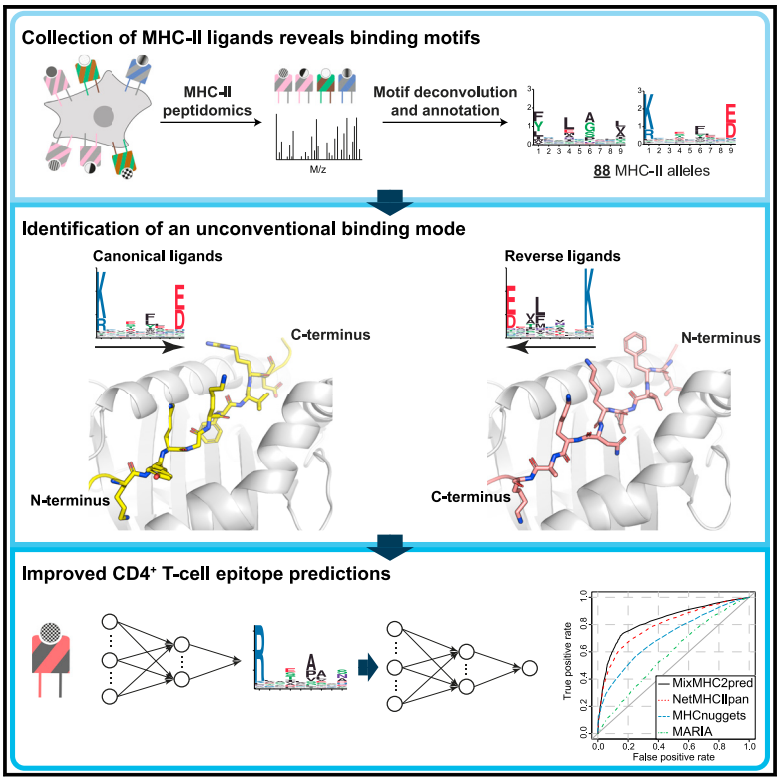
Article

Immunity

Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes

Graphical abstract



Highlights

- Analysis of MHC-II peptidomics data identifies >600,000 MHC-II ligands
- Motif deconvolution determines high-resolution binding motifs for 88 MHC-II alleles
- Structural analysis uncovers alternative binding modes of MHC-II ligands
- MixMHC2pred improves predictions of CD4$^+$ T cell epitopes

Authors

Julien Racle, Philippe Guillaume, Julien Schmidt, ..., Michal Bassani-Sternberg, Alexandre Harari, David Gfeller

Correspondence

julien.racle@unil.ch (J.R.), david.gfeller@unil.ch (D.G.)

In brief

CD4$^+$ T cells recognize peptides displayed on MHC-II molecules. Racle et al. curate >600,000 MHC-II ligands and derive high-resolution motifs for 88 MHC-II alleles. These motifs reveal a widespread reverse-binding mode for HLA-DP ligands and improve predictions of CD4$^+$ T cell epitopes with MixMHC2pred.

Racle et al., 2023, Immunity *56*, 1359–1375
June 13, 2023
https://doi.org/10.1016/j.immuni.2023.03.009

CellPress

Immunity

CellPress

Article

Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes

Julien Racle,[1,2,3,4,*] Philippe Guillaume,[1,4,5] Julien Schmidt,[1,4,5] Justine Michaux,[1,3,4,5,6] Amédé Larabi,[7] Kelvin Lau,[7] Marta A.S. Perez,[1,2,4] Giancarlo Croce,[1,2,3,4] Raphaël Genolet,[1,4,5] George Coukos,[1,3,4,5] Vincent Zoete,[1,2,4] Florence Pojer,[7] Michal Bassani-Sternberg,[1,3,4,5,6] Alexandre Harari,[1,3,4,5] and David Gfeller[1,2,3,4,8,*]

[1]Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[3]Agora Cancer Research Centre, Lausanne, Switzerland
[4]Swiss Cancer Center Leman (SCCL), Lausanne, Switzerland
[5]Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Lausanne, Switzerland
[6]Center of Experimental Therapeutics, Department of Oncology, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[7]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[8]Lead contact
*Correspondence: julien.racle@unil.ch (J.R.), david.gfeller@unil.ch (D.G.)
https://doi.org/10.1016/j.immuni.2023.03.009

SUMMARY

CD4$^+$ T cells orchestrate the adaptive immune response against pathogens and cancer by recognizing epitopes presented on class II major histocompatibility complex (MHC-II) molecules. The high polymorphism of MHC-II genes represents an important hurdle toward accurate prediction and identification of CD4$^+$ T cell epitopes. Here we collected and curated a dataset of 627,013 unique MHC-II ligands identified by mass spectrometry. This enabled us to precisely determine the binding motifs of 88 MHC-II alleles across humans, mice, cattle, and chickens. Analysis of these binding specificities combined with X-ray crystallography refined our understanding of the molecular determinants of MHC-II motifs and revealed a widespread reverse-binding mode in HLA-DP ligands. We then developed a machine-learning framework to accurately predict binding specificities and ligands of any MHC-II allele. This tool improves and expands predictions of CD4$^+$ T cell epitopes and enables us to discover viral and bacterial epitopes following the aforementioned reverse-binding mode.

INTRODUCTION

CD4$^+$ T cells are key components of the adaptive immune system. They are implicated in priming and modulating the immune response to pathogens and cancer. CD4$^+$ T cells also play an essential role in cancer immunotherapy,[1,2] as demonstrated by CD4$^+$ T cell responses following neoantigen-based cancer vaccines[3–5] and CD4$^+$ T cell-mediated regression of metastatic cancer following adoptive transfer of tumor-infiltrating lymphocytes.[6,7] CD4$^+$ T cell activation starts with the recognition of epitopes presented by the highly polymorphic class II major histocompatibility complex (MHC-II) on the surface of antigen-presenting cells. Despite their central role in infectious diseases, autoimmunity, and cancer, epitopes presented on MHC-II and targeted by CD4$^+$ T cells are still poorly described and difficult to predict. This represents an important bottleneck for fundamental immunology, cancer immunotherapy, and personalized cancer vaccines.

Peptides presented on MHC-II are processed by the class II antigen-presentation pathway. Most of these peptides come from extracellular proteins ingested and degraded in the endocytic pathway.[8] After cleavage, peptides—typically 12–25 amino acids (AAs) long—are loaded on MHC-II and the peptide-MHC-II complexes are displayed on the cell surface. The loading of peptides on MHC-II is facilitated by the action of chaperones, including HLA-DM and HLA-DO in humans.[8] The binding site of MHC-II molecules has been extensively characterized by X-ray crystallography, and MHC-II ligands adopt a conserved binding mode in these structures. This canonical binding mode consists of a linear 9-mer binding core, which makes most of the interactions with the MHC-II binding site and peptide flanking residues that extend on the N- and C-terminal parts of the binding core.[9] Specific pockets are known to accommodate residues at anchor positions (mainly P1, P4, P6, and P9) in the binding core of the MHC-II ligands.[9] Exceptions to this conserved binding mode have been reported in chickens where one MHC-II allele accommodates peptides with a 10-mer binding core.[10] In human, three peptides have been reported to bind in both the canonical and the reverse orientation (i.e., from N terminus to C terminus and from C terminus to N terminus).[11,12] It is, however, unclear how relevant and frequent this reverse-binding mode is for naturally presented MHC-II ligands and CD4$^+$ T cell epitopes.

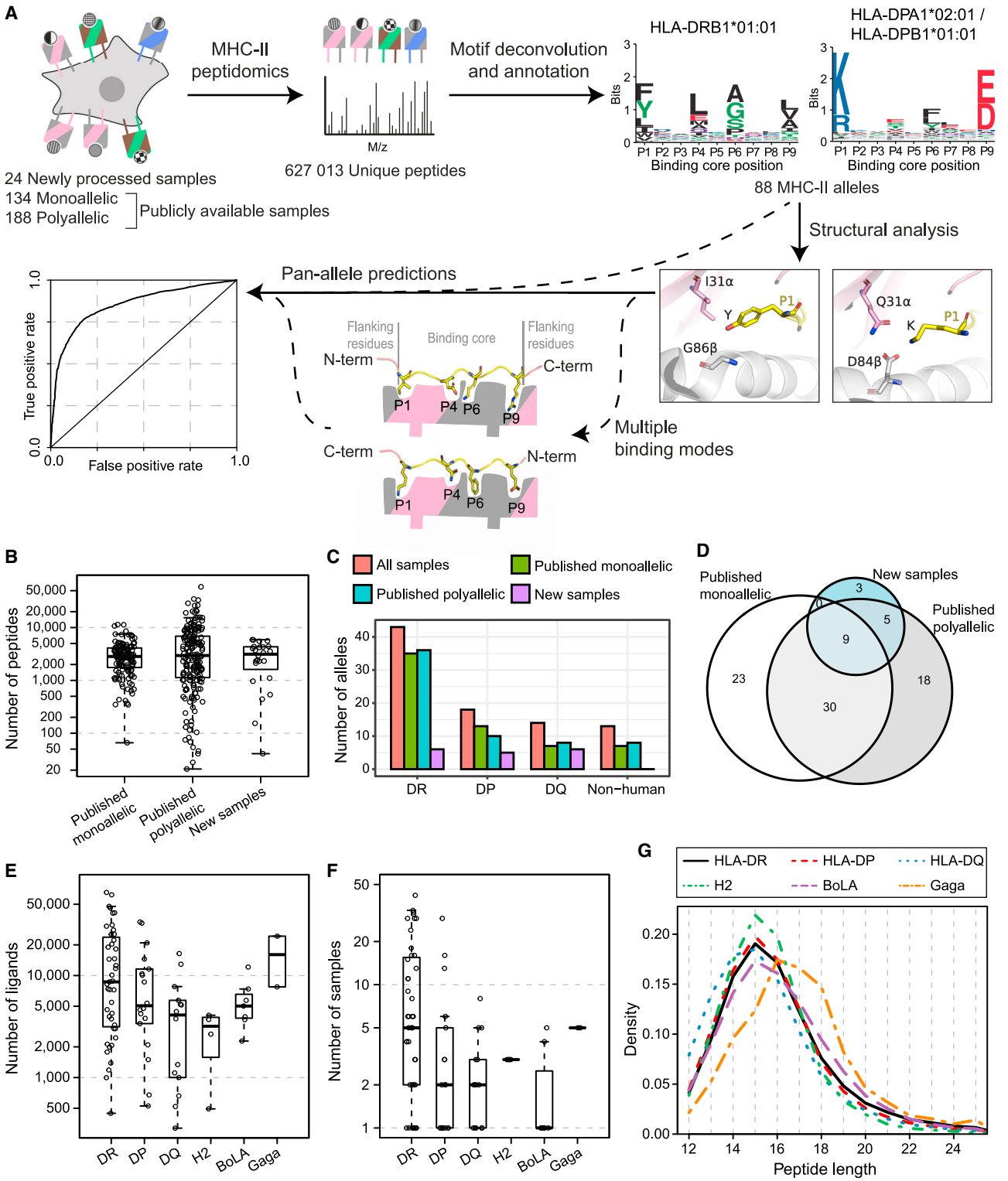


Figure 1. Curation of MHC-II peptidomics data reveals binding specificities for 88 MHC-II alleles

(A) Schematic view of the MHC-II motif analysis and class II epitope prediction pipeline. The main steps of our analyses consist of (1) collection of a large dataset of naturally presented MHC-II ligands identified in multiple MHC-II peptidomics samples, (2) motif deconvolution and annotation, (3) structural analysis of MHC-II binding specificities, and (4) development of a machine-learning predictor of MHC-II ligands and $CD4^+$ T cell epitopes. In the binding motifs, the x axis corresponds to the peptide-binding-core position and the y axis is the Shannon entropy measured in bits; for simplicity, these axes labels are omitted in the other figures of binding motifs.

(legend continued on next page)

In humans, MHC-II are also called class II human leukocyte antigens (HLA-II) and consist of three gene loci directly involved in presenting antigens to $CD4^+$ T cells: HLA-DR (including HLA-DRA1 and HLA-DRB1, -DRB3, -DRB4, and -DRB5 genes); HLA-DP (including HLA-DPA1 and HLA-DPB1 genes); and HLA-DQ (including HLA-DQA1 and HLA-DQB1 genes). Except for HLA-DRA1, these genes are highly polymorphic and more than 9,100 alleles have been identified in the IPD-IMGT/HLA database[13] as of 23.06.2022). Within each gene locus, MHC-II form heterodimers composed of an alpha chain (e.g., HLA-DPA1*02:01) and a beta chain (e.g., HLA-DPB1*01:01). Because of all the possible combinations between the alpha and beta chains, there is an even much higher number of potential MHC-II heterodimers (“MHC-II alleles”). The polymorphic residues mostly lie in the peptide-binding site,[14] resulting in highly allele-specific peptide-binding motifs[15–18] (graphically represented with sequence logos).

The polymorphism of MHC-II genes, the vast diversity of MHC-II binding motifs and the complexity of the class II antigen-presentation pathway represent important hurdles for reliable predictions of naturally presented MHC-II ligands and $CD4^+$ T cell epitopes. In recent studies, we and others have shown how high-throughput mass-spectrometry-based MHC-II peptidomics can be used to improve these predictions.[15–17,19,20] This was achieved through the identification of MHC-II motifs using either monoallelic samples[15] or motif deconvolution in polyallelic samples.[16,17] Beyond MHC-II binding motifs, MHC-II peptidomics also revealed specificity in the first and last AAs in peptide-flanking residues and a specific peptide length distribution (peaked at 15 AAs).[16,21–23] We have also observed that the peptide binding-core offset (defined as the position of the binding core relative to the middle of the peptide) is slightly shifted toward the C terminus of the MHC-II ligands.[16] The expression of both the epitope source proteins and the HLA-II molecules in antigen-presenting cells has also been found to correlate with antigen presentation,[15,19] although this notion has been recently challenged.[24] Influence of the source protein subcellular localization has also been reported.[24]

Today, several MHC-II ligand prediction tools are available. These include allele-specific (e.g., MixMHC2pred-1.2[16] and NeonMHC2[15]), pan-HLA-DR (e.g., MARIA[19]), and pan-allele predictors (e.g., NetMHCIIpan-4.0[17] and MHCnuggets[25]). The latter aim at capturing correlation patterns between MHC-II binding sites and binding specificities. However, the data used to train these predictors are limited to few alleles and consist mostly of HLA-DR ligands. As a result, epitope predictions for poorly characterized alleles, especially HLA-DP and HLA-DQ or alleles from other species, have limited accuracy.

Here, we collected and curated a dataset of MHC-II ligands and determined the binding specificities of more than 80 MHC-II alleles. This enabled us to improve our molecular understanding and prediction capability of MHC-II ligands. These results refine and expand our understanding of the universe of $CD4^+$ T cell epitopes that could be therapeutically targeted in infectious diseases, autoimmunity, and cancer immunotherapy.

RESULTS

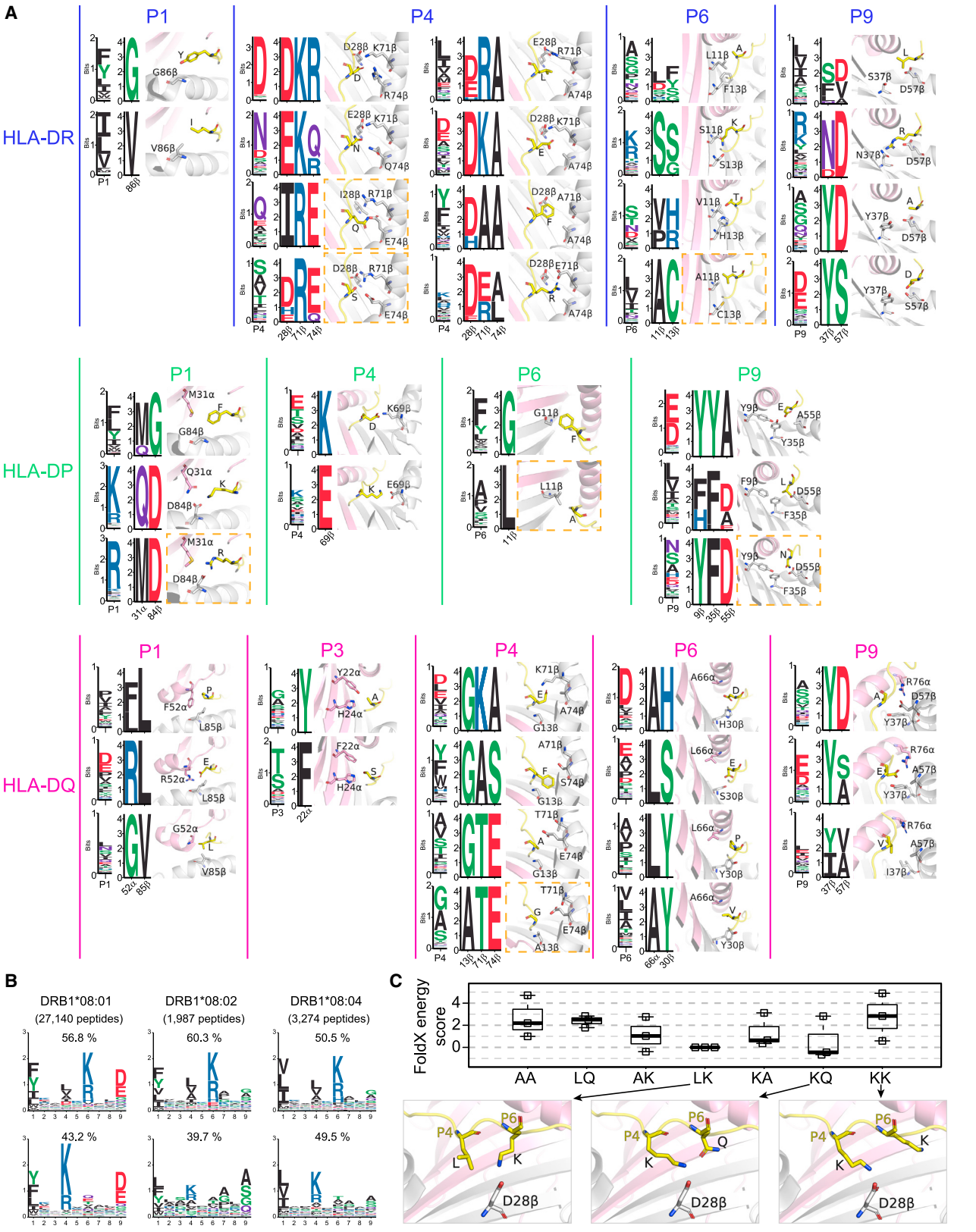
Curation of MHC-II peptidomics data reveals binding specificities for 88 MHC-II alleles

To improve our understanding of the specificity of class II antigen presentation, we developed a pipeline to infer MHC-II binding specificities and predict MHC-II ligands and $CD4^+$ T cell epitopes (Figure 1A). We first performed a thorough literature curation to search for available mass-spectrometry-based MHC-II peptidomics datasets and collected data from 30 published studies for a total of 322 samples and 615,361 unique peptides, including MHC typing of each sample (Tables S1 and S2A). Most of these samples were obtained from human cells using anti-HLA-DR or anti-pan-HLA-II antibodies. Other samples were obtained using anti-HLA-DP or anti-HLA-DQ antibodies[26–28] and cells transfected with tagged HLA-II allowing for the isolation of peptides bound to a single allele.[15] A few samples were obtained from mice,[29–31] cattle[32] and chickens.[10] To further enrich for HLA-DP and HLA-DQ ligands, we used mass-spectrometry-based MHC-II peptidomics to sequentially isolate peptides with anti-HLA-DR, anti-HLA-DP, anti-HLA-DQ, and anti-pan-HLA-II antibodies (see STAR Methods). Applying this strategy to six different cell lines or meningioma tissues enabled us to obtain 44,334 unique peptides, including 11,779 HLA-DP and 16,146 HLA-DQ ligands (Tables S1 and S2B). This was especially useful with respect to the limited number of publicly available HLA-DQ ligands (31,045 unique peptides).

