Footprints of antigen processing boost MHC class II natural ligand binding predictions

Carolina Barra[1]*, Bruno Alvarez[1]*, Massimo Andreatta[1], Søren Buus[2] and Morten Nielsen[1,3]

* These authors contibuted equally to the work

[1] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina

[2] Department of Immunology and Microbiology, Faculty of Health Sciences, University of Copenhagen, Denmark

[3] Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

Abstract

Major Histocompatibility complex class II (MHC-II) molecules present peptide fragments to T cells for immune recognition. Current predictors for peptide:MHC-II binding are trained on binding affinity data, generated *in-vitro* and therefore lacking information about antigen processing. For the first time, we here describe prediction models of peptide:MHC-II binding trained directly on naturally eluted peptides, and show that these, in addition to peptide binding to the MHC, incorporate identifiable rules of antigen processing. In fact, we observed detectable signals of protease cleavage at defined positions of the peptides. We also hypothesize a role of the length of the terminal ligand protrusions for trimming the peptide to the epitope presented. The results of integrating binding affinity and eluted ligand data in a combined model demonstrate improved performance for the prediction of MHC-II ligands, and foreshadow a new generation of improved peptide:MHC-II prediction tools of considerable importance for understanding and manipulating immune responses.

Introduction

Major Histocompatibility Complex Class II (MHC-II) molecules play a central role in the immune system of vertebrates. MHC-II present exogenous, digested peptide fragments on the surface of antigen presenting cells, forming peptide-MHC-II complexes (pMHCII). On the cell surface, these pMHCII complexes are scrutinized, and if certain stimulatory conditions are met, a T

helper lymphocyte may recognize the pMHC-II and initiate an immune response (Rudolph, Stanfield, & Wilson, 2006).

The precise rules of MHC class II antigen presentation are influenced by many factors including internalization and digestion of extracellular proteins, the peptide binding motif specific for each MHC class II molecule, and the transport and surface half-life of the pMHCIIs. The MHC-II binding groove, unlike MHC class I, is open at both ends. This attribute facilitates peptide protrusion out of the groove, thereby allowing longer peptides (and potentially whole proteins) to be loaded onto MHC-II molecules (Kim et al., 2014; Sette, Adorini, Colon, Buus, & Grey, 1989). Peptide binding to MHC-II is mainly determined by interactions within the peptide binding groove, which most commonly encompass a peptide with a consecutive stretch of 9 amino acids (Andreatta, Jurtz, et al., 2017). Ligand residues protruding from either side of the MHC binding groove are commonly known as Peptide Flanking Regions (PFRs). The PFRs are variable in length and composition, and affect both the peptide MHC-II binding (Lovitch, Pu, & Unanue, 2006), and the subsequent interaction with T cells (Arnold et al., 2002; Carson, Vignali, Woodland, & Vignali, 1997; Godkin et al., 2001). The open characteristic of the MHC-II binding groove does not constrain the peptides to a certain length thereby increasing the diversity of sequences that a given MHC-II molecule can present. Also, MHC-II molecules are highly polymorphic, and their binding motifs have appeared to be more degenerate than MHC-I motifs (Andreatta, Schafer-Nielsen, Lund, Buus, & Nielsen, 2011; Hammer et al., 1993; Sturniolo et al., 1999).

Considering all the aspects mentioned above, MHC-II motif characterization and rational identification of MHC-II ligands and epitopes is a highly challenging and costly endeavor. Because MHC-II is a crucial player in the exogenous antigen presentation pathway, considerable efforts have been dedicated in the past to develop efficient experimental techniques for MHC-II peptide binding quantification. The traditional approach to quantify peptide-MHC-II binding relies on measuring binding affinity, either as the dissociation constant Kd of the complex (Roche & Cresswell, 1990; Hall et al., 2002), or in terms of IC50 (concentration of the query peptide which displaces 50% of a bound reference peptide) (Buus, Sette, Colon, Miles, & Grey, 1987). To date, data repositories such as the Immune Epitope Database (IEDB) (Vita et al., 2015) have collected more than 150,000 measurements of peptide-MHC-II binding interactions. Such data have been used during the last decades to develop several prediction methods with the ability to predict binding affinities to the different

alleles of MHC class II. While the accuracy of these predictors has increased substantially over the last decades due the development of novel machine learning frameworks and a growing amount of peptide binding data being available for training (Andreatta, Trolle, et al., 2017), state-of-the-art methods still fail to accurately predict accurately MHC class II ligands and T cell epitopes (Gowthaman & Agrewala, 2008; Wang et al., 2008).

Recent technological advances in the field of mass spectrometry (MS) have enabled the development of high-throughput assays, which in a single experiment can identify several thousands of peptides eluted of MHC molecules (reviewed in Caron et al., 2015). Large data sets of such naturally presented peptides have been beneficial to define more accurately the rules of peptide-MHC binding (Abelin et al., 2017; Bassani-Sternberg & Gfeller, 2016; Bergseng et al., 2015; Chong et al., 2017; Clement et al., 2016; Jurtz et al., 2017; Ooi et al., 2017). For several reasons, analyses and interpretation of MS eluted ligand data is not a trivial task. Firstly, because any given individual constitutively expresses multiple allelic variants of MHC molecules; thus, the ligands detected by MS are normally a mixture of specificities, each corresponding to a different MHC molecule. Secondly, as mentioned above, MHC-II ligands can vary widely in length, and identification of the binding motifs requires a sequence alignment over a minimal binding core. Finally, data sets of MS ligands often contain contaminants and false spectrum-peptide identifications, which add a component of noise to the data. We have earlier proposed a GibbsCluster method capable of dealing with all these issues, allowing the characterization of binding motifs and the assignment of probable MHC restrictions to individual peptides in such MS ligand data sets (Andreatta, Alvarez, & Nielsen, 2017; Andreatta, Lund, & Nielsen, 2013).

Because naturally eluted ligands incorporate information about properties of antigen presentation beyond what is obtained from *in-vitro* binding affinity measurements, large MS-derived sets of peptides can be used to generate more accurate prediction models of MHC antigen presentation (Abelin et al., 2017; Bassani-Sternberg & Gfeller, 2016; Jurtz et al., 2017). As shown recently, generic machine learning tools, such as NNAlign (Andreatta et al., 2011; Nielsen & Andreatta, 2017), can be readily applied to individual MS data sets, which in turn can be employed for further downstream analyses of the immunopeptidome (Alvarez, Barra, Nielsen, & Andreatta, 2018). The amount of MHC molecules characterized by MS eluted ligand data is, however, still limited. This has led us to suggest a machine learning framework where peptide binding data of both MS and *in-vitro* binding assays are merged in the training of the

prediction method (Jurtz et al., 2017). This approach has proven highly powerful for MHC class I, but has not, to the best of our knowledge, been applied to MHC class II.

Undoubtedly, antigen processing plays a critical role in generating CD4+ T cell epitopes presented by MHC class II molecules. It is assumed that endo- and exo-peptidase activities, both before and after binding to the MHC-II molecule, play a key role in the generation and trimming of MHC class II ligands (Blum, Wearsch, & Cresswell, 2013; Lippolis et al., 2002). However, the precise rules of MHC class II antigen processing are poorly understood. Kropshofer et al. identified patterns of protein cleavage e.g. the presence of proline at the penultimate N and C terminal position of HLA-DR ligands (Kropshofer et al., 1993). In contrast, Bird et al. suggested that endolysosomal proteases have a minor and redundant role in peptide selection leading to the conclusion that the effect of processing on the generation of antigenic peptides is “relatively non-specific” (Bird, Trapani, & Villadangos, 2009). Given this context, it is perhaps not surprising that limited work has been aimed at integrating processing signals into a prediction framework for MHC-II ligands.

In this work, we have analysed large data sets of MS MHC-II eluted ligands obtained from different research laboratories covering two HLA-DR molecules with the purpose of investigating the consistency in the data; quantifying the differences in binding motifs contained with such MS eluted data compared to traditional *in-vitro* binding data; defining a machine-learning framework capable of integrating information from MS eluted ligand and *in-vitro* binding data into a prediction model for MHC-II peptide interaction prediction; and finally evaluating if inclusion of potential signals from antigen processing are consistent between different data sets and can be used to boost performances of the developed prediction models.

Results

1. Data filtering and motif deconvolution

We first set out to analyse the different MS data sets of eluted ligands. Data were obtained from two recent publications; Ooi et al., 2017 (termed P) and Clement et al., 2016 (termed S) covering the HLA-DRB1*01:01 and HLA-DRB1*15:01 MHC class II molecules. Data were obtained from either human (termed h) or HLA-DR transfected mice (termed m) cell lines. Using this syntax, DRB1 Ph corresponds to the HLA-DRB1*01:01 data from the human cell in the study by Ooi et al. (for more details see Materials and Methods). Here, we applied the GibbsCluster method with default parameters for MHC class II to both filter out potential noise

and to identify the binding motif(s) contained in each data set. The result of this analysis is shown in figure 1, and confirms the high quality of the different ligand data sets. In all data sets less than 7% of the peptides were identified as noise (assigned to the trash cluster), and in all cases does GibbsCluster find a solution with the number of clusters matching the number of distinct MHC specificities present in a given data set. In this context, the DR15-Ph is of particular interest, since this data set was obtained from a heterozygous cell line expressing two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (shortened here as DR15/51). Consequently, this data set contains a mixture of peptides eluted from both of these HLA-DR molecules. As apparent from figure 1, the GibbsCluster method was able to handle this mixed data set, and correctly identified two clusters with distinct amino acid preferences at the anchor positions P1, P4, P6 and P9. Further, a comparison of the motifs identified from the different data sets sharing the exact same HLA-DR molecules revealed a very high degree of overlap, again supporting the high accuracy of both the MS eluted ligand data and of the GibbsCluster analysis tool.

