Predicting CD4 T-cell epitopes based on antigen cleavage, MHCII presentation, and TCR recognition

Dina Schneidman-Duhovny [1,2¤]*, Natalia Khuri[1,2,3], Guang Qiang Dong[1,2], Michael B. Winter[2], Eric Shifrut[4], Nir Friedman[4], Charles S. Craik[2,5], Kathleen P. Pratt[6], Pedro Paz[7]*, Fred Aswad [7]*, Andrej Sali[1,2,3]*

1 Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States of America, **2** Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, United States of America, **3** Graduate Group in Biophysics, University of California at San Francisco, San Francisco, CA, United States of America, **4** Department of Immunology, Weizmann Institute of Science, Rehovot, Israel, **5** California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, CA, United States of America, **6** Uniformed Services University of the Health Sciences, Bethesda, MD, United States of America, **7** Bayer HealthCare, San Francisco, CA, United States of America

¤ Current address: School of Computer Science and Engineering, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
* sali@salilab.org (AS); dina@salilab.org (DS); pedro.paz@bayer.com (PP); fred.aswad@bayer.com (FA)

Abstract

Accurate predictions of T-cell epitopes would be useful for designing vaccines, immunotherapies for cancer and autoimmune diseases, and improved protein therapies. The humoral immune response involves uptake of antigens by antigen presenting cells (APCs), APC processing and presentation of peptides on MHC class II (pMHCII), and T-cell receptor (TCR) recognition of pMHCII complexes. Most *in silico* methods predict only peptide-MHCII binding, resulting in significant over-prediction of CD4 T-cell epitopes. We present a method, ITCell, for prediction of T-cell epitopes within an input protein antigen sequence for given MHCII and TCR sequences. The method integrates information about three stages of the immune response pathway: antigen cleavage, MHCII presentation, and TCR recognition. First, antigen cleavage sites are predicted based on the cleavage profiles of cathepsins S, B, and H. Second, for each 12-mer peptide in the antigen sequence we predict whether it will bind to a given MHCII, based on the scores of modeled peptide-MHCII complexes. Third, we predict whether or not any of the top scoring peptide-MHCII complexes can bind to a given TCR, based on the scores of modeled ternary peptide-MHCII-TCR complexes and the distribution of predicted cleavage sites. Our benchmarks consist of epitope predictions generated by this algorithm, checked against 20 peptide-MHCII-TCR crystal structures, as well as epitope predictions for four peptide-MHCII-TCR complexes with known epitopes and TCR sequences but without crystal structures. ITCell successfully identified the correct epitopes as one of the 20 top scoring peptides for 22 of 24 benchmark cases. To validate the method using a clinically relevant application, we utilized five factor VIII-specific TCR sequences from hemophilia A subjects who developed an immune response to factor VIII replacement therapy. The known HLA-DR1-restricted factor VIII epitope was among the

six top-scoring factor VIII peptides predicted by ITCall to bind HLA-DR1 and all five TCRs. Our integrative approach is more accurate than current single-stage epitope prediction algorithms applied to the same benchmarks. It is freely available as a web server (http://salilab.org/itcell).

Introduction

Adaptive immunity involves cellular and humoral responses [1]. Cellular immunity is primarily mediated by cytotoxic CD8[+] T cells which recognize peptide antigens presented by Major Histocompatibility Complex class I molecules (MHCI) while humoral immunity requires CD4[+] T helper cells responding to peptide-MHC class II complexes (pMHCII) to support antibody production by B cells. The humoral immune response pathway proceeds through multiple stages (Fig 1) [2]. First, an antigenic protein is endocytosed by antigen presenting cells (APCs). For the MHCII pathway, the protein can then be cleaved in the endosome by acid-dependent proteases into peptides of ~10–30+ residues. The invariant chain of the MHCII receptor blocks the peptide binding site in the nascent MHCII protein in the endoplasmic reticulum (ER) and facilitates the export of MHCII receptors without peptide ligands from the ER to a vesicle. Next, the vesicle fuses with the endosome that contains the whole antigen or its peptides. The invariant MHCII chain is then cleaved, leaving only a non-covalently bound small fragment (CLIP) that continues to block the MHCII peptide binding site. CLIP is subsequently removed by human leukocyte antigen DM (HLA-DM, with an MHCII-like structure), allowing for subsequent interactions with antigenic peptides in the vesicle. Additional cutting and trimming of peptides can take place after the formation of a stable peptide-MHCII (pMHCII) complex [3]. Next, stable pMHCII complexes are presented on the APC surface. If the pMHCII complex is recognized by a T-cell receptor (TCR) on a CD4[+] T cell, the cell becomes activated, producing helper cytokines that support B-cell activation, differentiation to plasma cells, and finally antibody generation.