Combining all these data led to a total of 627,013 unique peptides (1,540,995 peptides when counting duplicates across samples) coming from 346 samples corresponding to 201 different cell lines or tissues with full MHC-II typing (Figure 1B; Tables S1 and S2). These numbers compare favorably with existing databases, such as the Immune Epitope Database (IEDB),[33] which contains 508,070 unique MHC-II ligands derived from different experimental methods, with 472,162 unique MHC-II ligands obtained from mass-spectrometry-based experiments (as of 29.08.2022).

We then performed motif deconvolution using MoDec on each sample and identified shared motifs across samples sharing the same alleles, following the procedure described by Racle et al.[16] (see STAR Methods). The motif identification and annotation were manually verified in each sample. In total, we could confidently describe the binding specificities of 88 MHC-II alleles, including 43 HLA-DR, 18 HLA-DP, 14 HLA-DQ, 4 mouse H-2, 7 cattle BoLA-DR, and 2 chicken Gaga-BLB alleles (Figures 1C

(B) Number of unique peptides per sample that were collected in this study, grouped by the type of study of origin of the sample.
(C) Number of MHC-II alleles for which we could determine the binding specificity with our motif deconvolution analysis pipeline.
(D) Euler diagram of the alleles for which we could determine the binding specificity with respect to the different types of samples of origin.
(E) Number of ligands identified for each allele, grouped by genes and species of origin.
(F) Number of samples in which the motif of a given allele was identified, grouped by genes and species of origin.
(G) Peptide length distributions per genes and species of origin.
Box plots in (B), (E), and (F) indicate the medians and upper and lower quartiles. See also Figure S1 and Tables S1 and S2.



(legend on next page)

and S1A). These binding motifs are supported by 637,821 unique peptide/MHC-II interactions, which almost doubles the number of unique peptide/MHC-II interactions with fully resolved MHC-II typing available in IEDB (346,154). Binding specificities for 39 MHC-II alleles are supported by peptides from both monoallelic and polyallelic samples—23 MHC-II alleles only by monoallelic data and 26 MHC-II alleles only by polyallelic data (Figure 1D). This demonstrates the importance of integrating monoallelic and polyallelic data to reach the best allelic coverage. The binding specificities of most alleles were supported by thousands of peptides (Figure 1E). Motifs describing these alleles were predominantly identified in multiple samples sharing a common allele (Figure 1F; see also Racle et al.[16]), demonstrating the robustness of our approach. Our data also show that the distributions of peptide lengths and binding-core offsets are broadly conserved in human and nonhuman MHC-II alleles (Figures 1G and S1B),[16,34] with the main discrepancy observed in samples transfected with tagged HLA-II (Figure S1C).

Our framework allowed us to describe the binding specificity of virtually all HLA-DR alleles available in our dataset, as well as most HLA-DP and HLA-DQ at the beta-chain level (Figure S1D). 67% of the peptides corresponded to motifs unambiguously annotated to unique MHC-II alleles; these peptides were used to build the final motifs. 9% of the peptides had length shorter than 12 AAs and were not considered here. The remaining 24% of the peptides could not be fully annotated; these represent, for example, peptides assigned to ambiguous motifs describing multiple alleles with similar binding specificities in a sample or to unspecific motifs (putative contaminants).

MHC-II binding specificities reflect biochemical properties of the MHC-II binding pockets

Our MHC-II binding motifs provide a unique opportunity to better understand the characteristics of MHC-II binding specificities. Consistent with previous studies, we observed that the HLA-DR, HLA-DP, mouse, and cattle alleles usually have four clear anchors at positions P1, P4, P6, and P9. HLA-DQs have slightly weaker binding specificities with main anchors at P3, P4, and P6 and sub-anchors at P1 and P9, in general (Figure S1A). The binding specificities for all human, mouse, and cattle alleles can be described by 9-mer motifs, suggesting that the bulging mechanism observed in MHC-I alleles is rare for most MHC-II alleles (see previous analysis in Racle et al.[16]). For the only exception known to accommodate a longer binding core (chicken allele Gaga-BLB2*002:01),[10] we additionally used MoDec to search for a 10-mer motif. We observed that the binding specificity of this allele could be well described by two motifs, with almost half of ligands possessing a binding core of 9 AAs and the other half a binding core of 10 AAs (Figure S2A; see also STAR Methods).

To investigate the molecular determinants of MHC-II binding specificities, we performed unsupervised clustering of the binding motifs for all human alleles for each HLA locus (i.e., HLA-DR, -DP, and -DQ) and for each anchor position (see STAR Methods). We could observe different classes of specificities (Figure 2A, left). We then retrieved the sequences of the most variable residues in the MHC-II binding pockets interacting with the anchor residue in the ligand for all alleles found in each cluster (Figure 2A; Tables S3A–S3C; see STAR Methods). As expected, alleles found in distinct specificity clusters (i.e., rows within each column of Figure 2A) had differences in their binding pockets. To interpret these different clusters structurally, we used available crystal structures of representative alleles in each cluster (Figure 2A). For a few cases, no crystal structure was available, and we used structural modeling (see STAR Methods). This revealed a clear correspondence between the MHC-II binding motifs and the binding pockets of MHC-II alleles. For example, for HLA-DR alleles, large and bulky AAs (mainly F or Y) are observed at P1 when glycine (G) is found at 86β (P1 binding pocket), while less bulky hydrophobic AAs (I, L, or V) are observed at P1 in HLA-DR alleles when valine (V) is found at 86β (Figure 2A; Table S3A), recapitulating previous findings obtained from specific HLA-DR alleles.[35,36] This mutual exclusivity reflects the steric clash that would happen between F or Y in the ligand and V in the binding pocket. For HLA-DP alleles, a small AA at 84β (G) correlated with bulky AA at P1 (F, L, Y, or I), and a negatively charged AA at 84β (D) correlated with positively charged AA at P1 (K or R) (Figure 2A; Table S3B). Furthermore, a long polar AA at 31α (Q) correlated with K at P1, while a long nonpolar AA at 31α (M) correlated with R at P1 (Figure 2A; Table S3B). K at P1 can simultaneously engage into polar or charged interactions with D84β and Q31α, whereas the two nitrogens of R at P1 would preferentially face the two oxygens of the carboxyl group of D84β (Figure S2B). This conformation is more favorable if M is found at 31α instead of Q. Similar analyses for other anchor positions are detailed in Table S3D. Most of the observations in Figure 2A could be explained by steric hindrance or polar and charged interactions. This demonstrates a clear correspondence between our MHC-II motifs and the sequences of MHC-II binding sites. Several of these observations have already been made based on structural analyses of limited sets of HLA-II alleles.[35,37,38] In those cases, our results provide stronger statistical evidences and larger allelic coverage, as well as a unified picture of the determinants of HLA-II specificity across all anchor positions in HLA-DR, -DP, and -DQ alleles.

Figure 2. MHC-II binding specificities reflect biochemical properties of the MHC-II binding pockets

(A) For each type of human MHC-II (HLA-DR, -DP, and -DQ) and each anchor position, MHC-II alleles were clustered based on the binding specificity at this position and the different clusters are shown as different rows for each position. On each row, the first motif (e.g., P1) represents the average peptide-binding specificity of the MHC-II alleles in a cluster. The second motif (e.g., 86β) represents the sequence of the MHC-II residues making important contacts with the residue in the ligand at the specified anchor position. The structural arrangement of the residues in the MHC-II binding site (pink for alpha chain, gray for beta chain) and the ligand (yellow) is shown on the right based on existing X-ray structures. Cases where structural modeling was used are indicated by a dashed orange rectangle.

(B) Multiple binding specificities for HLA-DRB1*08 alleles. Percentage above the motifs indicate the fraction of peptides assigned to each sub-specificity.

(C) Molecular interpretation of the multiple specificities observed in HLA-DRB1*08:01. The top box plot shows the calculated change in the FoldX energy score for various peptides with 0, 1, or 2 positively charged AAs at P4 and P6 (see also Table S3E). The bottom image shows a model of HLA-DRB1*08:01, which contains one single negatively charged residue (D28β) that is able to interact with positively charged sidechains at either P4 or P6, but not both simultaneously.

See also Figure S2 and Table S3.

For most alleles, a single binding specificity was observed. Yet, for several HLA-DRB1*08 we observed two binding motifs (Figure 2B). The two motifs suggest that a positively charged AA (K or R) is favorable either at P4 or P6, but not at both positions at the same time. To understand the molecular mechanism of this bi-specificity, we calculated the FoldX energy score[39] of several peptides with a charged residue either at P4 or P6, two charged residues at these positions or no charged residue (see STAR Methods). Our calculations indicate that peptides with two charged residues have higher energy scores (i.e., weaker predicted binding affinity) than those with only one charged residue (Figure 2C). This can be understood since the central part of the binding site of HLA-DRB1*08 alleles contains one negatively charged residue (D28β), which can interact with positively charged sidechains either at P4 or P6, but not at both P4 and P6 (Figure 2C). This mutual exclusivity of charged residues at P4 and P6 appears to be restricted to HLA-DRB1*08 alleles and may require G at 13β (Figures S2C and S2D).

MHC-II binding specificities reveal a widespread reverse-binding mode in HLA-DP ligands

Another type of bi-specificity was observed for 7 out of 18 HLA-DP alleles (Figure S1A) across multiple samples. These include several monoallelic samples, indicating that both motifs indeed describe the binding specificity of a single allele. For these alleles, superimposing the motifs revealed a clear symmetry between the first and the second binding motifs (Figure 3A). It is unlikely that residues with opposite biochemical properties (e.g., K versus E at P1) could fit at the same position. Therefore, we hypothesized that the second motif corresponded to a reverse-binding mode where the peptides are bound from the C terminus to the N terminus. This hypothesis was further supported by the fact that the distribution of binding-core offsets for peptides following the second motifs was skewed toward the N terminus, unlike all other MHC-II ligands (Figure S3A).

To validate our hypothesis, we first tested the binding of different peptides to HLA-DPA1*02:01-DPB1*01:01 (see STAR Methods). Starting with two peptides predicted to bind in the canonical orientation, we observed that replacing with alanine the predicted P1 (K) and P9 (E) anchor residues abrogated the binding, while replacing K at P1 with R did not affect the binding (Figures 3B and S3B; Table S4A). We then reversed the sequence in order to obtain peptides following the second binding motif. Figure 3B shows that the binding was preserved. In general, the binding was stronger for peptides following the first motif (i.e., the predicted canonical binders) than for those with the reversed sequence (the predicted reverse binders). We confirmed this observation with 39 peptides synthesized in both directions (Figure 3C; Table S4B). These results were consistent with the weights of the motifs in Figure 3A. As a sidenote, one should not conclude that we can reverse the sequence of any ligand and that it would still bind, as demonstrated in Figure 3C.

We then attempted to crystallize peptides predicted to bind in either the canonical or the reverse orientation (see STAR Methods). We first obtained the X-ray crystal structure (at a resolution of 1.62 Å) of a peptide which matches the first motif (KNLEKYKGKFVREID, core underlined). This peptide binds in the canonical orientation to HLA-DPA1*02:01-DPB1*01:01, with K5 filling the P1 binding pocket and E13 filling the P9 binding pocket (Figures 3D, left and S3C). We then crystallized a peptide compatible with the second motif (IEFVFKNKAKEL, resolution of 2.9 Å). We observed that the binding happens in the reverse orientation with the first residue of the core (E2) filling the P9 binding pocket, and the last residue of the core (K10) filling the P1 binding pocket (Figures 3D and S3C). Most of the interactions mediated by each anchor residue were preserved (Figure 3D). Overlaying the two structures further demonstrated a remarkable alignment not only of the sidechains (Figure S3D) but also of the backbone N–H and C=O groups (Figure S3E), as well as a conservation of most backbone H-bonds (Figure S3F). These results reveal how HLA-DP alleles can accommodate different peptides' binding in different orientations without extensively remodeling their binding site.

During class II antigen presentation, MHC-II ligands are processed by different proteases. Footprints of this process are visible in the N- and C-terminal contexts of MHC-II ligands.[22,40] However, the timing and the impact of the positioning of peptides in the MHC-II binding site are still unclear.[41,42] MHC-II ligands binding in the reverse orientation provide an opportunity to shed light on this process. Figure S3G shows that the motifs of the N- and C-terminal contexts of the reverse-binding peptides were very similar to those of the canonical binding peptides. Consistent with previous predictions,[15] this observation supports a model in which the cleavage and trimming takes place first, independently of the positioning of the peptides in the MHC-II binding site.

To investigate the detection limit of our approach to identify reverse ligands, we simulated the presence of ligands reversely bound to HLA-DR or HLA-DQ alleles at different fractions and determined if the reverse motif could be identified with MoDec (see STAR Methods). Our results indicate that in a monoallelic sample of 3,000 ligands, the reverse motif was found when 5%–10% of the ligands were reverse ligands. In polyallelic samples with 3,000 ligands, a minimum of 20% of reverse ligands of the allele that included this binding mode was needed (Figure S3H). In terms of absolute numbers, we observed a limit of 150–200 reverse ligands to be detectable by MoDec (Figure S3I). Considering that monoallelic samples with 3,000–6,000 peptides cover a large fraction of our MHC-II alleles (Figures 1B and 1C; Table S1), we can provide an upper bound of roughly 5% for the fraction of ligands that could be bound in the reverse orientation to a given allele without being detected by our approach.

MHC-II binding specificities can be accurately predicted for alleles without known ligands

The high polymorphism of MHC-II genes prevents the experimental determination of the binding specificity for all alleles. Our collection of MHC-II motifs provides an opportunity to train an accurate predictor of MHC-II ligands for any allele (referred to as pan-allele predictor). To this end, we designed a machine-learning framework composed of two distinct successive blocks (Figure 4A). In the first block, the aim was to predict the MHC-II binding motifs, defined as position probability matrices (PPMs) (see STAR Methods) directly from the MHC-II sequences. In the second block, the aim was to predict actual MHC-II ligands based on their sequence and the PPM of the corresponding MHC-II allele.

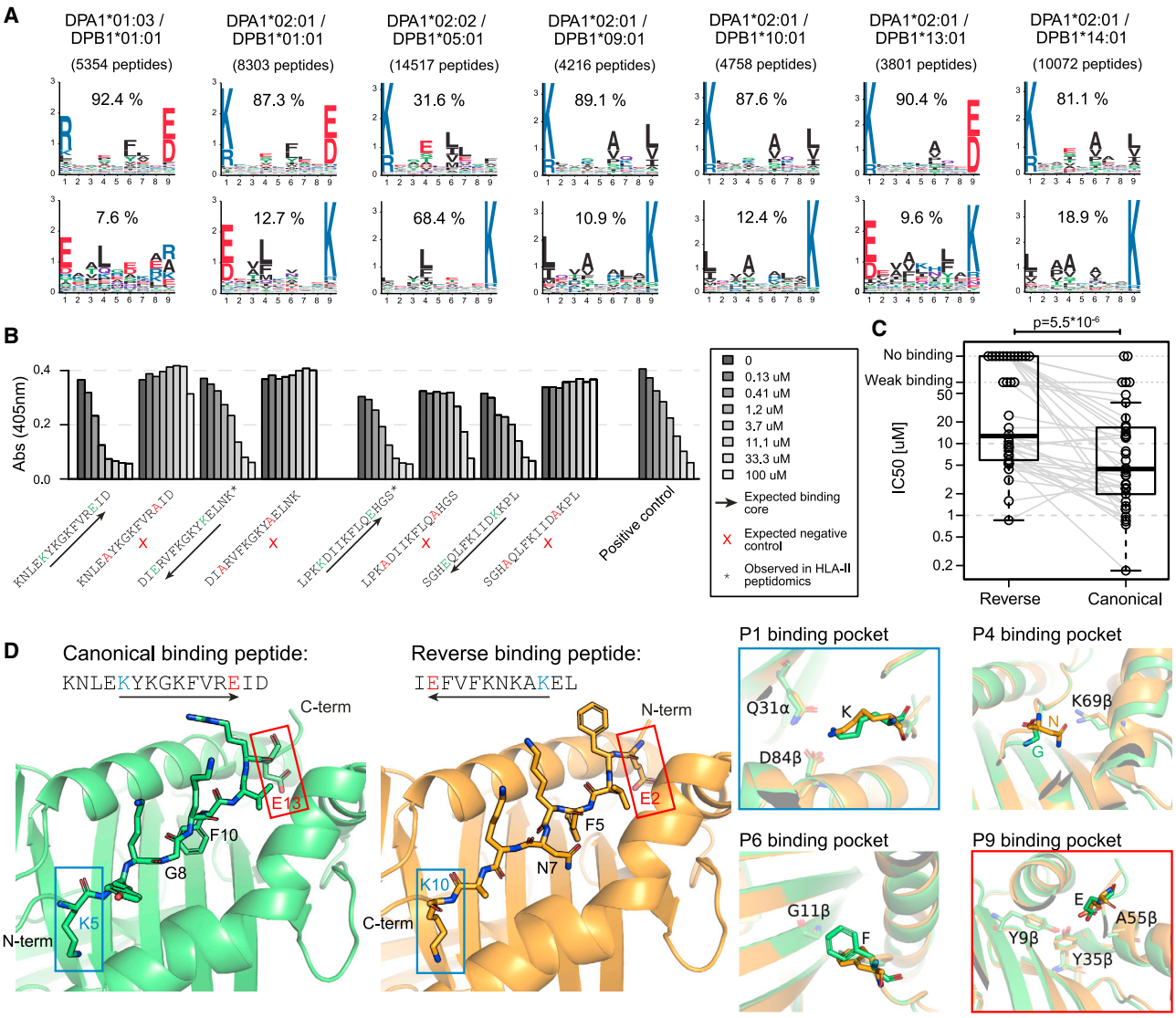


Figure 3. MHC-II binding specificities reveal a widespread reverse-binding mode in MHC-II ligands

(A) HLA-DP alleles with symmetrical multiple binding specificities. Percentage above the motifs indicate the fraction of peptides assigned to each sub-specificity.

(B) Binding competition assays of peptide variants bound to HLA-DPA1*02:01-DPB1*01:01. The different peptides correspond to (1) predicted canonical binders, (2) alanine mutations at anchor residues P1 and P9, (3) predicted reverse binders (reversed sequences), and (4) alanine mutations at P1 and P9 in the reversed sequence. Additional peptides are shown in Figure S3B. Stars indicate peptides that were seen in MHC-II peptidomics data.

(C) Half-maximal inhibitory concentrations (IC_{50}) in binding-competition assays for 39 peptides fitting the reverse-binding motif and the corresponding 39 inverse sequences (fitting therefore the canonical binding motif). Average IC_{50} between 2 repetitions are shown. Box plots indicate the medians and upper and lower quartiles; result of a paired two-sided Wilcoxon signed rank test is indicated.

