Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data

Birkir Reynisson[1], Carolina Barra[1], Bjoern Peters[2,3], Morten Nielsen[1,4],*

1) Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark
2) Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 92037
3) Department of Medicine, University of California, San Diego, CA 92093
4) Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Argentina

*) Corresponding author - morni@dtu.dk

Abstract

Major Histocompatibility Complex II (MHC II) molecules play a key role in the onset and control of cellular immunity. In a highly selective process, MHC II presents peptides derived from exogenous antigens on the surface of antigen presenting cells for T cell scrutiny. Understanding the rules defining this presentation holds key insights into the regulation and potential manipulation of the cellular immune system. In our work for furthering rational prediction of peptide epitopes, we have integrated data from eluted MHC ligand mass spectrometry assays, that inherently contains signals of the events leading to antigen presentation. The full integration of this data type is a fruit of recent developments of the NNAlign_MA machine learning framework: two output-neuron architecture, encoding of ligand context and logical handling of ligands with multiple potential allele annotation. In large-scale benchmarking, we demonstrate here how these features synergistically improve ligand and CD4 T cell epitope prediction performance beyond that of current state-of-the-art predictors. We observe motifs in ligand termini consistently across MHC alleles and we put forth a strategy of targeting these motifs for MHC II agnostic de-immunization.

Introduction

Major Histocompatibility Complex class II (MHCII) molecules play a pivotal role in the adaptive immune system. Antigen Presenting Cells (APCs) display MHC II molecules in complex with peptides[1] on their surface. These peptides are products of extracellular proteins internalized by APCs and proteolytically digested in endocytic compartments. During protein degradation, peptide fragments of the antigen are loaded into the binding cleft of MHC II and peptides that bind stably (forming a pMHCII complex) are shuttled to the cell surface for presentation to T-helper (Th) cells of the immune system[2]. Th cells scrutinize the surface of APCs and if the T cell receptor (TCR) recognizes a pMHCII complex, Th cells can become activated. Peptides that cause such T cell activation are termed T cell epitopes. Regulation of Th cell activation is critical since they coordinate the activation of effector cells. Peptide-MHC presentation is a necessary and highly selective step in the process of T cell activation and understanding the rules defining this presentation holds key insights into pathogen recognition of the cellular immune system.

Given this, large efforts have been dedicated to characterizing the rules of MHC II peptide presentation. MHC II is a heterodimer, the alpha and beta chains of which together form the peptide binding cleft. In humans, the Human Leukocyte Antigen (HLA) of MHC class II is encoded by three different loci (HLA-DR, -DQ and DP)[3]. The corresponding HLA genes have numerous allelic variants with polymorphisms that are mostly clustered around the residue locations forming the peptide binding cleft, resulting in a wide range of distinct peptide binding specificities. MHC II has an open binding cleft, allowing it to interact with peptides of a broad length range, most commonly of 13-25 amino acids[4]. The binding cleft of MHC II interacts predominantly with a 9-mer register of the interacting peptide, termed the binding core[5]. The placement of this binding core varies between peptides, resulting in peptide flanking residues (PFRs) of differing length protruding

out from the binding cleft of the MHC II molecule. PFRs have been shown to influence both the interaction with MHC II and the activation of T cells[6]. These facts together, makes the study and prediction of MHC II peptide interaction highly challenging.

Traditionally, in vitro peptide-MHC Binding Affinity (BA) assays have been used to generate data to characterize the specificity of MHC-II molecules[7], and a range of machine-learning prediction models have been developed from this data to identify the rules of peptide MHC binding: NetMHCII[8], NetMHCIIpan[9,10], TEPITOPE[11], TEPITOPEpan[12], ProPred[13]. Such models have in turn great potential to guide epitope discovery experiments[14], vaccine design[15,16] and deimmunization of biotherapeutics[17].

However, peptide-MHCII binding affinity has in several studies been demonstrated to be a relatively poor correlate of MHC antigen presentation[18] and peptide immunogenicity[19]. Likewise, several studies have demonstrated that MHC-II peptide binding prediction models can benefit from being trained on so-called immunopeptidome data obtained by liquid chromatography coupled mass spectrometry(LC-MS/MS)[20,21]. In a typical MHC II immunopeptidome Eluted Ligand (EL) assay, antigen presenting cells are lysed and MHC molecules are purified via immunoprecipitation. The bound peptide ligands are next chromatographically eluted from MHC molecules and sequenced via MS/MS[22,23]. The result of such an assay is a list of peptide sequences restricted to at least one of the MHC molecules that the interrogated cell line expresses.

The biological relevance of EL data is a major advantage over BA data. EL data implicitly contains signal from steps of MHC II antigen presentation, such as antigen digestion, MHC loading of ligands and cell surface transport. Amino acid preferences in termini of immunopeptidome ligands is clear evidence of proteolytic digestion and ligand prediction has been improved by explicitly

encoding this signal in prediction models[20,24,25]. Ligand length preferences of MHC II is also inherent to EL data but not BA data and this preference can hence be learned by models trained on EL data[20,21].

While most prediction methods are trained either on binding affinity (BA) or MHC MS eluted ligand (EL) data, the work by Jurtz et al.[26] proposed an architecture where the two data types could effectively be integrated into one machine learning framework. This framework has later with high success been applied to train models with state-of-the-art performance for prediction of MHC class I antigen presentation and CD8 T cell epitopes[26,27]. Recently, the framework was further refined to also cover MHC class II data, and demonstrate how signals of antigen processing contained within and flanking the MS EL data could be effectively integrated to boost the predictive power[20].

The immense polymorphism of MHC combined with the high experimental cost burden associated with characterizing the specificity of individual MHC molecules, makes specificity characterization of all MHC molecules a prohibitively expensive undertaking. Given this, pan-specific prediction models have been proposed, which are trained on peptide interaction data covering large and diverse sets of MHCs with the purpose of and learning the associations between the MHC protein sequence and its peptide specificity. The value of these models lies in their ability to predict peptide-MHC binding for all alleles of known sequence including those characterized with limited or even no peptide binding data[28].

Ligands eluted from cell lines expressing only one MHC molecule can be unambiguously annotated to that allele and are termed single allele (SA) ligands. Such cell lines can be generated through genetic engineering[29] or careful experimental design, i.e. matching an immunopurification antibody specific to an MHC loci with a cell line homozygous to said loci (as done most often when

conducting an HLA-DR specific LC-MS/MS experiment). However, this scenario is the exception in immunopeptidomics studies. EL data from patient samples will more often be composed of peptides of mixed specificities corresponding to the different MHC molecules expressed in the given antigen presenting cell. Such data is termed multi allele (MA) ligands and annotating MA ligands to their respective MHC in such data is not a trivial task. Ligands from MA data can be deconvoluted into separate motifs in an unsupervised manner with tools like GibbsCluster[32], but this still leaves the task of assigning motifs to their respective alleles[30]. Solving the problem of MA ligand annotation holds the promise of extending EL data into allele specificities hitherto not characterized by SA data. A solution to this critical challenge of MA data has recently been proposed by an extension of the NNAlign algorithm: NNalign_MA[30].

