







Improved prediction of HLA antigen presentation hotspots: Applications for immunogenicity risk assessment of therapeutic proteins

Anders Steenholdt Attermann,¹ 
Carolina Barra,¹ 
Birkir Reynisson,¹ 
Heidi Schiøler Schultz,²
Ulrike Leurs,² 
Kasper Lamberth² 
and
Morten Nielsen^{1,3} 

¹Department of Health Technology, Technical University of Denmark, Lyngby, Denmark,

²Assay, Analysis & Characterisation, Global Research Technologies, Novo Nordisk A/S, Måløv, Denmark and ³Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

doi:10.1111/imm.13274

Received 17 July 2020; revised 15 September 2020; accepted 16 September 2020.

Attermann and Barra are regarded as joint First Authors.

Funding information

This work was supported in part by the 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).

Correspondence

Morten Nielsen, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark.
Email: morni@dtu.dk
Senior author: Morten Nielsen

INTRODUCTION

Therapeutic proteins such as antibodies and coagulation factors are an essential part of modern medicine that provides treatments for complex diseases including cancer, rheumatic diseases and haemophilia. Inherently, therapeutic proteins bear the risk that the patient's immune

system responds to it as foreign. A key example of this is the generation of antidrug antibodies (ADAs). ADAs have the potential to neutralize the functional activity of the drug and to increase drug clearance, ultimately affecting its pharmacokinetic and pharmacodynamic profiles. In addition, ADAs can potentially cross-react with non-redundant endogenous proteins and, in rare cases, elicit

Summary

Immunogenicity risk assessment is a critical element in protein drug development. Currently, the risk assessment is most often performed using MHC-associated peptide proteomics (MAPPs) and/or T-cell activation assays. However, this is a highly costly procedure that encompasses limited sensitivity imposed by sample sizes, the MHC repertoire of the tested donor cohort and the experimental procedures applied. Recent work has suggested that these techniques could be complemented by accurate, high-throughput and cost-effective prediction of *in silico* models. However, this work covered a very limited set of therapeutic proteins and eluted ligand (EL) data. Here, we resolved these limitations by showcasing, in a broader setting, the versatility of *in silico* models for assessment of protein drug immunogenicity. A method for prediction of MHC class II antigen presentation was developed on the hereto largest available mass spectrometry (MS) HLA-DR EL data set. Using independent test sets, the performance of the method for prediction of HLA-DR antigen presentation hotspots was benchmarked. In particular, the method was showcased on a set of protein sequences including four therapeutic proteins and demonstrated to accurately predict the experimental MS hotspot regions at a significantly lower false-positive rate compared with other methods. This gain in performance was particularly pronounced when compared to the NetMHCIIpan-3.2 method trained on binding affinity data. These results suggest that *in silico* methods trained on MS HLA EL data can effectively and accurately be used to complement MAPPs assays for the risk assessment of protein drugs.

Keywords: HLA antigen presentation; HLA eluted ligands; immunogenicity assessment; prediction; protein immunogenicity.

Abbreviations: ADAs, antidrug antibodies; BA, binding affinity; EL, eluted ligand; MAPPs, MHC-associated peptide proteomics; MCC, Matthews correlation coefficient; MHC-II, major histocompatibility complex class II; MS, mass spectrometry

anaphylactic reactions, both having detrimental complications for the patients.¹

Failure to identify such immunogenicity risks early in preclinical stages has critical economical and potential patient health implications for protein drug development. A striking example of this includes vatreptacog alfa (VA), a modified recombinant activated human factor VIIa (rFVIIa) analogue with improved procoagulant activity. The VA variant differed from the WT-rFVIIa in 3 amino acid substitutions, V158D, E296V and M298Q. Due to an incidence of 11% ADAs in phase 3 of clinical trials, the development of VA was discontinued. In contrast, there have been no reports of ADAs with the WT-rFVIIa, which has been used clinically for more than two decades.^{2,3}

In view of this, effective means to assess the risk of therapeutic protein immunogenicity are highly on demand. ADAs are most often T-cell-dependent and generated through T-cell activation of B cells. The most selective step defining T-cell activation is the major histocompatibility complex class II (MHC-II) antigen presentation of peptides to T cells,⁴ and MHC-II presentation is often utilized as a proxy measure of immunogenicity.^{1,3} Advancements in mass spectrometry (MS) technologies have greatly enhanced the sensitivity of MHC-associated peptide proteomics (MAPPs) assays. The more general MS HLA ligand elution technique^{5–7} has been used to identify large volumes of HLA ligands (so-called eluted ligand or EL data sets) casting renewed light on the rules of HLA antigen presentation^{8–11} and HLA binding motifs,^{12,13} and has served as a tool for the identification of pathogen and cancer neoepitopes.^{14,15}

MHC-associated peptide proteomics is a powerful tool to identify immunogenic hotspots of biotherapeutics (reviewed in Ref. 16), and the main goals of using MAPPs have been to understand the causes of immunogenicity, compare biotherapeutic candidates, guide deimmunization and investigate immunogenicity observed in the clinic (reviewed in Ref. 17). In contrast to technologies based on HLA binding only, MAPPs can identify an MHC-II peptide repertoire from proteins who have been taken up, processed and presented by antigen-presenting cells – thereby identifying more ‘real’ potential T-cell epitopes with a lower degree of overpredictiveness. However, while natural uptake, processing and HLA presentation of peptides are a prerequisite for the development of high-affinity ADAs, MAPPs data should not be considered as a direct prediction of neither T-cell epitopes nor ADA and immunogenicity towards a given biotherapeutic. Development of ADAs is dependent on additional multiple factors including recognition by T and B cells, immune status of the patient, route of administration and induction of tolerance.^{18,19} Therefore, the ability to trigger T-cell and ADA responses has to be addressed experimentally subsequently.

Despite the clear benefit of using MAPPs assays to evaluate the risk of immunogenicity, this technique is highly cost-intensive and time-consuming, and the outcome is limited in sensitivity imposed by sample sizes, the MHC repertoire of the donor cohort tested and the experimental procedures applied. On that basis, it would be highly attractive if accurate high-throughput and cost-effective prediction models could be developed to complement these experimental techniques.

Historically, MAPPs data have been difficult to use for developing MHC-II predictors, as the obtained peptidome data are multi-allelic, and therefore ‘concealing’ the peptide-to-allele mapping information, which is needed for training a supervised model. Most in silico MHC-II prediction tools used to predict therapeutic protein immunogenicity have therefore been trained using in vitro peptide binding affinity (BA) data.^{20,21} These data fail to capture antigen-processing effects.²² In addition, in vitro peptide BA has proven to be a relative poor correlate to peptide T-cell immunogenicity²³ and MHC-II BA predictors have been demonstrated to share a relatively limited specificity leading to a large proportion of false-positive predictions and resulting in an overprediction of potential immunogenicity.^{10,24–26} In contrast, MAPPs data (and MS MHC EL data set in general) are a more rich data source, as information is obtained from all steps of the antigen processing and presentation including natural peptide length-preference distribution and signals of proteolytic cleavage in the termini of the MAPPs ligands.^{10,12,27–29} Several studies have also demonstrated how such EL data can be utilized to train improved prediction methods for HLA antigen presentation.^{11,12,29–31}

Recently, Barra et al.³² illustrated how neural networks could be trained on MS data, which could be effectively applied to predict HLA-DR antigen presentation hotspots and T-cell immunogenicity for infliximab and rituximab. This method showed superior performance to conventional HLA binding prediction methods trained on in vitro BA data. However, this study was limited in the amount of data available as only seven donors were included for the experimental assessment of the infliximab HLA hotspot regions, and no donor HLA-typing information was available for the rituximab data limiting the performance evaluation.