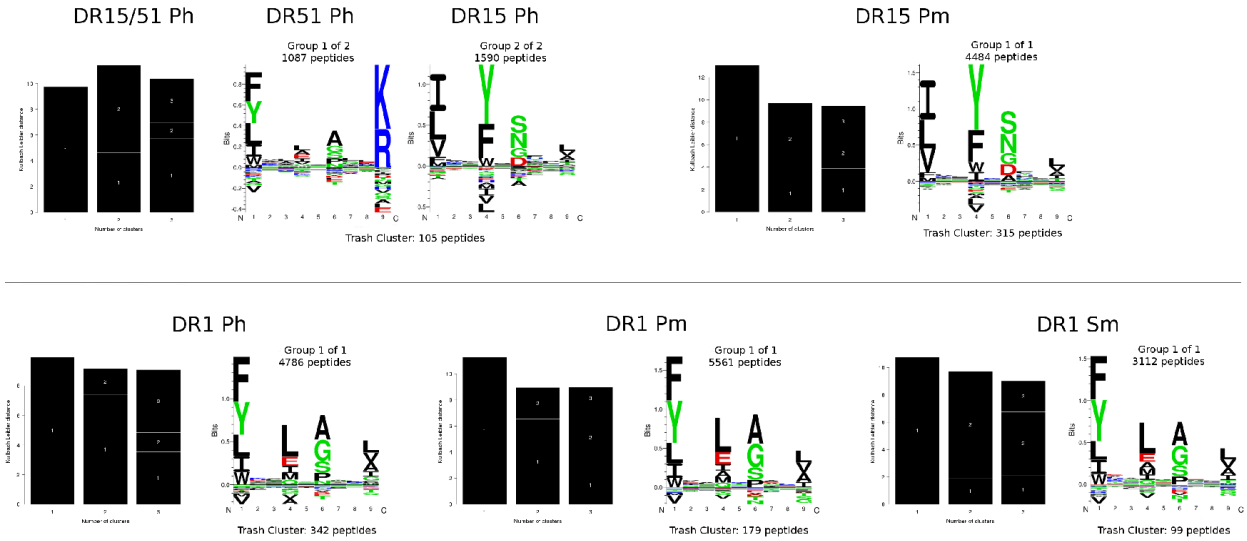


Figure 1. GibbsCluster output for the different ligand data sets. For each data set, we show the Kullback-Leibler distance (KLD) histogram, and the motif of the GibbsCluster solution corresponding to the maximal KLD value. The upper row gives the results for the DR15/51 data set and the lower row the results for the DR1 data set. Note that DR15-Ph data set was obtained from a cell line which expresses two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (DR15/51).

2. Training prediction models on MHC class II ligand data

After filtering and deconvolution with GibbsCluster, MHC peptide binding prediction models were constructed for each of the 5 data sets corresponding to the majority clusters in figure 1. Since only one data set was available for HLA-DR51, this molecule was not included in the subsequent analyses. Models were trained using the NNAlign framework as described in materials and methods. The eluted ligand data sets (EL) were enriched with random natural peptides labeled as negatives, as described in Material and Methods. Models were likewise trained on relevant and existing data sets of peptide binding affinities (BA) obtained from the IEDB. Prior to training, both the eluted ligand and binding affinity data sets were filtered to include only peptides of length 11 to 19 amino acids, each data set was split into 5 partitions using a 9mer common-motif clustering, and models were trained and evaluated using 5 fold cross-validation as described in Material and Methods. These analyses revealed a consistent and high performance for the models trained on the different eluted ligand data sets (Table 1). In accordance with what has been observed earlier for MHC class I (Jurtz et al., 2017), the overall cross-validated performance of models trained on binding affinity data is lower than that of models trained on eluted ligand data.

Table 1: Cross-validated performance of NNAlign models generated on binding affinity (BA) and ligand elution (EL) data sets.

	BA single models			EL single models	
Allele	AUC 0.1	AUC	Dataset	AUC 0.1	AUC
DRB1*01:01	0.374	0.840	DR1 Ph	0.819	0.960
			DR1 Pm	0.888	0.986
			DR1 Sm	0.851	0.977
DRB1*15:01	0.287	0.843	DR15 Ph	0.901	0.987
			DR15 Pm	0.917	0.989

The binding motifs captured by the different models are shown in figure 2. As evidenced by identical anchor positions (P1, P4, P6 and P9) and virtually identical anchor residues, highly consistent motifs were obtained from the same HLA-DR molecules irrespective of the source of the peptide (i.e. whether they were obtained from human or mouse cells, or from one laboratory or another). This even extended to the motifs obtained when using binding data, although we do observe subtle, but consistent, differences between the binding motifs derived from eluted ligand and peptide binding affinity data, exemplified by the preference for E at P4 and for D at

P6 in the eluted ligand motifs for DR1 and DR15, respectively. Such preferences are absent from the motifs derived from the peptide binding affinity data.

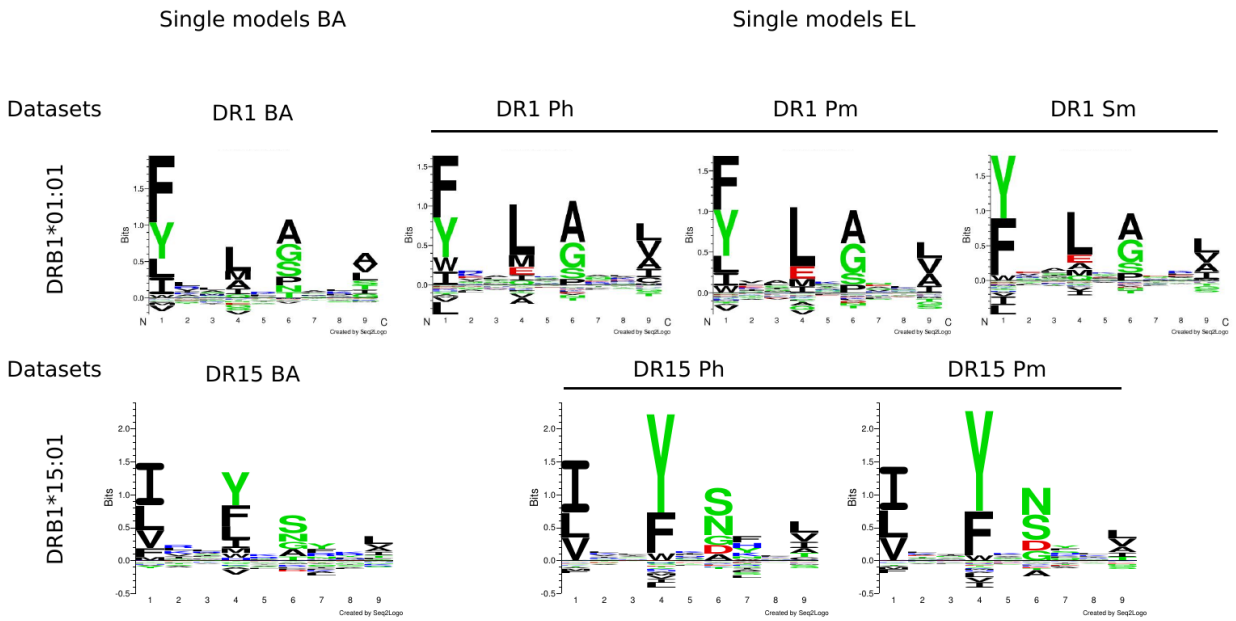


Figure 2. Logos for models trained using the NNAlign method (Nielsen & Andreatta, 2017) with eluted ligand or binding affinity data for the molecules DRB1*01:01 and DRB1*15:01. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 900.000 random natural peptides.

3. Training a combined prediction model on MHC-II binding affinity and ligand elution data

Earlier work on MHC class I has demonstrated that the information contained within eluted ligand and peptide binding affinity data is complementary, and that a prediction model can benefit from being trained in an integral way using both data types (Jurtz et al., 2017). Here we investigate if a similar observation could be made for MHC class II. As proposed by Jurtz et al, we extended an artificial neural network architecture with two output neurons to the NNAlign framework. In short, the model architecture takes as input a mixed set of two data types - in this case, eluted ligands and peptide binding affinity data. All model parameters except the parameters connected to the output prediction value are shared between the two data types. This allows the model to leverage shared information between the two data types, potentially leading to improved predictive performance (for details refer to Jurtz et al., 2017). As the datasets of eluted ligands are much larger than the available binding affinity datasets, the training was performed in a balanced way so that on average the same number of data points of each data type are used for training in each training iteration. Models were trained with the same model parameters as used for the single data type model, in a 5 fold cross-validation

manner, and performance was reported in terms of AUC and AUC0.1 (Table 2). Comparing the performance of the single data type to the multiple data type models for the different data sets, a consistent improvement in predictive performance was obtained when the two data types where combined. This improvement is similar to what we have previously observed for MHC class I predictions (Jurtz et al. 2017).