NNAlign_MA handles MA data naturally by annotating MA data during training in a semi-supervised manner based on MHC co-occurrence, MHC exclusion, and pan-specific binding prediction. In the work by Alverez et al., the ability of NNalign_MA to automatically deconvolute and learn binding motifs from complex EL MA data was showcased in several scenarios including MHC I, BoLA-I and MHC II, and the algorithm demonstrated high potential for building an accurate pan-specific predictor for MHC II antigen presentation with broad allelic coverage.

In this work, we extend this earlier work and apply the NNalign_MA framework to construct the first pan-specific MHC II antigen presentation predictor including MS data. We do this by integrating three recent updates of the NNAlign algorithm; namely the two output neuron architecture allowing for training on both BA and EL data, explicit encoding of PFRs and ligand context to learn the signal of proteolytic antigen processing, and lastly benefitting from the potential of the NNAlign_MA method to train the prediction model on both SA and MA ligands from an extensive EL training set. Taken together, the goal is to maximally utilize the information gained from mass spectrometry

eluted ligands in the task of prediction of peptide-MHC II antigen presentation. The three outlined features will be shown to constructively contribute to the goal of ligand and ultimately epitope prediction.

Results

Accurate prediction of MHC class II antigen presentation has proven to be highly challenging. Early work has suggested that training models on EL data improves prediction performance. Here, we set out to validate this and build a model that incorporates recent advances to handle and fully benefit from extensive EL data sets. The underlying methodology is based on NNAlign_MA; a semi-supervised algorithm that annotates ligands and learns the associated MHC binding motifs from MA data sets[30]. In large-scale benchmarks, we investigate the ability of NNAlign_MA to consistently deconvolute motifs from data sets of mixed peptide specificities without diluting its applicability on simpler SA data. Furthermore, looking into the ligand context, we validate earlier findings suggesting an improved predictive performance empowered by signal of antigen processing, and investigate if this signal is consistent across MHC loci. To quantify the benefit of including EL data in MHC II interaction models, the developed model finally is benchmarked against the current state of the art predictor NetMHCIIpan-3.2 in terms of both MHC class II ligand and CD4+ epitope prediction.

NNAlign_MA algorithm

MA data are numerous and diverse, and more than half of ligands gathered for this project stem from MA data sets (see Supplementary Table 1 for a summary of EL data sets). When dealing with EL data from MHC II heterozygous cell lines, the ligands are of mixed specificities and it is not trivial to assign ligands to their respective restricting MHC II molecule. Using semi-supervised learning, NNAlign_MA leverages information from SA data to annotate MA data. This is achieved

with a burn-in period in which only SA data is used for training, after which MA data is introduced, annotated and used for training. The annotation is achieved by predicting binding to all MHC molecules assigned to the given MA data set, and assigning the restriction from the highest prediction value. With the allele assignment in place, the MA data becomes equivalent to SA data and can be used for training. Note, that this annotation step is repeated in every training iteration.

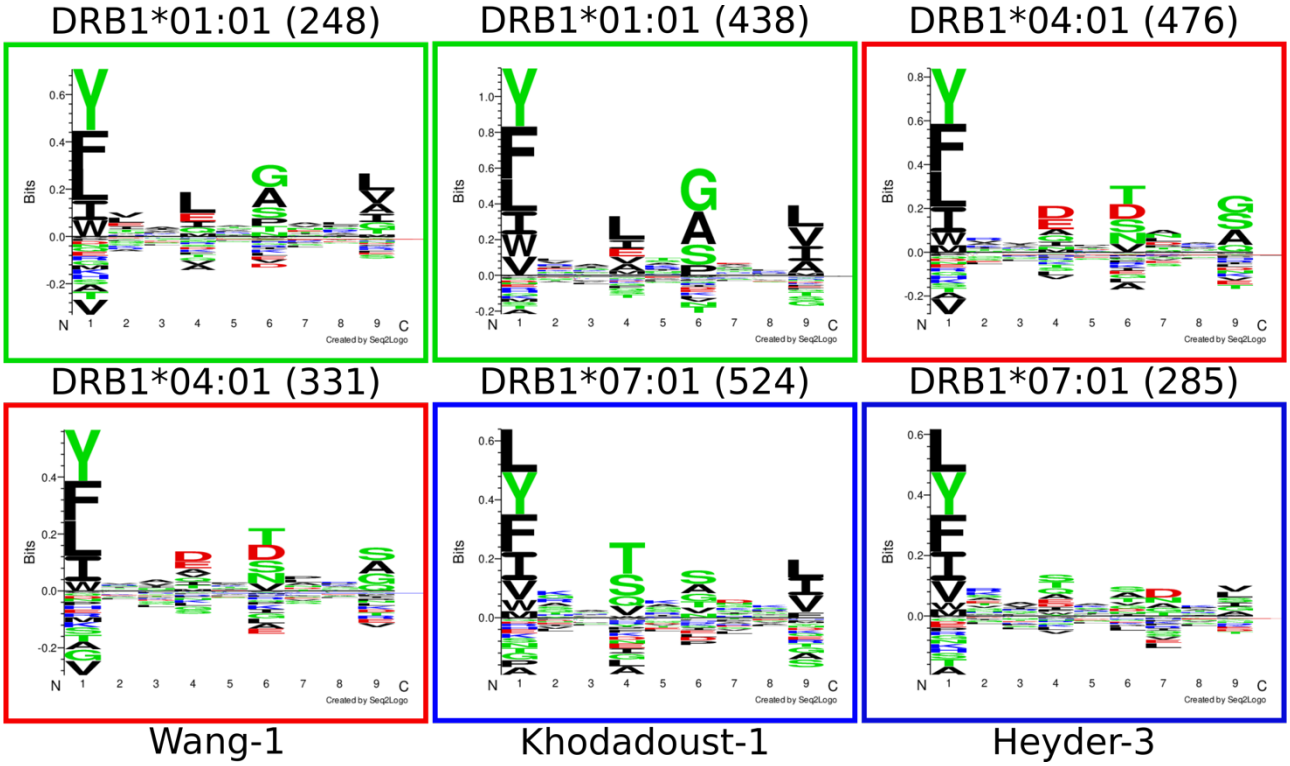

Figure 1) Multi-allele (MA) data motif deconvolution of 3 different cell lines sharing overlapping alleles. Each column of logos represents binding motifs for all alleles present in each cell line. Matching alleles across cell lines share common colours (green HLA-DRB1*01:01, red: HLA-DRB1*04:01, blue: HLA-DRB1*07:01). The figure shows the ability of NNAlign_MA to implicitly leverage allele overlap and the exclusion principle to annotate ligands from MA data and learn their binding motifs. The logos were generated with Seq2Logo, using only ligands with a prediction score of >0.01.