In this study, we sought to resolve these limitations and performed a more profound performance evaluation of methods for the prediction of HLA antigen presentation hotspot regions, donor-specific MAPPs experiments and risk assessment of protein drug immunogenicity. Here, an extensive novel MAPPs data set combined with publicly available MS HLA eluted ligand and peptide BA data was utilized to train a predictor of MHC-II ligand presentation. The predictor was trained using the machine-learning framework NNAlign_MA.³³ Briefly, the

framework allows both for the integration of BA and MS HLA eluted ligand data in the training procedure and for the motif deconvolution of HLA eluted ligand data from cell lines expressing multiple HLA molecules. This latter feature allows the framework to greatly expand the volume of the accessible training data and the HLA coverage of the developed predictor.²⁹

To assess the performance of the developed predictor, it was benchmarked against NetMHCIIpan-3.2²⁴ and MixMHC2pred¹² on several independent data sets including a set of four protein therapeutics to demonstrate to what degree the different methods can accurately predict the experimentally observed MS hotspot regions while maintaining a low false-positive rate.

MATERIALS AND METHODS

Data sets

Three different data types were used in this study: peptide BA measurements; single-allele (SA) MS MHC eluted ligands (EL); and multi-allele (MA) MS MHC EL. Here, SA EL refers to data obtained from cell lines or experimental settings where only one single MHC molecule is expressed on the cell, and MA EL refers to data obtained from cells expressing multiple MHCs.

BA and SA EL data were obtained from Reynisson et al.²⁹. The BA data included HLA-DR, HLA-DP and HLA-DQ peptides within a length range of 13–21 amino acids with IC50 (in nM) transformed to a value within 0–1 range by $1 - \log(\text{IC50}) / \log(50\,000)$. The SA EL data included HLA-DR, HLA-DP and HLA-DQ ligands. MA EL data were obtained from two sources: partly collected in Reynisson et al.²⁹ and partly generated in-house. The Reynisson et al.²⁹ data set included HLA-DR, HLA-DP and HLA-DQ ligands. Note that the Reynisson data set is to a very high degree obtained from the publication and data underlying the MixMHC2pred method.¹² The in-house MA EL data were generated using a standard MAPPs MS-LC/LC set-up (for details, refer to Supporting information). In short, monocytes were isolated from PBMC obtained from healthy volunteer donors and differentiated in vitro into immature DCs. DCs were pulsed with the benchmarked protein of interest and matured overnight with LPS. Next, DCs were lysed, and HLA-DR molecules were recovered using L243 mAb immunoprecipitation. Peptides were eluted from the HLA-DR molecules and sequenced using LC-MS/MS. Raw MS spectra were analysed searching against the human reference proteome using a false discovery rate of (FDR) 5% for ligands derived from therapeutic proteins and FDR of 1 otherwise. This in-house data set is available in Tables S1 and S2.

The EL data set was filtered for ligands within length range of 13–21 amino acids. Donor-specific MA data sets were merged and duplicates removed. EL data are

positive by nature. For proper neural network training, it is necessary to train on both positive and negative data examples. To resolve this, artificial negative data were generated for each cell line/donor data set as described in,²⁷ by randomly sampling peptides from the human proteome, uniformly for each length between 13 and 21 amino acids for five times the amount of the most represented length of the positive ligands. Peptide context information included three downstream N-terminal amino acids, three upstream C-terminal amino acids of the source protein and three N-terminal and three C-terminal amino acids of the ligand (as described in Ref. 27). A summary of the different data sets is given in Table 1.

Three test protein data sets were defined for model development and benchmarking (for details, refer to Supporting information). Test set 1 and test set 2 consisted of self-proteins and were used to tune classification threshold (test set 1, available in Table S3) and perform benchmark evaluation (test set 2, available in Table S4). The third test constructed from four therapeutic proteins was used to showcase the predictive power of the developed model. In all cases, redundant training instances sharing a 9-mer common motif or more with one of the test set proteins were removed prior to training resulting in removal of 5.7% of the training data. The impact of the redundancy removal in the training data of each test set is detailed in Table S5. The high overlap between the test sets and the training data (40–80%) highlights the importance of performing this redundancy reduction step in order to use the three test sets for reliable performance assessment.

NNAlign_MA training

The model for prediction of HLA antigen presentation was trained as described earlier²⁹ using the NNAlign_MA machine-learning framework. To limit the effect of performance overestimation and model overfitting, the training data were partitioned with a common motif clustering algorithm (described in Ref. 34) with a 9-mer motif length corresponding to the MHC-II binding core. BA and EL training data were clustered simultaneously and then separated, resulting in five partitions for BA and five partitions for EL training data. In short, the NNAlign_MA framework is based on a three-layer neural network architecture with two output values (one for BA and one for EL likelihood) allowing integration of mixed data types.³³ Initially, NNAlign_MA is pretrained exclusively with fully annotated peptide data (BA and SA MS ligand data). The pretrained NNAlign_MA network is used to annotate the most likely HLA restriction of ligands in the MA EL based on the donor-typing information, casting the MA data into a SA format, allowing for training also on the MA data. The annotation step is performed in

Table 1. Training data without exclusion of redundancy to benchmark self-proteins and therapeutic proteins

Type	Pos.	Rand. neg.	Total	Donors	DR	DQ	DP
BA	106 429	–	106 429	–	35	27	9
SA EL	88 862	804 465	893 327	33	14	5	2
MA EL	269 506	2 678 445	2 947 951	93	33	52	21
MA EL in-house	372 586	3 456 360	3 828 946	40	29	0	0

every training iteration allowing the model to converge to an optimal annotation of HLA restriction while also learning the associated HLA-specific binding motif (for further details on the NNAlign_MA training, refer to the Supporting information).

MAPPs predictions and benchmark evaluation

To benchmark the performance for a given predictive method, a given benchmark protein was *in silico*-digested into all possible peptides in the length range of 13–21 amino acids. Next, prediction score was computed for all digested peptides for each HLA of the given donor. In case a donor lacked HLA-DRB3, HLA-DRB4 and HLA-DRB5 typing, DRB3*02:02, DRB3*03:01 and DRB5*01:01 typing was imputed when possible based on their known linkage disequilibria to DRB1*14:01, DRB1*13:02 and DRB1*15:01, respectively. The prediction scores were transformed to percentile ranks by aligning the prediction score to prediction scores for 100 000 random peptides uniformly distributed within length range from 13 to 21 for the respective HLA-DR alleles, and afterwards, the peptides were classified as binders or non-binders depending on whether the minimal rank across all tested HLAs was lower or greater than a given threshold. This threshold was defined by assessing the performance on test set 1. The performance was estimated from the Matthews correlation coefficient (MCC) between the experimental MAPPs-identified ligands and MAPPs-predicted binders for each donor in the cohort. To account for situations where a predicted binder is not identical to but shares a binding core with a MAPPs ligand, an MCC-core performance value was also calculated per donor. Here, all predicted binders with predicted binding cores shared with one or more of the donor's target ligands were counted as true-positive predictions. For both MCC and MCC-core, the median donor MCC is reported for each benchmark protein from the donor MCCs.

For visualization, MAPPs profiles were generated for both target MAPPs and predicted MAPPs ligands. MAPPs profiles were produced by mapping and stacking ligands to the benchmark protein sequence obtaining a count for each position. The position counts were normalized to a value between 0 and 1 by dividing with the maximal position count.

NetMHCIIpan-3.2²⁴ and MixMHC2pred (v1.1.3)¹² were included to establish a baseline and a state-of-the-art benchmark, respectively. In both cases, a default setting of the predictor together with a tuned percentile rank classification threshold was used to generate MAPPs predictions. The HLA-DR allele coverage of MixMHC2pred is limited to 24 HLA-DR alleles, not covering all the HLA-DR alleles in the donor cohorts of the benchmark proteins.