Table 2. Cross-validated performance measured in terms of AUC and AUC0.1 for models trained on combined binding affinity and ligand elution data.

		Combined models _BA		Combined models _EL	
Allele	Dataset	AUC 0.1	AUC	AUC 0.1	AUC
DRB1*01:01	DR1 Ph	0.385	0.845	0.823	0.966
	DR1 Pm	0.376	0.843	0.893	0.987
	DR1 Sm	0.381	0.843	0.867	0.980
DRB1*15:01	DR15 Ph	0.288	0.844	0.908	0.987
	DR15 Pm	0.294	0.846	0.917	0.990

Combined _BA and Combined _EL refer to output neuron for “binding affinity” or “eluted ligand” prediction, respectively.

Constructing the binding motif captured by the different models confirms the findings from the single data types model, with clearly defined and consistent binding motifs in all cases, and with subtle differences in the preferred amino acids at the anchor positions between motifs derived from the binding affinity and eluted ligand output value of the models.

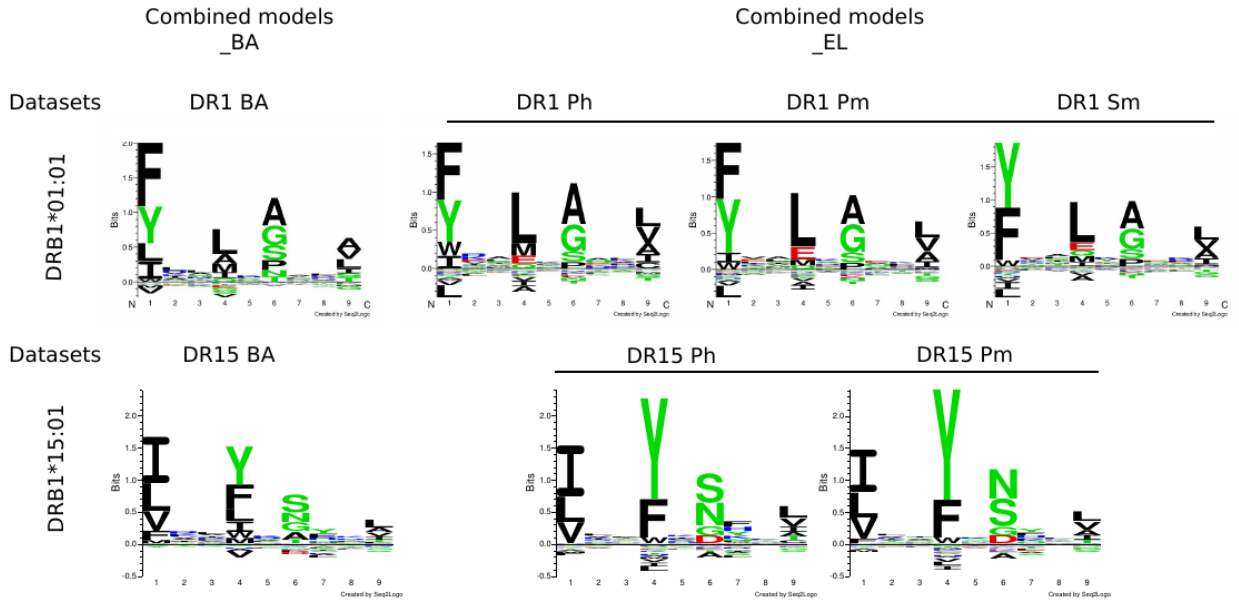


Figure 3. Logos for combined models trained with binding affinity and eluted ligand data for the molecules DRB1*01:01 and DRB1*15:01. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 900.000 random natural peptides.

We next turned to the issue of accurately predicting the preferred length of peptides bound to the different HLA-DR molecules. The MS eluted ligand data demonstrates a length preference for the two MHC class II molecules centered on a length around 14-16. Earlier prediction models have not been able to capture this length preference, and have in general had the bias of assigning higher prediction values to longer peptides (data not shown). We have earlier demonstrated that including information about the peptide length in a framework integrating MS eluted ligand and peptide binding affinity data, allows the model to capture the length preference of the two data sets (Jurtz et al., 2017). Applying a similar approach to the MHC class II data, we obtain the results shown in figure 4, confirming that also for class II the model is capable of approximating the preferred length preference of each molecule.

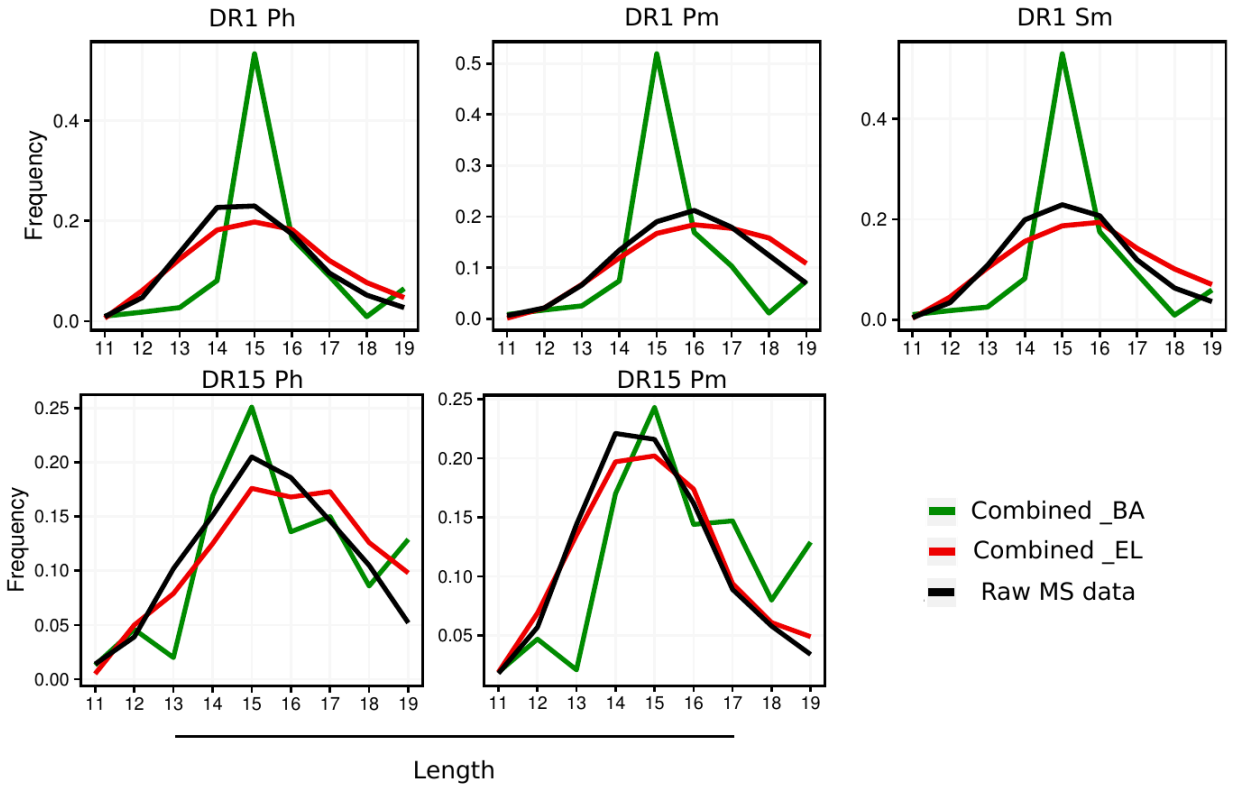


Figure 4. Length preference learned by combined models trained with both binding affinity and eluted ligand data. Histograms represent the top 0.1 and 1% scoring predictions for _EL and _BA respectively, of 1 million random peptides of lengths 11 to 19 amino acids extracted from natural occurring proteins. Green and red lines represent predictions obtained using the “binding affinity” output neuron and the “eluted ligand” output neuron, respectively, in the five combined models. Black lines represent the observed length distribution of the ligand training set.

Lastly, we performed an evaluation across data sets to confirm the robustness of the results obtained, and to limit any unforeseen signal of performance overfitting. For each data set, we used the two-output model trained above to predict the other ligand data sets of the same allotype. Prior to evaluation, all data with a 9mer overlap between training and evaluation sets were removed. We observed that, in all cases, models trained on a specific data set retain a high predictive performance for the prediction of ligands of the same allotype derived from a different experiment (Table 3). These results confirm the high reproducibility of the motifs across different cell lines, as well as the robustness of the prediction models derived from individual data sets.

Table 3. Independent evaluation on other datasets from the same allele.