A great strength of NNAlign_MA is that MA data annotation is integral to the training process, circumventing the need for offline motif deconvolution and annotation. The fully automated process implicitly leverages data set allele overlap and exclusion principles to learn motifs and annotate ligands. This functionality is visualized in Figure 1 which shows the motifs identified from three MA

EL data sets that pairwise share one MHC allele. This figure illustrates how the allele annotation can be achieved by comparing motifs and alleles shared between multiple data sets. Note, also how a consistent motif is identified for DRB1*07:01 despite this molecule not being represented by SA EL data.

Comparison of models trained on SA and MA data

NNalign_MA's extends the training data by its ability to automatically annotate MA data. To access if NNalign_MA can learn specificities uniquely covered by MA data without diluting learning from SA data, we compared three models in terms of cross-validation performance. One model is trained exclusively on SA data (SA EL data and BA data) and another is trained on all data (SA EL, MA EL and BA data). We term these the SA- and MA-models, respectively. Finally, we included NetMHCIIpan-3.2 in the comparison as a representative model that is only trained on BA data. Figure 2 displays the performance of the different models on the two EL SA and MA data subsets (the performance of the SA and MA models evaluated using cross-validation). As with the motif deconvolution logos, only AUC performance for data sets with more than 100 ligands are presented.

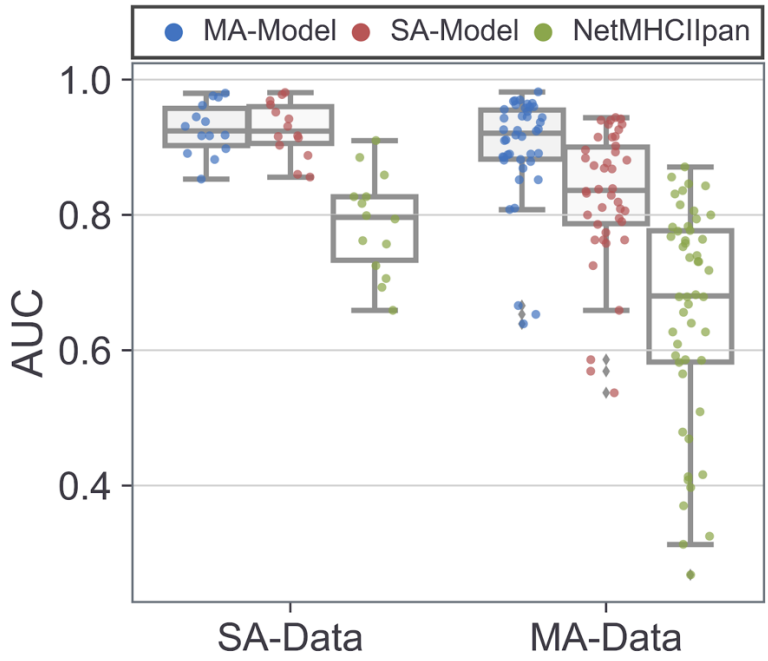

Figure 2) Comparing the performance of a model trained on SA data only and MA data, SA-model and MA-model respectively. Both models were evaluated on the whole data set by predicting each 5 test-sets in a cross-validation setup, and next were concatenating each test set and calculating data set wise AUC scores. Every point in the figure represents an AUC value for a single data set. The two models are compared separately for SA and MA EL data sets. The equivalency of the models is tested by counting the number of data sets for which each model has higher performance and performing a binomial test with p=0.5 and excluding ties.

These results demonstrated that the SA and MA models share comparable predictive performance when evaluated on SA data (p=0.774, binomial test excluding ties), supporting that NNAlign_MA maintains performance on SA data when including MA data in training. Moreover, the MA-model significantly outperformed the SA-model when evaluated on MA-data (p-value < 10^-12 in a binomial test). A closer look the data sets for which the MA-model outperformed the SA-model shed light on the source of the improved performance of the MA-model. By assigning to each MA data set a distance corresponding to the maximal nearest neighbour distance from its set of alleles to the set of SA alleles (for details on the allelic distance measure refer to Materials and Methods), we observe a strong correlation between this distance and the gain in

performance of the MA-model to that of the SA-model (see Supplementary Figure 2, PCC: 0.703). That is, the performance gain of MA-model was largest for data sets with one or more alleles very dissimilar to the SA data. This result thus demonstrates the power of NNAlign_MA to integrate information in MA data to effectively increase the allelic coverage of the training data and boost performance for alleles not characterized by SA data. By way of example does the EL-SA data set cover 11 MHC molecules, and when introducing the MA data, this number is increased to 41 (32 of which are assigned more than 100 ligands by NNAlign_MA). Finally, are both the SA and the MA models demonstrated to significantly outperform NetMNHCpan-3.2 on both EL data sets ($p<0.001$ in all cases). Taken together, these results demonstrate that NNAlign_MA can successfully deconvolute motifs from MA data and use this to boost the predictive power beyond that of methods trained on SA data. Likewise, these results confirm the earlier finding that methods trained on EL data outperform methods trained on BA data for the task of predicting MHC eluted ligands[20].

Source protein context boosts ligand prediction

Immunopeptidome data inherently contains signal from steps leading to antigen presentation and patterns of proteolytic digestion have been described in the terminal and context regions of ligands[20]. Earlier work has suggested that models that encode this context information have superior performance when predicting ligand data[20,24,25]. Here, we set out to validate this observation on the extensive data set gathered for this work.

The encoding of context data was performed as previously described[20] (for details refer to Materials and Methods), and models with and without encoding of ligand context were trained and evaluated using cross-validation. The results of the benchmark is shown in Figure 3, and demonstrate a significantly improved performance of the model trained including the ligand context

(MAC-Model, p<10^-13). This result thus confirms the earlier finding conducted on a data set covering only a limited set of HLA-DR alleles[20]. In Figure 2 and 3, three data sets have AUC scores below 0.7 for the MA model (Khodadoust-2, Heyder-6 and Ritz-11). All these data sets contain alleles that are only represented by MA data and all have relatively few ligands. This scenario complicates the task for NNAlign_MA, resulting in the observed low predictive performance, and high-lights the key general challenge associate with extracting meaningful information from data sets of limited size (we will discuss this further below).