Statistical tests

Statistical significance was evaluated using one-sided binomial tests excluding ties if not otherwise specified.

RESULTS

Here, we developed a method for prediction of the outcome of MAPPs assays from a large set of novel MS MHC eluted ligand data combined with data sets of BA measurements and MS MHC EL data sets from previous publications.²⁹ The method is based on an advanced prediction model for peptide MHC antigen presentation, and its performance was benchmarked against other state-of-the-art methods and showcased on a panel of therapeutic proteins (Figure 1).

Training of NNAlign_MA

The essential component of the proposed prediction framework is a method for accurate prediction of MHC peptide antigen presentation. This method was developed using the NNAlign_MA machine-learning modelling framework to train a predictor on a large MHC peptidome dataset. NNAlign_MA utilizes a semi-supervised learning structure to initially leverage information from SA EL and BA data to annotate MA EL data, and consecutively iterates this process to allow for a complete and accurate specificity deconvolution.³³ Using this approach, Reynisson et al.²⁹ demonstrated that NNAlign_MA was able to expand allelic coverage and boost predictive performance for HLA-II alleles not present in the SA EL data supporting the semi-supervised learning structure of NNAlign_MA.

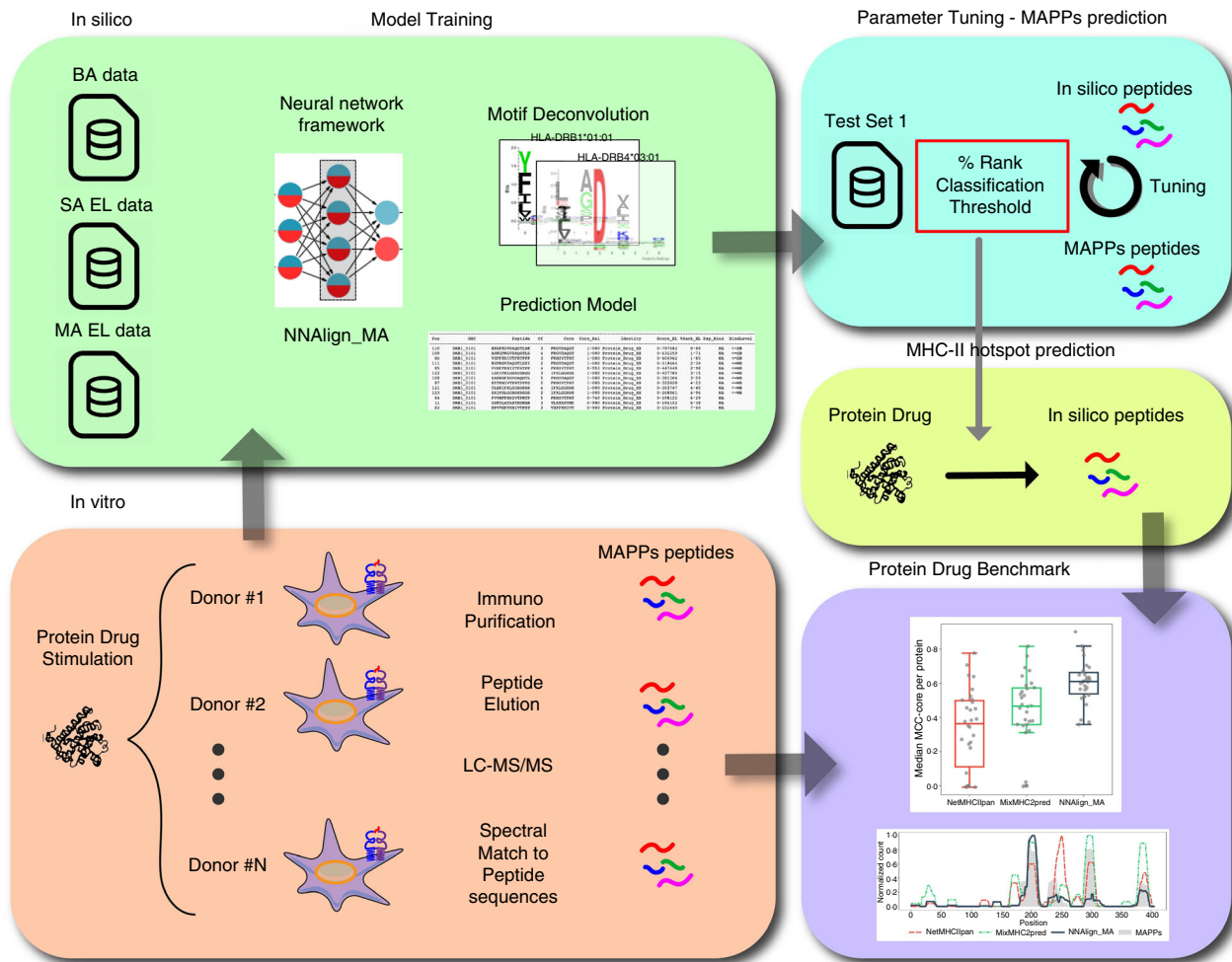


Figure 1. Overview of modelling and benchmark pipeline. Upper panel: An in silico model was developed, trained on MHC-II peptide binding affinity data (BA), and SA and multi-allele (MA) mass spectrometry EL. The NNAlign_MA model performs HLA motif deconvolution and predicts peptide HLA antigen presentation likelihood. The model is applied to perform in silico identification of HLA ligands for a protein of interest (test set 1), where a % rank classification threshold is defined. With the given threshold, peptides derived from proteins and protein therapeutics are predicted to bind to a set of HLA-II alleles (in silico peptides). Lower panel: MAPPs experiments are performed for a protein of interest in a panel of HLA-typed donors, and donor-specific MS ligand data sets are identified. Identified peptides from self-proteins constitute MA EL data sets, while peptides derived from protein and protein therapeutics (MAPPs peptides) are used for benchmarking, and ligand hotspot profiles are generated and compared with in silico profiles

NNAlign_MA was trained for integrating an in-house collection of EL data consisting of close to 375 000 HLA-DR ligands obtained from 221 samples covering 40 distinct HLA-DR molecules, combined with ~100 000 peptide BA measurements, close to 90 000 SA, and ~270 000 MA data points from a Reynisson et al.²⁹ In total, this data set consists of close to 840 000 MHC-peptide interaction data points covering a potential of 42 distinct HLA-DR molecules. Three independent evaluation data sets were constructed and used for model optimization (one data set) and performance evaluation (two data sets) (for details on the different datasets, refer to Materials and methods). The MHC-II antigen presentation prediction model was trained using a fivefold cross-

validation scheme (for details of the hyperparameters, refer to Materials and methods). The cross-validation performance for EL data sets was a median area under the ROC curve (AUC) of 0.95, demonstrating the high ability of NNAlign_MA to classify MHC-II ligands (Table S6).