Here, we focus on predicting T-cell epitopes within an input antigen sequence, restricted to a given human MHCII (Human Leukocyte Antigen, HLA) and to specific TCR sequences. Knowledge of immunodominant epitopes is helpful in designing vaccines [4], improving cancer immunotherapy [5], studies of autoimmune diseases [6, 7], and engineering of protein replacement therapies [8]. Unfortunately, most current *in vitro* and *in vivo* experimental methods for predicting immunogenicity in pre-clinical settings are resource intensive and relatively slow [5, 9, 10]. Therefore, an accurate, rapid, and inexpensive computational approach for assessment of protein immunogenicity would be valuable.

Current computational methods that are informative about T-cell epitopes predict mainly peptide-MHCII affinity [11]. They can be divided into machine learning [12] and *ab initio* approaches [13]. Machine learning methods train predictors by relying on a large number of experimentally measured binding affinities for multiple MHCII alleles [14, 15], such as those available in the Immune Epitope Database (IEDB) [16]. State-of-the-art predictors have an average area under the receiver-operating characteristic (ROC) curve (AUC) of ~0.8, as shown by independent validation [14, 15]. The drawback of machine learning methods is that sufficient experimental datasets are not always available for an MHCII allele of interest, resulting in lower accuracy [12]. In addition, because the length of an MHCII-binding peptide can vary from approximately 10 to 30+ residues, mistakes can be made in registering the 9-mer peptide subsequence in the core of an MHCII binding groove [17, 18]. It is also hard to account for
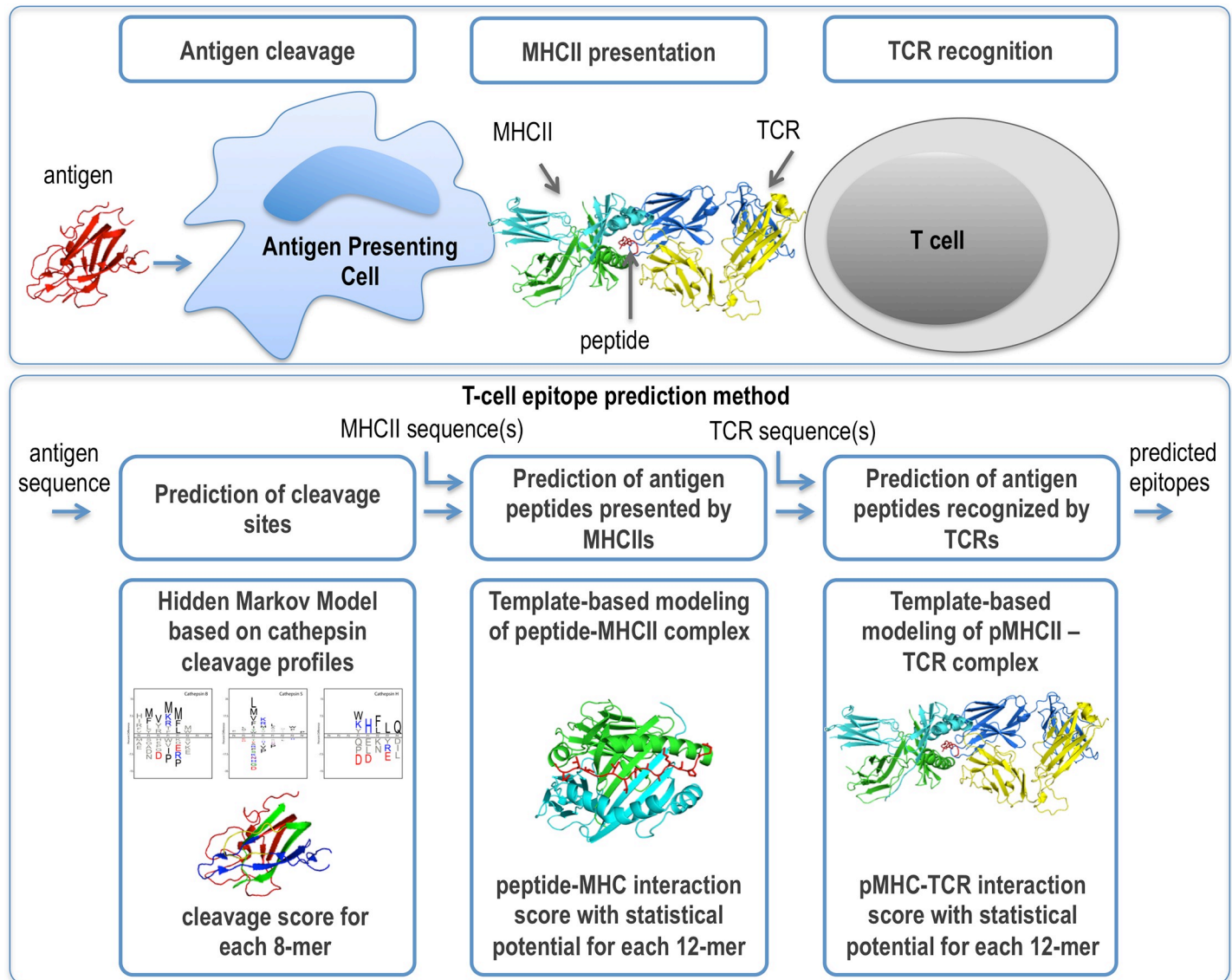
Fig 1. Overview of the 3 steps of antigen processing that are modeled using the current approach.