(D) Crystal structures of the canonical binder KNLEKYKGKFVREID (PDB: 7ZAK) and the reverse binder IEFVFKNKAKEL (PDB: 7ZFR) bound to HLA-DPA1*02:01-DPB1*01:01. The four panels on the right show the overlap of the two structures at the four binding pockets.

See also Figure S3 and Table S4.

Technically, the first block consists of a set of fully connected neural networks for each binding-core position (Figure S4A, left; see STAR Methods). The binding sites corresponding to each position and used as inputs of the neural networks were determined based on existing crystal structures (see STAR Methods). This block also incorporates the different multiple specificities and was trained on our set of PPMs (see STAR Methods). The second block of the predictor consists of a fully connected neural network (Figure S4A, right). It takes as input the score of the peptide against the PPMs of the corresponding MHC-II allele together with other features linked to antigen presentation (i.e., peptide length, binding-core offset, and peptide processing and cleavage features) (see STAR Methods). This block was trained on our dataset of MHC-II ligands and random negative peptides (see STAR Methods).

We first benchmarked how accurate our predictor (MixMHC2pred-2.0) was in predicting the MHC-II binding motifs for alleles without known ligands. We performed a

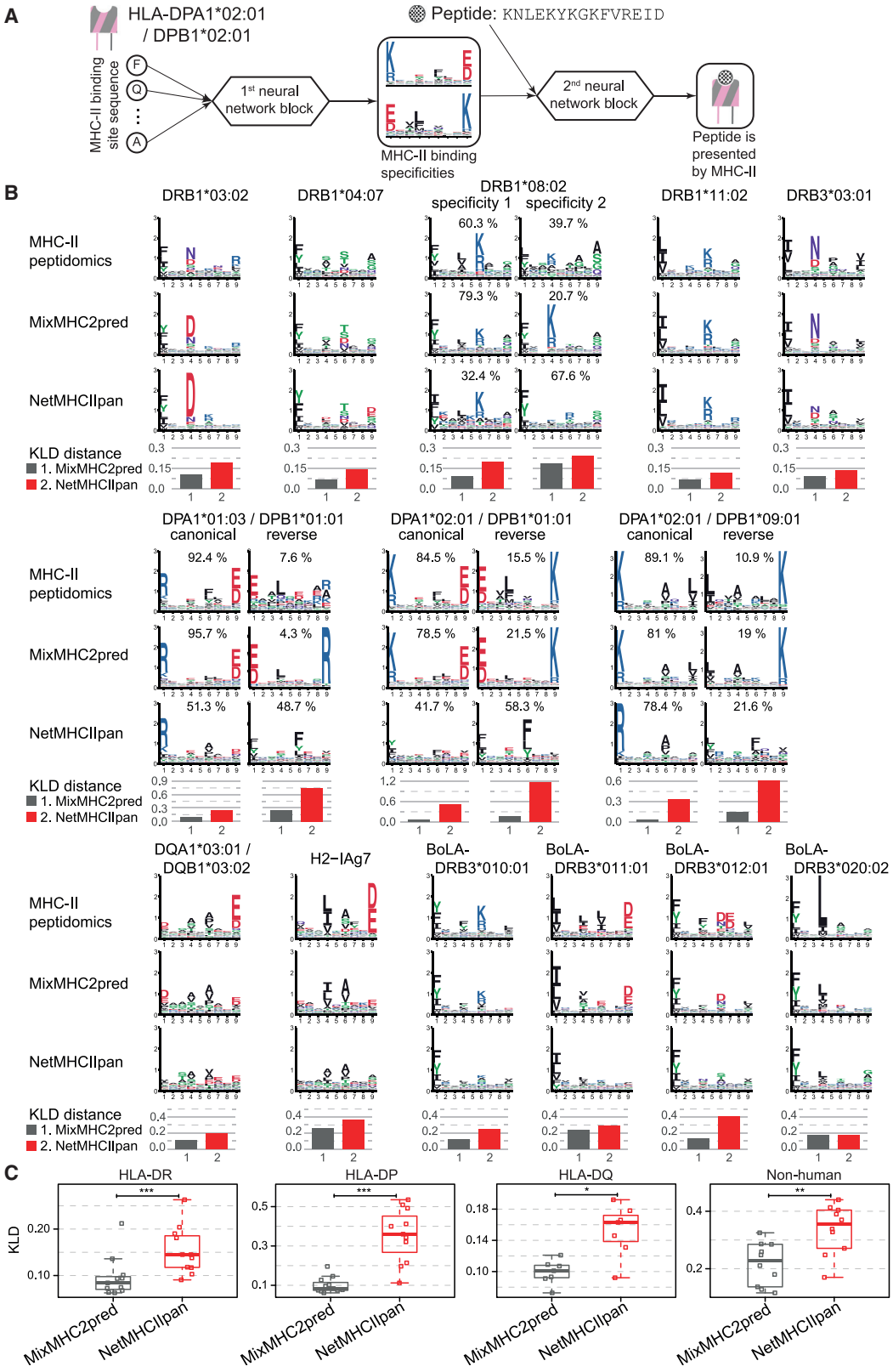


Figure 4. MHC-II binding specificities can be accurately predicted for alleles without known ligands

(A) Schematic description of the pan-allele predictor comprising two consecutive blocks of neural networks. (See Figure S4A and STAR Methods for the full details of the model.)

(legend continued on next page)

leave-one-allele-out cross-validation, where all the data from one allele are removed from the training (see STAR Methods). To compare with the current state-of-the-art pan-allele predictor NetMHCIIpan-4.0, we focused on MHC-II alleles absent from its training set. For both predictors, predicted PPMs were built by considering 100,000 random human peptides and selecting the top 1% best predicted peptides (see STAR Methods). The resulting PPMs were compared with those derived from MHC-II peptidomics studies using the Kullback-Leibler divergence (KLD) (see STAR Methods). Our results showed that MixMHC2pred better inferred the binding specificities of alleles without known ligands (Figures 4B and 4C; Table S5A). The multiple specificities (including the reverse-binding specificity) could be well predicted by MixMHC2pred, while they were not detectable with NetMHCIIpan (Figure 4B) (see STAR Methods). Results were similar when directly predicting the ligands of each allele instead of the PPMs (Figure S4B; Table S5B; see STAR Methods).

To quantify the impact of the similarity between alleles on their binding-specificity predictions, we compared the sequence similarity of the 88 alleles available in our training data and the prediction accuracy in the leave-one-allele-out cross-validation (see STAR Methods). A clear correlation was observed (Figure S4C). We then computed for all alleles in multiple species the sequence similarity to the 88 alleles with known ligands (Figure S4D; Table S3F; STAR Methods). Our results showed that all human HLA-DR and -DP alleles had a high sequence similarity to some allele with known ligands. This suggests that obtaining MHC-II peptidomics data for these alleles would not dramatically improve the prediction accuracy. A fraction of HLA-DQ alleles showed low sequence similarity with alleles with known ligands. The majority of these cases corresponded to unstable HLA-DQ heterodimers like DQA1*01:03-DQB1*02:02,[43] but some other HLA-DQ alleles may benefit from MHC-II peptidomics data (e.g., DQA1*04:01-DQB1*04:01 or DQA1*01:03-DQB1*06:01) (Table S3F). DR genes from MHC-II alleles in other species often have an intermediate sequence similarity to alleles with known ligands (Figure S4D), suggesting that a decent accuracy would be reached. Regarding MHC-II DQ genes from other species, we saw more evolutionary divergence with lower similarity to alleles with known ligands. For these alleles, predictions should have limited accuracy and would benefit from additional MHC-II peptidomics data.

We finally explored the impact of different choices for the parameters of our neural networks (see STAR Methods). Overall, the number of hidden nodes had only a minimal impact on the prediction accuracy. The model architecture chosen for MixMHC2pred was also the most accurate in predicting the binding specificities (Figure S4E).

MixMHC2pred improves predictions of MHC-II ligands and CD4$^+$ T cell epitopes

We then benchmarked the accuracy of MixMHC2pred to predict naturally presented MHC-II ligands, using the leave-one-allele-out cross-validation setting and the area under the curve of the receiver operating characteristic curve (ROC AUC) as a measure of prediction accuracy (see STAR Methods). Results showed improved predictions for MixMHC2pred compared with other pan-allele predictors (Figures 5A–5C; Table S5C). We then determined the relative impact of the different input parameters of MixMHC2pred (see STAR Methods). Results showed that the full model was significantly better than any variant (Figure S5A). These results also showed that the binding score was the most important input to the model, followed by the N- and C-terminal contexts of the peptide that lie within the peptide. The context outside of the peptide had a smaller impact, as well as the binding-core offset and peptide length (Figure S5A).

As MixMHC2pred incorporated the reverse-binding mode, we could then use it to estimate the fraction of ligands bound in the reverse orientation in different contexts. Our analysis indicated that the presence of reverse-binding ligands was not influenced by the origin of the sample (Figure S5B), nor by the expression of HLA-DM (Figure S5C), although the number of samples available for this analysis was limited.

HLA-DP and HLA-DQ are polymorphic on both alpha and beta chains. This can give rise to *cis*- and *trans*-heterodimers, respectively (i.e., alpha and beta chains from the same, respectively different chromosomes). For HLA-DQ, two groups of genotypes are known.[43] In the first group (G1), alpha chains come from HLA-DQA1*02, 03, 04, 05, and 06 and beta chains come from HLA-DQB1*02, 03, and 04. In the second group (G2), alpha chains come from HLA-DQA1*01 and beta chains come from HLA-DQB1*05 or 06. Heterodimers consisting of two chains from the same group bind stably.[44] Conversely, *trans*-heterodimers with alpha and beta chains coming from different groups (i.e., G1α-G2β or G2α-G1β, or G1G2 for simplicity) are unstable.[43,44] Following recent studies about HLA-DQ *trans*-heterodimers,[43,45] we investigated the fraction of ligands coming from *cis*- or *trans*-HLA-DP and HLA-DQ heterodimers. For HLA-DQ samples with two different genotypes, our results suggest that on average less ligands are presented by G1G2 HLA-DQ *trans*-heterodimers compared with *cis*-heterodimers (Figures S6A and S6B; see STAR Methods). Motifs annotated through our deconvolution pipeline also never corresponded to such HLA-DQ *trans*-heterodimers (Figure S1A). Regarding HLA-DQ samples with the four alleles from the same group (either G1 or G2) and HLA-DP samples, our analysis showed that the two heterodimers with the highest fraction of ligands often share a common alpha or beta chain (Figures S6A and S6B). This suggests that both *cis*- and *trans*-heterodimers can

(B) Comparison between actual and predicted motifs for alleles observed in monoallelic MHC-II peptidomics samples and absent from NetMHCIIpan training set. Multiple specificities, when present, are shown and the fraction of peptides observed and predicted per specificity is indicated above each motif (for NetMHCIIpan the multiple specificities were analyzed with MoDec, see STAR Methods). The average distance, measured with Kullback-Leibler divergence (KLD) per peptide core position, between the motifs observed in MHC-II peptidomics and the predicted ones is shown below each allele.
(C) KLD between the specificities observed in MHC-II peptidomics and predicted by MixMHC2pred (leave-one-allele-out) and NetMHCIIpan for all MHC-II alleles absent from NetMHCIIpan training. Box plots indicate the median, upper, and lower quartiles; results of a paired two-sided Wilcoxon signed rank test are indicated (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).
See also Figure S4 and Tables S5A and S5B.

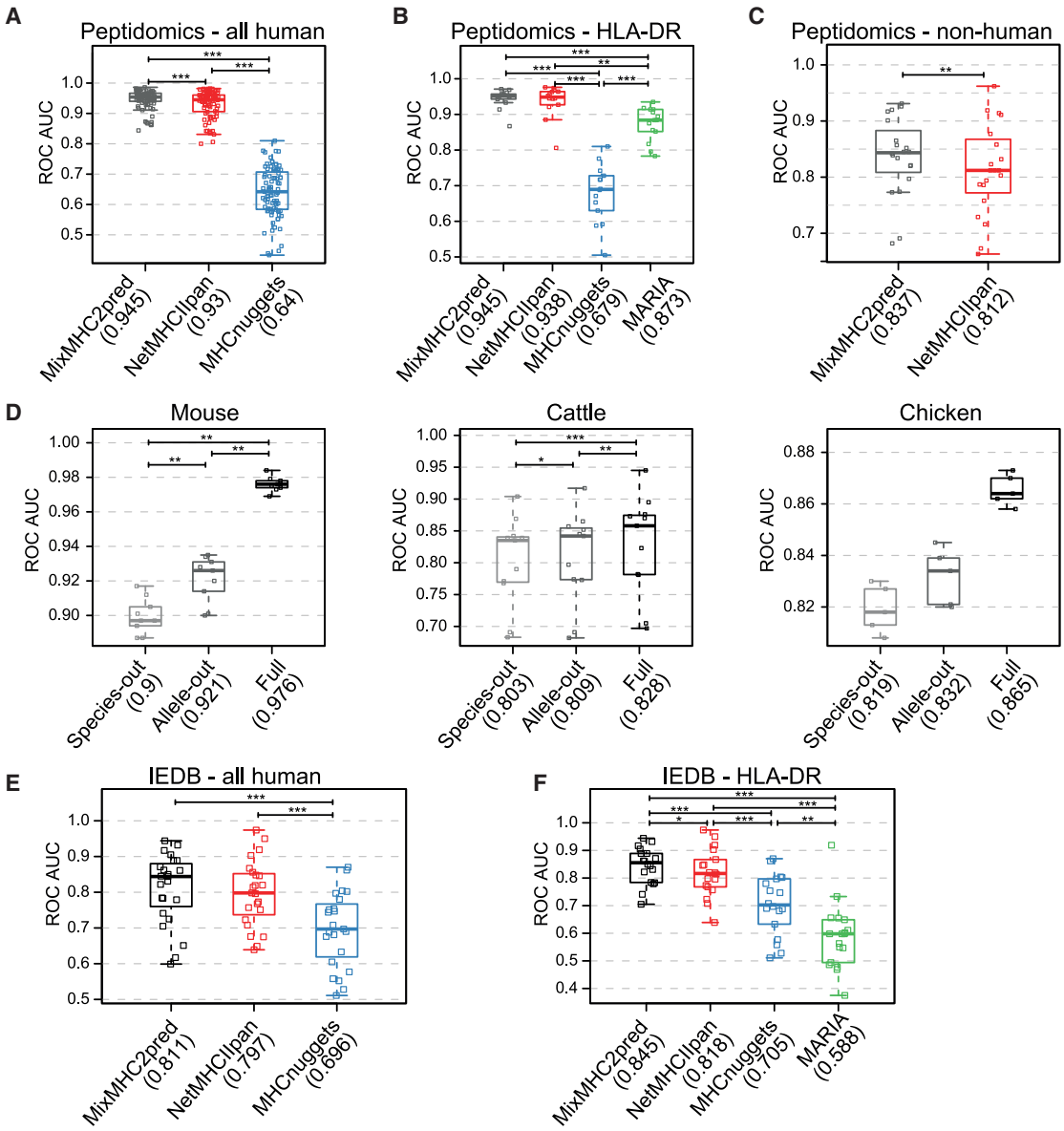


Figure 5. MixMHC2pred improves predictions of MHC-II ligands and CD4$^+$ T cell epitopes

(A–C) ROC AUC for predictions of peptides presented by MHC-II. Only samples that are absent from the training set of all predictors are included. MixMHC2pred results were obtained in a leave-one-allele-out context. (A) All human samples; (B) human HLA-DR only samples; (C) all nonhuman samples (mouse, cattle, and chicken samples). MARIA could only be applied on HLA-DR and MHCnuggets only on human and mouse alleles.

(D) ROC AUC for predictions of peptides presented by MHC-II in mouse, cattle, and chicken, using our prediction framework trained on (1) all data except those from the species where predictions are made (leave-one-species-out), (2) all data except those containing the allele for which predictions are made (leave-one-allele-out), or (3) all data (full model).

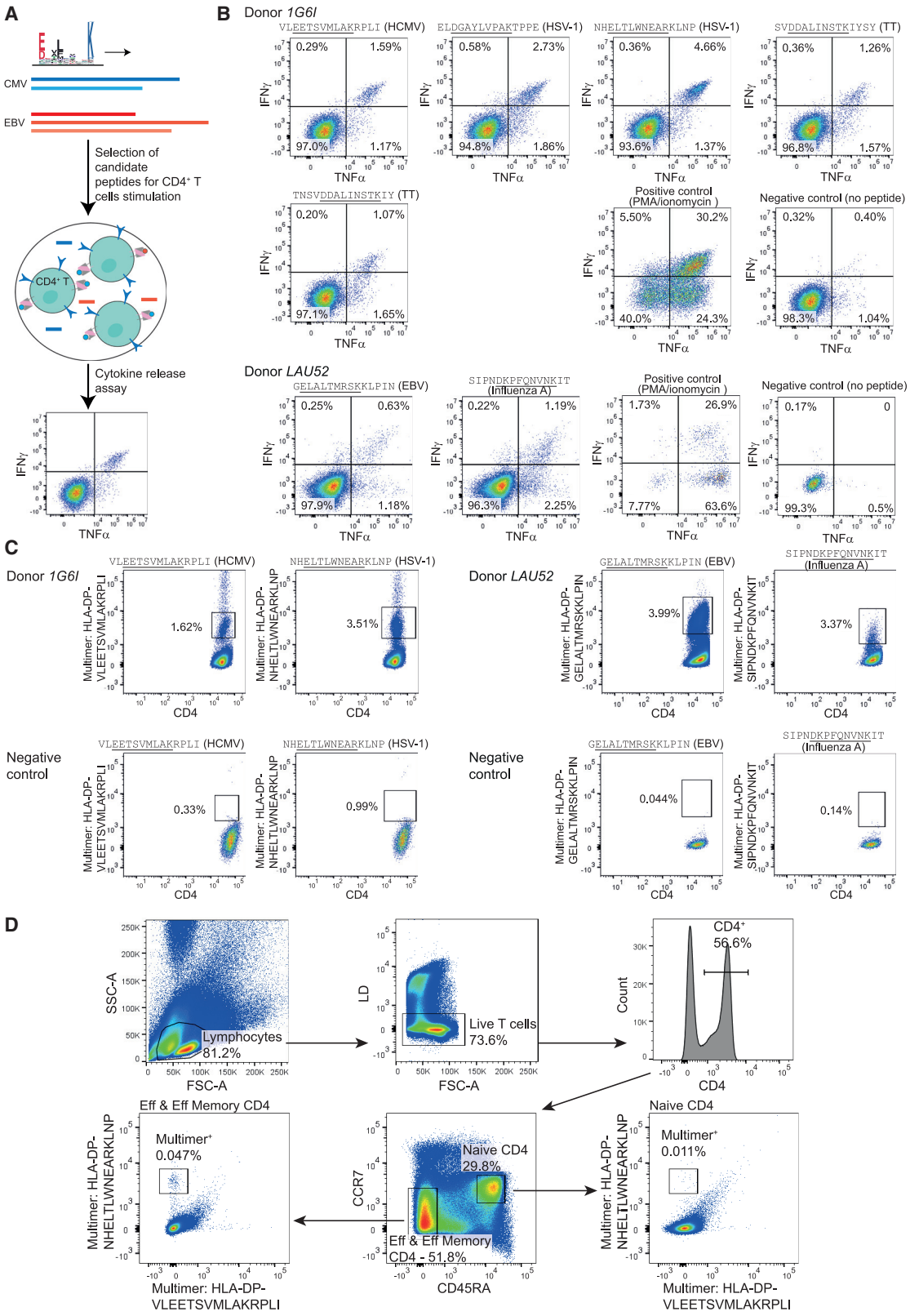
(E and F) ROC AUC for predictions of CD4$^+$ T cell epitopes found in IEDB for (E) all human data, (F) only HLA-DR alleles.