		Models					
Training set	Evaluation set	Single BA	Combined _BA	Single EL	Combined _EL	NetMHCIIpan*	NetMHCII*
DR1 Ph	DR1 Pm	0.570	0.644	0.839	0.849	0.562	0.585
	DR1 Sm	0.573	0.648	0.810	0.813	0.576	0.578
DR1 Pm	DR1 Ph	0.498	0.557	0.757	0.754	0.478	0.469
	DR1 Sm	0.514	0.549	0.665	0.669	0.519	0.511
DR1 Sm	DR1 Ph	0.506	0.556	0.704	0.707	0.481	0.478
	DR1 Pm	0.568	0.636	0.824	0.835	0.540	0.570
DR15 Ph	DR15 Pm	0.584	0.672	0.869	0.869	0.427	0.580
DR15 Pm	DR15 Ph	0.615	0.710	0.888	0.889	0.459	0.583

Single BA: model trained on binding affinity only; Combined BA: model trained on both binding affinity and elution data, using the binding affinity output neuron; Single El: model trained on eluted ligands only; Combined EL: model trained on both binding affinity and elution data, using the eluted ligand output neuron. *NetMHCIIpan-3.2 and NetMHCII-2.3 were trained on IEDB binding affinity data. Performance values are expressed in terms of AUC 0.1.

4. Signals of Ligand Processing

Having developed improved models for prediction of MHC class II ligand binding, we next analysed whether the models could be used to identify signals of antigen processing in the MS eluted ligand data sets. We hypothesized that information concerning antigen processing should be present in the regions around the N and C terminus of the ligand. These regions comprise residues that flank the MHC binding core called Peptide Flanking Regions (PFRs), and residues from the ligand source protein sequence located outside the ligand (see figure 5 for a schematic overview).

We speculate that the signals of antigen processing will depend, to some degree, on the length of the PFRs on each side of the binding core. MHC-II ligands are cut and trimmed by exopeptidases, which operate according to specific motifs in prioritizing cleavage sites. However, in the case of short PFRs, the MHC hinders access to the protease and prevents trimming of the residues in close proximity to the MHC (Larsen, Pedersen, Buus, & Stryhn, 1996; Mouritsen, Meldal, Werdelin, Hansen, & Buus, 1992) For this reason, we expect to observe cleavage motifs only in peptides with sufficiently large PFRs, where the end-of-the-trimming signal is given by the peptide sequence rather than by MHC hindrance. To validate this hypothesis, we identified the PFRs of the ligands in the DR15 Pm EL dataset, as well as three “context” residues found immediately upstream and downstream of the ligand in its source protein. To avoid over-estimation of the performance, the binding core was identified from the cross-validated eluted ligand predictions of the two output model. The ligands were split into

groups depending on the length of the C and N terminal PFRs, and sequence logos were generated for each ligand subset using Seq2Logo (Fig. 5).

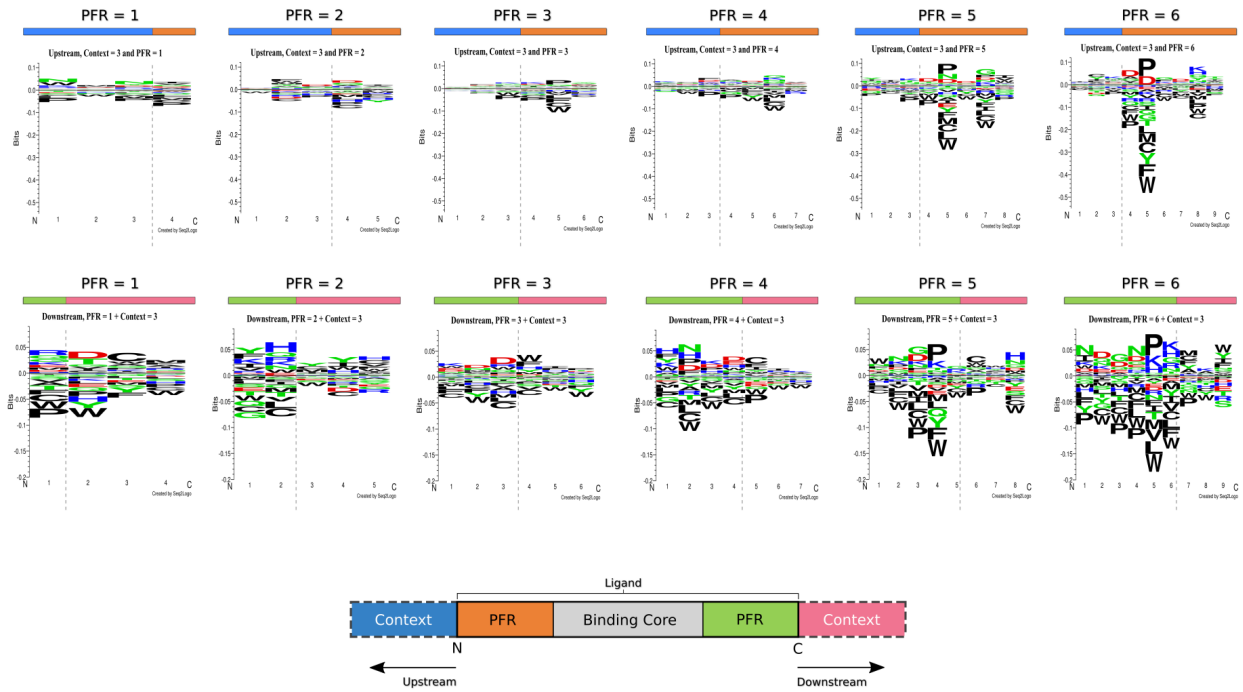


Figure 5. Logo plots for the upstream (top row) and downstream (bottom row) processing regions present in the HLA-DRB1*15:01 Pm dataset. Context length is always fixed at three residues, while Peptide Flanking Regions (PFRs) length vary from one to six amino acids (left to right). All logos of the same terminal region, upstream and downstream were estimated from the same number (down-sampled is necessary) of peptides, 111 and 44 sequences respectively. Amino acid background frequencies were calculated using the antigenic source protein of all the ligands present in the dataset. Motifs were generated as using Seq2logo as described in Materials and Methods.

The results displayed in figure 5 clearly confirm the important role of the MHC in shaping the processing signal. For both the N and C terminal data sets, we observe a clear enrichment of proline (P) at the second position from the ligand terminals only for data sets where the PFR is longer than 2 amino acids. This observation is confirmed from the reanalysis of a structural data set of peptide:HLA-DR complexes from the Protein Data Bank (PDB) previously assembled for benchmarking the accuracy for MHC-II binding core identification (Andreatta et al., 2015). On this PDB dataset, 29% of the entries with a N-terminal PFR longer than 2 amino acids contain a proline at the second position from the N terminal, and 38% of the entries with a C terminal PFR longer than 2 amino acids contain a proline at the second position from the C terminal (data not shown). Given this, we next combined all ligands with PFR length larger than 2, together with

the three source protein context residues at either C or N terminal side of the ligand. Then, we used these to construct a motif of the processing signal (see Figure 6).

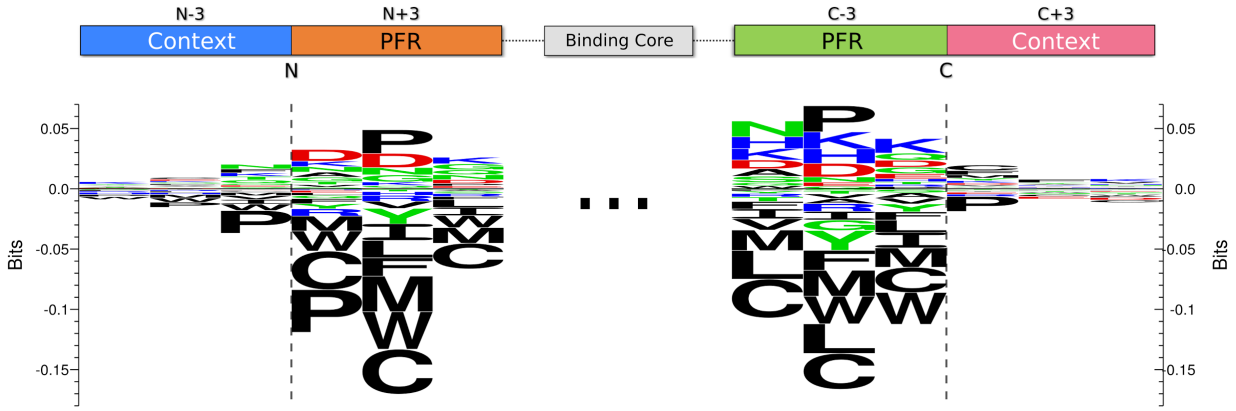


Figure 6. Logo plots for the DR15-Pm dataset, upstream and downstream processing signals. For each logo, the closest three PFR and context residues neighboring the N and C terminus are shown.

The processing motif confirms the strong preference for prolines at the second but last position in the ligand at both N and C terminal sides, as well as a clear signal of depletion of other hydrophobic amino acid types towards the terminals of the ligand. Note that the cysteine depletion in the PFR is likely to be a technological artifact, as cysteines have previously been shown to be underrepresented in MS-derived peptide data sets (Abelin et al., 2017; Bassani-Sternberg et al., 2017). From this figure it is also clear that processing signals contained outside the ligand are very weak.