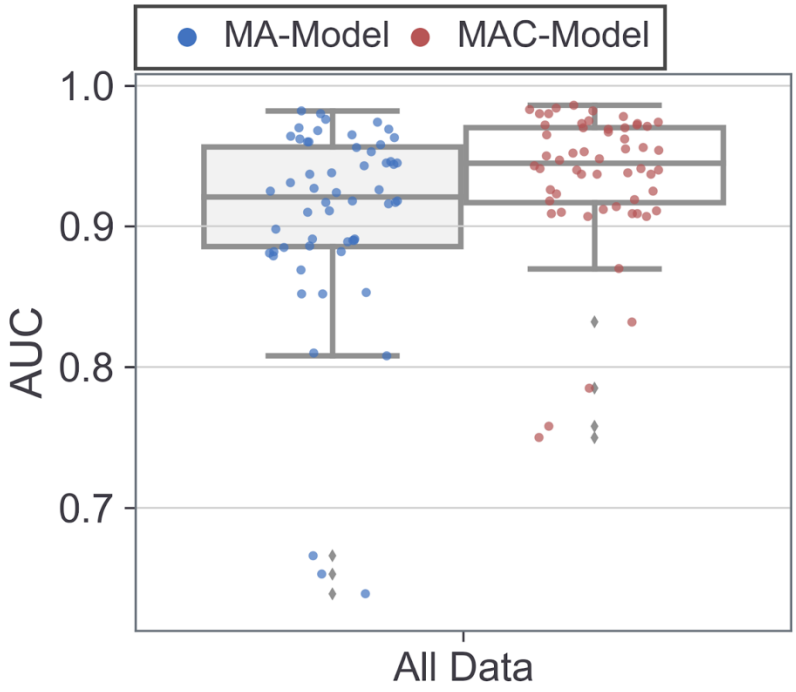

Figure 3) Comparison of the cross validation evaluation of MA models trained and evaluated with (MAC-Model) and without context (MA-Model). Each point in the figure represents a data set wise AUC. The statistical comparison of the two model was performed using binomial test, resulting in a p-value < 10^-13.

Consistent motif deconvolution from MA data

In the previous analyses, we have demonstrated the high predictive power of NNAlign_MA, and suggested that this performance was driven by accurate deconvolution of binding motifs in MA data. To further support this claim, we presented in Supplementary Figure 1, binding motifs generated from the predicted binding cores of ligands in each MA data set when deconvoluted by

NNAlign_MA. By visual inspection, this figure confirms that for a vast majority of cases, NNAlign_MA is able to successfully and consistently deconvolute motifs from MA data. For most DR data sets, motifs with sharp enrichment at positions P1, P4, P6 and P9 are captured. However, the figure also reveals that this ability can be compromised if the number of ligands assigned to an allele is low and/or the quality of the data is low. This is exemplified by the motifs for DRB1*08:01 - Heyder-6, DRB1*11:04 - Khodadoust-2, DRB1*13:01 - Heyder-4, Heyder-5 and Heyder-6, all defined from small ligand data sets, and the latter from the noisy motif for HLA-DRB3*01:01 in the Mommen-1 data set constructed from more than 6000 (80%) of the ligands. Note, also that motifs for all these alleles are defined sharply and consistent in other data sets, indicating that NNAlign_MA does learn the binding preference of these alleles, and that the poor resolution of the listed motifs is a results of poor data quality. We will discuss these issues further below.

A more quantitative analysis of deconvolution consistency was achieved by calculating pairwise correlations between PSSMs representing the motifs for MHC molecules shared between multiple data sets. That is, for every MHC molecule shared between 3 or more data sets, PSSMs were generated from the binding cores of ligands assigned to said allele for each data set, and the PSSMs pairwise compared in terms of a simple correlation analysis. The result of this analysis is visualized as heatmaps in Supplementary Figure 3. The median consistency score (for details on this measure refer to Materials and Methods) over the set of 11 molecules is 0.835 with most observations distributed tightly around this value (for an explanation of the consistency score, refer to Materials and Methods). DRB1*11:04 and DRB1*13:01 are two exceptions from this with consistency scores of 0.494 and 0.605, respectively. In line with what we have shown earlier, the source of these low consistency scores is in both cases grounded in the quantity and quality of the underlying MA data. Looking at the consistency heatmap for DRB1*13:01, it is clear that not all sources of data are of equal quality. The motifs from the three Heyder data sets[31] all have low

correlation to the remaining motifs and are all based on less than 100 ligands. Excluding these data sets from the analysis results in a consistency score of 0.783. As for DRB1*11:04, all motifs for this allele are based on less than 130 ligands, giving overall noisy results.

A further quantification of the consistency of the individual motif deconvolutions obtained by NNAlign_MA was obtained by calculating a predictive positive value (PPV) for each set of peptides (positive and negative) assigned to the given MHC molecule within each data set. In short, the PPV is calculated as the proportion of true positives within the top N predictions. To account for noise in the data, N was set to 90% of the number of ligands assigned to the given allele. The result of this analysis is presented in Supplementary Table 2. The median PPV0.9 was 0.770 and most molecules cluster around this value. Examples of low PPV0.9 (<0.60) molecules align with the earlier findings, including DRB1*11:04 and DRB1*13:01 from the three Heyder data sets, and DRB3*01:01 from the Mommen-1 data set, all mentioned earlier as being low quality and/or quantity data sets. The remaining molecules with PPV0.9 less than 0.60 are likewise characterized few ligands (<~200).

Finally, we tested if the motif for a given MHC molecule generated from MA data matched the motif obtained from SA data. The results of this analysis are visualized in Supplementary Figure 4. Here, motifs as obtained by the MA and SA data for DR molecules represented by both data types. Comparing the motifs in terms of the correlation between the two PSSMs, we in all cases find a correlation above 0.9. This observation may seem trivial since the network has learned SA motifs during pre-training. However, the result confirms that the inclusion of MA data and the successful annotation of MA-exclusive molecules does not dilute the signal of the SA data.

Combined, these results support the claim that NNAlign_MA can accurately and consistently deconvolute binding motifs in EL MA data.

Signal of proteolytic antigen processing

We have previously demonstrated how the predictive power of NNAlign_MA is boosted when incorporating information from the ligand source protein context (Figure 3). To further investigate the source of this, and seek to relate the findings to specific proteolytic signals of antigen processing, we in Figure 4, show sequence logos representing the N and C terminal context signal contained within the EL data set. Here, all ligands were mapped to their antigen source protein to extract ligand context of 6 residues (3 upstream and 3 downstream of the ligand). Along with the context, 3 residues at each termini were extracted. A few observations can be made from this figure: firstly, the signal is predominantly contained within the ligand rather than the context (Figure 4A). Secondly, the data show a pronounced enrichment of Proline in positions N+2 and C-2 in agreement with earlier findings[20]. Finally, there is an enrichment for charged amino acids at both termini. The two latter observations are supported by an unsupervised deconvolution of the context motif using GibbsCluster[32] (Figure 4B). This deconvolution clearly shows the strong preference for P at positions N+2 and C-2, and displays further well-defined motifs supporting the notion that proteolytic enzymes of more than one specificity are at play in the degradation of MHC II antigens in agreement with earlier findings[24]. Analysing the data separately for DR, DQ and H2 molecules revealed very similar results (see Supplementary Figures 5, 6 and 7), again suggesting that the observed signal and identified motifs are related to antigen processing and not MHC binding.