"flanking" peptide residues on either side of the core 9-mer, which can modulate binding affinity, and only the NetMHCII method currently takes this into account [14]. In contrast, current *ab initio* methods rely on the MHCII structure for predicting the affinity of peptide binding [13]. These methods do not require experimentally derived peptide binding datasets and can be applied to any MHCII allele. However, their accuracy (AUC from 0.62 to 0.68) tends to be lower than that of the machine learning methods [13].

While there are methods that consider preprocessing of class I peptides [19, 20], the current computational approaches for class II consider only a single step (*ie*, peptide-MHCII binding) in the multiple steps of the adaptive immune response. However, many peptides that are capable of binding to a given MHCII will not actually be presented on the MHCII of an APC (*eg*, because the peptide is not excised from the antigen during cleavage). Moreover, many peptide-MHCII complexes that are presented will not be recognized by a TCR (*eg*, self-peptides

that are constitutively presented on APCs). There are six MHC class II genes in humans (HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1) of which all but DRA are polymorphic. While peptide presentation by MHCII is essential for immunogenicity, it is clearly not sufficient, as immunogenicity also requires engagement of antigenspecific TCRs from the diverse T cell repertoire (estimated to contain over 1,000,000 unique sequences [21] with a potential diversity on the order of 10^{15} [22]) generated during T cell development. Therefore, ignoring pMHC-TCR interactions leads to overprediction of potential immunogenicity. Including predicted protease cleavage sites in the antigen sequence and assessing the feasibility of ternary complex formation between pMHCII and the corresponding TCR, in addition to evaluating the compatibility between the peptide and MHCII, could significantly improve the accuracy of T-cell epitope prediction. Here, we propose such a combined approach for predicting epitopes, given the antigen, MHCII, and TCR sequences.

First, we designed an antigen cleavage predictor. The minimal cell-free antigen cleavage system requires cathepsins B, H, and S as well as HLA-DM and MHCII [23]. Therefore, we experimentally determined specificity profiles of cathepsins B, H, and S, using a global mass spectrometry-based approach [24]. These specificity profiles enabled us to design an accurate antigen cleavage site predictor. Second, we designed a structure-based predictor for peptide-MHCII recognition. Using structural modeling and affinity data for a large number of peptide-MHCII pairs in IEDB, we trained an atomic distance-dependent statistical potential [25] that was optimized for assessing the stability of peptide-MHCII complexes. Finally, we designed a structure-based predictor for pMHCII-TCR recognition, enabled by next generation sequencing platforms that can sequence a large number of short DNA sequences covering the diverse complementarity determining regions (CDRs) of the TCRs [26–28]. Our approach is modular, for example cleavage sites prediction to filter the list of peptides is optional. The cleavage sites predictor may be more useful for predicting epitopes in pathogen-derived antigens, which are more sensitive to digestion by cathepsins, compared to auto-antigens [29]. Finally, our method can also be used when only partial information, such as TCR β-variable regions only, is available, even though the antigen specificity of the complete TCR depends on both α and β chains.

We begin by presenting the integrative structure-based approach (Fig 1). We benchmark the method using 20 known structures of pMHCII-TCR complexes as well as four pMHCII-TCR complexes with defined HLA and TCR sequences only. Finally, we apply the method to predict a validated HLA-DR1-restricted T-cell epitope in blood coagulation factor VIII (FVIII). FVIII-specific T-cell clones were obtained from two hemophilia A subjects who developed a neutralizing antibody response to therapeutic FVIII infusions. The TCR β-variable regions of these clones were sequenced and the resulting sequences incorporated into the variable regions of the TCR models used in our prediction algorithm.

Results

Method

The input is an antigen sequence of interest, an MHCII allele sequence, and a TCR β or α+β variable region sequence. The output is a list of potential epitopes (*ie*, peptide sequences in the antigen that are predicted to form a complex with the given MHCII and TCR) as well as structural models of the pMHCII and pMHCII-TCR complexes. The method proceeds in three steps (Fig 1) (Materials and Methods). First, we predict cleavage sites in the antigen sequence based on cathepsins S, B, and H cleavage profiles. These sites can be used to filter the set of all possible 12-mer peptides for the next step. Second, we build a structural model of the peptide-MHCII complex and assess it with an optimized statistical potential. Third, we assess whether

or not any of the top-scoring peptide-MHCII complexes are likely to bind to a given T-cell receptor (TCR), based on the score of the best structural model of the ternary peptide-MHCII-TCR complex and the locations of the predicted cleavage sites.