Numbers in parentheses below each predictor's name correspond to the average ROC AUC values. Box plots indicate the medians and upper and lower quartiles; the results of a paired two-sided Wilcoxon signed rank test are indicated (*p < 0.05, **p < 0.01, ***p < 0.001).

See also Figures S5 and S6 and Tables S5C–S5E.

bind peptides. These results are consistent with what is currently known about HLA-DQ heterodimers and with the recent analysis of this phenomenon by Nilsson and colleagues.[45] From a practical point of view, our results support the use of all four isoforms when predicting HLA-DP or HLA-DQ ligands, except for the G1G2 HLA-DQ *trans*-heterodimers.

We next examined if MixMHC2pred would be amenable to predictions for species without known MHC-II ligands for any allele. To this end, we retrained our predictor by removing all data coming from one species and predicted the MHC-II ligands from this species (leave-one-species-out cross-validation) (see STAR Methods). We compared the predictions of this model



(legend on next page)

with the predictions obtained in the leave-one-allele-out setting and in the cases where all available data from all species were used (full model). As expected, the full model was more accurate, followed by the leave-one-allele-out model and then the leave-one-species-out model (Figure 5D). This was consistent with the results reported in Figures S4C and S4D. Still, AUC values were better than random when the predictor was not trained on any data from a given species, demonstrating that MHC-II ligand predictions can be extrapolated to other species, although with some loss in prediction accuracy (Figure 5D).

We then benchmarked the predictions for CD4$^+$ T cell epitopes, using data coming from IEDB[33] (see STAR Methods). Even though many of these epitopes had been selected based on exiting MHC-II ligand predictors (mainly NetMHCIIpan), we observed that the predictions of MixMHC2pred were more accurate than those of other tools (Figures 5E and 5F; Table S5D). We also explored the impact of the multiple specificity representing the reverse-binding mode in predictions of CD4$^+$ T cell epitopes. Our results show that MixMHC2pred was better at predicting HLA-DP CD4$^+$ T cell epitopes than a model including only canonical binders (Figure S6C).

Multiple specificities reveal reverse-binding CD4$^+$ T cell epitopes

To capitalize on the ability of our tool to model multiple binding specificities of MHC-II alleles, we scanned common viral and bacterial proteomes and selected 39 peptides that were predicted to follow only the reverse-binding mode of HLA-DPA1*02:01-DPB1*01:01 (Figure 6A; see STAR Methods). Binding-competition assays validated the binding of 26 of these peptides (Figure 3C; Table S4B). We then stimulated CD4$^+$ T cells isolated from peripheral blood mononuclear cells (PBMCs) of two HLA-DPA1*02:01-DPB1*01:01$^+$ donors using pools of these peptides and measured cytokine production. After deconvolving the responses, we could identify seven peptides eliciting TNF-α and IFNγ production (Figure 6B; Table S6A). We then built peptide-MHC-II multimers with four of those peptides. A clear multimer$^+$ population was found for each epitope, demonstrating that the responses originated from the peptides bound to HLA-DPA1*02:01-DPB1*01:01 (Figure 6C; STAR Methods). To gain insights into the clonality of these reactive CD4$^+$ T cells, we sequenced their T cell receptor (TCR). Oligoclonal responses were observed for each epitope (Table S6B), including a quasi-monoclonal recognition of the Epstein-Barr virus (EBV) epitope GELALTMRSKKLPIN (with a single TCRα and a dominating TCRβ). To investigate whether these TCR could be found in other donors, the alpha and beta-chain sequences were used to query separately TCRα and TCRβ repertoires through the iReceptor web platform[46] (see STAR Methods). Most of the alpha and beta CDR3 sequences of our study have been already observed in other donors, including 13 cases (out of 32) with exactly the same CDR3, V and J genes (Table S6B). Overall, these observations show that TCR chains recognizing reverse-binding epitopes can be found in multiple human TCR repertoires.

Next, we investigated whether these epitopes could have elicited a memory response. To this end, CD4$^+$ T cells from the two donors were stained directly *ex vivo* (i.e., without any prior stimulation) using the four multimers validated above. For one donor we could observe a direct *ex vivo* response that was mediated by effector and effector memory CD4$^+$ T cells (Figure 6D). This suggests that reverse-binding MHC-II ligands can elicit natural CD4$^+$ T cell recognition. Reverse-binding epitopes identified in this work had poor scores (% rank > 20) with other MHC-II ligand predictors or when considering only canonical binding orientation in our model (Table S6A). This demonstrates how thorough analysis of large datasets of MHC-II ligands, together with machine-learning algorithms, can improve and expand the scope of CD4$^+$ T cell epitope predictions.

DISCUSSION

CD4$^+$ T cells and MHC-II alleles play a central role in the immune recognition of infected or malignant cells and have been linked with multiple autoimmune diseases.[47,48] Here we capitalized on both public and in-house MHC-II peptidomics data to derive accurate MHC-II motifs for a large panel of MHC-II alleles (see also our MHC Motif Atlas [http://mhcmotifatlas.org][18]) and improve predictions of CD4$^+$ T cell epitopes. The fact that the different classes of binding specificities could be rationalized in terms of molecular interactions with residues in the different binding pockets provides an independent validation of our MHC-II motifs. Some of the correlation patterns between the MHC-II binding motifs and MHC-II binding pockets were already reported in structural analyses of a limited number of MHC-II alleles[35,38] and were used in the definition of MHC-II supertypes.[37] However, several other observations are specific to this work. For instance, HLA-DPB1*01:01 and HLA-DPB1*04:02 were assigned to the same supertype,[37] while both their specificity at P1 (K or R, respectively F, L, Y or I) and their P1 binding pockets (D, respectively G at 84β) are clearly different. This mainly reflects the limited numbers (<30 alleles) and lower resolution of MHC-II motifs used to define MHC-II supertypes. As such, our results provide both a refined view of the main MHC-II binding

Figure 6. Multiple specificities reveal reverse-binding CD4$^+$ T cell epitopes

(A) Schematics of the search strategy for reverse-binding epitopes restricted to HLA-DPA1*02:01-DPB1*01:01. CD4$^+$ T cells from HLA-DPA1*02:01-DPB1*01:01$^+$ donors were then stimulated with the selected peptides. Responses were evaluated through cytokine release assays deconvolving the response to the individual peptides.

(B) TNF-α and IFNγ response of the positive viral and bacterial peptides observed after deconvolving the response in two HLA-DPA1*02:01-DPB1*01:01$^+$ donors. (The predicted binding core of the peptide is underlined; EBV, Epstein-Barr virus; HCMV, human cytomegalovirus; HSV-1, herpes simplex virus 1; TT, tetanus toxin protein.)

(C) Validation with peptide-MHC-II multimers of the reverse-binding epitopes (negative controls based on irrelevant HLA-matched donors).

(D) FACS results directly on *ex vivo* CD4$^+$ T cells of donor *1G6I*, showing that recognition of the epitope NHELTLWNEARKLNP happens through effector and effector memory CD4$^+$ T cells, not naive CD4$^+$ T cells.

See also Table S6.

specificities, in line with the recent classification proposed for HLA-DP alleles,[49,50] and a robust incorporation of these observations into an accurate MHC-II ligand prediction tool (MixMHC2pred-2.0).

Multiple specificities were especially frequent in HLA-DP alleles, and we anticipate that a few multiple specificities may have been missed with our motif deconvolution approach for MHC-II alleles with fewer ligands (e.g., HLA-DPA1*02:01-DPB1*11:01 and HLA-DRB1*08:03). These observations were consistent with the previously reported multiple specificity of HLA-DPA1*02:02-DPB1*05:01.[26] For HLA-DP alleles, our work demonstrated that the two different motifs correspond to two different binding modes (i.e., canonical and reverse) of HLA-II ligands. We do not exclude that reverse binders may also be found among the ligands of other alleles, similar to the CLIP peptides observed to bind in both orientations to HLA-DRB1*01:01.[11] However, we could never detect multiple motifs with the same type of symmetry for HLA-DR or HLA-DQ alleles. Moreover, most alleles do not have the same specificity at P1 and P9, or at P4 and P6. For this reason, we can rule out the hypothesis that reverse binders would fit the same motif as canonical binders, which would make them difficult to detect by sequence analysis only. This absence of detectable reverse-binding ligands in HLA-DR (and HLA-DQ) could explain why the earlier observations of reverse binders based on some very specific peptides[11,12] have not been followed by other similar observations. Furthermore, the reverse-binding mode was not captured by existing predictors and was not particularly anticipated from a structural point of view, since several contacts between MHC-IIs and their ligands are mediated by the backbone atoms of the ligands and would not necessarily be conserved in the reverse orientation. This demonstrates the power of unbiased analyses of MHC-II ligands to unravel properties of MHC-II alleles.

Within humans, our work revealed that MHC-II binding motifs and MHC-II ligands could be accurately predicted even for alleles without known ligands. This is an important improvement of MixMHC2pred-2.0 compared with MixMHC2pred-1.2 and NeonMHC2, which were not applicable to most patients and could not be included in our benchmarks. When attempting to make predictions in species without any information about MHC-II ligands (i.e., leave-one-species-out cross-validation), we observed lower, though not random, accuracy. This is likely a limitation of all pan-allele MHC-II ligand predictors, which should be used with care in distant species like fish or birds. For instance, most tools would not be able to predict the 10-mer binding core of the chicken allele Gaga-BLB2*002:01, if not explicitly trained on these data.

Altogether, our work provides a high-quality dataset of MHC-II ligands; precise definition of MHC-II binding motifs; refined understanding of the molecular determinants of these motifs, including a widespread reverse-binding mode of HLA-DP ligands; and improved machine-learning predictions of $CD4^+$ T cell epitopes. The fact that the viral epitopes following the reverse-binding mode and eliciting responses in effector memory $CD4^+$ T cells could not have been identified with other MHC-II ligand predictors demonstrates the promise of machine-learning algorithms like MixMHC2pred to better characterize and expand the repertoire of $CD4^+$ T cell epitopes. The improved accuracy of $CD4^+$ T cell epitope predictions may contribute to accelerating personalized immunotherapy approaches in autoimmune diseases or cancer.

Limitations of the study

Our study demonstrated the presence of reversely bound HLA-DP ligands but did not observe a similar binding mode for alleles from other gene loci. We cannot exclude that such binding may happen for some other alleles, but that these ligands were too rare to reach the detection limit from our framework. Enhanced methods specifically searching for such ligands may be able to find some already in the currently available samples, or new experimental methods targeting these may be needed. Additional studies will further be essential to better characterize the impact of the reverse-binding mode on T cell recognition.

Regarding the predictions, our model incorporated steps of antigen processing and presentation as well as of peptide binding, thanks to the MHC-II peptidomics data used for training our model. This enabled accurate predictions of MHC-II ligands for most alleles. However, we cannot exclude the presence of some technical biases in such data, for instance, linked to mass spectrometry experiments.[20] As expected, accuracy was lower for predictions of $CD4^+$ T cell epitopes. Integrating additional features of peptide presentation or T cell recognition may therefore help to further strengthen epitope predictions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell lines
  - Patient material
- METHOD DETAILS
  - Curation of published MHC-II peptidomics data
  - HLA typing
  - Generation of antibody-crosslinked beads
  - Purification of HLA-II bound peptides
  - LC–MS/MS analyses of HLA-II peptides
  - Peptide identification
  - Deconvolution and annotation of MHC-II motifs
  - MHC-II sequences retrieval and alignment
  - Motifs clustering
  - Analysis of published MHC-II structures
  - Structural modeling
  - Production of HLA-DPA1*02:01-DPB1*01:01
  - Binding competition assays
  - Protein crystallization and structure determination
  - Motifs of the N- and C-terminal contexts
  - Simulation of reverse binding ligands
  - Pan-allele MHC-II ligand predictor development
  - Benchmarking MHC-II binding specificity predictions
  - Benchmarking MHC-II ligand predictions

- Proportion of reverse ligands in different samples
- Analysis of cis- and trans- heterodimers
- Benchmarking CD4$^+$ T-cell epitope predictions
- Selection of candidate epitopes following the reverse binding mode
- Peptides and peptide-MHC-II multimers
- Identification of antigen-specific CD4$^+$ T-cell responses
- Sorting of naive and effector memory CD4$^+$ T cells
- Peptide-MHC-II multimer validation and sorting of CD4$^+$ T cells
- Bulk TCR sequencing
- Analysis of TCR sequences

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.immuni.2023.03.009.

ACKNOWLEDGMENTS

We thank Camilla Jandus from the University of Geneva for providing us a patient-derived PBMC sample. We thank the Protein Modeling Unit of the University of Lausanne for the support in structural bioinformatics. We thank Peter A. van Veelen, Michel G.D. Kester, and their colleagues[26]; Ana Marcu and her colleagues[51]; and Birkir Reynisson, Morten Nielsen, and their colleagues[32] for sharing MHC-II peptidomics data related to their published studies. D.G. and J.R. acknowledge support from the Swiss Cancer Research Foundation (KFS-4961-02-2020). G. Croce is supported by the Marie-Curie Fellowship (H2020-MSCA-IF-2020, no. 101027973).

AUTHOR CONTRIBUTIONS

Conceptualization, J.R. and D.G.; methodology, J.R.; software, J.R.; investigation, J.R., P.G., J.S., J.M., A.L., K.L., M.A.S.P., G. Croce, R.G., F.P., and D.G.; resources, G. Coukos, V.Z., F.P., M.B.-S., A.H., and D.G.; data curation, J.R.; writing – original draft, J.R. and D.G.; writing – review & editing, J.R. and D.G.; visualization, J.R., J.S., and D.G.; supervision, J.R., G. Coukos, V.Z., F.P., M.B.-S., A.H., and D.G.; funding acquisition, D.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 24, 2022
Revised: November 9, 2022
Accepted: March 15, 2023
Published: April 5, 2023

REFERENCES

1. Alspach, E., Lussier, D.M., Miceli, A.P., Kizhvatov, I., DuPage, M., Luoma, A.M., Meng, W., Lichti, C.F., Esaulova, E., Vomund, A.N., et al. (2019). MHC-II neoantigens shape tumour immunity and response to immunotherapy. Nature *574*, 696–701.
2. Borst, J., Ahrends, T., Bąbała, N., Melief, C.J.M., and Kastenmüller, W. (2018). CD4+ T cell help in cancer immunology and immunotherapy. Nat. Rev. Immunol. *18*, 635–647. https://doi.org/10.1038/s41577-018-0044-0.
3. Hu, Z., Leet, D.E., Allesøe, R.L., Oliveira, G., Li, S., Luoma, A.M., Liu, J., Forman, J., Huang, T., Iorgulescu, J.B., et al. (2021). Personal neoantigen vaccines induce persistent memory T cell responses and epitope spreading in patients with melanoma. Nat. Med. *27*, 1–11.
4. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature *547*, 217–221. https://doi.org/10.1038/nature22991.
5. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrörs, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature *547*, 222–226.
6. Tran, E., Turcotte, S., Gros, A., Robbins, P.F., Lu, Y.-C., Dudley, M.E., Wunderlich, J.R., Somerville, R.P., Hogan, K., Hinrichs, C.S., et al. (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science *344*, 641–645. https://doi.org/10.1126/science.1251102.
7. Zacharakis, N., Chinnasamy, H., Black, M., Xu, H., Lu, Y.-C., Zheng, Z., Pasetto, A., Langhan, M., Shelton, T., Prickett, T., et al. (2018). Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. Nat. Med. *24*, 724–730.
8. Neefjes, J., Jongsma, M.L.M., Paul, P., and Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat. Rev. Immunol. *11*, 823–836. https://doi.org/10.1038/nri3084.
9. Holland, C.J., Cole, D.K., and Godkin, A. (2013). Re-directing CD4+ T cell responses with the flanking residues of MHC class II-bound peptides: the core is not enough. Front. Immunol. *4*, 172. https://doi.org/10.3389/fimmu.2013.00172.
10. Halabi, S., Ghosh, M., Stevanović, S., Rammensee, H.-G., Bertzbach, L.D., Kaufer, B.B., Moncrieffe, M.C., Kaspers, B., Härtle, S., and Kaufman, J. (2021). The dominantly expressed class II molecule from a resistant MHC haplotype presents only a few Marek's disease virus peptides by using an unprecedented binding motif. PLoS Biol. *19*, e3001057.
11. Günther, S., Schlundt, A., Sticht, J., Roske, Y., Heinemann, U., Wiesmüller, K.-H., Jung, G., Falk, K., Rötzschke, O., and Freund, C. (2010). Bidirectional binding of invariant chain peptides to an MHC class II molecule. Proc. Natl. Acad. Sci. USA *107*, 22219–22224.
12. Schlundt, A., Günther, S., Sticht, J., Wieczorek, M., Roske, Y., Heinemann, U., and Freund, C. (2012). Peptide linkage to the α-subunit of MHCII creates a stably inverted antigen presentation complex. J. Mol. Biol. *423*, 294–302.
13. Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., and Marsh, S.G.E. (2020). IPD-IMGT/HLA database. Nucleic Acids Res. *48*, D948–D955.
14. Unanue, E.R., Turk, V., and Neefjes, J. (2016). Variations in MHC Class II antigen processing and presentation in health and disease. Annu. Rev. Immunol. *34*, 265–297.
15. Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. Immunity *51*. 766-779.e17.
16. Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., et al. (2019). Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat. Biotechnol. *37*, 1283–1286.
17. Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., and Nielsen, M. (2020). Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J. Proteome Res. *19*, 2304–2315.
18. Tadros, D.M., Eggenschwiler, S., Racle, J., and Gfeller, D. (2023). The MHC Motif Atlas: a database of MHC binding specificities and ligands. Nucleic Acids Res. *51*, D428–D437. https://doi.org/10.1093/nar/gkac965.
19. Chen, B., Khodadoust, M.S., Olsson, N., Wagar, L.E., Fast, E., Liu, C.L., Muftuoglu, Y., Sworder, B.J., Diehn, M., Levy, R., et al. (2019). Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol. *37*, 1–12.
20. Gfeller, D., Liu, Y., and Racle, J. (2023). Contemplating immunopeptidomes to better predict them. Semin. Immunol. *66*, 101708. https://doi.org/10.1016/j.smim.2022.101708.

21. Barra, C., Alvarez, B., Paul, S., Sette, A., Peters, B., Andreatta, M., Buus, S., and Nielsen, M. (2018). Footprints of antigen processing boost MHC class II natural ligand predictions. Genome Med. *10*, 84.

22. Ciudad, M.T., Sorvillo, N., Alphen, F.P. van, Catalán, D., Meijer, A.B., Voorberg, J., and Jaraquemada, D. (2017). Analysis of the HLA-DR peptidome from human dendritic cells reveals high affinity repertoires and nonconventional pathways of peptide generation. J. Leukoc. Biol. *101*, 15–27. https://doi.org/10.1189/jlb.6HI0216-069R.

23. Falk, K., Rötzschke, O., Stevanovíc, S., Jung, G., and Rammensee, H.-G. (1994). Pool sequencing of natural HLA-DR, DQ, and DP ligands reveals detailed peptide motifs, constraints of processing, and general rules. Immunogenetics *39*, 230–242.