The deconvolution of HLA-DR restriction for MA EL data sets was evaluated by producing sequence logo representations of individual HLA-DR binding motifs as described in the Supporting information. Figure 2 displays sequence logo representations of the binding motifs from donors 6, 7 and 37. These donors were selected as representative examples, the complete deconvolution is shown in Figure S1. From this figure, it is apparent that NNAlign_MA performs sharp binding

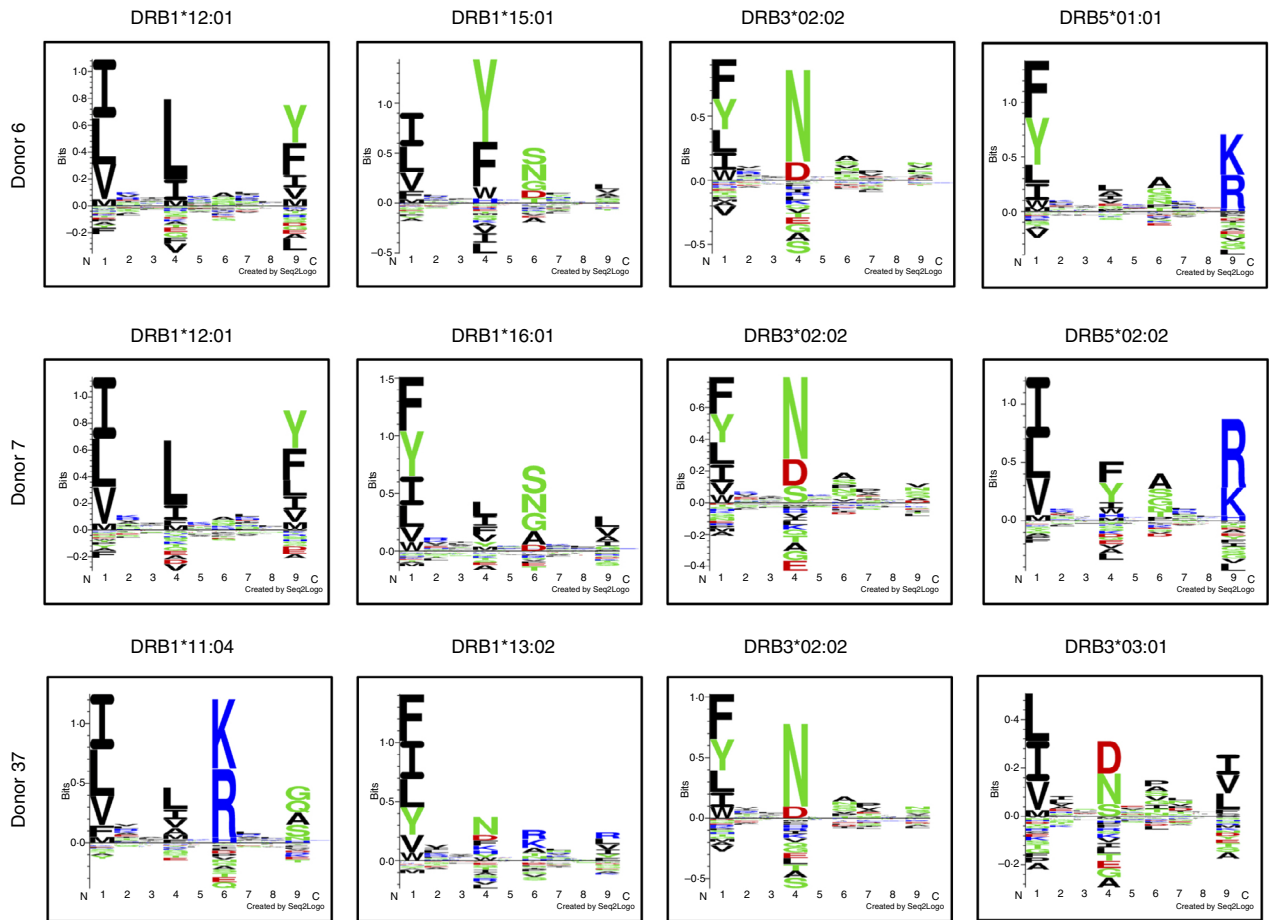


Figure 2. Deconvolution of HLA restriction by NNAlign_MA for 3 representative donors. Sequence logo representations of binding motifs obtained by NNAlign_MA for donors 6, 7 and 37 from the in-house MA EL data. Motif deconvolution was performed for each donor MA EL data using NNAlign_MA and cross-validation. To generate each allele logo, all 9 amino acid-core sequences assigned by NNAlign_MA to each of the HLA-DR molecules from a given donor MS data set were pooled, and the binding motif was plotted using Seq2Logo.⁴⁰ For additional information, refer to Supporting information. A %rank threshold of 20% was used to discard peptide contaminants that likely do not bind HLA-DR. Donors expressing the same HLA-DR alleles show logo representations with shared consistency among the anchor positions. The full deconvolution of the 40 donors is included in Figure S1

motif deconvolution across donors, with clearly defined anchor positions and consistent motifs for alleles shared between multiple donors (exemplified by the HLA-DR alleles shared between two or more of the donors included in the figure). This consistency in binding motifs of HLA-DR alleles expressed by multiple donors reflects both the high quality of the EL data included in this study and the ability of NNAlign_MA to accurately infer the motifs from MA EL data. The consistency among HLA-DR motifs across donors sharing the same alleles was quantified in a pairwise position-specific scoring matrix (PSSM) correlation analysis, resulting in a median deconvolution consistency score (refer to Materials and methods for details on this measure) of 0.88 (Figure S2). An estimation of the predicted positive value (PPV) evaluating the ‘cleanness’ of donor-specific HLA-DR allele predictions (for details, refer to

Supporting information) resulted in a median PPV of 0.786 (Table S7).

NNAlign_MA outperforms existing MHC-II predictors in MHC-II antigen presentation prediction

A set of 20 self-proteins was used for tuning the percentile rank classification threshold (test set 1) and a set of 30 self-proteins for benchmarking to NetMHCIIpan²⁴ and MixMHC2pred¹² (test set 2). The self-proteins were randomly selected from a pool of self-proteins with high ligand coverage (for details on the two test data sets, refer to Materials and methods and Tables S3,S4).

The proteins from test set 1 were in silico-digested into all overlapping peptides of length range from 13 to 21 amino acids, and predictions were generated for all donor HLAs with NNAlign_MA, NetMHCIIpan and

MixMHC2pred. Next, the performance of the different methods was evaluated using MCC and MCC-core measures (details in Materials and methods) capturing the overlap between the predicted and experimentally identified HLA ligands. In short, these measures are defined from classifying predicted binders and non-binders into true or false prediction depending on the overlap to the observed experimental ligands. For the MCC measure, the overlap was requested to be identical to the observed ligand in order for a positive prediction to be true, and for the MCC-core, a more relaxed overlap was requested, so that only the predicted binding core should be contained within the experimental ligand to be considered as a true prediction. This MCC-core measure hence allows slight mismatches between the predicted and the measured ligands to be counted as positive predictions. Classification of predicted binders/non-binders was defined according to percentile rank scores (detailed in Materials and methods). Optimal percentile rank classification thresholds were selected based on MCC and MCC-core performance values for test set 1, resulting in values of 1% rank for NNAlign_MA and MixMHC2pred, and 7% rank for NetMHCIIpan (Figure S3).

Next, the different methods were benchmarked on the proteins in test set 2. Here, NNAlign_MA was found to significantly outperform both other methods with a median MCC of 0.21 compared with 0.03 (NetMHCIIpan-3.2, P -value <0.001) and 0.14 (MixMHC2pred, P -value <0.001) (Figure 3A). The same pattern was observed when calculating the MCC-core performance. As expected, we observed an increased performance for all three methods with a median performance of 0.61 for NNAlign_MA, 0.36 for NetMHCIIpan-3.2 and 0.47 for MixMHC2pred. Also here, the performance of NNAlign_MA was significantly superior to the two other methods (P -value <0.001 and P -value = 0.01; Figure 3B). Note that the performance values of NetMHCIIpan-3.2 and in particular MixMHC2pred most likely are overestimated due to the overlap between the training data of MixMHC2pred and the test set 2 evaluation data (Table S5).