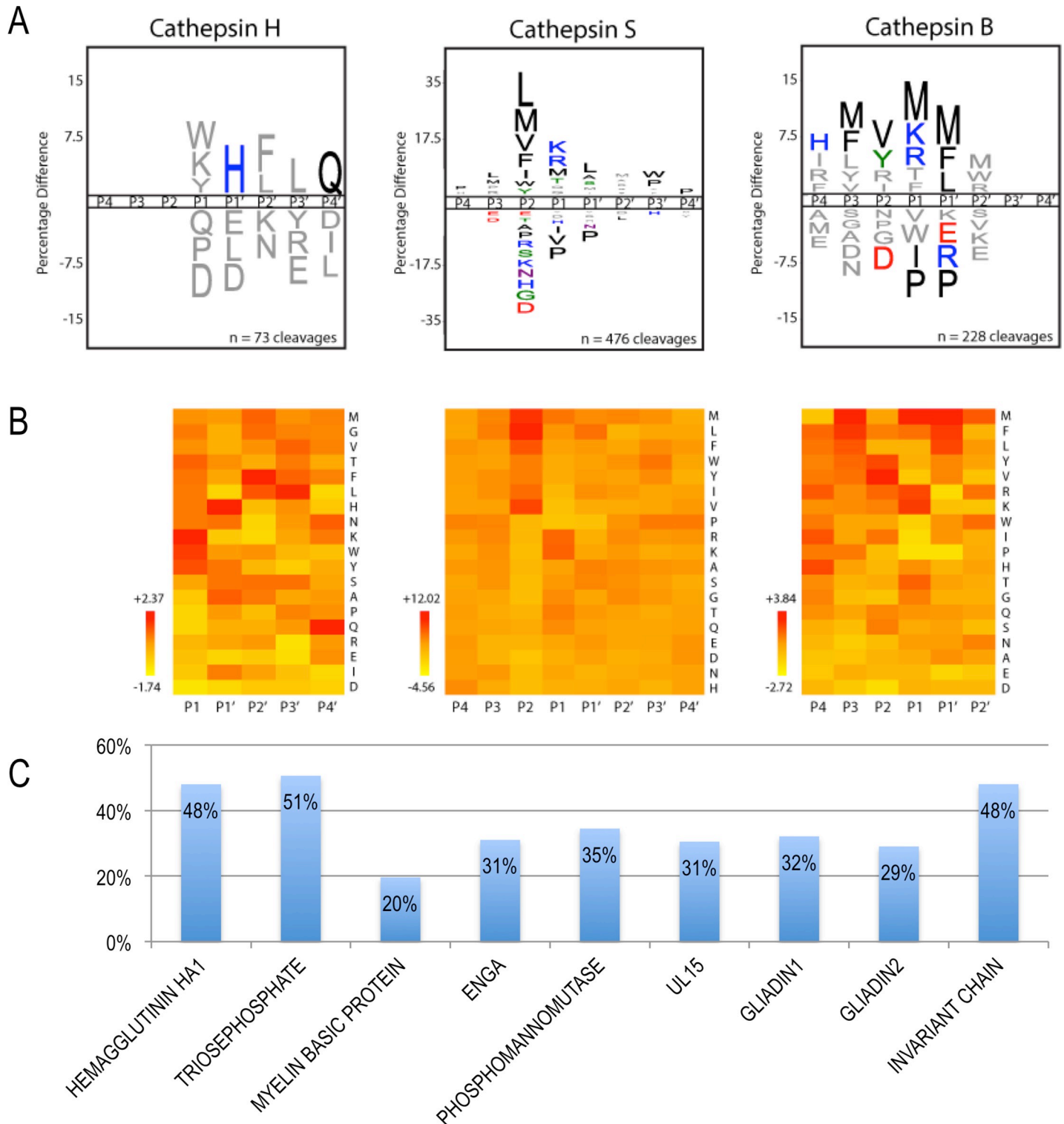## Substrate specificity profiles for cathepsins B, H, and S

In this step we predict the cleavage sites in the antigen sequence. In a cell-free (*in vitro*) antigen processing system, cathepsins B, H, and S were found to be sufficient for cleavage of model protein antigens [12, 23]. Cathepsin S is the major endoprotease involved in class II antigen processing outside the thymus. Carboxypeptidase (cathepsin B) and aminopeptidase (cathepsin H) activities are important for trimming longer fragments bound to MHCII molecules; they are constitutively expressed in all professional APCs. Therefore, we obtained peptide cleavage datasets [24], constructed specificity matrices for the three cathepsins, and designed a cleavage predictor (Fig 2). Cathepsin S has a preference for hydrophobic residues in the P2 position and for positively charged residues in the P1 position, consistent with the profiling of the P1-P4 positions using a combinatorial library [30]. The remaining positions have broad specificity, consistent with the major role of this enzyme in antigen processing. The profiles of cathepsins B and H clearly identify them as a predominant di-carboxypeptidase and a mono-aminopeptidase, respectively. Pathogen-derived epitopes have different processing pathways and are quite sensitive to digestion by cathepsins, compared to auto-antigens that are more resistant to digestion [12]. As a result, it is recommended to use the cleavage predictor as an initial step when analyzing pathogen-derived antigens, while it may be prudent to generate predictions for auto-antigens both with and without the cathepsin-cleavage prediction filter.