24. Ramarathinam, S.H., Ho, B.K., Dudek, N.L., and Purcell, A.W. (2021). HLA class II immunopeptidomics reveals that co-inherited HLA-allotypes within an extended haplotype can improve proteome coverage for immunosurveillance. Proteomics *21*, e2000160. https://doi.org/10.1002/pmic.202000160.

25. Shao, X.M., Bhattacharya, R., Huang, J., Sivakumar, I.K.A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., et al. (2020). High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. Cancer Immunol. Res. *8*, 396–408.

26. Balen, P. van, Kester, M.G.D., Klerk, W. de, Crivello, P., Arrieta-Bolaños, E., Ru, A.H. de, Jedema, I., Mohammed, Y., Heemskerk, M.H.M., Fleischhauer, K., et al. (2020). Immunopeptidome analysis of HLA-DPB1 allelic variants reveals new functional hierarchies. J. Immunol. *204*, 3273–3282. https://doi.org/10.4049/jimmunol.2000192.

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

28. Ritz, D., Sani, E., Debiec, H., Ronco, P., Neri, D., and Fugmann, T. (2018). Membranal and blood-soluble HLA Class II peptidome analyses using data-dependent and independent acquisition. Proteomics *18*, e1700246. https://doi.org/10.1002/pmic.201700246.

29. Draheim, M., Wlodarczyk, M.F., Crozat, K., Saliou, J.-M., Alayi, T.D., Tomavo, S., Hassan, A., Salvioni, A., Demarta-Gatsi, C., Sidney, J., et al. (2017). Profiling MHC II immunopeptidome of blood-stage malaria reveals that cDC1 control the functionality of parasite-specific CD4 T cells. EMBO Mol. Med. *9*, 1605–1621.

30. Sofron, A., Ritz, D., Neri, D., and Fugmann, T. (2016). High-resolution analysis of the murine MHC class II immunopeptidome. Eur. J. Immunol. *46*, 319–328. https://doi.org/10.1002/eji.201545930.

31. Wan, X., Vomund, A.N., Peterson, O.J., Chervonsky, A.V., Lichti, C.F., and Unanue, E.R. (2020). The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides. Nat. Immunol. *21*, 1–9.

32. Fisch, A., Reynisson, B., Benedictus, L., Nicastri, A., Vasoya, D., Morrison, I., Buus, S., Ferreira, B.R., Santos, I.K.F. de M., Ternette, N., et al. (2021). Integral use of immunopeptidomics and immunoinformatics for the characterization of antigen presentation and rational identification of BoLA-DR–Presented peptides and epitopes. J. Immunol. *206*, 2489–2497.

33. Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The immune epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

34. Kaabinejadian, S., Barra, C., Alvarez, B., Yari, H., Hildebrand, W.H., and Nielsen, M. (2022). Accurate MHC motif deconvolution of immunopeptidomics data reveals a significant contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome. Front. Immunol. *13*, 835454.

35. Davenport, M.P., Quinn, C.L., Chicz, R.M., Green, B.N., Willis, A.C., Lane, W.S., Bell, J.I., and Hill, A.V. (1995). Naturally processed peptides from two disease-resistance-associated HLA-DR13 alleles show related sequence motifs and the effects of the dimorphism at position 86 of the HLA-DR beta chain. Proc. Natl. Acad. Sci. USA *92*, 6567–6571. https://doi.org/10.1073/pnas.92.14.6567.

36. Verreck, F.A.W., van de Poel, A., Drijfhout, J.W., Amons, R., Coligan, J.E., and Konig, F. (1996). Natural peptides isolated from Gly86/Val86-containing variants of HLA-DR1,-DR 11, -DR13, and -DR52. Immunogenetics *43*, 392–397. https://doi.org/10.1007/BF02199809.

37. Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B., and Sette, A. (2011). Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics *63*, 325–335.

38. Kusano, S., Kukimoto-Niino, M., Satta, Y., Ohsawa, N., Uchikubo-Kamo, T., Wakiyama, M., Ikeda, M., Terada, T., Yamamoto, K., Nishimura, Y., et al. (2014). Structural basis for the specific recognition of the major antigenic peptide from the Japanese cedar pollen allergen cry j 1 by HLA-DP5. J. Mol. Biol. *426*, 3016–3027.

39. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382–W388. https://doi.org/10.1093/nar/gki387.

40. Paul, S., Karosiene, E., Dhanda, S.K., Jurtz, V., Edwards, L., Nielsen, M., Sette, A., and Peters, B. (2018). Determination of a predictive cleavage motif for eluted major histocompatibility complex Class II ligands. Front. Immunol. *9*, 1795.

41. Bird, P.I., Trapani, J.A., and Villadangos, J.A. (2009). Endolysosomal proteases and their inhibitors in immunity. Nat. Rev. Immunol. *9*, 871–882. https://doi.org/10.1038/nri2671.

42. Sercarz, E.E., and Maverakis, E. (2003). MHC-guided processing: binding of large antigen fragments. Nat. Rev. Immunol. *3*, 621–629.

43. Petersdorf, E.W., Bengtsson, M., Horowitz, M., McKallor, C., Spellman, S.R., Spierings, E., Gooley, T.A., and Stevenson, P.; International Histocompatibility Working Group in Hematopoietic Cell Transplantation (2022). HLA-DQ heterodimers in hematopoietic cell transplantation. Blood *139*, 3009–3017. https://doi.org/10.1182/blood.2022015860.

44. Tollefsen, S., Hotta, K., Chen, X., Simonsen, B., Swaminathan, K., Mathews, I.I., Sollid, L.M., and Kim, C.Y. (2012). Structural and functional studies of trans-encoded HLA-DQ2.3 (DQA1*03:01/DQB1*02:01) protein molecule. J. Biol. Chem. *287*, 13611–13619. https://doi.org/10.1074/jbc.M111.320374.

45. Nilsson, J.B., Kaabinejadian, S., Yari, H., Peters, B., Barra, C., Gragert, L., Hildebrand, W., and Nielsen, M. (2022). Machine learning reveals limited contribution of trans-only encoded variants to the HLA-DQ immunopeptidome by accurate and comprehensive HLA-DQ antigen presentation prediction. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.14.507934.

46. Corrie, B.D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F.M.W., Christley, S., Scott, J.K., et al. (2018). iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol. Rev. *284*, 24–41. https://doi.org/10.1111/imr.12666.

47. Latorre, D., Kallweit, U., Armentani, E., Foglierini, M., Mele, F., Cassotta, A., Jovic, S., Jarrossay, D., Mathis, J., Zellini, F., et al. (2018). T cells in patients with narcolepsy target self-antigens of hypocretin neurons. Nature *562*, 63–68.

48. Matzaraki, V., Kumar, V., Wijmenga, C., and Zhernakova, A. (2017). The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. *18*, 76. https://doi.org/10.1186/s13059-017-1207-1.

49. Laghmouchi, A., Kester, M.G.D., Hoogstraten, C., Hageman, L., de Klerk, W., Huisman, W., Koster, E.A.S., de Ru, A.H., van Balen, P., Klobuch, S., et al. (2022). Promiscuity of peptides Presented in HLA-DP Molecules from Different Immunogenicity Groups Is Associated With T-Cell Cross-Reactivity. Front. Immunol. *13*, 831822. https://doi.org/10.3389/fimmu.2022.831822.

50. Meurer, T., Crivello, P., Metzing, M., Kester, M., Megger, D.A., Chen, W., van Veelen, P.A., van Balen, P., Westendorf, A.M., Homa, G., et al. (2021). Permissive HLA-DPB1 mismatches in HCT depend on immunopeptidome divergence and editing by HLA-DM. Blood *137*, 923–928. https://doi.org/10.1182/blood.2020008464.

51. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D.J., Freudenmann, L.K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke,

M., Wagner, P., et al. (2021). HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J. Immunother. Cancer *9*, e002071. https://doi.org/10.1136/jitc-2020-002071.

52. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372. https://doi.org/10.1038/nbt.1511.

53. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647. https://doi.org/10.1093/bioinformatics/btx469.

54. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539. https://doi.org/10.1038/msb.2011.75.

55. Webb, B., and Sali, A. (2016). Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinform. *54*, 5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3.

56. Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132. https://doi.org/10.1107/S0907444909047337.

57. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. Sect. J. Struct. Biol. *75*, 861–877.

58. Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132. https://doi.org/10.1107/S0907444904019158.

59. Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., Bolotin, D.A., Staroverov, D.B., Putintseva, E.V., Plevova, K., et al. (2014). Towards error-free profiling of immune repertoires. Nat. Methods *11*, 653–655. https://doi.org/10.1038/nmeth.2960.

60. UniProt Consortium (2021). UniProt: the universal protein KnowledgeBase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

61. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552. https://doi.org/10.1093/nar/gkab1038.

62. Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980. https://doi.org/10.1038/nsb1203-980.

63. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210. https://doi.org/10.1093/nar/30.1.207.

64. Deutsch, E.W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J.J., Kundu, D.J., García-Seisdedos, D., Jarnuczak, A.F., Hewapathirana, S., Pullman, B.S., et al. (2020). The ProteomeXchange Consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res. *48*, D1145–D1152. https://doi.org/10.1093/nar/gkz984.

65. Cassotta, A., Paparoditis, P., Geiger, R., Mettu, R.R., Landry, S.J., Donati, A., Benevento, M., Foglierini, M., Lewis, D.J.M., Lanzavecchia, A., et al. (2020). Deciphering and predicting CD4+ T cell immunodominance of influenza virus hemagglutinin. J. Exp. Med. *217*, e20200206.

66. Clement, C.C., Becerra, A., Yin, L., Zolla, V., Huang, L., Merlin, S., Follenzi, A., Shaffer, S.A., Stern, L.J., and Santambrogio, L. (2016). The dendritic cell major histocompatibility complex II (MHC II) peptidome derives from a variety of processing pathways and includes peptides with a broad spectrum of HLA-DM sensitivity. J. Biol. Chem. *291*, 5576–5595. https://doi.org/10.1074/jbc.M115.655738.

67. Collado, J.A., Alvarez, I., Ciudad, M.T., Espinosa, G., Canals, F., Pujol-Borrell, R., Carrascal, M., Abian, J., and Jaraquemada, D. (2013). Composition of the HLA-DR-associated human thymus peptidome. Eur. J. Immunol. *43*, 2273–2282. https://doi.org/10.1002/eji.201243280.

68. Dheilly, E., Battistello, E., Katanayeva, N., Sungalee, S., Michaux, J., Duns, G., Wehrle, S., Sordet-Dessimoz, J., Mina, M., Racle, J., et al. (2020). Cathepsin S regulates antigen processing and T cell activity in non-Hodgkin lymphoma. Cancer Cell *37*, 674–689.e12. 10/ggvwds.

69. Forlani, G., Michaux, J., Pak, H., Huber, F., Joseph, E.L.M., Ramia, E., Stevenson, B.J., Linnebacher, M., Accolla, R.S., and Bassani-Sternberg, M. (2021). CIITA-transduced glioblastoma cells uncover a rich repertoire of clinically relevant tumor-associated HLA-II antigens. Mol. Cell. Proteomics *20*, 100032.

70. Garde, C., Ramarathinam, S.H., Jappe, E.C., Nielsen, M., Kringelum, J.V., Trolle, T., and Purcell, A.W. (2019). Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. Immunogenetics *71*, 445–454.

71. Goncalves, G., Mullan, K.A., Duscharla, D., Ayala, R., Croft, N.P., Faridi, P., and Purcell, A.W. (2021). IFNγ modulates the immunopeptidome of triple negative breast cancer cells by enhancing and diversifying antigen processing and presentation. Front. Immunol. *12*, 645770.

72. Graciotti, M., Marino, F., Pak, H., Baumgaertner, P., Thierry, A.C., Chiffelle, J., Perez, M.A.S., Zoete, V., Harari, A., Bassani-Sternberg, M., et al. (2020). Deciphering the mechanisms of improved immunogenicity of hypochlorous acid-treated antigens in anti-cancer dendritic cell-based vaccines. Vaccines *8*, 271.

73. Kalaora, S., Nagler, A., Nejman, D., Alon, M., Barbolin, C., Barnea, E., Ketelaars, S.L.C., Cheng, K., Vervier, K., Shental, N., et al. (2021). Identification of bacteria-derived HLA-bound peptides in melanoma. Nature *592*, 1–6.

74. Khodadoust, M.S., Olsson, N., Wagar, L.E., Haabeth, O.A.W., Chen, B., Swaminathan, K., Rawson, K., Liu, C.L., Steiner, D., Lund, P., et al. (2017). Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. Nature *543*, 723–727.

75. Marino, F., Semilietof, A., Michaux, J., Pak, H.-S., Coukos, G., Müller, M., and Bassani-Sternberg, M. (2020). Biogenesis of HLA ligand presentation in immune cells upon activation reveals changes in peptide length preference. Front. Immunol. *11*, 1981.

76. Nelde, A., Kowalewski, D.J., Backert, L., Schuster, H., Werner, J.-O., Klein, R., Kohlbacher, O., Kanz, L., Salih, H.R., Rammensee, H.-G., et al. (2018). 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.

77. Newey, A., Griffiths, B., Michaux, J., Pak, H.S., Stevenson, B.J., Woolston, A., Semiannikova, M., Spain, G., Barber, L.J., Matthews, N., et al. (2019). Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. J. Immunother. Cancer *7*, e000440.

78. Ooi, J.D., Petersen, J., Tan, Y.H., Huynh, M., Willett, Z.J., Ramarathinam, S.H., Eggenhuizen, P.J., Loh, K.L., Watson, K.A., Gan, P.Y., et al. (2017). Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. Nature *545*, 243–247. https://doi.org/10.1038/nature22329.

79. Ting, Y.T., Petersen, J., Ramarathinam, S.H., Scally, S.W., Loh, K.L., Thomas, R., Suri, A., Baker, D.G., Purcell, A.W., Reid, H.H., et al. (2018). The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. J. Biol. Chem. *293*, 3236–3251.

80. Wang, Q., Drouin, E.E., Yao, C., Zhang, J., Huang, Y., Leon, D.R., Steere, A.C., and Costello, C.E. (2017). 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. https://doi.org/10.1021/acs.jproteome.6b00386.

81. Maccari, G., Robinson, J., Ballingall, K., Guethlein, L.A., Grimholt, U., Kaufman, J., Ho, C.-S., de Groot, N.G., Flicek, P., Bontrop, R.E., et al. (2017). IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res. *45*, D860–D864. https://doi.org/10.1093/nar/gkw1050.

82. Afrache, H., Tregaskes, C.A., and Kaufman, J. (2020). A potential nomenclature for the Immuno Polymorphism Database (IPD) of chicken MHC genes: progress and problems. Immunogenetics *72*, 9–24.

83. Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics *20*, 1388–1397. https://doi.org/10.1093/bioinformatics/bth100.

84. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

85. Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. *45*, D271–D281. https://doi.org/10.1093/nar/gkw1000.

86. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA *89*, 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

87. Ritz, C., Baty, F., Streibig, J.C., and Gerhard, D. (2015). Dose-response analysis using R. PLoS One *10*, e0146021. https://doi.org/10.1371/journal.pone.0146021.

88. Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics *25*, 1293–1299. https://doi.org/10.1093/bioinformatics/btp137.

89. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., et al. (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One *2*, e796. https://doi.org/10.1371/journal.pone.0000796.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-pan-HLA-II (IVA12)	ATCC	Cat# HB-145; RRID: CVCL_G223
Anti-HLA-DR (LB3.1)	ATCC	Cat# HB-298; RRID: CVCL_G667
Anti-HLA-DP	Leinco Technologies	Cat# H127; RRID: AB_2737511
Anti-HLA-DQ	Biorad	Cat# MCA379G; RRID: AB_322104
Anti-HLA-DQ	MyBioSource	Cat# MBS570226
Anti-CD4	BD	Cat# 562970; RRID: AB_2744424
Anti-IFNγ	BD	Cat# 554702; RRID: AB_398580
Anti-TNFα	BD	Cat# 340512; RRID: AB_400435
Anti-CCR7	BioLegend	Cat# 353227; RRID: AB_11219587
Anti-CD45RA	BioLegend	Cat# 304108; RRID: AB_314412
Biological samples		
PBMC from healthy donor, melanoma patient and snap-frozen meningioma tissues	Biobank of the Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.	Protocol F-25/99, 2017-00305, and F-42/92
Chemicals, peptides, and recombinant proteins		
HLA-DPA1*02:01-DPB1*01:01	Peptide & Tetramer Core Facility, University of Lausanne	N/A
Peptides from various viral and bacterial origins	Peptide & Tetramer Core Facility, University of Lausanne	N/A
RPMI	GIBCO	61870-010
MEM NEAA	GIBCO	11140-035
2-Mercaptoethanol	GIBCO	31350-010
Sodium Pyruvat	GIBCO	11360-033
HEPES	BioConcept	5-31F00H
Pen/Strep	BioConcept	4-01F00H
Human Serum	Biowest	S419H-100
FBS	Biowest	S-1810-500
IL2	Novartis	PZN02238131
Near-IR Dead staining kit	ThermoFischer	L10119
Fix/Perm	BioLegend	426803
PMA/ionomycin	ThermoFischer	00-4975-93
Protein Transport Inhibitor	eBiosciences	00-4980-93
DAPI	Sigma-Aldrich	10236276001
Critical commercial assays		
DNeasy kit	Qiagen	69504
TruSight HLA v2 Sequencing Panel kit	CareDx	#20000215
Dynabeads mRNA DIRECT purification kit	ThermoFisher	61011
MessageAmp II aRNA Amplification Kit	Ambion	AMB17515
CD4 T cell isolation kit	Miltenyi	130-045-101
Deposited data		
MHC-II peptidomics data	This paper	PRIDE: PXD034773
MHC-II peptidomics data from other studies	Multiple studies	See Table S1 for the list of studies
X-ray structure of a peptide bound to HLA-DP in canonical orientation	This paper	PDB: 7ZAK

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
X-ray structure of a peptide bound to HLA-DP in reverse orientation	This paper	PDB: 7ZFR
TCR sequencing data	This paper	GEO: GSE205588
Experimental models: Cell lines		
JY	ATCC	77441
CM467 and RA957	In house	N/A
Software and algorithms		
MixMHC2pred-2.0	This paper	https://doi.org/10.5281/zenodo.7737217; https://github.com/GfellerLab/MixMHC2pred
Assign TruSight HLA v.2.1	CareDx	https://labproducts.caredx.com/software/assign/trusight
MaxQuant platform v.1.5.5.1	Cox and Mann[52]	https://www.maxquant.org/; RRID: SCR_014485
MoDec v1.2	Racle et al.[16]	https://github.com/GfellerLab/MoDec
ggseqlogo	Wagih[53]	https://github.com/GfellerLab/ggseqlogo
Clustal Omega v.1.2.2	Sievers et al.[54]	http://www.clustal.org/omega/
PyMOL v.2.3.3	Schrödinger	https://pymol.org/; RRID: SCR_000305
Modeller software v10.1	Webb and Sali[55]	https://salilab.org/modeller/; RRID: SCR_008395
FoldX v5.0	Schymkowitz et al.[39]	https://foldxsuite.crg.eu/; RRID: SCR_008522
XDS program package	Kabsch[56]	https://xds.mr.mpg.de/; RRID: SCR_015652
drc v.3.0.1	R package	https://cran.r-project.org/web/packages/drc/
Phenix Suite v1.19.2-4158	Liebschner et al.[57]	http://www.phenix-online.org/; RRID: SCR_014224
Coot v0.9.4	Emsley and Cowtan[58]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/; RRID: SCR_014222
Keras v2.7	R package	https://cran.r-project.org/package=keras
Tfdatasets v2.7	R package	https://cran.r-project.org/package=tfdatasets
NetMHCIIpan-4.0	Reynisson et al.[17]	https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-4.0
MHCnuggets v2.3.2	Shao et al.[25]	https://karchinlab.org/apps/appMHCnuggets.html
MARIA	Chen et al.[19]	https://maria.stanford.edu
FlowJo v10.7.1	BD	https://www.flowjo.com/; RRID: SCR_008520
MIGEC	Shugay et al.[59]	https://migec.readthedocs.io/en/latest/; RRID: SCR_016337
Other		
Human MHC-II sequences	IPD-IMGT/HLA	https://www.ebi.ac.uk/ipd/imgt/hla/; RRID: SCR_002971
MHC-II sequences from various species	IPD-MHC	https://www.ebi.ac.uk/ipd/mhc/; RRID: SCR_007749
UniProt database containing protein sequences	The UniProt Consortium[60]	https://www.uniprot.org/; RRID: SCR_002380
iReceptor web platform containing TCRα and TCRβ repertoires	Corrie et al.[46]	https://gateway.ireceptor.org/; RRID: SCR_022294

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to David Gfeller (david.gfeller@unil.ch).