To visualize the overlap between the predicted and MS-identified HLA-DR-enriched hotspots, ligand profiles were generated for the different methods. Figure 3C showcases one such profile for the self-protein P01911.2. Here, NNAlign_MA shared a higher agreement with the experimental MS EL data compared with both MixMHC2pred and NetMHCIIpan-3.2, overall predicting few false-positive peaks. For instance, NetMHCIIpan and MixMHC2pred both predict an additional strong peak towards the N-terminal that does not align with the MAPPs data. These observations are also reflected in the MCC performance values for P01911.2 of the different methods (MCC = 0.24, 0.10 and -0.01; MCC-core = 0.53, 0.36 and 0.06, for NNAlign_MA, MixMHC2pred and NetMHCIIpan, respectively). Overall,

these analyses demonstrate the high power of the NNAlign_MA method to predict MS-identified HLA-DR-enriched hotspots.

Predicting the outcome of MAPPs experiment in therapeutic proteins

To further test the developed tool, we evaluated its power to predict the outcome of a series of MAPPs experiments of therapeutic proteins and compared the performance with that of NetMHCIIpan-3.2 and MixMHC2pred. Therapeutic proteins included were vatreptacog alfa (coagulation factor VII analogue), coagulation factor X analogue (referred to as factor X), liraglutide peptide backbone (GLP-1 agonist) and infliximab (data obtained from Barra et al.³² and Karle et al.³⁵).

This benchmark confirmed the overall superior performance of NNAlign_MA and MixMHC2pred compared to NetMHCIIpan-3.2 with average MCC and MCC-core values over the four therapeutic proteins (five protein chains) of MCC 0.26, 0.20 and 0.10 and MCC-core 0.72, 0.67 and 0.53 for NNAlign_MA, MixMHC2pred and NetMHCIIpan-3.2, respectively (Table S8). Note that also here, NetMHCIIpan-3.2 and in particular MixMHC2pred are expected to show an overestimated predictive power due to the overlap between their respective training data and the protein drug test data (Table S5). Figure 4 showcases the results for vatreptacog alfa. MAPPs data for vatreptacog alfa were generated from a cohort of 38 donors, where a unique set of 929 donor-specific ligands were identified. The donor cohort covered 29 HLA-DR alleles. Performance was evaluated as described for test set 2. In terms of the strict MCC evaluation, NNAlign_MA and MixMHC2pred exhibited a comparable performance with donor median MCC scores of 0.34 and 0.32, respectively. Both predictors demonstrated an improved performance compared to NetMHCIIpan with a donor median MCC of 0.11 (P -value <0.001) (Figure 4A). Evaluating the predictors in terms of MCC-core, NNAlign_MA and MixMHC2pred again demonstrated a comparable performance with donor median MCC binding core scores of 0.76 and 0.71 (P -value = 0.44), respectively. Moreover, both predictors showed an improvement compared with NetMHCIIpan (donor median MCC-core performance of 0.53, P -value <0.001) (Figure 4B). Visually assessing MAPPs profiles (Figure 4C) exhibited that all predictors captured the four experimental HLA-DR-enriched hotspot regions. However, both NetMHCIIpan and MixMHC2pred predicted an additional pronounced false-positive peak at position ~175 that NNAlign_MA did not. Results for factor X, liraglutide peptide backbone, and infliximab HC and LC are provided in Figures S4–S7. Also here, NNAlign_MA demonstrated a higher median MCC and MCC-core performance across donors compared with NetMHCIIpan for all protein drugs except for liraglutide

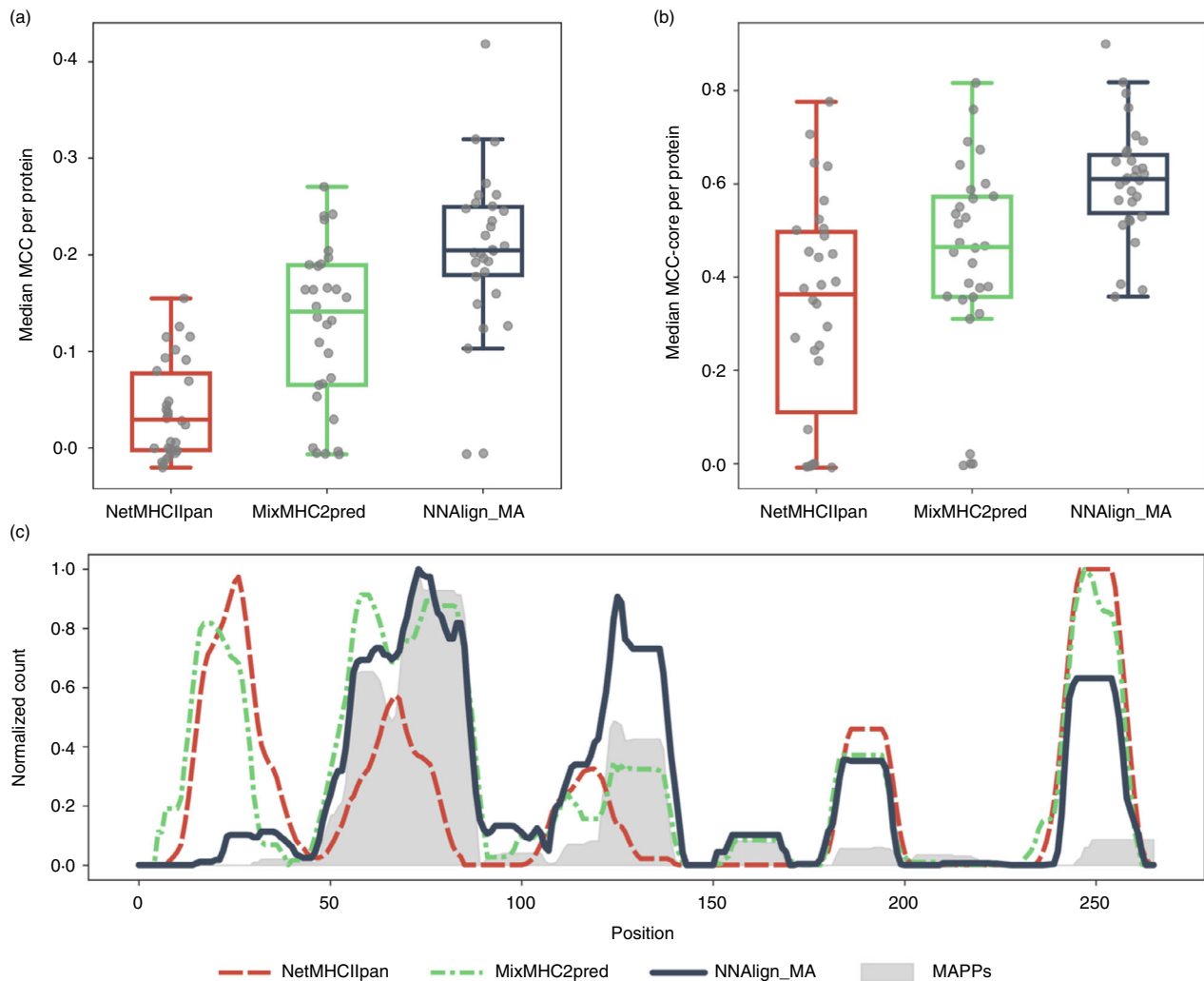


Figure 3. Benchmark of MAPPs prediction for test set 2. (A) Each data point shows the median of the donor-specific MCC performance of each method per benchmark protein. Median performances: NNAlign = 0.21; NetMHCIIpan-3.2 = 0.3; and MixMHC2 = 0.14. NNAlign_MA performance was significantly higher compared with the two other methods (P -value <0.001). (B) Each data point shows the median of the donor-specific MCC-core performance of each method per benchmark protein. Median performances: NNAlign_MA = 0.61; NetMHCIIpan-3.2 = 0.36; MixMHC2pred = 0.47. NNAlign_MA performance was significantly higher (P -value <0.001 and P -value = 0.01, respectively). (C) Predicted and MS HLA-DR ligands are mapped to generate ligand profiles for self-protein P01911.2 as described in the text

backbone (Figure S5). This increase is significant for infliximab HC and LC ($n = 13$, Figure S6 and S7), but not for factor X (P -value = 0.063, $n = 7$, Figure S4). NNAlign_MA likewise showed a generally higher MCC and MCC-core performances compared with MixMHC2pred. This performance gain was, however, only significant for MCC (except for factor X, P -value = 0.5; Figure S5). Overall, these additional test case examples support the conclusions from vatreptacog alfa and consolidate the high performance of the developed in silico method to accurately predict the experimentally identified HLA antigen presentation hotspot regions at a reduced false-positive rate compared with the other methods included in the benchmark.