Next, we investigated to what degree the signal observed in the HLA-DRB1*15:01 Pm data set could be consistently identified in the other four data sets. To do this, we repeated the above analysis for each data set and computed the corresponding processing Position Specific Scoring Matrix (PSSM). A similarity matrix was constructed by calculating the similarity between all pairs of PSSMs in terms of the Pearson's correlation coefficient between the two vectors of the 6*20 elements (6 positions and 20 amino acid propensity scores at each position) (see Figure 7, the PSSMs from each data are included in supplementary figure S1).

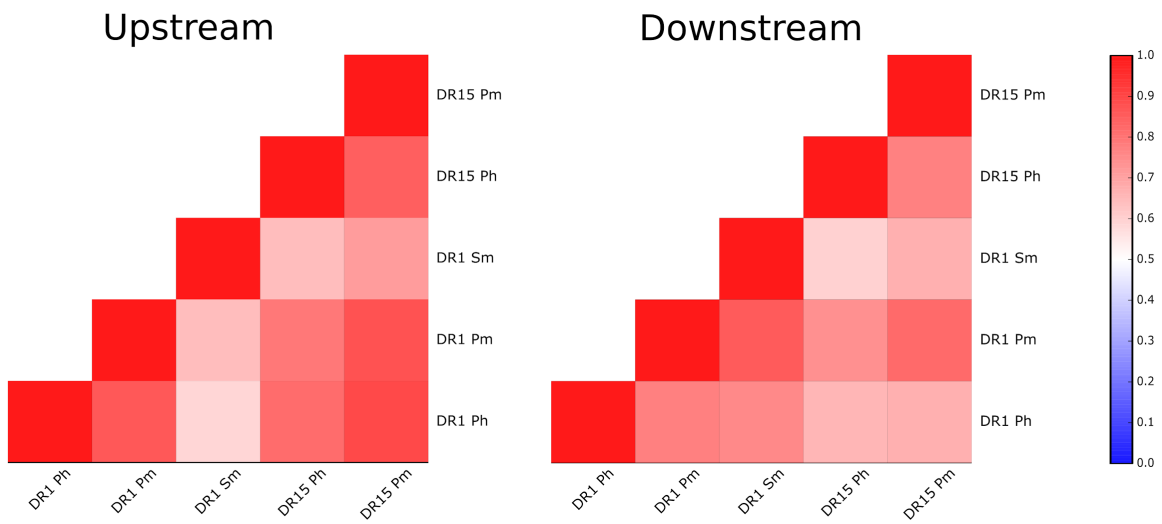


Figure 7. Pearson Correlation Coefficient (PCC) matrices of the processing signals in the five datasets, for both upstream and downstream processing regions.

Figure 7 exhibits a clear positive correlation between the processing motif in all the datasets involved. The mean PCC score for the matrices in Fig. 7 was 0.77 for upstream and 0.73 for downstream, with the lowest PCC=0.59 (for the DR1Sm-DR1Ph pair, upstream) and the maximum PCC=0.89 (for DR15Pm-DR1Ph, upstream). These results suggest that there are components of the processing signals captured that are MHC- and even species-independent: the correlation between the two human and mouse data sets is as high as the correlation between two data sets within the same species. These results strongly suggest that the observed signal is related to antigen processing.

5. Incorporating ligand processing into a combined predictor

Having identified a consistent signal associated with antigen processing, we next investigated whether this signal could be integrated in a prediction model to boost predictive performance. To conduct this analysis, we used the data sets and data partitions described earlier. The processing signal was integrated into the machine-learning framework by complementing the encoding of each ligand with the 3 N terminal context, 3 N terminal peptide, 3 C terminal context, and 3 C terminal peptide residues (see figure 5). For peptide binding affinity data, the context information was presented with three wildcard amino acids “XXX”. The result in terms of AUC0.1 of the 5 fold cross-validation evaluation is shown in table 4, and the results of the cross-dataset evaluation in table 5. In both cases, we observe a substantial and consistent improved performance of the eluted ligand prediction of the model when including the context information. As expected, this gain is not observed for the binding affinity predictions (data not shown).

Because our analysis supports that all the eluted peptides of the same allele, irrespective of their source, have similar peptide signals, we merged all eluted ligands from the same allotype and trained a merged model. Evaluating the cross-validation performance of the global model on the data from the individual sets showed consistently higher performance for all datasets (column "With context merged datasets" in Table 4). Further, as control for performance overestimation, we trained a control model with shuffled context residues. As expected, we for this model due to the loss of the antigen processing signal observed a performance comparable to the merged model trained without context (Table 4, "With shuffled context" column).

Table 4. Cross-validation performance of prediction models trained with vs without context information.

Allele	Evaluation dataset	Without context	With context	With context merged datasets	With shuffled context
DRB1*01:01	DR1 Ph	0.823	0.840	**0.861**	0.816
	DR1 Pm	0.893	0.913	**0.923**	0.897
	DR1 Sm	0.867	0.881	**0.912**	0.866
DRB1*15:01	DR15 Ph	0.908	0.917	**0.940**	0.910
	DR15 Pm	0.917	0.928	**0.932**	0.918

Models trained on both binding affinity and ligand elution data, with and without context information, and using the eluted ligand output neuron. "With context" columns indicate models trained with 3 amino acids from the source protein and 3 amino acids from PFR sequences for the N- and C-terminus of the ligand. Merged datasets refer to a model using all datasets combined for the same allele. Shuffled context refers to a model trained with context but where the order of the amino acids was randomized. Performance values are expressed in terms of AUC 0.1.

Table 5. Cross-dataset performance of prediction models trained with vs without context information

Training set	Evaluation set	Without context	With context
DR1 Ph	DR1 Pm	0.849	**0.866**
	DR1 Sm	0.813	**0.843**
DR1 Pm	DR1 Ph	0.754	**0.789**
	DR1 Sm	0.669	**0.711**
DR1 Sm	DR1 Ph	0.707	**0.752**
	DR1 Pm	0.835	**0.860**
DR15 Ph	DR15 Pm	0.869	**0.884**
DR15 Pm	DR15 Ph	0.889	**0.910**

Performance values are expressed in terms of AUC 0.1.

As a final illustration of the signal captured by the model trained including the context information, we constructed sequence motifs of the top 1% highest scoring peptides from a list of 1 million random natural peptides of length 10-25 and their context (supplementary figure 2). As expected, the motif contained within the N and C terminal peptide flanks and context resembles strongly the motif described in figure 6.

Discussion

Peptide binding to MHC II is arguably the most selective step in antigen presentation to CD4+ T cells. The ability to measure (and predict) specific CD4+ responses is crucial for the understanding of pathological events, such as infection by pathogens or cancerous transformations. Recent studies have also highlighted a potential role for CD4+ T cells for the development of cancer immunotherapies (Kreiter et al., 2015; Tran et al., 2014; Zanetti, 2015). Characterizing peptide:MHC-II binding events has been a focal point of research over the last decades. Large efforts have been dedicated in conducting high-throughput, in-vitro measurements of peptide MHC II interactions (Justesen, Harndahl, Lamberth, Nielsen, & Buus, 2009; J. Sidney et al., 2001; John Sidney et al., 2013), and these data have been used to develop methods capable of accurately predicting the interaction to MHC II from the peptide sequence alone (Andreatta et al., 2015; Nielsen & Andreatta, 2017; Nielsen, Lundegaard, & Lund, 2007; Singh & Raghava, 2001). While these approaches have proven highly successful as guides in the search for CD4 epitopes (Gfeller, Bassani-Sternberg, Schmidt, & Luescher, 2016; Nielsen, Lund, Buus, & Lundegaard, 2010), a general conclusion from these studies is that MHC II *in-vitro* binding affinity (whether measured or predicted) is a relatively poor correlate of immunogenicity (Backert & Kohlbacher, 2015). In other words, peptide binding to MHC II is a necessary but not sufficient criterion for peptide immunogenicity. The same situation holds for MHC class I presented epitopes. Here, however, peptide binding to MHC I is a very strong correlate to peptide immunogenicity, and can be used to discard the vast majority (99%) of the irrelevant peptide space while maintaining an extremely high (>95%) sensitivity for epitope identification (Jurtz et al., 2017). For MHC II, recent studies suggest the corresponding numbers to fall in the range 80% (specificity), 50% sensitivity (Jensen et al., 2018). For these reasons, we suggest that features other than MHC II in-vitro binding affinity may be critical for MHC II antigen presentation. Based on five MS MHC II eluted ligand data sets, we have here attempted to address and quantify this statement.

Firstly, we have demonstrated that MS MHC II eluted ligand data sets generated by state-of-the-art technologies and laboratories are of very high quality, comprising low noise levels and allowing very precise determination of MHC II binding motifs. Overall, the obtained binding motifs show overlap with the motifs identified from in-vitro binding affinity data, with subtle differences at well-defined anchor positions.

Secondly, we demonstrated that high accuracy prediction models for peptide MHC II interaction can be constructed from the MS derived MHC II eluted ligand data; that the accuracy of these models can be improved by training models integrating information from both binding affinity and eluted ligand data sets; and that these improved models can be used to identify eluted ligands in independent data sets at an unprecedented level of accuracy.