Testing on input antigen sequences with known ternary complex structures

To test the accuracy of the method, we used a benchmark dataset of 20 human pMHCII-TCR complexes with crystal structures available in the Protein Data Bank (PDB [31]; Table 1). For each complex, we identified and utilized the sequence of the parent protein of the peptide epitope. The benchmark dataset contains antigen sequences from eight proteins (hemagglutinin HA1, triosephosphate isomerase, myelin binding protein (MBP), engA, phosphomannomutase, UL15, and two gliadin proteins). The 20 complexes have some redundancy, since some involved the same MHCII allele, TCR, or peptide epitope sequence. However, the dataset includes 14 non-redundant pMHCII complexes.

For additional benchmarking to test the accuracy of pMHCII complex predictions only, we included an additional 18 pMHCII complexes from the PDB without a known cognate TCR. This combined benchmarking set consists of 32 pMHCII complexes (14 non-redundant pMHCII-TCR complexes + 18 pMHCII complexes).

Accuracy of predicting antigen cleavage sites.   We applied the cleavage prediction protocol to the sequences of the eight protein antigens in our benchmark set and to the MHCII invariant chain, which is also cleaved by cathepsins in the endosome (S1 Table). The method predicted 165 cleavage sites in these nine sequences. In the absence of knowing the true cleavage sites for these nine proteins, we can only check whether or not the predicted cleavage sites lie outside the actual epitopes, as they should. Only 5 of the predicted cleavage sites (three in MBP and two in EngA) lie within the known epitopes. The three MBP epitopes are in fact cleaved in healthy individuals, who thus avoid an immune response, in agreement with the prediction. In contrast, they are not cleaved in patients with multiple sclerosis for whom the epitopes were defined. In these patients, MBP, a short disordered protein, binds to the recycled MHCII molecules without undergoing endosomal cleavage [32]. In summary, as expected, the protease cleavage prediction significantly reduced the number of peptide candidates for

Fig 2. A and B) Cleavage profiles for cathepsins H, S, and B. A) iceLogo representations of substrate specificity. Residues above the line are favored at a given position; residues below the line are disfavored. Statistically significant residues (p < 0.05) are colored according to their physicochemical properties. B) Heat map representations of residue preference clustered and colored by Z-score at each position. Favored residues have Z > 0 and disfavored residues have Z < 0. Cleavage profiles are provided from P1-P4' for cathepsin H, P4-P4' for cathepsin S, and P4-P2' for cathepsin B based on their predominant experimentally-determined cleavage preferences. C) Percentage of 12-mer peptides in the benchmark proteins remaining, following filtering by the cleavage sites predictor.

Table 1. Benchmark of epitope predictions based on crystal structures. The epitope core residues within the peptides are bolded.

PDB	Protein	Confirmed epitope peptide	Antigen length	MHCII type	TCR	ITCell Rank	Rank MHC	Rank TCR	Rank NetMHCII pan(core)	Combined rank
1j8h	HEMAGGLUTININ HA1	**PKYVKQNTLKLA**T	317	DRB1*0401	HA1.7	1	4	3	1	1
1fyt	HEMAGGLUTININ HA1	**PKYVKQNTLKLA**T	317	DRB1*0101	HA1.7	1	7	1	12	1
4e41	TRIOSEPHOSPHATE ISOMERASE	G**ELIGILNAAKVP**AD	249	DRB1*0101	G4	3	1	72	2	1
2iam	TRIOSEPHOSPHATE ISOMERASE	G**ELIGILNAAKVP**AD	249	DRB1*0101	E8	1	1	16	2	1
2ian	TRIOSEPHOSPHATE ISOMERASE	G**ELIGTLNAAKVP**AD	249	DRB1*0101	E8	1	1	9	2	1
1ymm	MYELIN BASIC PROTEIN	EN**PVVHFFKNIVTP**	169	DRB1*1501	OB.1A12	13	5	61	1	1
2wbj	ENGA	DF**ARVHFISALHGS**G	304	DRB1*1501	OB.1A12	254	155	288	1	3
3pl6	MYELIN BASIC PROTEIN	N**PVVHFFKNIVTP**R	169	DQA1*0102-DQB1*0502	Hy.1B11	1	2	4	1	1
4grl	PHOSPHOMANNOMUTASE	R**LLMLFAKDVVSR**N	463	DQA1*0102-DQB1*0502	Hy.1B11	20	12	74	6	7
4may	UL15	**QLVHFVRDFAQL**	735	DQA1*0102-DQB1*0502	Hy.1B11	3	3	24	10	1
3o6f	MYELIN BASIC PROTEIN	**FSWGAEGQRPGF**GSGG	169	DRB1*0401	MS2-3C8	93	152	20	99	113
1zgl	MYELIN BASIC PROTEIN	V**HFFKNIVTPRTP**P	169	DRB5*0101	3A6	18	6	52	1	1
4ozf	GLIADIN	**APQPELPYPQP**GS	291	DQA1*0501-DQB1*0201	JR5.1	20	33	27	218	185
4ozg	GLIADIN	**APQPELPYPQP**GS	291	DQA1*0501-DQB1*0201	D2	1	33	2	218	144
4ozh	GLIADIN	**APQPELPYPQP**GS	291	DQA1*0501-DQB1*0201	S16	11	33	15	218	175
4ozi	GLIADIN	**QPFPQPELPYP**	291	DQA1*0501-DQB1*0201	S2	12	8	45	244	83
4gg6	GLIADIN2	**SGEGSFQPSQEN**P	307	DQA1*0301-DQB1*0302	SP3.4	1	1	81	53	1
4z7u	GLIADIN2	P**SGEGSFQPSQEN**PQ	307	DQA1*0301-DQB1*0302	S13	1	1	11	53	1
4z7v	GLIADIN2	**SGEGSFQPSQEN**P	307	DQA1*0301-DQB1*0302	L3-12	1	1	22	53	1
4z7w	GLIADIN2	P**SGEGSFQPSQEN**PQ	307	DQA1*0301-DQB1*0302	T316	1	1	17	53	1
Number of cases out of total with the correct epitope ranked #1						10/20	3/14	1/20	5/14	13/20
Number of cases out of total with the correct epitope among top20						18/20	11/14	10/20	10/14	15/20