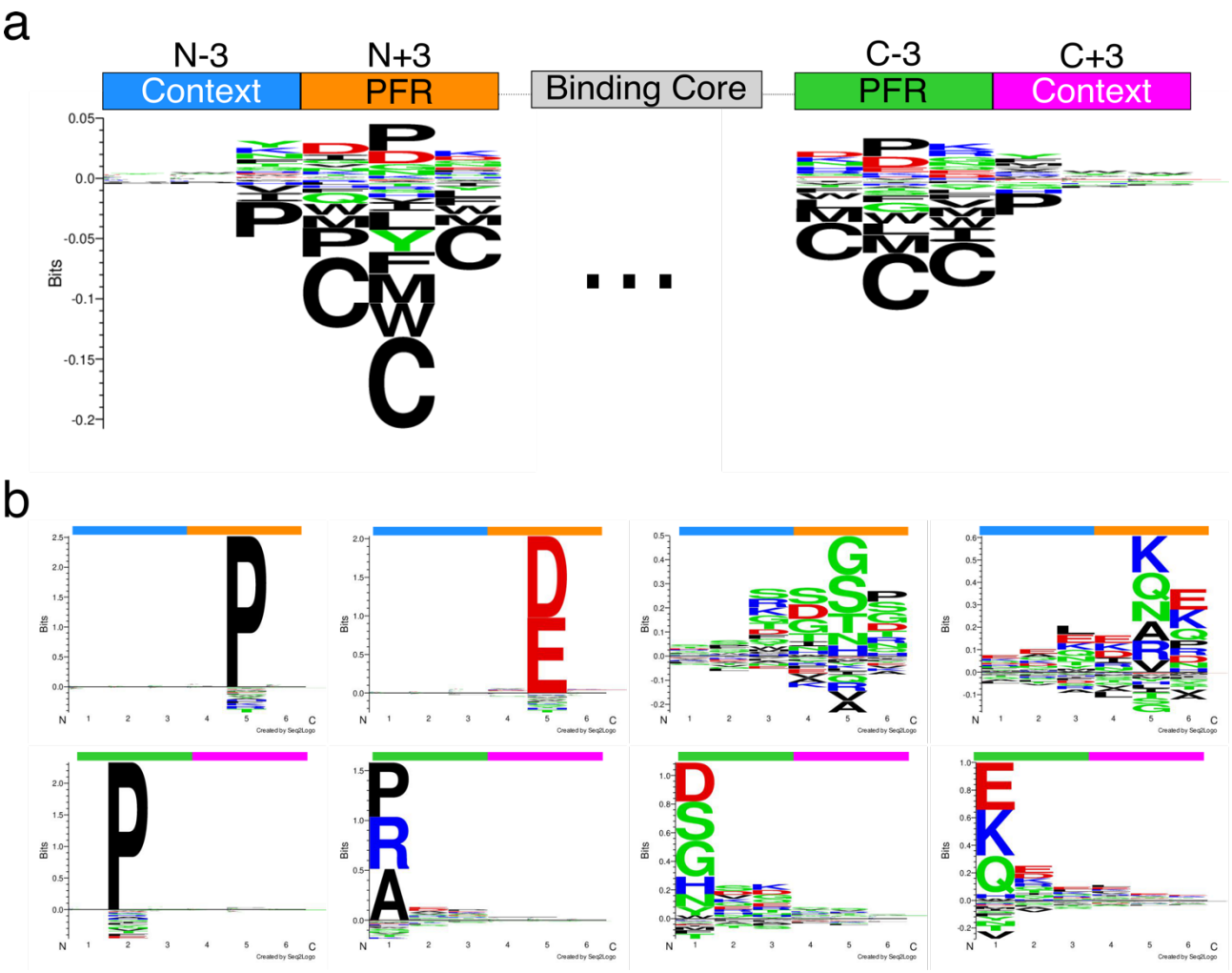

Figure 4) Sequence logos for N- (top left) and C-terminal (top right) and context sequences for all ligands in the data set. The top panel shows separate logos for the N- and C-terminus of all ligands (with PFRs of length >2). The diagram indicates placement of residues relative to the N- and C-terminus of ligands. The bottom two panels show the GibbsCluster deconvolution of N- and C-terminal contexts, respectively.

CD4+ epitope evaluation

Prediction of CD4+ epitopes is the ultimate benchmark of peptide-MHC II binding predictors. Earlier work has suggested that the gained performance for prediction of MS ligands observed when including context information only to a minor degree is transferred to epitope identification[20,24]. Here, we set out to test if the models developed here would align with this finding. That is, we compared the predictive performance of models trained with and without

context encoding (MAC-model/MA-Model) on a large set epitopes from the IEDB. As a reference both methods were further benchmarked against the state of the art method NetMHCIIpan-3.2. In the benchmark performance was evaluated using the Frank score. In short, Frank is the proportion of false-positive predictions within a given epitope source protein, i.e. percentage of peptides with a prediction score higher than that of the epitope. Using this measure, a value of 0 corresponds to a perfect prediction (the known epitope is identified with the highest predicted binding value among all peptides found within the source protein) and a value of 50.0 to random prediction. In a first experiment (Figure 5), Frank values were obtained by *in silico* digesting the epitope source protein into overlapping peptides of the length of the epitopes. This benchmark demonstrated a significantly improved performance of the MA-model compared to NetMHCIIpan-3.2 ($p<0.01$, binomial test excluding ties, median Frank values for the two methods of 2.454 and 3.921, respectively). The results however also confirm the earlier finding that signals of antigen processing contained within peptide context did not benefit the predictive power (performance of MAC-Model was comparable to that of NetMHCIIpan).

As discussed earlier[20,24], this observation might not come as a surprise, since peptides tested for T cell immunogenicity most commonly are generated as overlapping peptides scanning a source protein, and hence are not expected to follow any rules of antigen processing. To account for this bias, and to try to bring out the full potential of modelling context, we next applied a scoring scheme where the score of a given peptide was assigned from the sum of the individual prediction values from all 13-21mer peptides with predicted binding cores overlapping the given peptide (for details see Materials and Methods). However, even though this scoring scheme levelled out the difference between the MA- and MAC-models, the results demonstrated an overall loss in predictive power compared to using the simpler scoring scheme in Figure 5 (data not shown).