Materials availability

Materials generated in this study are available upon request from the lead contact, David Gfeller (david.gfeller@unil.ch).

Data and code availability

- Mass spectrometry-based MHC-II peptidomics data generated for this study have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository[61] (PRIDE: PXD034773). The models of the crystal structures resolved in this work have been deposited in the worldwide Protein Data Bank[62] (PDB: 7ZAK [canonical binder] and PDB: 7ZFR [reverse binder]). TCR sequencing data have been deposited in NCBI's Gene Expression Omnibus[63] (GEO: GSE205588).
- MixMHC2pred-2.0 has been deposited at Zenodo (https://doi.org/10.5281/zenodo.7737217); it is also available at https://github.com/GfellerLab/MixMHC2pred and through a webserver http://mixmhc2pred.gfellerlab.org.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines

Epstein–Barr-virus-transformed human B-cell lines JY (ATCC,77441), CM467, RA957, (a gift from P. Romero (Ludwig Institute for Cancer Research Lausanne), were maintained in RPMI-1640+GlutaMAX medium (Life Technologies) supplemented with 10% heat-inactivated FBS (Dominique Dutscher) and 1% penicillin–streptomycin solution (BioConcept). Cells were grown to the required cell numbers, collected by centrifugation at 1,200 rpm for 5min, washed twice with ice cold PBS and stored as dry cell pellets at −20 °C until use.

Patient material

PBMCs from donor 1G6I, PBMCs from melanoma patient LAU52 and snap-frozen meningioma tissues from patients (3865-DM, 3947-GA, 4021) were obtained from the bio-bank of the Centre Hospitalier Universitaire Vaudois (CHUV, Lausanne, Switzerland). Informed consent of the participants was obtained following requirements of the Institutional Review Board (Ethics Commission, CHUV). Protocol F-25/99 has been approved by the local ethics committee and the biobank of the Lab of Brain Tumor Biology and Genetics. Protocol 2017-00305 for antigen and T cell discovery in tumors has been approved by the local ethics committee. Protocol F-42/92 has been approved by the local ethics committee.

METHOD DETAILS

Curation of published MHC-II peptidomics data

We searched in the literature and on ProteomeXchange (http://www.proteomexchange.org/)[64] for available high-throughput mass spectrometry-based MHC-II peptidomics datasets in which the MHC-II typing was also known. This led us to consider the following studies[10,15,16,19,24,26–32,51,65–80]: as well as the study PRIDE: PXD019466 that is available on ProteomeXchange but without any corresponding publication.

We downloaded the sequences of different reference proteomes from EMBL-EBI (https://ftp.ebi.ac.uk/pub/databases/reference_proteomes/): human (UP000005640_9606), mouse (UP000000589_10090), cattle (UP000009136_9913) and chicken (UP000000539_9031), obtaining both the canonical and additional sequences in fasta format (from release 2020_04). The peptides obtained from MHC-II peptidomics datasets were then mapped to these proteomes, in order to identify their protein of origin and determine their N- and C-terminal contexts (3 residues upstream of the peptide + 3 N-terminal residues of the peptide (N-terminal context); 3 C-terminal residues of the peptide + 3 residues downstream of the peptide (C-terminal context)).

HLA typing

Genomic DNA was extracted using DNeasy kit from Qiagen and 500ng of genomic DNA was used for the typing. High-resolution 4-digit HLA typing was performed with the TruSight HLA v2 Sequencing Panel kit from CareDx according to the manufacturer instruction. Briefly, class I and class II genes were amplified by PCR. Illumina adapters were added by tagmentation. After normalization and purification, the samples were sequenced on a MiSeq instrument (Illumina). Sequencing data were analyzed with the Assign TruSight HLA v.2.1 software (CareDx).

Generation of antibody-crosslinked beads

Anti-pan-HLA-II and anti-HLA-DR monoclonal antibodies were purified from the supernatant of HB145 (ATCC, HB-145) and HB298 cells (ATCC, HB-298), respectively, grown in CELLLine CL-1000 flasks (Sigma-Aldrich) using protein A-sepharose 4B beads (pro-A beads; Invitrogen) while anti-HLA-DP (Leinco Technologies) and anti-HLA-DQ (from either Biorad or MyBioSource) were purchased from the respective providers. Antibodies were cross-linked separately to pro-A beads at a concentration of 1 to 2 mg of antibodies per milliliter of beads following incubation with pro-A beads for 1h at room temperature. Chemical crosslinking was performed by addition of dimethyl pimelimidate dihydrochloride (Sigma-Aldrich) in 0.2M sodium borate buffer, pH 9 (Sigma-Aldrich) at a final concentration of 20mM for 30min. The reaction was quenched by incubation with 0.2M ethanolamine, pH 8 (Sigma-Aldrich) for 2h. Crosslinked antibodies were kept at 4 °C in PBS until use.

Purification of HLA-II bound peptides

Cells were lysed in PBS containing 0.25% sodium deoxycholate (Sigma-Aldrich), 0.2mM iodoacetamide (Sigma-Aldrich), 1mM EDTA, 1:200 protease inhibitors cocktail (SigmaAldrich), 1mM phenylmethylsulfonylfluoride (Roche) and 1% octyl-beta-dglucopyranoside (Sigma-Aldrich) at 4 °C for 1h. The lysis buffer was added to cells at a concentration of 1×10^8 cells per milliliter. Cell lysates were cleared by centrifugation with a table-top centrifuge (Eppendorf Centrifuge) at 4 °C at 14,200 rpm for 50min. Meningioma tissues were placed in tubes containing the same lysis buffer and homogenized on ice in three to five short intervals of 5 s each using an Ultra Turrax homogenizer (IKA) at maximum speed. For 1 g of tissue, 10–12ml of lysis buffer was required. Cell lysis was performed at 4 °C for 1h. Tissue lysates were cleared by centrifugation at 20,000 rpm in a high-speed centrifuge (Beckman Coulter, JSS15314) at 4 °C for 50min.

Tissue cleared lysates were loaded first on affinity purification columns (BioRad, 731-1550) containing pro-A beads (pre-clear column, to remove non-specific antibodies). Tissues and cells lysates were loaded sequentially on columns containing cross-linked beads in the following order: CM467 samples on anti-DQ, DP, DR, pan-HLA-II antibodies, 3865 and JY on anti-DR, DQ, DP, pan-HLA-II antibodies, and samples RA957, 3947-GA, 4021 on anti-DR,DP,DQ, pan-HLA-II antibodies. The affinity columns were then washed with 2 column volumes of 150 mM sodium chloride (NaCl; Carlo-Erba) in 20 mM Tris-HCl pH 8, 2 column volumes of 400 mM NaCl in 20 mM Tris-HCl pH 8, and again 2 column volumes of 150 mM sodium chloride in 20 mM Tris-HCl pH 8. Finally, the beads were washed in 1 column volume of 20 mM Tris-HCl pH 8. HLA complexes and the bound peptides were eluted by adding twice 1% trifluoroacetic acid (TFA) or 4 times Acetic acid 0.1N at a volume equivalent to or slightly higher than the volume of beads present in the column. HLA peptides were purified and concentrated with by loading into Sep-Pak tC18 96-well plates (Waters) preconditioned with 1 mL of 80% acetonitrile (ACN) in 0.1% TFA and then with 2 mL of 0.1% TFA. The C18 wells were then washed with 2 mL of 0.1% TFA. The HLA peptides were eluted with 500 μL of 32% ACN in 0.1% TFA into Eppendorf tubes. Recovered peptides were dried using vacuum centrifugation (Thermo Fisher Scientific) and stored at −20°C.

LC–MS/MS analyses of HLA-II peptides

HLA-II, HLA-DR, HLA-DP and HLA-DQ peptide samples were resuspended in 10 μl of 2% ACN in 0.1% formic acid (FA) and aliquots of 3 μl were used for each MS analysis. The LC-MS/MS system consisted of an Easy-nLC 1200 (Thermo Fisher Scientific) coupled to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). Peptides were separated on a 450-mm analytical column (8-μm tip, 75-μm inner diameter, PicoTip emitter, New Objective) packed with ReproSil-Pur C18 (1.9-μm particles, 120 Å pore size, Dr Maisch GmbH). The separation was performed at a flow rate of 250 nl/min by a gradient of 0.1% formic acid in 80% acetonitrile (solvent B) and 0.1% formic acid in water (solvent A). HLA-II peptides were eluted by the following gradient: 0 to 80 min (2 - 32% B); 80 to 84 min (32 – 45% B); 84 to 85 min (45 – 100% B); and 85 to 95 min (100% B). Data were acquired using a data-dependent acquisition (DDA) method. Full-scan MS spectra were acquired in the Orbitrap at a resolution of 60,000 (at 200 *m/z*) with an auto gain control (AGC) target value of 3×10^6 ions. For Tandem mass spectrometry (MS/MS), ten most abundant precursor ions were sequentially isolated, activated by higher-energy collisional dissociation (NCE = 27) and accumulated to an AGC target value of 2×10^5 with a maximum injection time of 120 ms. In the case of assigned precursor ion charge states of one, and from six and above, no fragmentation was performed. MS/MS resolution was set to 15,000 (at 200 *m/z*). Selected ions were dynamically excluded for additional fragmentation for 20 s. The peptide match option was disabled. The raw files and MaxQuant output tables have been deposited to the ProteomeXchange Consortium via the PRIDE[61] partner repository with the dataset identifier PRIDE: PXD034773.

Peptide identification

We employed the MaxQuant platform v.1.5.5.1[52] to search the peak lists against a fasta file containing the human proteome (Homo_sapiens UP000005640_9606, the reviewed part of UniProt, including 21,026 entries downloaded in March 2017) and a list of 247 frequently observed contaminants. Peptides with a length between 8 and 25 amino acids were allowed. The second peptide identification option in Andromeda was enabled. The enzyme specificity was set as unspecific. A false-discovery rate of 1% was required for peptides and no protein false-discovery rate was set. The initial allowed mass deviation of the precursor ion was set to 6ppm and the maximum fragment mass deviation was set to 20ppm. Methionine oxidation and N-terminal acetylation were set as variable modifications.

Deconvolution and annotation of MHC-II motifs

To search for motifs describing the binding specificities of the alleles present in our compiled MHC-II peptidomics dataset (Table S1), we used our motif deconvolution tool MoDec.[16] MoDec uses a probabilistic framework to search for common motifs of size L (L=9 in general for MHC-II ligands) present anywhere along the sequence of the MHC-II ligands identified by mass spectrometry in a given sample. MoDec does not rely on any prior knowledge of the potential allele binding specificities from the sample, and it will find these motifs in an unsupervised manner. The different motifs identified per sample typically correspond to binding specificities of the different MHC-II alleles. Both the mapping of each peptide to a specific motif and the identification of the binding core are derived from the maximum responsibility value returned by MoDec. Potential contaminant peptides are also identified during the deconvolution. MoDec-1.2 was run using the recommended options ‘‘–MHC2 –makeReport -r 50’’, searching for motifs of length 9 AAs, and searching between 1 and up to 12 different motifs per sample (depending on the sample). For the chicken samples, when looking only at 9-mers motifs, one of the motifs that we find does not contain anchors at the end of the motif (Figure S2A inset). Searching instead

for 10-mers motifs (i.e., 10-mer binding core in MoDec), we see a clear motif with anchor residues at the beginning and the end (Figure S2A, 2nd motif). This corresponds to the bulging mode that had been previously described for this allele.[10]

Following our previously established procedure,[16] all samples were manually reviewed, and few samples that were of too low quality to be included in the analyses were not included in Tables S1 and S2 (e.g., samples with too few peptides present, or incorrect MHC-II typing, where clear motifs are present but are describing other alleles than the alleles supposed to be present in the given sample). MHC-II motifs were then annotated to their respective MHC-II allele by identifying shared motifs across samples sharing the same MHC-II allele. In this way, we could identify the motifs of 88 different alleles. Further, peptides clearly assigned to motifs corresponding to alleles not supposed to be in the sample were considered as contaminants and were removed from further analyses. Peptides that could not confidently be assigned to a specific allele were not considered for the analyses of binding specificities. These include cases with too few peptides to build a clear motif, peptides assigned to a very flat non-specific motif or the flat motif of MoDec, peptides assigned to a clear motif but without a known allele because MHC-II typing was incomplete (e.g. only the HLA-DR and HLA-DQ were known but not the HLA-DP), or cases when two alleles of similar binding specificity are present in a sample but a unique motif describing these alleles was obtained. These peptides were nevertheless kept for the prediction benchmarks of MHC-II ligands to prevent any potential biases in our validation datasets. Multiple specificities appear when two (or more) clearly different motifs from a sample are identified as coming from the same allele (either because the sample is monoallelic or because the same multiple specificity motifs are appearing in multiple samples sharing only the given allele). Motifs shown in Figure S1A were obtained by grouping together the peptides assigned to each allele from all samples and aligning them based on the binding core identified by MoDec. ggseqlogo[53] was used to plot these motifs in all the figures (with the height of the motifs corresponding to Shannon entropy measured in bits). After motif deconvolution, MoDec also returns the binding core offset of each peptide and these values were used for Figure S1B. Following our previous definition,[16] this binding core offset is symmetric around 0 for each peptide (i.e., a binding core offset is equal to 0 when the binding core of the peptide is perfectly at the middle of the peptide, a negative value means the binding core is towards the peptide N-terminus and a positive value means the binding core is towards the peptide C-terminus). The list of all peptides assigned to each allele in each sample can be found in Table S2.

MHC-II sequences retrieval and alignment

Human MHC-II sequences were retrieved from IPD-IMGT/HLA database.[13]

Mouse MHC-II sequences (called H2-IAx and H2-IEx, where *x* gives the name of the allele) were manually retrieved from NCBI's Protein database (https://www.ncbi.nlm.nih.gov/protein/), searching for MHC class II sequences of Mus musculus.

Cattle MHC-II sequences and MHC-II sequences from multiple other species (Figure S4D; Table S3F) were retrieved from IPD-MHC database.[81]

Chicken MHC-II sequences (Gaga-BL) are not yet part of IPD-MHC database. We found the accession numbers corresponding to the different alleles in the *Online Resource 2* of Afrache et al.,[82] searched for these accession on NCBI's Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore) and obtained the translated CDS sequences of some BLB1 and BLB2 genes.

The MHC-II sequences of human, mouse, cattle (BoLA-DR) and chicken were aligned together using Clustal Omega v.1.2.2.[54] The MHC-II sequences from the other species were then aligned against these with Clustal Omega using the results from the first alignment as a *profile1*. Table S3F lists all these aligned sequences.

Motifs clustering

For each allele for which we could obtain a motif, we started by computing a position probability matrix, $PPM^{a}_{l,i}$, (*a*: the allele, *l*: the binding core position, *i*: the amino acid identity), using all the identified binding cores of all the peptides assigned to this allele (Table S2). We further included a pseudocount based on the BLOSUM62 substitution matrix with a parameter β=200.[83] The Kullback-Leibler divergence (KLD) was then computed between all pairs of alleles from a same gene locus (HLA-DR, HLA-DP or HLA-DQ), for each binding core position *l*:

$$KLD^{a,b}_{l} = -\sum_{i=1}^{20} PPM^{a}_{l,i} \cdot \log\left(\frac{PPM^{b}_{l,i}}{PPM^{a}_{l,i}}\right) \quad \text{(Equation 1)}$$

These KLD were then clustered through hierarchical clustering (using *hclust* function from R with the *average* clustering method). Thresholds to define the different clusters based on the hierarchical clustering were manually defined. Resulting clusters are given in Tables S3A–S3C. The binding specificities plotted in Figure 2A (first motifs to the left in each column) correspond to the average $PPM^{a}_{l,i}$ between the alleles of each cluster. Clusters containing a single allele were not considered in the analysis. We also note that the HLA-DP ligands corresponding to the reverse binding mode (see section "MHC-II binding specificities reveal a widespread reverse binding mode in HLA-DP ligands") were not included in the analyses of Figure 2A, since our results subsequently revealed that AAs at P1, resp. P4, actually fit in the P9, resp. P6, binding pocket, and vice versa.

Analysis of published MHC-II structures

Crystal structures in PDB format where obtained from the RCSB PDB (https://www.rcsb.org/).[84,85] Structures containing human MHC-II alleles were retrieved based on the following *Sequence Motif*: "WRLEEFGRFASFEAQGALANIAVDKANLEIMTKRSNYTPITN" for HLA-DR, "FYVDLDKKETVWHLEEFG" for HLA-DP and "CLVDNIFPPVVNIT" for HLA-DQ. A custom script was used to determine

to which MHC-II alpha and beta chains each PDB file corresponded (based on the chain A and chain B sequences in the PDB files and the MHC-II sequences obtained above).

We then manually reviewed these structures to determine the peptide binding cores. Residues in the MHC-II alpha and beta chains that were in close contact to each peptide binding core position (distance < 5 Å) were determined and kept in our list of contact positions if the same residue passed the distance threshold for at least two different MHC-II alleles and if the AA at this residue was not conserved among the MHC-II alleles for which we had obtained a binding motif. These residues in MHC-II alleles are those most likely to influence the binding specificity at a given binding core position in MHC-II ligands.

Sequence logos of the most important allele contact positions in each cluster (see above) were drawn with ggseqlogo[53] (Figure 2A; Tables S3A–S3C includes all contact positions). The numbering of the contact position residues follows the numbering found in X-ray structures for the alpha and beta chains. For the alleles from the first cluster of HLA-DQ at P1, we manually renumbered the amino acids F or L found at the residue 51α to exchange them with the gap found at 52α as they structurally better align to the residues at position 52α from the other HLA-DQ alleles.

We used PyMOL (https://pymol.org) to show representative images of the MHC-II and peptide contacts (Figures 1A and 2A), obtained from the following PDB IDs: 1BX2, 1DLH, 1KLU, 1S9V, 1UVQ, 2NNA, 3C5J, 3LQZ, 3PL6, 3WEX, 4IS6, 4MAY, 4MDJ, 4OZF, 4P57, 4P5M, 6BIR, 6CPL, 6DIG, 6PX6, 7N19 and 7ZAK.