Thirdly, we analyzed signals potentially associated with antigen processing. We postulate that if a processing signal exists, it should be influenced by the relative location of the peptide binding core compared to the N and C terminal of the given ligand. This is because the MHC II molecule may hinder the access of the protease, thus preventing trimming of the residues in close proximity to the MHC (Stryhn et al, PMID 8691132). Investigating the data confirmed this hypothesis, and a strong processing signal (with a preference for Prolines at the second amino acid position from the N and C terminal of the ligand), was observed only for ligands where the length of the region flanking the binding core was three amino acids or more. This observation was found consistently in all 5 data sets independent of MHC II restriction and host-species (human or mouse).

Lastly, we integrated this information associated with antigen processing into a machine-learning framework, and demonstrated a consistently improved predictive performance not only in terms of cross-validation but also when applied to independent evaluation data sets.
In conclusion, we have demonstrated how integrating MHC class II in-vitro binding and MS eluted ligand data can boost the predictive performance for both binding affinity and eluted ligand likelihood predictions. To the best of our knowledge, we have also demonstrated for the first time how MHC II eluted ligand data can be used to extract signals of antigen processing, and how these signals can be integrated into a model with improved predictive performance.

The work here has been limited to two HLA-DR molecules, but the framework is readily extendable to additional molecules and eventually, once sufficient data becomes available, a

pan-specific predictor as has been shown earlier for MHC class I (Jurtz et al., 2017) may be achievable.

Materials and Methods

Datasets

HLA class-II peptidome data were obtained from two recent MS studies. Three datasets corresponding to the HLA-DRB1:0101: DR1-Ph, DR1-Pm (Ooi et al., 2017), and DR1-Sm (Clement et al., 2016), and two to DRB1*15:01: DR15-Ph and DR15-Pm (Ooi et al., 2017).
Here, the data sets with subscript h correspond to data obtained from human cell lines and data sets with the subscript m to data obtained from human MHC-II molecules transfected into MHC-II deficient mice cell lines. Note, that DR15-Ph data set was obtained from a heterozygous EBV transformed B lymphoblastoid cell line, IHW09013 (also known as SCHU), which express two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (shortened here with the name DR15/51).

MHC class II peptide binding affinity data was obtained from previous publications (Jensen et al., 2018) for the alleles DR1 (DRB1*01:01, 10,406 peptides) and DR15 (DRB1*15:01, 4,850 peptides).

The ligand datasets were analysed using the GibbsCluster-2.0 method as described earlier (Alvarez et al., 2018) using default settings to remove potential noise and biases imposed by some data sharing multiple binding specificities. The details of the two data sets are described in table 6.

Table 6. Summary of peptide binding and eluted ligand datasets references and its corresponding data points used for training the different models.

BINDERS

Reference	Source	Allele	L11–19
DR1–BA	Jensen et al., 2018	DRB1*01:01	9987
DR15–BA	Jensen et al., 2018	DRB1*15:01	4466

LIGANDS

Reference	Source	Cell	Allele	Unique	GC	L11–19	Random
DR1–Ph	Ooi et al., 2017	human	DRB1*01:01	5131	4616	3992	38115
DR1–Pm	Ooi et al., 2017	mouse	DRB1*01:01	5744	5549	5385	55710
DR1–Sm	Clement et al., 2016	mouse	DRB1*01:01	3216	3128	2963	30510
DR15–Ph	Ooi et al., 2017	human	DRB1*15:01	2782	1587	1390	12870
DR15–Pm	Ooi et al., 2017	mouse	DRB1*15:01	4810	4487	4229	42030

Number of points after filtering with Gibbs clustering (GC), amino acid length (L11-19) and number of Random peptides (Random) added to each dataset used for training.

NNAlign Modeling

Models predicting peptide-MHC interactions were trained as described earlier using NNAlign (Alvarez et al., 2018; Nielsen & Andreatta, 2017). Only ligands of length 11-19 amino acids were included in the training data. Random peptides of variable lengths derived from the non-redundant UniProt database were used as negatives. The same amount of random negatives was used for each length (11 to 19), and consisted of five times the amount of peptides for the most represented length in the positive ligand dataset. Positive instances were labeled with a target value of 1, and negatives with a target value of 0. Prior to training, the datasets were clustered using the common motif Hobohm algorithm 1 described earlier (Hobohm, Scharf, Schneider, & Sander, 1992) with a motif length of 9 amino acids to generate five partitions for cross validation.

Two types of model were trained; one with single data type (eluted ligand or binding affinity) input, and one with a mixed input of two data types. Single models per each dataset and allele were trained as described earlier with either binding affinity or eluted ligand data as input (Alvarez et al., 2018). All models were built as an ensemble of 250 individual networks generated with 10 different seeds, 2, 10, 20, 40, and 60 hidden neurons and 5 partitions. Models were trained for 400 iterations, without use of early stopping. Additional settings in the architecture of the network were used as previously described for MHC class II defaults in (Alvarez et al., 2018).

Some modifications were introduced to NNAlign to better account for specific challenges associated with MHC class II ligand data. For the network to be able to learn peptide length preferences, a “binned” encoding of the peptide length was introduced, consisting of a one-hot input vector of size nine (one neuron for each of the lengths 11 to 19). In order to guide binding core identification, a burn-in period was introduced with a limited search space for the P1 binding core position. During the burn-in period, consisting of a single learning iteration, only hydrophobic residues were allowed at the P1 anchor position. Starting from the second iteration, all amino acids were allowed at the P1 position (Supplementary figure 3).

Sequence logos

Sequence logos for binding motifs were constructed using Seg2Logo tool with default parameters (Thomsen & Nielsen, 2012). Logos for context information were generated as weighted Kulback-Leibler logos using Seq2Logo excluding sequence weighting. Amino acids were grouped by negatively charged (red), positively charged (blue), polar (green) or hydrophobic (black).

Acknowledgments

This work was supported in part by Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200010C and by the Science and technology council of investigation (CONICET-Argentina).

Competing interests

The authors have declared that no competing interests exist.

References

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., … Wu, C. J. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, *46*(2), 315–326. https://doi.org/10.1016/j.immuni.2017.02.007

Alvarez, B., Barra, C., Nielsen, M., & Andreatta, M. (2018). Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. *Proteomics*, *In Press*. https://doi.org/DOI:10.1002/pmic.201700252

Andreatta, M., Alvarez, B., & Nielsen, M. (2017). GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx248

Andreatta, M., Jurtz, V. I., Kaever, T., Sette, A., Peters, B., & Nielsen, M. (2017). Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology*, *152*(2), 255–264. https://doi.org/10.1111/imm.12763

Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S., & Nielsen, M. (2015). Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics*, *67*(11–12), 641–650. https://doi.org/10.1007/s00251-015-0873-y

Andreatta, M., Lund, O., & Nielsen, M. (2013). Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics (Oxford, England)*, *29*(1), 8–14. https://doi.org/10.1093/bioinformatics/bts621

Andreatta, M., Schafer-Nielsen, C., Lund, O., Buus, S., & Nielsen, M. (2011). NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PloS One*, *6*(11), e26781. https://doi.org/10.1371/journal.pone.0026781

Andreatta, M., Trolle, T., Yan, Z., Greenbaum, J. A., Peters, B., & Nielsen, M. (2017). An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics (Oxford, England)*. https://doi.org/10.1093/bioinformatics/btx820

Arnold, P. Y., La Gruta, N. L., Miller, T., Vignali, K. M., Adams, P. S., Woodland, D. L., & Vignali, D. A. A. (2002). The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. *Journal of Immunology (Baltimore, Md.: 1950)*, *169*(2), 739–749.

Backert, L., & Kohlbacher, O. (2015). Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Medicine*, *7*, 119. https://doi.org/10.1186/s13073-015-0245-0

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., … Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Computational Biology*, *13*(8), e1005725. https://doi.org/10.1371/journal.pcbi.1005725

Bassani-Sternberg, M., & Gfeller, D. (2016). Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. *Journal of Immunology (Baltimore, Md.: 1950)*, *197*(6), 2492–2499. https://doi.org/10.4049/jimmunol.1600808

Bergseng, E., Dørum, S., Arntzen, M. Ø., Nielsen, M., Nygård, S., Buus, S., … Sollid, L. M. (2015). Different binding motifs of the celiac disease-associated HLA molecules DQ2.5, DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of endogenous peptide repertoires. *Immunogenetics*, *67*(2), 73–84. https://doi.org/10.1007/s00251-014-0819-9

Bird, P. I., Trapani, J. A., & Villadangos, J. A. (2009). Endolysosomal proteases and their inhibitors in immunity. *Nature Reviews. Immunology*, *9*(12), 871–882. https://doi.org/10.1038/nri2671

Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). Pathways of antigen processing. *Annual Review of Immunology*, *31*, 443–473. https://doi.org/10.1146/annurev-immunol-032712-095910

Buus, S., Sette, A., Colon, S. M., Miles, C., & Grey, H. M. (1987). The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. *Science (New York, N.Y.)*, *235*(4794), 1353–1358.