subsequent testing (by 30% for the benchmark proteins, Fig 2C, S1 Table), while eliminating few of the true epitopes.

Additional benchmarking was performed using sequences of 1,000 MHC binding peptides from 280 proteins in the systeMHC dataset [33], which comprise the immunopeptidome of these proteins as determined by mass spectrometry of peptides eluted from antigen-presenting cells. This Hidden Markov Model (HMM)-based cleavage predictor reduced the number of candidate peptide epitopes for MHCII presentation by ~30%, with only ~3% of these 1,000 true MHCII binders predicted to be cleaved.

Accuracy of predicting pMHCII complexes. Fourteen unique peptide-MHCII pairs from the 20 pMHCII-TCR structures in our benchmark were used for validating prediction of peptide side-chain orientations and peptide registration in the MHCII binding cleft. Input peptide-MHCII templates were taken from the MHCII allele dataset (S2 Table) and contained a peptide different from the one that appears in the benchmark pMHCII-TCR complex. Consideration of each peptide was limited to its core 12 amino acid residues, although in principle shorter or longer peptides can be considered by the method. This 12-mer core starts two positions before the traditional 9-mer core peptide register. All possible 12-mer peptides in the antigen sequences were modeled into the complex, followed by an assessment of the resulting 3D model. The rank of the actual epitope peptide was determined and compared to that computed by the *NetMHCIIpan-3.1* predictor [15] (Tables 1 and S3). In three cases, the correct epitope was the top scoring peptide predicted using the SOAP score. In eight additional cases, the correct epitope was among the top 20 scoring peptides. For MBP, the correct peptide-MHCII

complex (PDB 3o6f) was ranked in the 152nd position, and thus was not identified correctly. A possible rationalization is provided by the low affinity of this epitope for HLA-DR4 and its "loose" accommodation in the MHCII binding cavity [34]. In the remaining two cases, we find the correct peptide was ranked in the top 30 solutions, but with an incorrect register of the peptide core in the MHCII cavity compared to the reference X-ray structure. Similar results were obtained with *NetMHCIIpan-3.0* (Table 1). In summary, ranking of the pMHCII models was insufficient to accurately predict epitopes, although it is clearly an essential step for the overall modeling and scoring including other aspects of the system (below).

We performed further independent benchmarking of peptide-MHCII recognition using PDB structures of pMHCII complexes that do not include the TCR. This enabled us to increase the number of non-redundant test cases from 14 to 32 (S3 Table). Here, all possible 12-mer peptides in the antigen sequences were used without filtering of predicted cleaved sequences. On average there were 440 12-mer peptides tested for each known MHCII binding peptide (S3 Table). In this benchmarking set the correct pMHCII binding register was the top-ranked prediction for 5 (16%) and 4 (13%) of the 32 benchmark cases evaluated by our statistical potential (SOAP) and by NetMHCIIpan, respectively (Fig 3, S3 Table). Because these two scoring approaches are based on different information, we tested whether the combined score could improve prediction accuracy. Indeed, using a combined score (sum of normalized scores), we were able to improve prediction accuracy, resulting in the correct pMHCII register receiving the top score for 9 (28%) of the 32 cases.

Finally, additional benchmarking to predict peptide-MHCII binding was performed using 29,700 MHCII binding peptides with measured affinities from the IEDB dataset. For each peptide, we identified and utilized the sequence of the parent protein. Our goal was to identify the MHCII binding peptide among all possible 12-mers in the parent protein sequence. Because there are no structures for these pMHCII complexes, their core registers in the MHCII cavities could not be validated. Again, all peptides were used without filtering of predicted cleaved sequences. The correct peptide binding register (overlap of at least 12 peptide residues) was found in the 10 top-scoring peptides for 14% and 22% of the 29,700 benchmark cases using the SOAP statistical potential and NetMHCIIpan, respectively (Part A of S1 Fig). The combined score improved the accuracy to 24%. We further checked the performance for peptides with high, intermediate, and low experimentally measured binding affinity. For high affinity peptides, NetMHCIIpan had the best performance: 34% of the cases included the correct MHCII binding peptide among 10 top–scoring peptides (Part B of S1 Fig). The combined score improved the accuracy to 36%. For peptides with intermediate affinity, SOAP and NetMHCIIpan had the same accuracy of 15%, while the combined score increased the accuracy to 20% (Part C of S1 Fig). For low affinity peptides, the accuracy was 12% and 6% for SOAP and NetMHCIIpan, respectively (Part D of S1 Fig). The results show that SOAP, which is a structure-based statistical potential, has the same accuracy over the entire affinity range, while NetMHCIIpan is more accurate for high affinity binders. Moreover, the combined score improves the accuracy for high and medium affinity binders.

