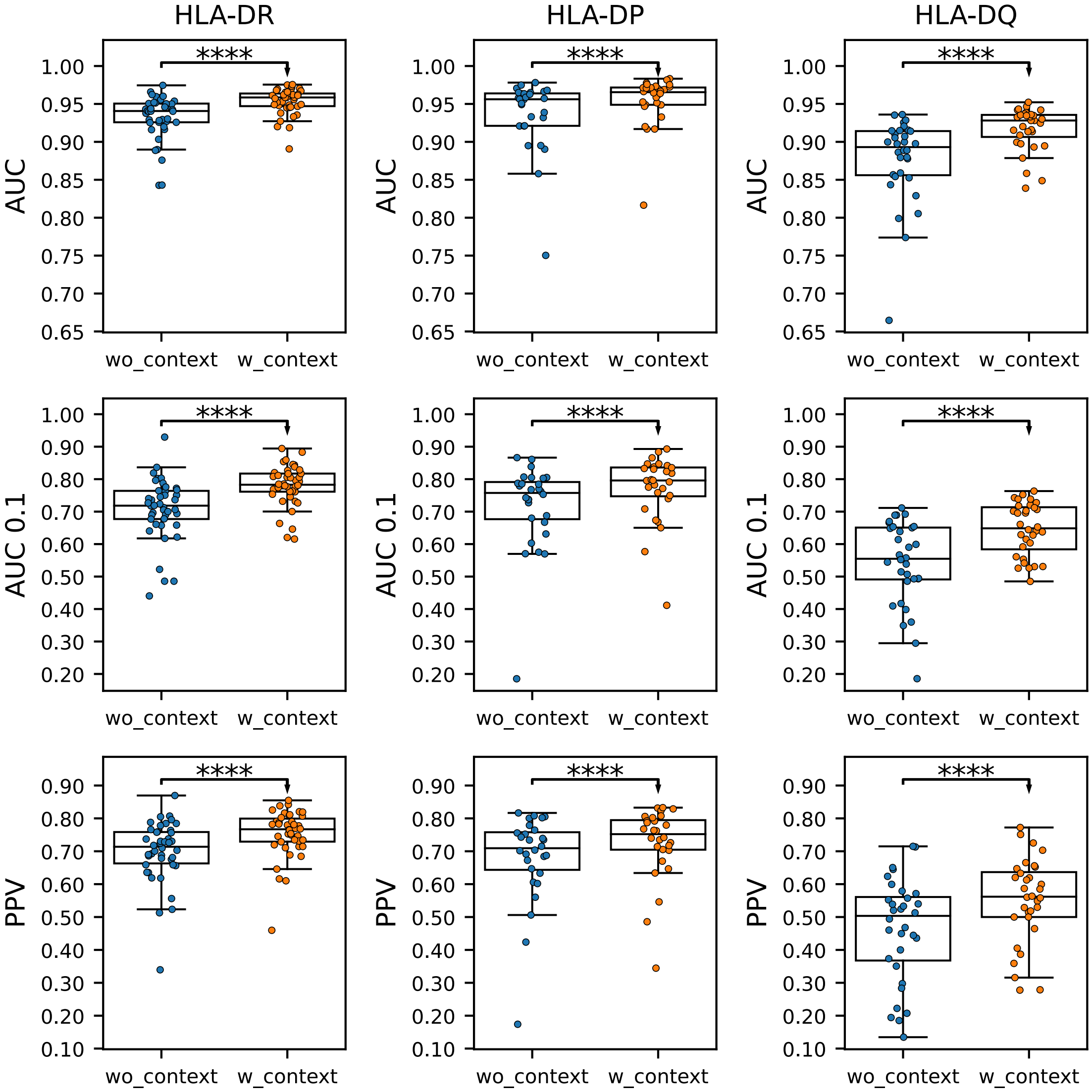

Figure S9: Performance across HLA class II loci of models trained without (wo_context) and with (w_context) context encoding. Each point is the performance in terms of AUC, AUC 0.1 or PPV for a given molecule. In both methods, peptide inversion was included during training and prediction. Results of one-tailed binomial tests are shown, with arrowheads indicating the direction of the test (*N* = 42, *N* = 28 and *N* = 32 for HLA-DR, HLA-DP and HLA-DQ, respectively, ****: $p < 0.0001$).