Caron, E., Kowalewski, D. J., Chiek Koh, C., Sturm, T., Schuster, H., & Aebersold, R. (2015). Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Molecular & Cellular Proteomics: MCP*, *14*(12), 3105–3117. https://doi.org/10.1074/mcp.O115.052431

Carson, R. T., Vignali, K. M., Woodland, D. L., & Vignali, D. A. (1997). T cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. *Immunity*, *7*(3), 387–399.

Chong, C., Marino, F., Pak, H.-S., Racle, J., Daniel, R. T., Müller, M., … Bassani-Sternberg, M. (2017). High-throughput and sensitive immunopeptidomics platform reveals profound IFNγ-mediated remodeling of the HLA ligandome. *Molecular & Cellular Proteomics: MCP*. https://doi.org/10.1074/mcp.TIR117.000383

Clement, C. C., Becerra, A., Yin, L., Zolla, V., Huang, L., Merlin, S., … Santambrogio, L. (2016). The Dendritic Cell Major Histocompatibility Complex II (MHC II) Peptidome Derives from a Variety of Processing Pathways and Includes Peptides with a Broad Spectrum of HLA-DM Sensitivity. *The Journal of Biological Chemistry*, *291*(11), 5576–5595. https://doi.org/10.1074/jbc.M115.655738

Gfeller, D., Bassani-Sternberg, M., Schmidt, J., & Luescher, I. F. (2016). Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology*, *5*(7), e1177691. https://doi.org/10.1080/2162402X.2016.1177691

Godkin, A. J., Smith, K. J., Willis, A., Tejada-Simon, M. V., Zhang, J., Elliott, T., & Hill, A. V. (2001). Naturally processed HLA class II peptides reveal highly conserved immunogenic

flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. *Journal of Immunology (Baltimore, Md.: 1950)*, *166*(11), 6720–6727.

Gowthaman, U., & Agrewala, J. N. (2008). In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. *Journal of Proteome Research*, *7*(1), 154–163. https://doi.org/10.1021/pr070527b

Hall, F. C., Rabinowitz, J. D., Busch, R., Visconti, K. C., Belmares, M., Patil, N. S., … Sonderstrup, G. (2002). Relationship between kinetic stability and immunogenicity of HLA-DR4/peptide complexes. *European Journal of Immunology*, *32*(3), 662–670. https://doi.org/10.1002/1521-4141(200203)32:3<662::AID-IMMU662>3.0.CO;2-5

Hammer, J., Valsasnini, P., Tolba, K., Bolin, D., Higelin, J., Takacs, B., & Sinigaglia, F. (1993). Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell*, *74*(1), 197–203.

Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Science: A Publication of the Protein Society*, *1*(3), 409–417. https://doi.org/10.1002/pro.5560010313

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., … Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. https://doi.org/10.1111/imm.12889

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., & Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology (Baltimore, Md.: 1950)*, *199*(9), 3360–3368. https://doi.org/10.4049/jimmunol.1700893

Justesen, S., Harndahl, M., Lamberth, K., Nielsen, L.-L. B., & Buus, S. (2009). Functional recombinant MHC class II molecules and high-throughput peptide-binding assays. *Immunome Research*, *5*, 2. https://doi.org/10.1186/1745-7580-5-2

Kim, A., Hartman, I. Z., Poore, B., Boronina, T., Cole, R. N., Song, N., … Sadegh-Nasseri, S. (2014). Divergent paths for the selection of immunodominant epitopes from distinct antigenic sources. *Nature Communications*, *5*, 5369. https://doi.org/10.1038/ncomms6369

Kreiter, S., Vormehr, M., van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., … Sahin, U. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*, *520*(7549), 692–696. https://doi.org/10.1038/nature14426

Kropshofer, H., Max, H., Halder, T., Kalbus, M., Muller, C. A., & Kalbacher, H. (1993). Self-peptides from four HLA-DR alleles share hydrophobic anchor residues near the NH2-terminal including proline as a stop signal for trimming. *Journal of Immunology (Baltimore, Md.: 1950)*, *151*(9), 4732–4742.

Larsen, S. L., Pedersen, L. O., Buus, S., & Stryhn, A. (1996). T cell responses affected by aminopeptidase N (CD13)-mediated trimming of major histocompatibility complex class II-bound peptides. *The Journal of Experimental Medicine*, *184*(1), 183–189.

Lippolis, J. D., White, F. M., Marto, J. A., Luckey, C. J., Bullock, T. N. J., Shabanowitz, J., … Engelhard, V. H. (2002). Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. *Journal of Immunology (Baltimore, Md.: 1950)*, *169*(9), 5089–5097.

Lovitch, S. B., Pu, Z., & Unanue, E. R. (2006). Amino-terminal flanking residues determine the conformation of a peptide-class II MHC complex. *Journal of Immunology (Baltimore, Md.: 1950)*, *176*(5), 2958–2968.

Mouritsen, S., Meldal, M., Werdelin, O., Hansen, A. S., & Buus, S. (1992). MHC molecules protect T cell epitopes against proteolytic destruction. *Journal of Immunology (Baltimore, Md.: 1950)*, *149*(6), 1987–1993.

Nielsen, M., & Andreatta, M. (2017). NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx276

Nielsen, M., Lund, O., Buus, S., & Lundegaard, C. (2010). MHC class II epitope predictive algorithms. *Immunology*, *130*(3), 319–328. https://doi.org/10.1111/j.1365-2567.2010.03268.x

Nielsen, M., Lundegaard, C., & Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, *8*, 238. https://doi.org/10.1186/1471-2105-8-238

Ooi, J. D., Petersen, J., Tan, Y. H., Huynh, M., Willett, Z. J., Ramarathinam, S. H., … Kitching, A. R. (2017). Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature*, *545*(7653), 243–247. https://doi.org/10.1038/nature22329

Roche, P. A., & Cresswell, P. (1990). High-affinity binding of an influenza hemagglutinin-derived peptide to purified HLA-DR. *Journal of Immunology (Baltimore, Md.: 1950)*, *144*(5), 1849–1856.

Rudolph, M. G., Stanfield, R. L., & Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology*, *24*, 419–466. https://doi.org/10.1146/annurev.immunol.23.021704.115658

Sette, A., Adorini, L., Colon, S. M., Buus, S., & Grey, H. M. (1989). Capacity of intact proteins to bind to MHC class II molecules. *Journal of Immunology (Baltimore, Md.: 1950)*, *143*(4), 1265–1267.

Sidney, J., Southwood, S., Moore, C., Oseroff, C., Pinilla, C., Grey, H. M., & Sette, A. (2013). Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Current Protocols in Immunology*, *Chapter 18*, Unit 18.3. https://doi.org/10.1002/0471142735.im1803s100

Sidney, J., Southwood, S., Oseroff, C., del Guercio, M. F., Sette, A., & Grey, H. M. (2001). Measurement of MHC/peptide interactions by gel filtration. *Current Protocols in Immunology*, *Chapter 18*, Unit 18.3. https://doi.org/10.1002/0471142735.im1803s31

Singh, H., & Raghava, G. P. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics (Oxford, England)*, *17*(12), 1236–1237.

Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., … Hammer, J. (1999). Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature Biotechnology*, *17*(6), 555–561. https://doi.org/10.1038/9858