Generic Prediction of HLA presentation hotspot regions

Finally, we tested the developed predictor in a more realistic setting where the objective was to predict HLA presentation hotspots rather than the detailed outcome of individual MAPPs experiments. For that purpose, the NNAlign_MA-trained predictor was applied to construct HLA antigen presentation profiles for each of the protein sequences in test set 3 using a reference set of prevalent HLA-DR molecules in the population (for details on this list, refer to Table S9) rather than the individual HLAs of MAPPs-tested donors as done previously. Using each of the HLA-DR molecules in said reference set, all possible

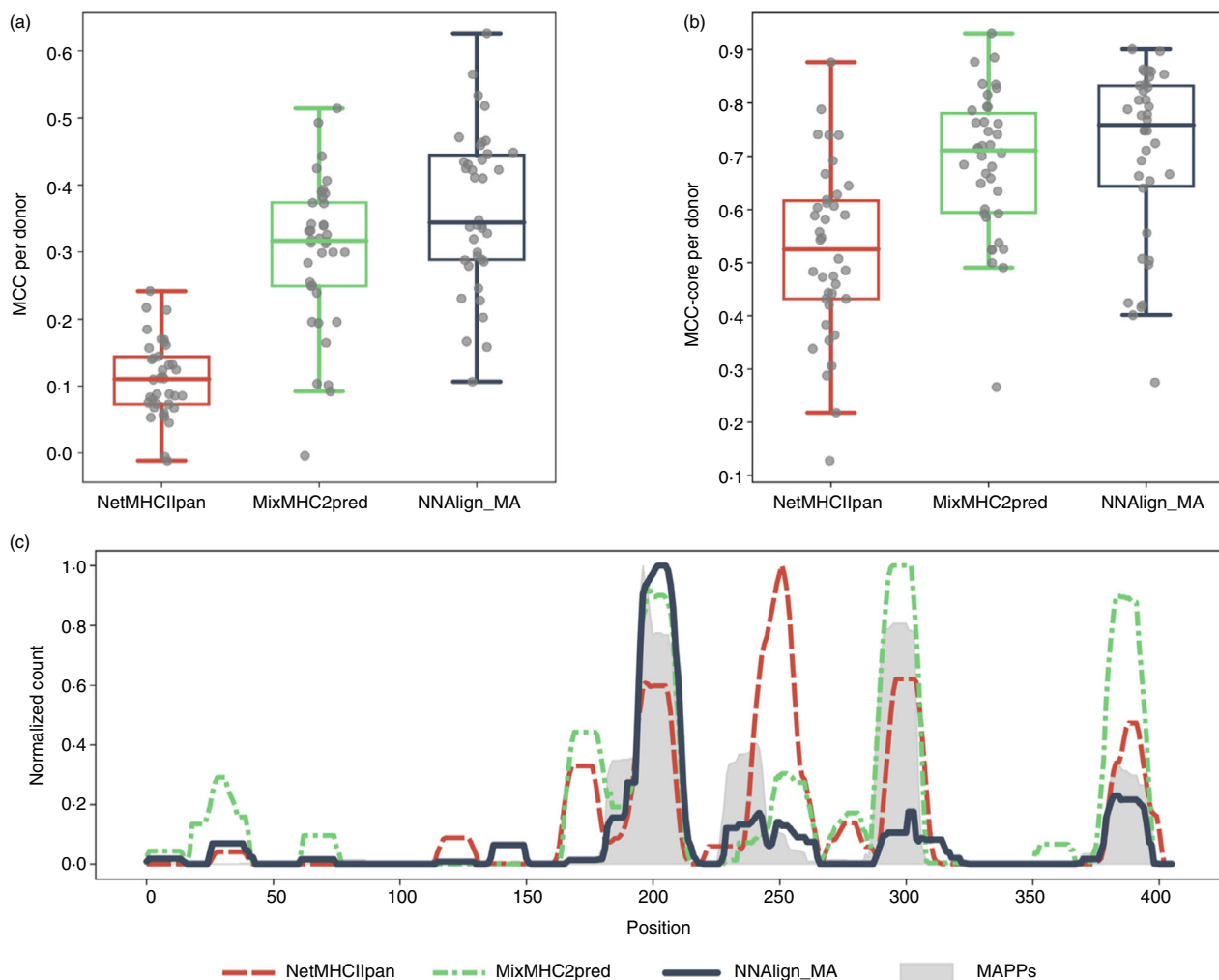


Figure 4. Performance evaluation of NNAlign_MA, MixMHC2pred and NetMHCIIpan-3.2 on MAPPs data for vatreptacog alfa. (A) Each data point shows the donor-specific MCC performance of each method. Median performances: NNAlign = 0.34; NetMHCIIpan-3.2 = 0.11; and MixMHC2 = 0.32. NNAlign_MA performance was significantly higher than NetMHCIIpan (P -value <0.001). (B) Each data point shows the donor-specific MCC-core performance of each method. Median performances: NNAlign = 0.76; NetMHCIIpan-3.2 = 0.53; and MixMHC2 = 0.71. NNAlign_MA performance was significantly higher than NetMHCIIpan only (P -value <0.001). (C) MAPPs and predicted ligand profiles for the different methods for vatreptacog alfa

peptides derived from the protein drugs in test set 3 were predicted and the binding threshold value defined earlier was used to classify the positive predicted ligands. Next, *in silico* profiles were constructed by mapping the pool of predicted ligands back to the source protein sequence and calculating the total number of ligands overlapping each position (normalizing the max count to 1). Figure 5 shows a comparison of the experimental and *in silico* MAPPs-derived profile for vatreptacog alfa (details for the other proteins are included in Figure S8 and Table S8). Investigating these figures demonstrates that this ‘generic’ hotspot identification approach shares a very high predictive overlap to the donor-specific approach described hereto. The mean and median of Spearman’s correlation between the experimental and *in*

silico-predicted HLA ligand profiles were 0.612/0.621 and 0.632/0.627 for the ‘donor-specific’ and ‘generic’ hotspot identification approaches. These results confirm that also this approach to a very high degree and at a low rate of false-positive predictions captures all the experimentally identified HLA ligand hotspot regions.