Structural modeling

HLA-DR alpha chain, HLA-DR beta chain and peptide sequences were used as starting points for the modeling. Homology models of the HLA-II-peptide complexes were generated using Modeller software v10.1.[55] Template structures were retrieved from Protein Data Bank.[85] Top matching templates were identified from the template library using an internal database with annotated alleles. The closest template was determined using the BLOSUM62 scoring function.[86] A total of 2,000 models were produced for each HLA-II-peptide complex. These models were subsequently ranked based on the sum of the Discrete Optimized Potential Energy (DOPE) calculated using Modeller over the peptide residues, as well as the HLA residues within 6 Å from the peptide. For each HLA-II-peptide complex, molecular interactions were analyzed in the top 5 ranked models over the 2,000. The final HLA-II-peptide structural model corresponds to the one with best score and highest number of favorable interactions.

The effect of amino acid mutations on HLA-II-peptide structural stability was estimated using FoldX software v5.0 after modeling the mutation using its buildmodel function.[39] Changes in FoldX energy score (DG mutant – DG wild-type) were calculated for each mutant using an ensemble of 10 conformations. A change > 1kcal/mol means a destabilizing effect while a change < -1 kcal/mol means a stabilizing effect. In Figures 2C and S2C, differences in FoldX energy scores are relative to the ''LK'' case.

Production of HLA-DPA1*02:01-DPB1*01:01

The extracellular region of the HLA-DPA1*02:01 and HLA-DPB101:01 chains were separately cloned into pMT\BiP\V5-His A (ThermoFisher scientific). The alpha chain construct harbors the acidic leucine zipper and terminates by a 6His-tag. The beta chain construct contains the basic leucine zipper and terminates with AviTag sequence. To generate cell lines expressing HLA-DPA1*02:01-DPB1*01:01, the two plasmids with a third plasmid conferring puromycin resistance, were cotransfected into Drosophila S2 cells using Cellfectin (ThermoFisher Scientific) according to the manufacturer protocol. Protein expression was induced by addition of 1 mM CuSO4. MHC class II molecules were purified from supernatants with Chelating Sepharose FF (Merck). Peptide loading was performed in citrate saline buffer (100 mM citrate, pH 6.0, 0.2% β-octyl-glucopyranoside (Calbiochem), 1×complete protease inhibitors (Roche)) with 100 μM peptide at 37°C for 24 hours, buffer-exchanged on a HiPrep 26/10 desalting column (Merck) into AviTag buffer and subsequently biotinylated with the BirA enzyme according to the manufacturer instructions (Avidity, Denver, Colorado, USA). Biotinylated MHC class II-peptide complexes were purified on a HisTrap HP column (Merck) and kept at -80°C until multimerization with streptavidin conjugates.

For protein crystallization ''empty'' HLA-DPA1*02:01-DPB1*01:01 was purified on a Superdex 75 10/300 GL column (Merck) into 20mM Tris pH 8.0, 150mM NaCl and concentrated at 10mg/ml.

Binding competition assays

Four peptides following well one of the two observed binding specificities of HLA-DPA1*02:01-DPB1*01:01 and present in multiple MHC-II peptidomics samples containing this MHC-II allele were selected. These sequences were additionally reversed in order to have four peptide sequences fitting well the observed canonical binding specificity and four sequences fitting well the reversed binding specificity. In addition, we also added the same sequences but where the AA present in the predicted P1 binding pocket was replaced by arginine (or lysine if the WT sequence had arginine), and we also added expected negative binders where the predicted peptide binding anchors P1 and P9 were replaced by alanine.

In addition to this first set of 24 human-derived peptides, we selected a second set of 2*39 peptides from viral and bacterial origins. For these, we downloaded viral proteomes found in UniProt[60] (https://www.uniprot.org/), from EBV (Ebstein-Barr virus), HCMV (human cytomegalovirus), HSV-1 (herpes simplex virus type 1), HSV-2 (herpes simplex virus type 2), Influenza A virus (only from the HA and NA proteins), SARS2 (SARS-CoV-2) and VZV (Varicella-zoster virus), considering only the reviewed proteins, potentially coming from multiple strains of the given viruses. We also downloaded from UniProt the tetX gene of tetanus toxin protein (TT), which is produced by the bacteria clostridium tetani. These proteomes were then cut in all overlapping 15-mer peptides and we selected 39 peptides (4-5 peptides per proteome) whose sequence fitted well to the reverse binding motif of HLA-DPA1*02:01-DPB1*01:01

and not to the canonical binding motif of this allele. These 39 sequences were also reversed to have peptides predicted to bind in the canonical orientation.

All these peptides were chemically synthesized using standard fmoc chemistry, purified by RP-HPLC (>80 % purity) and analyzed by UPLC-MS. Peptides were kept lyophilized at -80°C. To test the binding of these peptides to HLA-DPA1*02:01-DPB1*01:01, competition assays were performed by mixing in v-bottom 96-well plate (Greiner Bio-One) in 50 μl of citrate saline buffer (described above) 1 μg of the biotinylated empty allele with a FLAG-tagged peptide (IKTEKKTVQFSDDVQ for the 1st set of 24 peptides; TGVKIGEMPLTDSIL for the 2nd set of 78 peptides) at fixed concentration of 2 μM and candidate peptides were added to each well to a final concentration of 0, 0.13, 0.41, 1.2, 3.7, 11.1, 33.3, and 100 μM. For the control, untagged peptide (IKTEKKTVQFSDDVQ, respectively TGVKIGEMPLTDSIL) was used at the respective concentrations to the mix of allele and FLAG-tagged peptide. After incubation at 37 °C overnight. The binding of the tagged peptides to HLA-II molecule was measured by ELISA. The mix was transferred to a plate coated with avidin and the FLAG-peptide was detected with an anti-FLAG-alkaline phosphatase conjugate (Merck), developed with pNPP SigmaFAST (Merck) substrate and absorbance was read with a 405-nm filter (named Abs405 below).

Half maximal inhibitory concentrations (IC50) for the binding competition assays were computed in R with the functions *drm* and *ED* from the *drc* package v.3.0.1,[87] fitting the data with a four-parameter log-logistic function. Non binders were defined as the peptides with "0.8 * Abs405(@0 μM) < Abs405(@100 μM)", while weak binders were defined as peptides with "0.5 * Abs405(@0 μM) < Abs405(@100 μM) < 0.8 * Abs405(@0 μM)" (i.e., peptides that partially replaced the FLAG-tagged peptide at the higher tested concentration, but not sufficiently to reach their half maximal inhibitory concentration).

Protein crystallization and structure determination

HLA-DP (HLA-DPA1*02:01-DPB1*01:01) protein at around 10 mg/ml was mixed with peptides at a final concentration of 10 mM, and co-crystallized by hanging drop vapor diffusion method. Crystals of HLA-DP- canonical binder (sequence: KNLEKYKGKFVREID) formed in a couple of weeks in 15% w/v PEG 4000, 0.2 M Magnesium chloride hexahydrate, 0.1M Sodium cacodylate pH 6.5 and crystals of HLA-DP-reverse binder (sequence: IEFVFKNKAKEL) in 8% w/v PEG 20K, 8% v/v PEG 500 MME, 0.2M Potassium thiocyanate, 0.1M Sodium acetate pH 5.5. The crystals were cryoprotected with 25% glycerol. Diffraction data were collected at the Paul Scherrer Institute (SLS, Villigen) at PXIII beamline. Data were processed with the XDS Program Package.[56] Structures were solved by molecular replacement using Phaser-MR and PDB: 3WEX as template model. Manual model building and structure refinement were carried out in Phenix Suite[57] using coot software[58] and phenix-refine, respectively. After validation, the models were deposited in the PDB database under identifiers PDB: 7ZAK (canonical binder) and PDB: 7ZFR (reverse binder). Data collection and refinement statistics are summarized (Table S4C). The structures were displayed with PyMOL (https://pymol.org). The polar and charged interactions between the peptide and residues in the MHC-II binding site (Figures S2B and S3F) were determined with PyMOL, using default parameters.

Motifs of the N- and C-terminal contexts

Only peptides that were annotated as coming from an allele with the observed reversed binding mode were considered for this analysis (i.e., corresponding to the ligands shown in Figure 3A). ggseqlogo[53] was then applied on the N-terminal and C-terminal context sequences of these peptides, separately for the peptides following the canonical or reverse orientation (Figure S3G).

Simulation of reverse binding ligands

To study the detection limits of reversely bound ligands using MoDec, we built *in silico* samples containing different fractions of canonical and simulated reverse ligands. These samples were based on ligands from different alleles (DRB1*01:01, DRB1*07:01, DRB1*15:01, DRB4*01:03, DQA1*05:01-DQB1*02:01 and DQA1*05:01-DQB1*03:01), where the ligands from the given allele were randomly selected among all the ligands identified in our data for this allele (Table S2 – only ligands from these alleles were included, no contaminant peptides). Simulated samples were either monoallelic or contained a mix of 2-5 of these alleles (using a same number of ligands per allele in polyallelic samples). We then reversed the sequence from a fraction of the ligands of one of the alleles from the sample to have only one allele possessing simulated reverse ligands per sample (e.g., in a sample with 3 alleles and 3,000 total ligands, each allele had 1,000 ligands, and a fraction of 0.1 reverse binders means that 100 of these ligands corresponded to reverse ligands). Such *in silico* samples were built for different total number of ligands and different fraction of simulated reverse ligands.

We then ran MoDec on each sample, searching for 1-8 motifs, and we manually verified if the results contained a motif corresponding to the simulated reverse ligands. When a reverse binding motif was evident, we indicated in Figures S3H and S3I that reverse ligands could be found in the sample. We indicated that the sample contained only a weak motif of reverse binders when this motif was less clear (weak specificities at the anchor positions, such that if we did not know that an allele had reverse ligands, we would likely have considered this motif as not representing an allele's binding specificity).

Pan-allele MHC-II ligand predictor development

Our pan-allele predictor, MixMHC2pred-2.0 is composed of 2 successive blocks of neural networks with distinct tasks (Figures 4A and S4A). We implemented these neural networks in R, using the packages *keras* (version 2.7) and *tfdatasets* (version 2.7), relying on TensorFlow (version 2.6).

The first block describes the binding specificity. The idea of this block is to determine the binding specificity of an allele based on the sequence of its binding site residues (i.e., residues that are near of the peptide) (Figure S4A left part). This is a similar idea than the PickPocket method that was developed for HLA-I predictions,[88] but PickPocket gives identical importance to all binding site residues, while our neural networks can learn the relative importance of each residue. In our model, this block consists of independent neural networks for each peptide binding core position, l (hereafter we refer to these independent neural networks as NN^1_l). The input of NN^1_l is the sequence of the contact AA residues in the P_l binding pocket of the MHC-II allele a, and the output is the PPM at this binding core position for this allele (i.e. $PPM^a_{l,j}$, including the BLOSUM62 pseudocount, described above, corresponding therefore to a vector of the frequency of the 20 AAs at this binding core position). The contact AA residues used as input are the joint set of close contact residues from HLA-DR, -DP and -DQ explained above (Tables S3A–S3C), after renumbering these based on the alignment of all the MHC-II sequences together, including human and non-human MHC-II (Table S3F). Each input AA was encoded to a numerical vector of size 21 equal to the sum of the one-hot encoding of the given AA plus the row corresponding to this AA in the BLOSUM62 probability matrix[86]; the 21st element of this vector represents a gap or absent AA (this 21st element has value 0 except when the given "residue" in the allele is a gap instead of a true AA, in which case this last element has a value of 2). Each NN^1_l is composed of one fully connected hidden layer with 100 hidden nodes, based on rectified linear unit (ReLU) activation function and a gaussian noise of std 0.1. A dropout of 0.2 was added after these hidden nodes during the model training, and a softmax activation function was used for the output layer. The loss optimized corresponded to the Kullback Leibler divergence, and it was optimized using Adam optimizer with a learning rate of 0.005, a decay 0.005/250 and other parameters at default values. A maximum of 250 epochs was set and optimization stopped after no improvement of the loss during 30 epochs.

In order to account for the multiple specificities that are observed for some alleles, we replicated these NN^1_l three times: a first time including only all the canonical binders of specificity 1, a second time where the canonical binders of specificity 2 are used instead when the given allele had two canonical binders specificity (i.e. for the DRB1*08 alleles at present, but could accommodate additional if observed), and a third time where the reverse binders are used instead of the corresponding canonical binders when a reverse binding specificity was observed for the given allele (the sequence of the reverse binding peptides was inverted there, to have the correct peptides' AA in the P1 binding pocket of the allele for example). A last independent neural network of the 1st block is implemented to predict the fraction of peptides in these different binding specificities. The input here is the full list of contact residues from the MHC-II alleles, encoded in the same way as above and the output is the fraction of peptides observed in each of the 3 sub-specificities (canonical 1, canonical 2 or reverse) for the given allele. This neural network has similar structure than above's NN^1_l except that 50 hidden nodes are used, with a learning rate of 0.0025, a maximum of 500 epochs and the loss corresponds to the categorical cross entropy. To avoid cases where multiple specificities would be present but with too few ligands to be observed, we restricted the training of this part to MHC-II alleles with more than 3,000 observed ligands. When predicting allele specificities, an MHC-II allele is assumed to possess multiple specificities only if this last neural network predicts a fraction of canonical specificity 2 or reverse specificity of at least 1% for the given allele. The training of all these neural networks of the 1st block is repeated 5 times with same parameters and the final outputs are the average between these.

After having trained the first block, we can give the sequence of any MHC-II allele as input and this first block will return the predicted binding specificities of this allele ($PPM^{a,s}_{l,j}$) (with s for the specificity: canonical 1, canonical 2 and reverse), as well as the relative fraction of peptides that are predicted to be bound to this allele in each of the three specificities (w_s). The second neural network block, NN^2, combines these with other features directly linked to a given peptide sequence (its sequence, length, binding core offset, peptide's N- and C-terminal contexts), in order to predict if the peptide is presented by the given allele (Figure S4A right). First, a PPM-based binding score is determined based on the given MHC-II allele specificities and peptide sequence:

$$B = \left(\sum_{s=1}^{3} w_s \cdot \max_{c \in CS} \left(w_c \prod_{l=1}^{L} \frac{PPM^{a,s}_{l,x_{l \oplus c}}}{f_{x_{l \oplus c}}} \right) \right) \quad \text{(Equation 2)}$$

Where w_c is the relative weight of the binding core offset c (similar to Figure S1B) and the best (maximum) value among all potential peptide binding core offsets is used for the inner parenthesis; L is the binding core length (i.e. 9 AAs); x_j indicates which amino acid is found in the peptide at the position j; f_i is a normalization factor, equal to the frequency of amino acid i in the human proteome. The "$l \oplus c$" in $x_{l \oplus c}$ is the "special sum" previously defined,[16] which makes that the binding core offsets are symmetric around 0 for each peptide. The binding score B of the peptide is then transformed to a percentile rank B_{rank} based on the scores of 10,000 random human peptides of the same size. The 1st input of NN^2 corresponds to a min-max scaling between 0 and 1 of the $log(B_{rank} + 10^{-4})$ (where the min B_{rank} is 0 and the max is 100, and 10^{-4} avoids $log(0)$). The 2nd input of NN^2 is a one-hot encoding of the best binding core offset c (determined from Equation 2 above), with values considered between -6 and 6. The peptide length is also one-hot encoded, for sizes between 12 and 21. The last set of inputs corresponds to the 12 AAs of the N- and C-terminal contexts, which were encoded following the same procedure as described above for the NN^1_l (in case of an unknown AA ("X"), the value 1/20 is used for the corresponding elements of this input vector).

Following these encoded input features, the NN^2 consists of a fully connected neural network with 1 hidden layer of 200 hidden nodes following a ReLU activation function with a gaussian noise of std 0.1. A dropout of 0.2 was added after these hidden nodes during the model training, and a sigmoid activation function was used for the single output node (with a value 1 if the given input peptide is presented by the given MHC-II alleles and 0 if not). Adam optimizer was used, with a learning rate of 0.001, a decay

of 0.001/150 and other parameters at default values. The binary cross entropy loss was optimized. A validation split of 1/5 was used and early stopping after 50 epochs without validation loss improvement was set (or a maximum of 150 epochs otherwise).

To train this NN^2, we used as positives the peptides observed in the MHC-II peptidomics samples (Table S2). We did not consider samples with missing MHC-II typing (e.g. samples obtained through anti-pan-HLA-II peptidomics when the HLA-DP had not been typed, but if the sample was obtained through anti-HLA-DR it was sufficient to have the HLA-DR typing), nor samples from chicken or cattle origin or containing a high fraction of predicted contaminant peptides, nor samples obtained from experiments where cells had been transfected with tagged HLA-II[15] due to the observed bias towards longer peptides (Figure S1C). Only peptides of sizes 12-21 AAs long were kept, and peptides whose context could not be determined were also removed. We then downsampled the training set to keep a maximum of 200,000 positive peptides. In multiallelic samples, all potential MHC-II alleles were kept (i.e. we did not use the allele annotation from MoDec): Equation 2 was applied to each allele of this sample and the best B_{rank} score was used for the inputs of NN^2. We finally removed peptides with best $B_{rank} > 30$ (better binders have lower values), which likely correspond to contaminant peptides observed in MHC-II peptidomics. For negative peptides we used four times more random human peptides than positives, with a uniform length distribution between 12 and 21 AAs.

After its training, the scores of NN^2 are transformed to percentile ranks (%Rank) based on the scores of 10^6 random human peptides and making that these follow the peptide length distribution observed in MHC-II peptidomics. For our final model, MixMHC2pred-2.0, the training of NN^2 is repeated 5 times and results correspond to the average of these repetitions. When running MixMHC2pred-2.0 in multiallelic samples, the %Rank against each allele is returned and the score of this sample is taken as the best (lowest) %Rank.

Benchmarking MHC-II binding specificity predictions

To compare the MHC-II binding specificities predicted by MixMHC2pred and NetMHCIIpan, we used 100,000 random human peptides of size 12-21 AAs and scored these using the different predictors against each allele of interest. For each allele, the best scoring 1% peptides were considered as the ligands to this allele, and their binding cores (returned by the predictor) were used to determine the frequency of each amino acid at each binding core position. We then compared these frequencies with the frequencies observed in MHC-II peptidomics data of the given allele by computing the KLD for each peptide binding core position between these frequencies (Equation 1). These KLD were averaged between all binding core positions. Lower KLD values mean that the predicted frequencies are closer to the frequencies observed in MHC-II peptidomics data. In Figure 4B, the comparison is performed against MHC-II peptidomics data coming only from monoallelic samples, while in Figure 4C it includes all MHC-II peptidomics data based on above's annotation of the motifs after the deconvolution using MoDec. MARIA and MHCnuggets were not included in these analyses as their output only consists of the predicted peptide presentation score, but they do not predict the binding core, which prevents inferring the motifs. Likewise, MixMHC2pred-1.2 and NeonMHC2 were not included in these analysis as they are allele-specific predictors and therefore cannot do any prediction for an allele that would be left-out from their training set in a leave-one-allele-out context.

To determine how accurate these specificities are for MHC-II alleles without known ligands, only alleles that were absent from NetMHCIIpan training were considered here. In this respect, we trained MixMHC2pred in a stringent leave-one-allele-out cross-validation setting: when doing the predictions for allele *A*, no peptide annotated as coming from allele *A* is used during the training of the first predictor block NN^1, and all peptides coming from all samples containing this allele *A* are removed from the second predictor block NN^2 (i.e., even if the peptide is annotated as coming from another allele in this sample, as long as the allele *A* is part of the list of alleles from this sample).