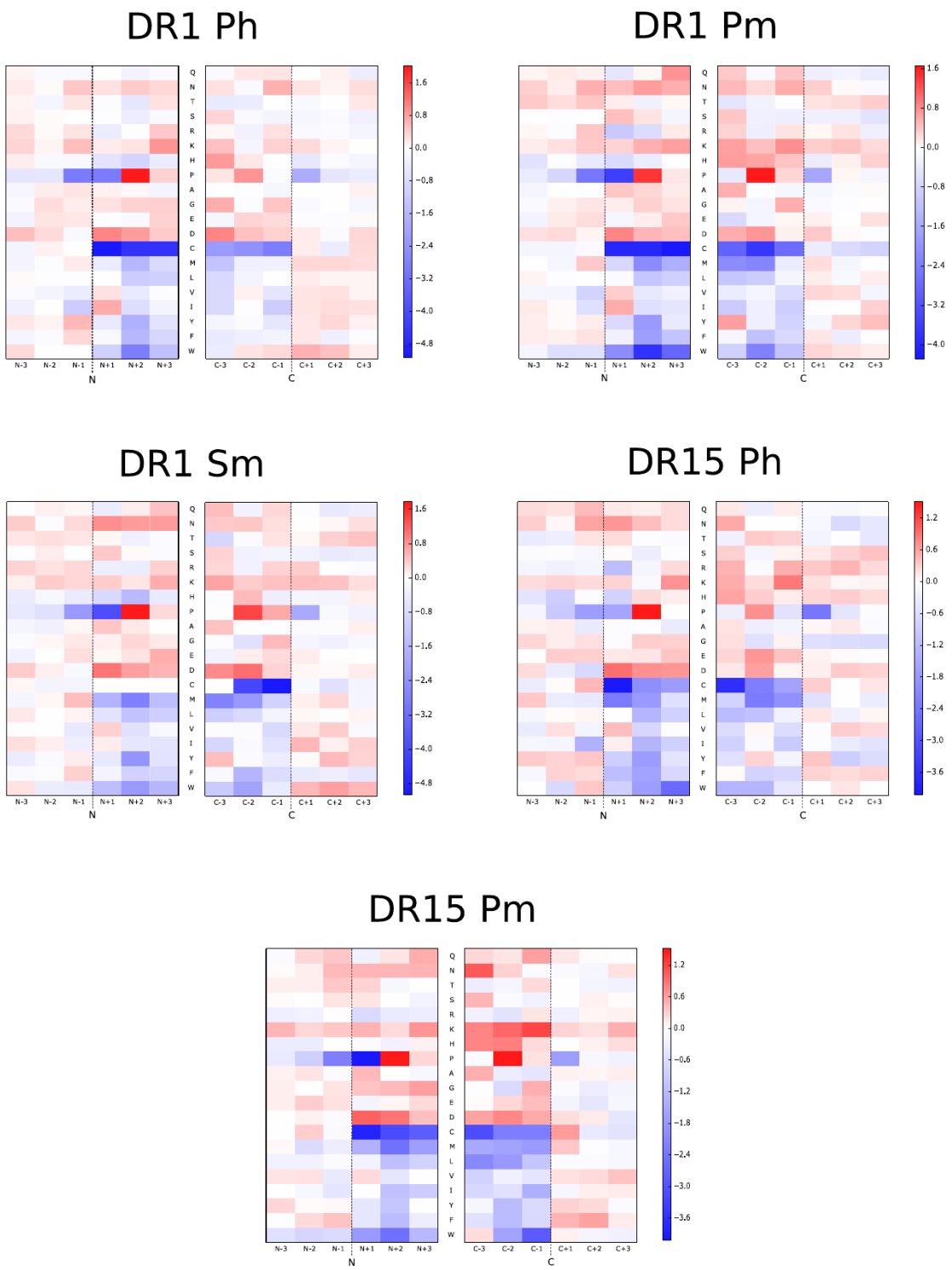
Thomsen, M. C. F., & Nielsen, M. (2012). Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Research*, *40*(Web Server issue), W281-287. https://doi.org/10.1093/nar/gks469

Tran, E., Turcotte, S., Gros, A., Robbins, P. F., Lu, Y.-C., Dudley, M. E., … Rosenberg, S. A. (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science (New York, N.Y.)*, *344*(6184), 641–645. https://doi.org/10.1126/science.1251102

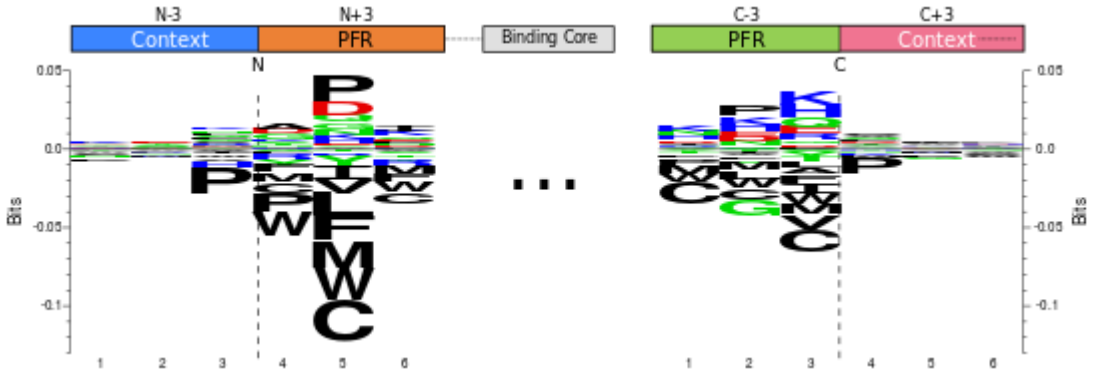
Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., … Peters, B. (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Research*, *43*(Database issue), D405-412. https://doi.org/10.1093/nar/gku938

Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., & Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Computational Biology*, *4*(4), e1000048. https://doi.org/10.1371/journal.pcbi.1000048

Zanetti, M. (2015). Tapping CD4 T cells for cancer immunotherapy: the choice of personalized genomics. *Journal of Immunology (Baltimore, Md.: 1950)*, *194*(5), 2049–2056. https://doi.org/10.4049/jimmunol.1402669

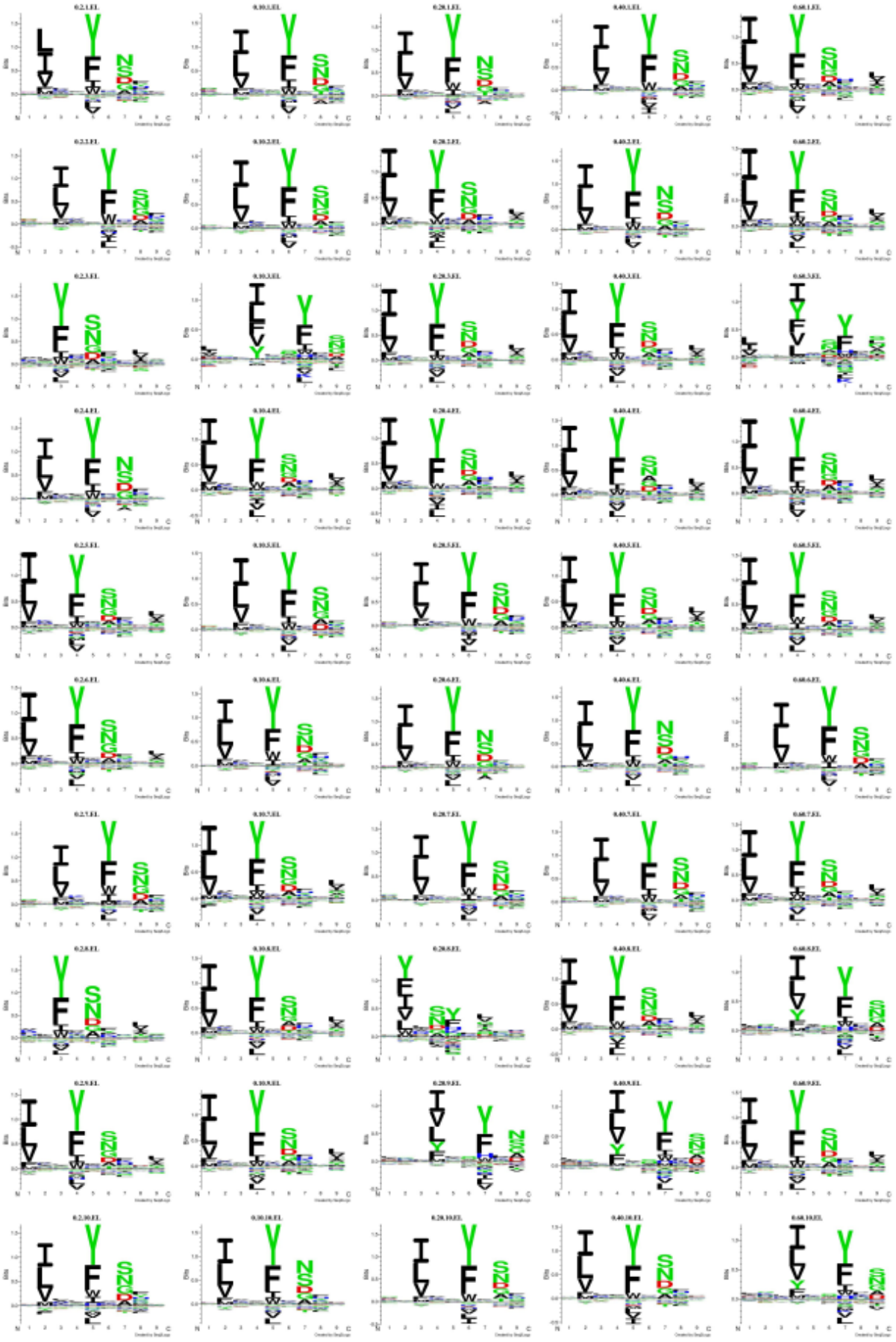


Supplementary Figure 1. Heatmap representation of the processing signal PSSMs for all five datasets. Color code correspond to the information content present in each position of the PSSM, for upstream and downstream. Colorbar values indicate the power of two to which information is increased or decreased (i.e. a value of -3 indicates a 2^3 fold position-specific information depletion, meaning eight times more probability to find such particular depletion than random).



Supplementary figure 2. Logo plots of the context residues learnt by the NNAlign with context model trained on DR15-Pm dataset. For each logo, three amino acids of the PFR and context residues neighbouring the N and C terminus are shown. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 1 million random natural peptides.

A



B

Supplementary figure 3. Logos from the core learnt by each network from NNAlign models trained with DR15-Ph eluted ligand dataset without A) and with P1 burn in implementation B). Each different logo represent the core learnt by a single network generated with a seed (1 to 10) and a set of hidden neurons (2, 10, 20, 40, 60) for partition 1 out of 5. Logos were constructed from the predicted cores in the top 1% scoring predictions of 1 million random natural peptides.