As a further benchmark evaluation, we next compared NNAlign_MA with two earlier methods for prediction of MHC class II antigen presentation developed by us. The first method (termed Barra) was described in Barra et al.³² and consists of a model trained on a limited set of MS EL data and benchmarked on MAPPs ligands and T-cell epitope responses in infliximab. The second model is the most recent version of NetMHCIIpan (version 4.0).^{29,36} This model was trained in a manner similar to

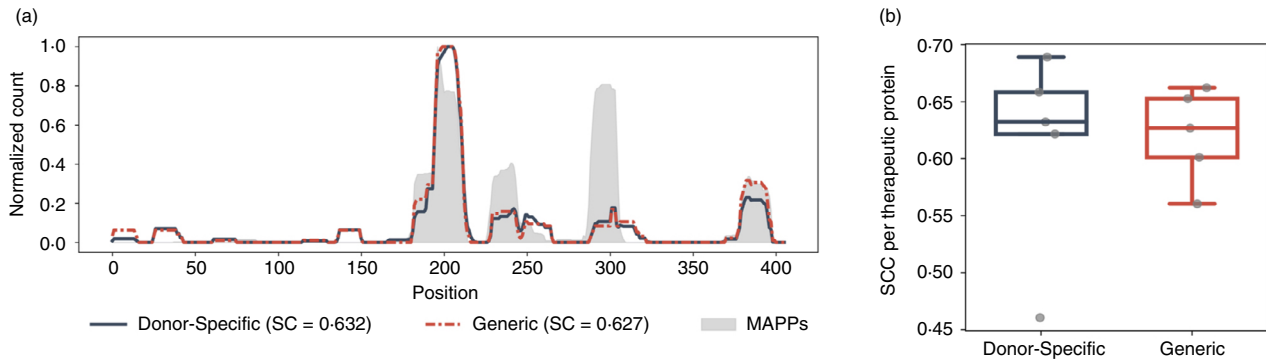


Figure 5. Comparison of MAPPs-identified and MAPPs-predicted HLA ligand hotspot regions for vatreptacog alfa. (A) ‘Donor-specific’ predicted profile was generated using the donor-specific prediction approach described for Figures 3 and 4. The ‘generic’ predicted profile was constructed using a general non-donor-specific list of world population prevalent HLA-DR molecules (Table S9). All profiles were constructed from the MAPPs or predicted HLA-DR ligands as described in the text. (B) Spearman’s correlation coefficients (SCCs) between the experimental MAPPs profiles and the ‘donor-specific’ and ‘generic’ predicted profiles for the five therapeutic proteins in test set 3

what has been described here. Only did the HLA-DR MA data set not include the novel MS data generated here and more importantly was the overlap to the three test data sets used in the current study maintained in training data. These two methods were used to predict the MAPPs profiles for the 4 therapeutic proteins using the ‘generic prediction of HLA presentation hotspot regions’ implementation. The results of the benchmark are shown in Figure S9. As expected, due to the large overlap between the infliximab MAPPs data and the HLA cohort used to develop the method by Barra et al., this method performs better for infliximab and substantially worse for the other protein drugs. On the other hand, NNAlign_MA and NetMHCIIpan-4.0 show an overall comparable performance. In conclusion, these analyses confirm the overall high power of prediction methods trained on MS EL data to predict the outcome of MAPPs experiments and likewise underline the importance of developing these prediction methods in the context of large and HLA diverse data sets.

DISCUSSION

MHC-associated peptide proteomics assays are frequently used during biopharmaceutical development for assessing and managing preclinical therapeutic protein immunogenicity, supporting deimmunization efforts and lead candidate selection. As it is a labour-intensive, low-throughput and costly assay, the industry would benefit greatly from an *in silico* tool that could predict presented MHC-II peptides. Barra et al.³² recently demonstrated that *in silico* models trained on a small MS MHC-II elution ligand data set could predict the outcome of MAPPs assays on therapeutic proteins. However, this study was limited by a small donor cohort and a very limited set of investigated proteins, which complicated the performance

evaluation. Here, we have addressed these shortcomings and expanded the proof of concept to cover a broader set of therapeutic proteins each characterized with MAPPs data for a larger donor cohort. This allowed us to define, in a robust manner, the optimal settings translating peptide MHC antigen predictions into assessment hotspot regions for potential T-cell epitopes. Altogether, these results showed an increased capability of the trained *in silico* model, termed NNAlign_MA, to predict MAPPs experimental data on a donor level.

The developed NNAlign_MA method was benchmarked against NetMHCIIpan-3.2 and MixMHC2pred. NetMHCIIpan-3.2 represents a traditional MHC-II predictor trained exclusively with peptide BA data that do not contain information about antigen processing and presentation. In contrast, the recently published MixMHC2pred is trained exclusively with MS ligand data that should capture at least in part this information. Ideally, the benchmark analysis should have been performed with MAPPs targets for an extensive set of therapeutic proteins. However, as only a highly limited set of MAPPs targets with HLA-DR donor-typing data were available, the benchmark was extended to include a set of self-proteins with mapped MS MHC-II ligands. In this benchmark, NNAlign_MA demonstrated a significant improvement in classifying MAPPs target ligands and binding cores compared with both NetMHCIIpan-3.2 and MixMHC2pred. Also, MixMHC2pred significantly outperformed NetMHCIIpan-3.2. This result confirmed earlier findings that methods trained on MS HLA EL perform better than methods trained on BA data for prediction of EL.^{27–29} Likewise, the result aligned with the earlier findings demonstrating a superior performance of NNAlign_MA compared with MixMHC2pred^{29,36} potentially driven by the improved motif deconvolution power of NNAlign_MA imposed by its pan-specific prediction power enabling leveraging of

information between data sets and expansion of predictive power to also cover allele characterized by limited (or even no) ligand data.²⁹

Focusing on the protein drug benchmark, a similar result was found with NNAlign_MA and MixMHC2pred achieving the highest performance both outperforming NetMHCIIpan-3.2. Investigating the overlap between the observed and predicted HLA-DR hotspot regions revealed that all methods in the vast majority of cases were capable of correctly predicting the observed hotspot regions, but that NetMHCIIpan-3.2 and to a less extent MixMHC2pred predicted a larger proportion of false-positive hotspots compared with NNAlign_MA.

Investigating further the performance of NNAlign_MA on the test set of therapeutic proteins, it was found that the overlap between the predicted and measured HLA ligands was augmented as the number of measured MAPPs ligands identified per protein sequence was increased. That is, a strong correlation (Spearman's correlation of 0.90, data from Table S8) was observed between the number of ligands per residue in the investigated protein and performance of NNAlign_MA estimated in terms of the MCC-core. This observation strongly suggests that at least part of the additional predicted hotspot regions could be true and that these were missed due to the limited size of donor cohort analysed and the general limited sensitivity of a single MAPPs experiment, as has been suggested earlier.³²

In line with earlier work,^{29,32} it was likewise observed that the predictive accuracy of the models was increased when including the full HLA-DR information for each donor rather than limiting this to HLA-DRB1. This is an important observation arguing that full and high-resolution HLA typing is essential for accurate interpretation and prediction of HLA antigen presentation in general and MAPPs experiments in particular. Here, we have only focused on HLA-DR motif deconvolution and antigen presentation; however, earlier work has demonstrated that complete high-resolution HLA-II typing including DQ and DP is beneficial also for MS HLA eluted data sets generated using HLA-DR immunoprecipitation.^{12,29} Future work will tell whether such higher resolution HLA data can likewise improve the power for prediction of MAPPs experiments.

The work here has been limited to the analysis of antigen presentation for prevalent HLA-DR alleles (as defined by the donor or world population). The developed machine-learning framework and in silico prediction models are, however, not limited to these HLAs but accurately cover a much wider range of HLA-DR, HLA-DP and HLA-DQ molecules.²⁹ This wide HLA coverage enables accurate predictions also for HLA molecules whose specificity is characterized with limited or even no experimental data.²⁹ That

makes the NNAlign_MA tool an ideal complement to experiment immunogenicity assessment in the context of, for instance, rare HLA alleles challenged by donor scarcity.

Several other methods have recently been published claiming improved predictive power for MHC class II antigen presentation including NeonMHC2,¹¹ MARIA³⁰ and EDGE.³¹ While these methods all proclaim very high prediction power, they all suffer from high restrictions in their public availability. For instance, NeonMHC2 only allows max 20 predictions per day for any given user, and MARIA allows only 5000 predictions per submission, making it impractical to include these tools in a large-scale benchmark analysis. Given these constraints, we do not here make any claims of superiority of the NNAlign_MA model to these methods. Future work is needed to perform such benchmark evaluations.