Multiple specificities were considered, and the fraction of peptides observed in each sub-specificity (when present) in MHC-II peptidomics data is indicated above the corresponding motifs (Figure 4B). One of the outputs of MixMHC2pred tells the predicted sub-specificity for each peptide and we directly used this. NetMHCIIpan does not return any information about from which sub-specificity a peptide would be coming. To allow having multiple specificities for NetMHCIIpan as well, we applied MoDec on the binding cores of the top 1% best predicted peptides from NetMHCIIpan, (running MoDec with the options "–nruns 50 –makeReport –specInits –no_flat_mot"). For alleles possessing multiple or reverse specificities, we show in Figure 4B the two motifs determined in this way for NetMHCIIpan, while for the alleles possessing a single specificity the motifs are directly obtained from NetMHCIIpan without applying MoDec.

In Figure S4B, we further benchmarked the predictions of the ligands observed for each allele instead of predicting the binding motifs. This was done by using the peptides from MHC-II peptidomics data annotated to each allele as above for the positives (keeping only peptides of sizes 12–21 AAs) and adding four times more random peptides as negatives (with a uniform length distribution). Here the predictions of the full peptide sequences were done (without needing to know or predict the binding cores); therefore, MARIA and MHCnuggets could be included in this benchmark. The area under the curve of the Receiver Operating Characteristic curve (ROC AUC) was computed for each predictor and for each allele separately.

The sequence similarity between two alleles *a* and *b* was computed similarly than in Nielsen et al.[89] It was obtained through the following equation:

$$\mathrm{Sim}^{a,b} = \frac{S(a,b)}{\sqrt{S(a,a) * S(b,b)}} \qquad \text{(Equation 3)}$$

Where *S(a,b)* is the BLOSUM50 alignment score[86] computed between the sequences of the contact AA residues in any of the 9 P_I binding pockets of the MHC-II alleles *a* and *b* (Table S3F). Based on this we defined the similarity of allele *a* to the alleles in the training set as

$$\mathrm{Sim}^{a}_{\mathrm{training}} = \max_{b \in \text{ training alleles}} \left(Sim^{a,b} \right) \quad \text{(Equation 4)}$$

Where the alleles in the training set are those with peptidomics data (the 88 alleles present in the training of MixMHC2pred, or 87 alleles when in the leave-one-allele-out context).

To compare the impact of different model architectures on the binding specificity predictions, we trained different variant models in the leave-one-allele-out cross-validation setting described above. All data from one allele were removed, the models were then trained on the remaining data, and the binding specificity of the left-out allele was predicted by considering the best 1% scoring peptides among a set of 100,000 random peptides. The KLD distance between the predicted allele's binding specificity and the binding specificity observed in our data for this allele was next computed, and it was compared to the KLD distance obtained in the same way for the model corresponding to MixMHC2pred's architecture, in order to quantify the improvement or worsening of using another model than MixMHC2pred's one. This was repeated for the 88 different alleles from our data and results are shown in Figure S4E. For these analyses, only variants of the NN^1 part of our neural network were considered as this is the part where that is responsible for predicting the binding specificity of alleles. Following variants were considered: 1) using the same architecture than MixMHC2pred, but with different number of hidden nodes in the intermediate layer; 2) considering a similar architecture than MixMHC2pred, but considering a single specificity for each allele (by merging the peptides from each sub-specificity and thus not replicating the model three times); 3) considering one single neural network predicting directly the binding specificity of the 9 binding core positions (instead of 1 neural network per core position), where all input contact AA residues are connected to all nodes of the intermediate hidden layer; the intermediate layer consists of 9 groups of 100 hidden, nodes and each group of hidden node is connected to a separate output vector of size 20 (corresponding to the $PPM^{a}_{l,i}$ of one of the 9 binding core positions); this model was still also replicated three times to account for the multiple specificities; 4) this variant is similar to the variant Nr. 3, but all intermediate hidden nodes (either 100 or 9*100 nodes) are connected to all 9*20 output nodes. The loss optimized, activation function used and other parameters of the neural network were kept as in our original model.

Benchmarking MHC-II ligand predictions

The benchmark in Figures 5A–5C was performed using the data from the various MHC-II peptidomics samples (Tables S1 and S2). All peptides from a given sample were used together with the set of alleles describing this sample (while in Figure S4B the benchmark was done per allele, based on all ligands annotated to the given allele coming from multiple samples). Peptides of sizes 12-21 were considered. We did not include samples with missing MHC-II typing, the *AUT01_xx* samples from Marcu et al.[51] due to high predicted contamination, nor samples from experiments where cells had been transfected with tagged HLA-II[15] due to observed biases described above (Figure S1C). The positives were the peptides observed in each sample and we added four times more random peptides as negatives, taken from the same proteins as the proteins observed in the positive peptides and following a uniform length distribution between 12 and 21 AAs. In multiallelic samples, the scores of all these peptides for all alleles were computed and the best score among the sample's alleles was kept (lowest %Rank for MixMHC2pred, lowest %Rank_EL for NetMHCIIpan, lowest ic50 for MHCnuggets and highest score for MARIA). Using the predicted scores of each peptide, the ROC AUC was computed for each predictor and for each sample separately.

To avoid using the same peptides in testing and training of the predictors, we considered only samples that were absent from NetMHCIIpan and MARIA's training sets (MHCnuggets is not trained on any of these samples as it only considers binding affinity data). In this way, the tested samples were therefore absent from the training of NetMHCIIpan, MARIA and MHCnuggets, but the MHC-II alleles from these test samples were often still part of the training of these predictors (i.e., the same alleles were present in some other samples used in the training of these predictors, and therefore the specificity of these alleles could already be well described by these predictors). For our predictor we used the same stringent leave-one-allele-out cross-validation setting than described above, where the %Rank of each allele was obtained separately based on this leave-one-allele-out setting and then the best %Rank among those was used, ensuring that no peptide coming from any sample containing the test allele was present in its training set.

For MARIA, the gene expression of each protein from which a peptide is originating is needed as further input. The gene expression pre-defined in MARIA were used, based on our annotation of from which type of tissue each sample or cell line is originating (BRCA, COAD, K562, …).

The impact of the different peptide-related parameters of our model were studied using this same benchmark MHC-II ligands dataset. Two complementary analyses were performed. In the first analysis, we retrained the NN^2 part of our model, after removing one of the features from the model (e.g., the PPM binding score, or all the amino acids describing the peptide's N-terminal context). The model training and predictions were done in a 5-fold cross-validation setting, where the test data was split in 5 groups on a per sample basis (i.e., all peptides coming from a same sample were kept together in the same cross-validation test set) and all peptides present in a given test set were removed from the corresponding training set to ensure no overlap of the peptides between training and testing sets. ROC AUC were then computed per sample from the test set samples (i.e., computing one ROC AUC value for each sample from the test set), and this ROC AUC was compared to the value obtained by our full model trained in the same 5-fold cross-

validation setting. In the second analysis, we used our full model trained in the leave-one-allele-out cross-validation setting described above (without retraining the model), and we computed the ROC AUC values per sample when doing the predictions after randomly shuffling the values from one of the feature of the model between all the peptides from the given sample (e.g., after randomly reassigning the 6 AAs of the peptide's N-terminal context between the peptides (keeping the 6 AAs together, but shuffling these between peptides)). We again compared these ROC AUC values with the values obtained using our full model without any shuffling of the parameters. For these two analyses, we note that the binding core offset feature is appearing at two places in our model: as some input nodes in NN^2 and also in the binding score computation through Equation 2 above and we therefore tested the effect of this feature separately at each place (models Nr. 2 and 3 in Figure S5A (done by using a flat w_c in Equation 2 for model Nr. 3)) or simultaneously at the two places (model Nr. 4). Concerning the peptide's N- and C-terminal contexts, we considered different cases, where either only the part of the contexts that lies inside (or outside) of the peptide was removed or shuffled, or where only the context from one terminal was considered, or without any context. Finally, model Nr. 11 consisted of a full shuffling of all the input features, where the shuffling is done separately per feature (i.e., this corresponds to a model where the input values are fully random, within the ranged of allowed values).

To compare the prediction accuracy for species without known ligands (Figure 5D), the test sets were composed of all MHC-II peptidomics samples from the given species (mouse, cattle or chicken) and random negative peptides, as described above. In addition to the leave-one-allele-out model and to the full model (trained on all peptides from all samples described above), we also trained leave-one-species-out models, where all data from the given test species were removed from the training and the model was trained on the remaining data only.

Proportion of reverse ligands in different samples

We considered the samples containing HLA-DP alleles accommodating reverse ligands. These samples were obtained either with anti-pan-HLA-II or with anti-HLA-DP antibodies. We used a slightly modified version of MixMHC2pred which returned separately the %Rank predicted in the canonical or reverse orientations. Using this predictor, we computed the %Rank of each peptide from the sample towards all the alleles of the sample, and we kept only peptides with best %Rank < 20 and whose best %Rank was towards the HLA-DP of interest (independently of the binding orientation). To give more confidence that peptides were bound in a given orientation, we further filtered these to only keep those with $\%Rank_{best\text{-}orientation} < \%Rank_{other\text{-}orientation} - 5$. We then computed the fraction of these peptides who had a better score in the reverse orientation than canonical orientation. In samples containing multiple HLA-DP alleles with reverse ligands, this analysis was repeated for each allele, in order to compute separately the fraction of reverse ligands for each allele. Results are shown, either by grouping samples per type of experiments from which they were obtained (Figure S5B, annotation of the sample's type of experiment is given in Table S1) or shown separately for some samples containing a same allele (Figure S5C).

Analysis of cis- and trans- heterodimers

To study the fraction of ligands presented by cis- vs. trans-heterodimers, we considered all samples containing two different alpha chains and two different beta chains of the HLA-DP or HLA-DQ genes. We then used MixMHC2pred to predict the %Rank of each peptide with respect to all the alleles of the sample (considering all possible combinations of HLA-DP or -DQ alpha-beta chains, as well as HLA-DR alleles when the sample was obtained with anti-pan-HLA-II antibodies). To avoid potential contaminants, we only considered peptides with a %Rank < 20, and we only kept the peptides whose best score was obtained with an HLA-DP (respectively HLA-DQ) allele ("Case 1" in Figure S6A). We further considered a more stringent way to assign allelic restriction ("Case 2" in Figure S6A) by keeping only peptides with $\%Rank_{best\text{-}allele} < 20$ and $\%Rank_{best\text{-}allele} < \%Rank_{other\text{-}alleles} - 5$ (i.e., asking that the score of the given peptide towards the other alleles of the sample is not too much similar to the best allele's score). Only samples that contained at least 200 predicted HLA-DP (resp. HLA-DQ) ligands after this filtering were considered. For each sample, we then computed the percentage of ligands for each of the 4 possible HLA-DP (resp. HLA-DQ) alleles.

For samples with HLA-DQ genotypes of different groups (i.e., G1G2), we know which pairs of alleles are in cis and which are in trans. This enabled us to compare the fraction of peptides assigned to cis-heterodimers (referred to as cis1 and cis2 in Figure S6A, where cis1 was the cis-heterodimers with most ligands) or trans-heterodimers (referred to as trans1 and trans2 in Figure S6A). For other samples (i.e., HLA-DQ with the same genotype or HLA-DP), this information cannot be inferred from the HLA-II typing. We therefore named the heterodimer with the highest fraction of predicted ligands as A1-B1 in Figure S6A and named the other heterodimers accordingly. The motifs on the right of Figure S6A were built based on the binding core predicted for the ligands of each allele of the sample.

Benchmarking $CD4^+$ T-cell epitope predictions

All data for human $CD4^+$ T cells from the IEDB database were downloaded (as of 06.08.2021). We then filtered this data to keep the peptides of sizes 12-21 AAs which were observed in "multimer/tetramer", ICS and ELISPOT assays, and whose 4-digits MHC-II typing had been determined, considering the "Allele evidence codes": "MHC binding assay", "Single allele present", "T cell assay - Mismatched MHC molecules/ Biological process measured/ MHC subset identification/ T cell subset identification" and "Statistically inferred by motif or alleles present". This dataset included directly peptides annotated either as positives or negatives (Table S5E), and no artificial negatives were added for this analysis. The ROC AUC was computed for predictions made per allele separately, keeping only alleles with at least 3 positive and 3 negative peptides. As in the experiments from this dataset the short

peptides were usually directly tested, the antigen presenting cells did not need to cleave these peptides before presentation. Therefore, the part related to context encoding is not meaningful and we used here the option of not encoding this peptide context in NetMHCIIpan. MixMHC2pred includes a similar option, which consists in internally replacing the AAs from the context by "X"s and where the %Ranks are recomputed accordingly.

To study the impact of including the reverse binding mode into predictions of $CD4^+$ T cell epitopes, we considered epitopes found in IEDB data for HLA-DP alleles that accommodate reverse ligands. In this data, only positive epitopes were however available and we therefore added 10 random negative peptides per positive peptide, coming from the same protein as each positive, with a uniform peptide length distribution. The ROC AUC per allele was computed using MixMHC2pred, as well as for a same model where only the canonical binding specificity was included.

Selection of candidate epitopes following the reverse binding mode

The proteomes from various viral and bacterial proteins (see above, section about the binding competition assays) were cut in all overlapping 15-mer peptides, and we computed the binding scores of all these peptides with HLA-DPA1*02:01-DPB1*01:01, keeping separate the scores from the canonical and reverse specificity (i.e., scores from Equation 2 but keeping the scores from the index *s* separate instead of summing over them, corresponding to the canonical and reverse specificity). We then selected a set of 4-5 peptides per proteome that had a good score towards the reverse specificity and only a weak score towards the canonical specificity and we synthesized these peptides for experimental testing.

Peptides and peptide-MHC-II multimers

Peptides and peptide-MHC-II multimers were produced by the Peptide & Tetramer Core Facility of the University Hospital of Lausanne (CHUV). Peptides were chemically synthesized using standard fmoc chemistry, purified by RP-HPLC (>90 % purity) and analyzed by UPLC-MS. Peptides were kept lyophilized at -80°C. Biotinylated peptide-MHC-II monomers, loaded with peptides of interest were multimerized using streptavidin-PE (Cat# SA10044, Thermofisher Scientific) or streptavidin-APC (Cat# 405207, Biolegend) conjugates, then stored at 4°C and used within a week.

Identification of antigen-specific $CD4^+$ T-cell responses

Primary $CD4^+$ T cells were cultured in RPMI 1640 Glutamax media (GIBCO) supplemented with 8 % human serum (Biowest), non-essential amino acids (GIBCO), 2-mercaptoethanol (GIBCO), sodium pyruvate (GIBCO), HEPES (GIBCO), penicillin/streptomycin (BioConcept) and 100 IU.mL-1 of rhIL2 (Novartis).

$CD4^+$ T cells were isolated (ref 130-045-101, Miltenyi) from cryopreserved PBMC and co-incubated (10^6 mL^{-1}) with autologous irradiated CD4-depleted PBMCs (10^6 mL^{-1}) and pools of 3 to 4 peptides (2 μM) in RPMI supplemented with 8 % human serum and IL-2 (100 IU mL^{-1}). After 11 days, cells were put in RPMI supplemented with 8 % human serum without any IL2. At day 12, cells were washed with RPMI, diluted at 2.10^6 mL^{-1}, and 200,000 cells plated in 96w round bottom plates. Then 100 uL of peptide pools were added (2 μM final in R8) and cells were incubated 1 h at 37°C. Protein Transport Inhibitor (1/1000, eBioscience 00-4980-93) was added and cells incubated for additional 4 h at 37°C. Cells were then washed with PBS and stained for 15 min at RT with fixable Near-IR Dead Cell Staining Kit (Thermofisher L10119, 1/1000 in PBS). After three washes, cells were stained with CD4 antibody (BD 562970) for 20 min at 4°C. After additional three washes, cells were incubated with Fix/perm kit (Biolegend 426803) for 20 min at 4°C in the dark, and stained with anti TNFα (BD 340512) and anti IFNγ (BD 554702) for 30 min at 4°C.

After final washing, cells were resuspended in FACS buffer (PBS 0.5 % FBS 2 mM EDTA) and analyzed on a Cytoflex S1 flow cytometer. Data were analyzed using the FlowJo v10.7.1 software. Positive and negative controls were obtained by incubating cells with PMA/ionomycin (Thermofisher, Cat# 00-4975-93) or without peptide, respectively. For peptide pools leading to an immune response, experiment was repeated with single peptides.

Sorting of naive and effector memory $CD4^+$ T cells

Naive $CD4^+$ T cells and effector and effector memory $CD4^+$ T cells were isolated by Fluorescence-activated Cell Sorting (FACS) upon staining with anti-CD4 antibody (BD 562970), anti-CCR7 antibody (353227 BioLegend) and anti-CD45RA antibody (304108 BioLegend) for 30 min at 4°C. After three washes with FACS buffer (PBS 0.5 % FBS 2 mM EDTA) cells were incubated 10 min with DAPI (Sigma, Cat#10236276001) at 250 nM and washed again three times. Naïve ($CCR7^+$ and $CD45RA^+$) $CD4^+$ T cells and effector and effector memory ($CD45RA^-$) $CD4^+$ T cells were collected separately.

Peptide-MHC-II multimer validation and sorting of $CD4^+$ T cells

$CD4^+$ T cells were incubated with multimers (1/50 dilution) 45 min at 4°C in FACS buffer (PBS supplemented with 0.5 % FBS and 2 mM EDTA), isolated by FACS and either directly used for TCR sequencing or expanded with autologous irradiated CD4-depleted feeders in RPMI supplemented with 8 % human serum, phytohemagglutinin (Invivogen, 1 μg mL^{-1}) and IL2 (150 IU mL^{-1}).

Bulk TCR sequencing

mRNA was extracted using the Dynabeads mRNA DIRECT purification kit according to the manufacturer instructions (ThermoFisher) and was then amplified using the MessageAmp II aRNA Amplification Kit (Ambion) with the following modifications: *in vitro* transcription was performed at 37°C for 16 h. First strand cDNA was synthesized using the Superscript III (Thermofisher) and

a collection of TRAV and TRBV specific primers. Unique Molecular identifiers (UMI) of length 9 were added to each read. TCRs were then amplified by PCR (20 cycles with the Phusion from NEB) with a single primer pair binding to the constant region and the adapter linked to the TRAV and TRBV primers added during the reverse transcription. A second round of PCR (25 cycles with the Phusion from NEB) was performed to add the Illumina adapters containing the different indexes. The TCR products were purified with AMPure XP beads (Beckman Coulter), quantified and loaded on the MiniSeq instrument (Illumina) for deep sequencing of the TCRα or TCRβ chain.

Analysis of TCR sequences

The fastq files were processed with MIGEC,[59] using default parameters to demultiplex them and identify the TCRα and TCRβ clonotypes. For each sample, the frequency of each TCR chain was computed based on UMI corrected counts. Only TCRs with more than one UMI count and representing more than 1% of the total UMI counts were considered. TCRs with the same amino acid sequences were merged in Table S6B.

The CDR3 sequences of the alpha and beta chains were used to search TCRα and TCRβ repertoires through the iReceptor web platform,[46] which contains, as of June 2022, 7,111 repertoires for a total of 5.1 billion sequences. Hits were defined as those having the same CDR3 sequence (Table S6B). We further restrict our analysis by considering only TCRs with the same CDR3 and the same V, J genes (100% sequence identity).

QUANTIFICATION AND STATISTICAL ANALYSIS

P-values for the comparisons between IC50, KLD or AUC values in the different comparisons were computed with paired two-sided Wilcoxon signed rank-tests, as indicated in the corresponding figure legends. Statistical analyses were performed with R software.