Accuracy of predicting TCR recognition. We now evaluate whether testing of all possible 12-mer peptides for predicted binding to a known MHCII and a known TCR can identify the correct epitope based on the score of the pMHCII-TCR interface. Similar to the assessment of pMHCII complexes, the correct pMHCII-TCR complex was ranked among the 10 top-scoring models in five (25%) of the 20 benchmark cases (Table 1). For five additional cases, it was among the 20 top-scoring models. Therefore, the scoring of a ternary complex is informative for epitope prediction. Importantly, there was little if any correlation between the accuracy of pMHCII and pMHCII-TCR scoring. As a result, the separate scoring of binary and ternary
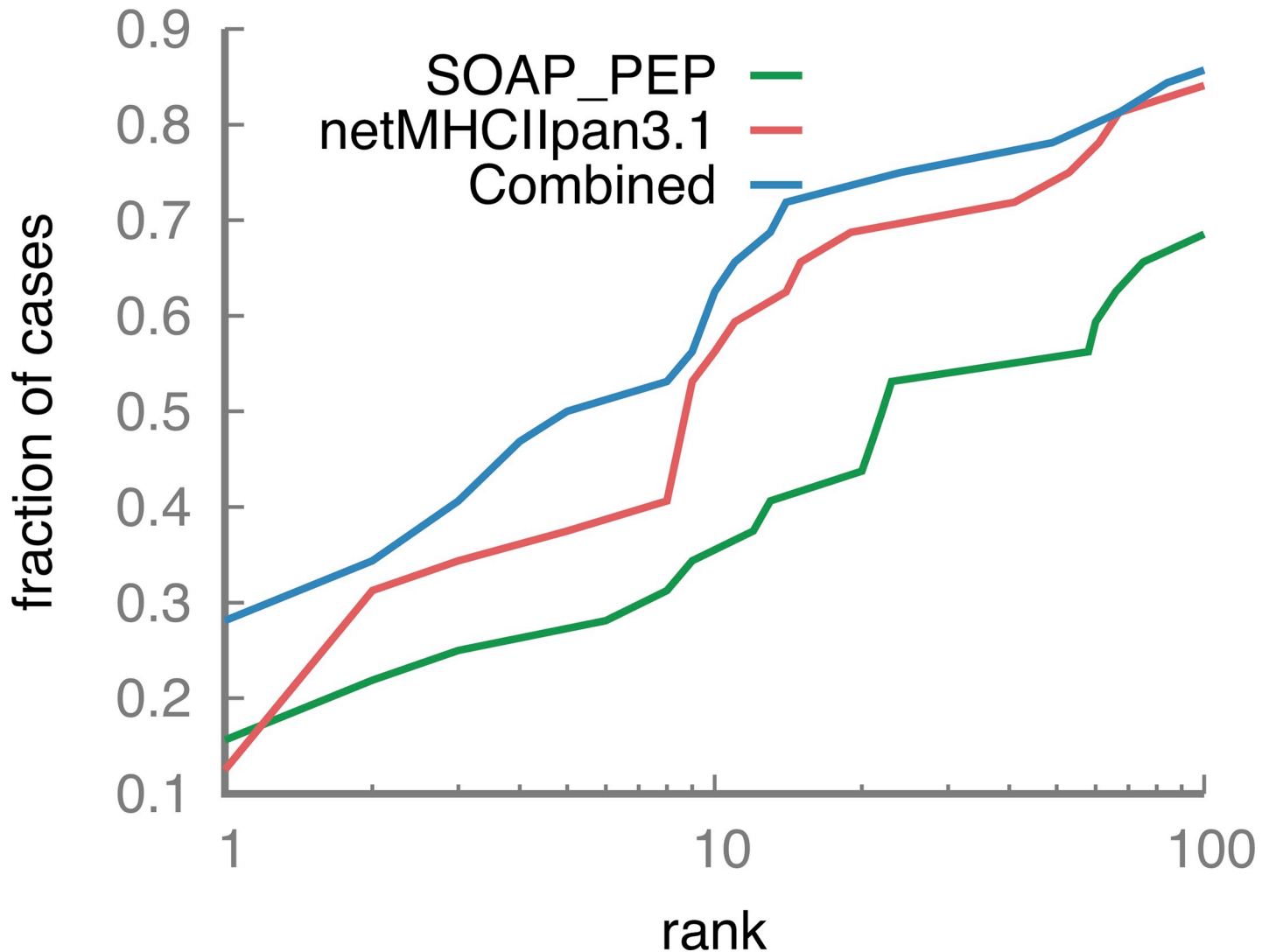
Fig 3. Success rate for SOAP_PEP, NetMHCIIpan3.1 and the two scoring functions combined.

complexes is complementary and thus increases the accuracy of the integrative approach (below).