It is important to underline that the current work and the developed in silico models are limited to the prediction of HLA antigen presentation. Neither the MAPPs experiment nor the in silico predictions make any assessment on the immunogenicity of the suggested HLA ligands and hotspot regions. Liraglutide backbone is a good example of that – as even though peptides are presented from liraglutide backbone, the frequency and levels of antidrug antibodies towards the current therapeutic version of liraglutide are low and do not impact glycaemic efficacy or safety.³⁷ However, it is possible that these presented peptides could initiate a higher frequency of antidrug antibodies in other versions of liraglutide with different quality attributes (such as impurities with additions and deletions in the primary sequence). Clinical trials would then be necessary to assess the immunogenicity risk of alternative versions of liraglutide.

With this reservation, we believe the results presented here strongly suggest that in silico prediction methods can and should serve as a guide, complementing MAPPs assays, for the general assessment of the immunogenicity of therapeutic proteins and the identification of HLA-II antigen presentation hotspot regions relevant for protein deimmunization.^{16,17,38,39}

ACKNOWLEDGEMENTS

Many thanks to Lise Wegner Thørner, Morten Bagge Hansen, Henrik Ullum and staff at Blodbanken/GivBlod (RH) for recruiting and handling the donors used in the study. We thank Linda Yim Ting Lam (Novo Nordisk technician) for assistance performing MS experiments.

CONFLICT OF INTEREST

HSS, UL and KL are employees of Novo Nordisk A/S and have stocks or stock options in Novo Nordisk A/S.

DATA AVAILABILITY STATEMENT

The authors declare that the data supporting the findings of this study are available within the article and its Supporting information.

REFERENCES

- Sauna ZE, Lagassé D, Pedras-Vasconcelos J, Golding B, Rosenberg AS. Evaluating and mitigating the immunogenicity of therapeutic proteins. *Trends Biotechnol.* 2018;**36**(10):1068–84.
- Mahlangu JN, Weldingh KN, Lentz SR, Kaicker S, Karim FA, Matsushita T, et al. Changes in the amino acid sequence of the recombinant human factor VIIa analog, vatreptacog alfa, are associated with clinical immunogenicity. *J Thromb Haemost.* 2015;**13**(11):1989–98.
- Lamberth K, Reedt-Runge SL, Simon J, Klementyeva K, Pandey GS, Padkjer SB, et al. Post hoc assessment of the immunogenicity of bioengineered factor VIIa demonstrates the use of preclinical tools. *Sci Transl Med.* 2017;**9**(372):eaag1286.
- Wieczorek M, Abualrous ET, Sticht J, Alvaro-Benito M, Stolzenberg S, Noé F, et al. Major histocompatibility complex (MHC) Class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front Immunol.* 2017;**8**:292.
- Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol Cell Proteomics.* 2015;**14**(12):3105–17.
- Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc.* 2019;**14**(6):1687–707.
- Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol.* 2016;**41**:9–17.
- McMurtrey C, Trolle T, Sansom T, Remesh SG, Kaever T, Bardet W, et al. *Toxoplasma gondii* peptide ligands open the gate of the HLA class I binding groove. *Elife.* 2016;**29**:5.
- Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J Immunol.* 2016;**196**(4):1480–7.
- Gfeller D, Bassani-Sternberg M. Predicting antigen presentation-what could we learn from a million peptides? *Front Immunol.* 2018;**9**:1716.
- Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* 2019;**51**(4):766–79.e17.
- Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol.* 2019;**37**:1283–1286.
- 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.* 2018;**17**(1):559–67.
- Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun.* 2020;**11**(1):1293.
- Vizcaino JA, Kubiniok P, Kovalchik KA, Ma Q, Duquette JD, Mongrain I, et al. The human immunopeptidome project: a roadmap to predict and treat immune diseases. *Mol Cell Proteomics.* 2020;**19**(1):31–49.
- Quarmany V, Phung QT, Lill JR. MAPPs for the identification of immunogenic hotspots of biotherapeutics; an overview of the technology and its application to the biopharmaceutical arena. *Expert Rev Proteomics.* 2018;**15**(9):733–48.
- Karle AC. Applying MAPPs assays to assess drug immunogenicity. *Front Immunol.* 2020;**11**:698.
- Sathish JG, Sethu S, Bielsky M-C, de Haan L, French NS, Govindappa K, et al. Challenges and approaches for the development of safer immunomodulatory biologics. *Nat Rev Drug Discov.* 2013;**12**(4):306–24.
- Schellekens H. Bioequivalence and the immunogenicity of biopharmaceuticals. *Nat Rev Drug Discov.* 2002;**1**(6):457–62.
- Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology* 2010;**130**(3):319–28.
- Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol.* 2008;**4**(4):e1000048.
- Rosenberg AS, Sauna ZE. Immunogenicity assessment during the development of protein therapeutics. *J Pharm Pharmacol.* 2018;**70**(5):584–94.
- Chaves FA, Lee AH, Nayak JL, Richards KA, Sant AJ. The utility and limitations of current web-available algorithms to predict peptides recognized by CD4 T cells in response to pathogen infection. *J Immunol.* 2012;**188**(9):4235–48.
- Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 2018;**154**(3):394–406.
- Birrueta G, Tripple V, Pham J, Manohar M, James EA, Kwok WW, et al. Peanut-specific T cell responses in patients with different clinical reactivity. *PLoS One* 2018;**13**(10):e0204620.
- Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, Nielsen M. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* 2018;**34**(9):1522–8.
- Barra C, Alvarez B, Paul S, Sette A, Peters B, Andreatta M, et al. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* 2018;**10**(1):84.
- Garde C, Ramarathinam SH, Jappe EC, Nielsen M, Kringelum JV, Trolle T, et al. Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics* 2019;**71**(7):445–54.
- Reynisson B, Barra C, Kaabinejadian S, Hildebrand WH, Peters B, Nielsen M. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res.* 2020;**19**(6):2304–15.
- Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol.* 2019;**37**:1332–43.
- Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, et al. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol* 2019; **37**:55–63.
- Barra C, Ackaert C, Reynisson B, Schockaert J, Jessen LE, Watson M, et al. Immunopeptidomic data integration to artificial neural networks enhances protein-drug immunogenicity prediction. *Front Immunol.* 2020;**23**(11):1304.
- Alvarez B, Reynisson B, Barra C, Buus S, Ternette N, Connelley T, et al. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T cell epitope predictions. *Mol Cell Proteomics.* 2019;**18**(12):2459–2477.
- Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 2007;**4**(8):238.
- Karle A, Spindeldreher S, Secukinumab KF. A novel anti-IL-17A antibody, shows low immunogenicity potential in human in vitro assays comparable to other marketed biotherapeutics with low clinical immunogenicity. *mAbs.* 2016;**8**(3):536–50.
- Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 2020;**48**(W1):W449–W454.
- Buse JB, Garber A, Rosenstock J, Schmidt WE, Brett JH, Videbæk N, et al. Liraglutide treatment is associated with a low frequency and magnitude of antibody formation with no apparent impact on glycemic response or increased frequency of adverse events: results from the Liraglutide Effect and Action in Diabetes (LEAD) trials. *J Clin Endocrinol Metab.* 2011;**96**(6):1695–702.
- Schubert B, Schärf C, Dönnies P, Hopf T, Marks D, Kohlbacher O. Population-specific design of de-immunized protein biotherapeutics. *PLoS Comput Biol.* 2018;**14**(3):e1005983.
- Dhanda SK, Grifoni A, Pham J, Vaughan K, Sidney J, Peters B, et al. Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity. *Immunology* 2018;**153**(1):118–32.
- Thomsen MCF, 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* 2012; **40**(W1):W281–W287.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Supplementary Material. Materials and methods.