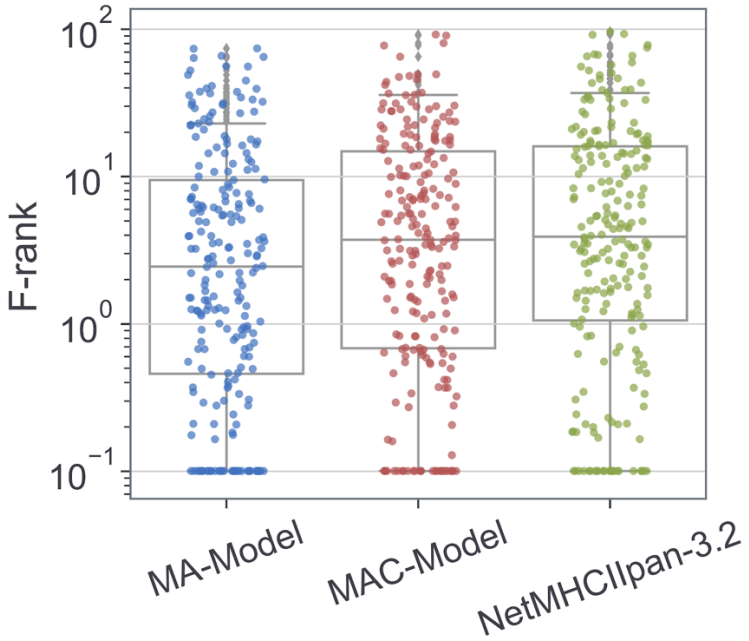

Figure 5) Frank evaluation of NNalign_MA and NetMHCIIpan-3.2 models on ICS epitopes from the IEDB. The MAC-Model (red) and MA-Model (blue) are trained with NNAlign_MA on MA data, with and without context encoding, respectively. The F-rank results for NetMHCIIpan-3.2 are shown in green. Each point in the plot represents the Frank of one epitope, the Frank being the percentage of false positive length matched peptides from the epitope source protein, as described in methods. Only results for epitopes for which at least one prediction method had Frank below 20 are presented, in total 228 epitopes. For visualization, Frank values of 0 are presented as 0.1005.

These findings thus align with earlier results[20,24] demonstrating an improved performance for CD4+ epitope identification of models trained including EL data compared to methods trained on binding affinity data only (NetMHCIIpan-3.2), but also somewhat surprisingly show that adding context information to such predictors does not boost performance for prediction of epitopes despite this being observed for prediction of MHC ligands.

A webserver implementation of the MA and MAC models is available at

http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/index_beta.php

Discussion

In this work, we have trained and evaluated a pan-specific MHC II predictor on MHC II immunopeptidome Eluted Ligand (EL) and binding affinity (BA) data. This model integrates recent advances proposed towards full utilization of EL data for prediction of MHC antigen presentation: two output neuron architecture, the encoding of proteolytic ligand context, and use of the NNAlign_MA machine learning framework allowing integration of multi-allele (MA) EL data. With an extensive EL data set, we have in a stepwise manner shown how these features come together to build model that significantly improves upon its predecessor: NetMHCIIpan-3.2 both when it comes to predicting ligands and CD4+ epitopes.

Firstly, we provide examples of how NNAlign_MA consistently deconvolutes MHC binding motifs from separate MA data sets, even for MHC molecules that are not represented by SA data. This confirms that NNAlign_MA provides an accurate solution to the challenge of annotating and learning from MA data. The model implicitly learns that data sets sharing alleles also share motifs and through the motifs learned from SA data and the principle of exclusion the model is able to assign motifs to individual MHC molecules. In few cases, we observed that data sets and/or alleles fared consistently poorly on all quantitative measures (cross-validation AUC, motif deconvolution, mean correlation of PSSMs and PPV0.9). All these data sets were characterized by low amounts of data and/or low data quality emphasizing a general issue; namely that the outcome of a machine-learning exercise is only as good as the quality of the input data. This however also hints to a direction of potential improvement of NNAlign_MA where an improved analytic power could be achieved by implementing a manner to handle noise along the lines for instance of the trash-bin suggested for instance in the GibbsCluster algorithm[32].

Secondly, MA-models were found to outperform both SA-models and NetMHCIIpan in cross-validation evaluation of MA data. MA models and SA models have comparable performance on SA data supporting the claim that the learning and performance of NNAlign_MA is not diluted by MA data. Relatedly, we observed that the largest gain of MA over SA models was observed for data sets that contained alleles distant to alleles covered by SA data. This result indicates that NNAlign_MA successfully increases the coverage of predictable alleles by training on MA data. In numerical terms, the number of close HLA-DR neighbours (distance of pseudo-sequences <0.05) to the EL data set is increased from 161 to 454 alleles by the inclusion of MA data (comparing to the set of all known MHC II DR). This observation is critical for our goal of training a pan-specific MHC II predictor on EL data, as the success of pan-specific predictors is contingent on training data with broad allelic coverage[33,34].

Encoding proteolytic context in models improved performance for prediction of ligands. We found that proteolytic context of ligands is a general feature that holds across MHC II loci and species. Clustering context information resulted in motifs with a strong signal of amino acid enrichment/depletion in the peptide flanking regions. These motifs are likely a signature of proteolytic cleavage processes taking place in the MHC II antigen presentation pathway[24]. Across loci and species, we observed a motif of Proline enrichment in the N+2 and C-2 positions of context. This is consistent with the hypothesis of blocked endopeptidase digestion by the secondary α-amine of Proline. Likewise, we consistently observed motifs with enrichment for aspartic and glutamic acid. We suggest this enrichment reflects Cathepsin disfavouring negatively charged amino acids near cleavage sites [35], leading to decreased digestion of ligands with said signature. This suggestion is further supported by the enrichment of negatively charged amino acids observed in EL data compared to BA binders (data not shown). With lesser information content, a motif of the N-terminal context deconvolution, displayed enrichment of glycine, serine

and threonine. A similar enrichment has been observed downstream of the cleavage site for Cathepsin S, a protease known to participate in the MHC II antigen presentation pathway[36]. The identified processing motifs are thus to a high degree associated with decreased potential for antigen processing. Given this, we speculate that they could serve as novel MHC II agnostic ways for engineering to enhance/reduce peptide and protein immunogenicity; that is amino acid mutations away from these motifs would lead to enhanced potential for cleavage, leading to a lower level of antigen presentation, and hence a reduction in potential immunogenicity. Likewise, could mutations into these motifs, lower the potential for antigen processing, inducing an increased immunogenicity.

Finally, the proposed model was found to improve CD4+ epitope prediction when compared to the state-of-the-art predictor NetMHCIIpan-3.2. However, in this benchmark, models trained with context encoding could at best be considered equivalent to models without context. This was observed even when accounting for bias in epitope data imposed by analysed T cell epitope peptides most often being generated synthetically and thus not reflecting rules of natural antigen processing. It is currently not clear to us what is the source of this surprising discrepancy between the performance of models trained with and without context encoding when evaluated on ligand and epitope data. This puzzling result has been observed before, and future work is needed to resolve if the discrepancy has a biological explanation or is imposed by biases in the epitope data (most epitope data are generated with a bias towards motives contained within current MHC-peptide binding algorithms, or assessed using multimer assays where the binding to MHC is performed in-vitro hence ignoring expression, processing, and the potentially important role of chaperones such as DM for the editing and formation of a stable peptide interactions).