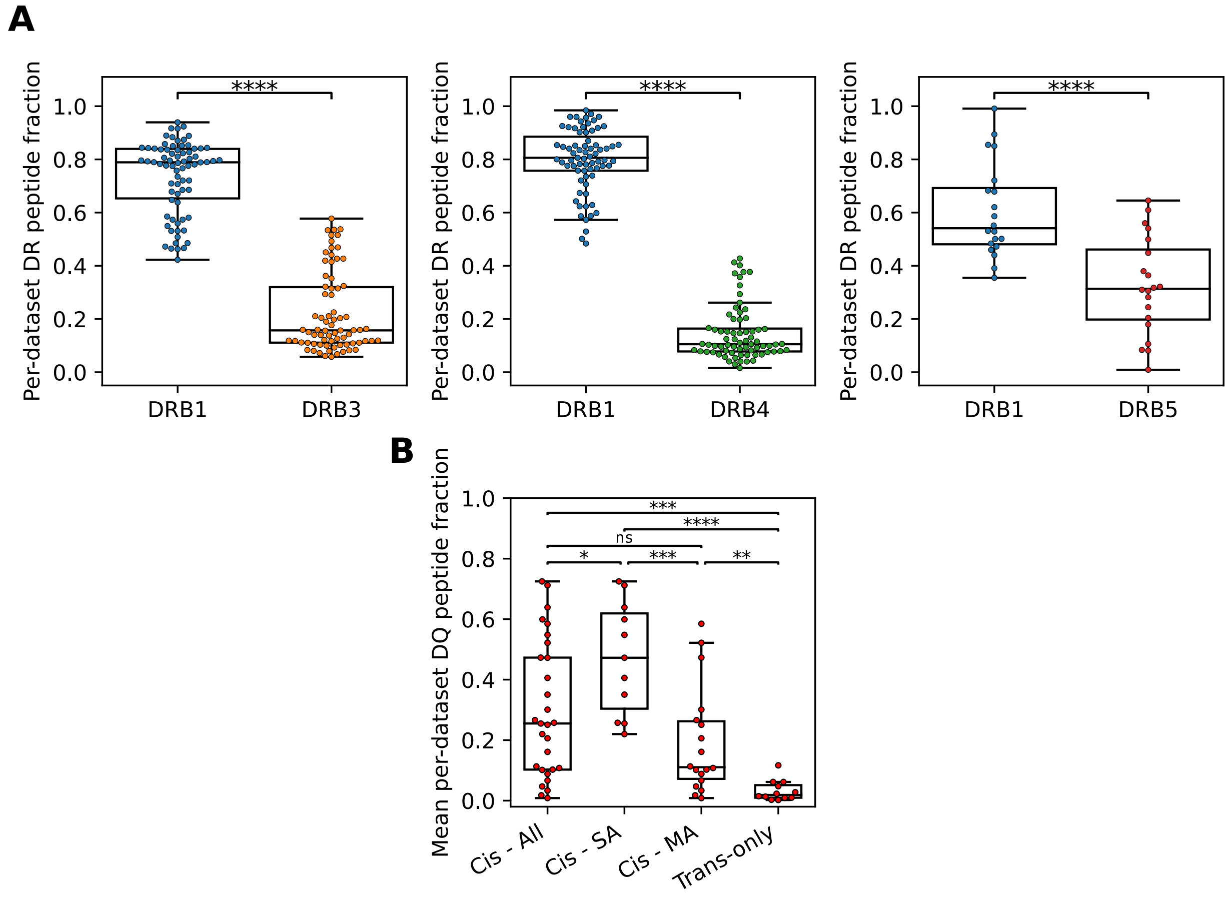

Figure S10: The roles of HLA-DR and HLA-DQ in shaping the immunopeptidome. A: Relative contribution of DRB3, 4 and 5 compared to DRB1 in the motif deconvolution. Each point corresponds to the DR peptide annotation fraction for a given DR molecule in a given dataset (*N* = 74, *N* = 71 and *N* = 20 for the DRB1 vs DRB3, DRB1 vs DRB4 and DRB1 vs DRB5 groups, respectively). **B:** Contribution of cis (N = 29) and trans-only (*N* = 12) DQ variants in DQ-heterozygous datasets. The cis variants are divided into three categories, namely all cis variants (Cis - All, *N* = 29), cis variants found in the DQ-SA training data (Cis - SA, *N* = 11), and cis variants found in the DQ-MA training data (Cis - MA, *N* = 18). Here, molecules found in the homozygous DQ datasets from Nilsson *et al.* 2023 (*25*) are grouped under the Cis - SA category. The results of two-sample unpaired t-tests are shown in panel A and B (****: $p < 0.0001$, ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, ns: not significant).

Supplementary Table Captions (tables provided in separate files)

Table S1: Overview of training data used in NetMHCIIpan-4.3. For three of the MA datasets from Balen *et al.* (*32-34*) (PvanBalen_DP_ASY_3436, PvanBalen_DP_AZP_2877 and PvanBalen_DP_UKM_7264), the DP typing was listed to have only two DP molecules, even though the cell lines are heterozygous on both the DP alpha and beta chain. As the cell lines should express four possible alpha-beta combinations, the typings for these cell lines were modified to include all four possible DP molecules, to allow for unbiased annotation of peptides during training and prediction.