The vast majority of current EL data has been generated for HLA-DR. By way of example are more than 75% of the data sets included in this work HLA-DR specific, and none of the data covers DP. The reasons for this are many and include the complicating factor that both the alpha and beta chains of HLA-DQ and HLA-DP are polymorphic making it non-trivial to assess which of the four possible combinations of alpha and two beta chains are presented as HLA molecules in a given cell. However, careful selection homozygous cell lines or cell lines with only one expressed alpha chain could help resolve this[37], and in this context EL data covering DQ and DP can without any further complication be integrated in NNAlign_MA modelling framework.

Properties other than antigen processing and HLA binding, such as protein expression, and differential access to the MHC-II presentation pathway, contribute to the likelihood of antigen presentation. In a recent paper Abelin et al.[37] have proposed a modelling framework integrating a panel of such properties, suggesting large improvement for prediction of HLA ligandomes. Further work remains to be done to validate the generality of these findings and their potential impact for general rational epitope discovery.

In conclusion, we have shown how the relatively simple NNAlign_MA machine learning framework can be applied to deconvolute binding motifs in MHC class II MA EL data sets, and how this deconvolution allows for identification of signals associated with antigen processing, and construction of a prediction model with significantly improved performance compared to state of the art for prediction of both eluted ligands and CD4 T cell epitopes.

Materials and Methods

Binding affinity data

Binding affinity data was gathered from the NetMHCIIpan-3.2 publication[10]. In line with the EL data set, only BA peptides of length 13-21 were included in the analysis. After filtering, the data set contained 131,077 peptides covering 59 HLA-II molecules. Peptide binding affinity data is quantitative and the measured binding affinity was transformed to fall in the range [0,1] as described in[38].

Eluted ligand data - Extraction

All MHC ligands in the IEDB[39] were downloaded (January 28th 2019th) and this table served as a guide to identify publications with MHC II EL data. Publications with full HLA class II typing and more than 1000 ligands were selected for data extraction. Ligand data was subsequently extracted from the individual publications [31,40-48]. Supplementary materials of the publications were processed and ligands extracted along with their source proteins and associated MHC molecules. Some additional data sets were found via literature search[49-52].

The resulting data sets were tables of ligands, their source proteins and a list of possible MHC II molecules from which the ligands could have been eluted (the MHC molecules expressed in the given cell line). All post-translationally modified peptides were excluded and only ligands of length 13-21 were included in the analysis (excluding 20.4% of the total extracted EL data), resulting in a total of 65 data sets, covering 102,584 MHC measurements, and 41 distinct MHC molecule. For a complete overview of the data refer to Supplementary Table 1.

Eluted ligand data - Negative peptide generation

Eluted ligand data contain only positive examples of MHC ligands. To successfully train a peptide-MHC interaction model, negative examples of ligands must be provided. These negative data were defined as described earlier[20,27] by randomly sampling peptides from the UniProt database[53]. Negatives were generated per data set and were made to follow a uniform length distribution in the range 13-21. For each length, the number of negatives was equal to 5 times the number of positives at the mode of the data set peptide length distribution.

Eluted ligand data - Context

Terminal regions inside and outside of ligands were encoded to capture signals of proteolytic digestion, as described previously[20]. A total of 12 residues were encoded for each ligand: six PFRs (three residues from the N-terminal and three from the C-terminal) and six ligand context residues from the source protein sequence (three residues upstream the ligand N-terminus and three residues downstream the C-terminus). Roughly 2% of ligands could not be mapped to a source protein and were encoded with a context of X's. BA data also received a context of X's. Negative EL data in training sets were also assigned context from the source protein from which they were sampled.

Data set partitioning

NNalign_MA was trained in a 5-fold Cross-Validation manner. Data was partitioned via common motif clustering[54] to ensure that no 9-mer subsequences were shared between partitions. BA and EL data were clustered simultaneously and next separated, resulting in a total of 10 partitions, 5 for EL data and 5 for BA data.

Epitopes

To evaluate the ability of prediction methods to prioritise epitope discovery, T Cell epitopes were downloaded from the IEDB (downloaded: March 21st 2019), and filtered to include only positive MHC II peptide epitopes of length 13-21. Furthermore, only epitopes measured by ICS were included in the benchmark.

Performance was evaluated using the F-rank measure. In short, epitopes are ranked by prediction scores amongst their length matched peptides from the antigen, and the F-rank score is calculated as the percentage of false positives, i.e. peptides with prediction scores higher than measured epitopes. To incorporate effects of antigen processing, an alternative scoring scheme was designed. Here, each antigen was *in silico* digested into kmers (peptides of length 13-21 amino acids) and scores predicted for each. Next, epitope length matched peptides from an antigen was assigned a score by summing prediction scores of all kmers which predicted binding core overlapping the given peptide. To remove noise in the data set, and filter out potential false positive epitopes, the data set was filtered to only include epitope with an F-frank values 20,0 for at least one of the methods benchmarked.

NNAlign_MA training

Network architectures with 2, 10, 20, 40 and 60 hidden neurons were used. For each training set, 10 random weight initializations were used. This resulted in an ensemble of 250 networks (5 folds, 5 architectures, and 10 seeds). All models were trained using backpropagation with stochastic gradient descent, for 400 epochs, without early stopping, and a constant learning rate of 0.05. Only SA data was included in training for a burn in period of 20 epochs, after which MA data was included in subsequent training cycles.

Sequence logos

Sequence Logos were generated with Seq2Logo[55]. The amino acid background frequency for all logos was constructed from the set of ligand source proteins. MHC II binding motifs detected in EL data sets were generated from ligand binding registers as predicted by NNAlign_MA, using default Seq2Logo setting. For context, separate logos were made for the N- and C-terminal using weighted Kullback-Leibler, excluding sequence weighting. Logos from GibbCluster deconvolution of context were generated with default settings of Seq2Logo. Note, that a condition for inclusion in context motif visualization was that the PFR of the ligand be 3 residue or longer, in agreement with earlier work[20].

Allele sequence distance measure

Sequence distance between alleles is calculated based on the alignment score of their pseudo sequences with the following relation: $d = s(A,B)/\sqrt{s(A,B) \cdot s(A,B)}$, where $s(A,B)$ represents the BLOSUM50 alignment score of the two pseudo sequences[56].

Deconvolution consistency score

As a measure of how consistent NNAlign_MA performs the task of MHC motif deconvolution, the motifs for a given MHC as obtained from different MA data sets were compared in terms of the Pearson correlation coefficient (PCC) of their PSSMs. Next, a consistency score for the deconvolution of an allele was defined as the average over these individual PCCs.

References

1. Unanue, E. R., Turk, V. & Neefjes, J. Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annu. Rev. Immunol.* **34**, 265–297 (2016).
2. Wieczorek, M. *et al.* Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front. Immunol.* **8**, 292 (2017).
3. Traherne, J. A. Human MHC architecture and evolution: implications for disease association studies. *Int. J. Immunogenet.* **35**, 179–192 (2008).
4. Chicz, R. M. *et al.* Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* **358**, 764–768 (1992).
5. Andreatta, M. *et al.* Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* **152**, 255–264 (2017).
6. Godkin, A. J. *et al.* Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. *J. Immunol.* **166**, 6720–6727 (2001).
7. Buus, S., Sette, A., Colon, S. M., Miles, C. & Grey, H. M. The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. *Science* **235**, 1353–1358 (1987).
8. Nielsen, M. & Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* **10**, 296 (2009).
9. Andreatta, M. *et al.* Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* **67**, 641–650 (2015).
10. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406 (2018).
11. Sturniolo, T. *et al.* Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* **17**, 555–561 (1999).

12. Zhang, L. *et al.* TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PLoS One* **7**, e30483 (2012).

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

14. Grifoni, A. *et al.* Characterization of Magnitude and Antigen Specificity of HLA-DP, DQ, and DRB3/4/5 Restricted DENV-Specific CD4+ T Cell Responses. *Front. Immunol.* **10**, 1568 (2019).

15. Rosa, D. S., Ribeiro, S. P. & Cunha-Neto, E. CD4+ T cell epitope discovery and rational vaccine design. *Arch. Immunol. Ther. Exp.* **58**, 121–130 (2010).

16. Lundegaard, C., Lund, O. & Nielsen, M. Predictions versus high-throughput experiments in T-cell epitope discovery: competition or synergy? *Expert Rev. Vaccines* **11**, 43–54 (2012).

17. Griswold, K. E. & Bailey-Kellogg, C. Design and engineering of deimmunized biotherapeutics. *Curr. Opin. Struct. Biol.* **39**, 79–88 (2016).

18. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404 (2016).

19. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med.* **7**, 119 (2015).

20. Barra, C. *et al.* Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* **10**, 84 (2018).

21. Garde, C. *et al.* Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* (2019). doi:10.1007/s00251-019-01122-z

22. Dudek, N. L., Croft, N. P., Schittenhelm, R. B., Ramarathinam, S. H. & Purcell, A. W. A Systems Approach to Understand Antigen Presentation and the Immune Response. *Methods Mol. Biol.* **1394**, 189–209 (2016).

23. Caron, E. *et al.* Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol. Cell. Proteomics* **14**, 3105–3117 (2015).

24. Paul, S. *et al.* Determination of a Predictive Cleavage Motif for Eluted Major Histocompatibility Complex Class II Ligands. *Front. Immunol.* **9**, 1795 (2018).

25. Racle, J. *et al.* Deep motif deconvolution of HLA-II peptidomes for robust class II epitope predictions. *bioRxiv* 539338 (2019). doi:10.1101/539338

26. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

27. Nielsen, M., Connelley, T. & Ternette, N. Improved Prediction of Bovine Leucocyte Antigens (BoLA) Presented Ligands by Use of Mass-Spectrometry-Determined Ligand and in Vitro Binding Data. *J. Proteome Res.* **17**, 559–567 (2018).

28. Nielsen, M. *et al.* NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* **2**, e796 (2007).

29. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).

30. Alvarez, B. *et al.* NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T cell epitope predictions. *bioRxiv* 550673 (2019). doi:10.1101/550673

31. Heyder, T. *et al.* Approach for Identifying Human Leukocyte Antigen (HLA)-DR Bound Peptides from Scarce Clinical Samples. *Mol. Cell. Proteomics* **15**, 3017–3029 (2016).

32. Andreatta, M., Alvarez, B. & Nielsen, M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research* **45**, W458–W463 (2017).

33. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13 (2009).

34. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177–186 (2012).

35. Choe, Y. *et al.* Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* **281**, 12824–12832 (2006).

36. Biniossek, M. L., Nägler, D. K., Becker-Pauly, C. & Schilling, O. Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *J. Proteome Res.* **10**, 5363–5373 (2011).

37. Abelin, J. G. *et al.* Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity* (2019). doi:10.1016/j.immuni.2019.08.012

38. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017 (2003).

39. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

40. Bergseng, E. *et al.* 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**, 73–84 (2015).

41. Sofron, A., Ritz, D., Neri, D. & Fugmann, T. High-resolution analysis of the murine MHC class II immunopeptidome. *Eur. J. Immunol.* **46**, 319–328 (2016).

42. Clement, C. C. *et al.* 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. *J. Biol. Chem.* **291**, 5576–5595 (2016).

43. Karunakaran, K. P. *et al.* Identification of MHC-Bound Peptides from Dendritic Cells Infected with Salmonella enterica Strain SL1344: Implications for a Nontyphoidal Salmonella Vaccine. *J. Proteome Res.* **16**, 298–306 (2017).

44. Ooi, J. D. *et al.* Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature* **545**, 243–247 (2017).

45. Wang, Q. *et al.* Immunogenic HLA-DR-Presented Self-Peptides Identified Directly from Clinical Samples of Synovial Tissue, Synovial Fluid, or Peripheral Blood in Patients with Rheumatoid Arthritis or Lyme Arthritis. *J. Proteome Res.* **16**, 122–136 (2017).

46. Nelde, A. *et al.* HLA ligandome analysis of primary chronic lymphocytic leukemia (CLL) cells under lenalidomide treatment confirms the suitability of lenalidomide for combination with T-cell-based immunotherapy. *Oncoimmunology* **7**, e1316438 (2018).

47. Ritz, D. *et al.* Membranal and Blood-Soluble HLA Class II Peptidome Analyses Using Data-Dependent and Independent Acquisition. *Proteomics* **18**, e1700246 (2018).

48. Ting, Y. T. *et al.* The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. *J. Biol. Chem.* **293**, 3236–3251 (2018).

49. Scally, S. W. *et al.* A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. *J. Exp. Med.* **210**, 2569–2582 (2013).

50. Khodadoust, M. S. *et al.* Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature* **543**, 723–727 (2017).

51. Mommen, G. P. M. *et al.* Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. *Mol. Cell. Proteomics* **15**, 1412–1423 (2016).

52. Álvaro-Benito, M., Morrison, E., Abualrous, E. T., Kuropka, B. & Freund, C. Quantification of HLA-DM-Dependent Major Histocompatibility Complex of Class II Immunopeptidomes by the Peptide Landscape Antigenic Epitope Alignment Utility. *Front. Immunol.* **9**, 872 (2018).

53. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

54. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).

55. Thomsen, M. C. F. & Nielsen, M. 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 Res.* **40**, W281–7 (2012).

56. Nielsen, M. *et al.* Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput. Biol.* **4**, e1000107 (2008).