Accuracy of the entire integrative approach. Our integrative approach ranks potential epitopes based on the pMHCII and pMHCII-TCR scores as benchmarked above (Materials and Methods). In 10 of 20 cases benchmarked against crystal structures of pMHCII-TCR complexes, the correct epitope was ranked as the best by the ITcell scoring function (Table 1). In eight additional cases, the correct peptide was among the 20 top-scoring peptides. The only clearly inaccurate predictions correspond to the MBP (PDB 3o6f) and ENGA epitopes (PDB 2wbj), which have outlying pMHCII scores as discussed above. Combining the ITcell score with the NetMHCIIpan score resulted in 13 out of 20 correct top scoring predictions. However, the combined score for gliadin epitopes was worse, since NetMHCIIpan does not rank these epitopes even among 100 top scoring ones. In summary, the benchmark demonstrates the synergistic contributions of the individual scores, resulting in accurate or nearly accurate

prediction of epitopes for all but two cases in the benchmark. Moreover, combining ITcell with NetMHCIIpan further improves the accuracy.

Testing on sequences without known ternary complex structures

Next, we test the method in a more realistic setting, where the sequences of the antigen, MHCII, and TCR β-chain are known [35], but the corresponding ternary complex crystal structure is not available. In each of these cases, a confirmed MHCII-restricted epitope has been identified by non-crystallographic methods, e.g. MHCII tetramers. Thus, only an assessment of the accuracy of the epitope prediction is performed in this benchmark; an epitope is considered correctly predicted if it overlaps by at least 5 residues with the correct epitope. This second benchmark contains 4 combinations of antigen, MHC, and TCR sequences from mice (Table 2). Because mouse protease cleavage specificities may differ from those of humans, we did not use the APC cleavage predictor in this benchmark. Despite this handicap, the correct epitopes were among the top 11 best scoring peptides in all cases.

Application to FVIII

To validate the method by applying it to a clinically relevant human immune response, potential *HLA-DRB1*01:01* (abbreviated DR1)-restricted epitopes in blood coagulation FVIII were predicted, using five known FVIII-specific TCRβ sequences as input to the algorithm. TCRβ sequences were determined using cDNA from FVIII-specific T-cell clones that were isolated from two hemophilia A subjects who developed an immune response to FVIII infused to treat their bleeding disorder [36]. Epitope mapping using peptide-loaded HLA-DR1 tetramers had identified a 20-mer peptide recognized by these clones as the FVIII peptide spanning residues 2194–2213 (SYFTNMFATWSPSKARLHLQ). The 9-residue HLA-DR1-binding register was subsequently identified experimentally using truncated and sequence-modified peptides [37]. The 20-mer peptide and its 10-residue binding register (FTNMFATWSP) were revealed to modelers only after the modeling predictions were shared with experimentalists.

The method was applied in this case without using the cleavage predictor. Based on the benchmark results (Table 1), we expected the actual epitope to be among the 100 top scoring peptides out of 2340 possible overlapping 12-mers. Several 12-mers fully or partially overlapping with the correct 20-mer peptide were among these 100 top scoring peptides. Complexes of the five TCR sequences, the HLA-DR1 receptor, and the core 12-mer (SYFTNMFATWSP) ranked 14, 19, 79, 72, and 52, respectively (Fig 4A–4C). This 12-mer was predicted as the most likely epitope for two reasons. First, it was the only consensus 12-mer from the 20-mer peptide since it ranked among the 100 top scoring peptides for all five TCR sequences. Second, it was the only 12-mer fully overlapping with the 20-mer peptide (Fig 4A). Additional 12-mers from the 20-mer peptide were among the 100 best scoring peptides, but they overlapped less (only 5–6 residues) and did not result from considering each of the five TCR sequences. Overall,

Table 2. Benchmark of murine T-cell epitope predictions based on comparative models.

Protein	Confirmed epitope peptide	Antigen length	MHCII type	TCR CDR3 sequence	TCR V	TCR J	ITcell Rank	Rank MHC	Rank TCR
Hsp60	VLGGGCALLRCIPALDSLTPANED	572	g7 (NOD)	CASSLGGNQDTQYF	V12	J2.5	11	72	28
MBP	ASQKRPSQR	250	U (B10.PL/J)	CASSGTDQDTQYF	V16	J2.5	2	14	7
MBP	ASQKRPSQR	250	U (B10.PL/J)	CASGDAGGSYEQYF	V8.2	J2.7	7	33	19
hMDM2	LLGDLFGV	490	HLA-A*0201	CASGDWGYEQYF	V8.2	J2.7	2	1	89
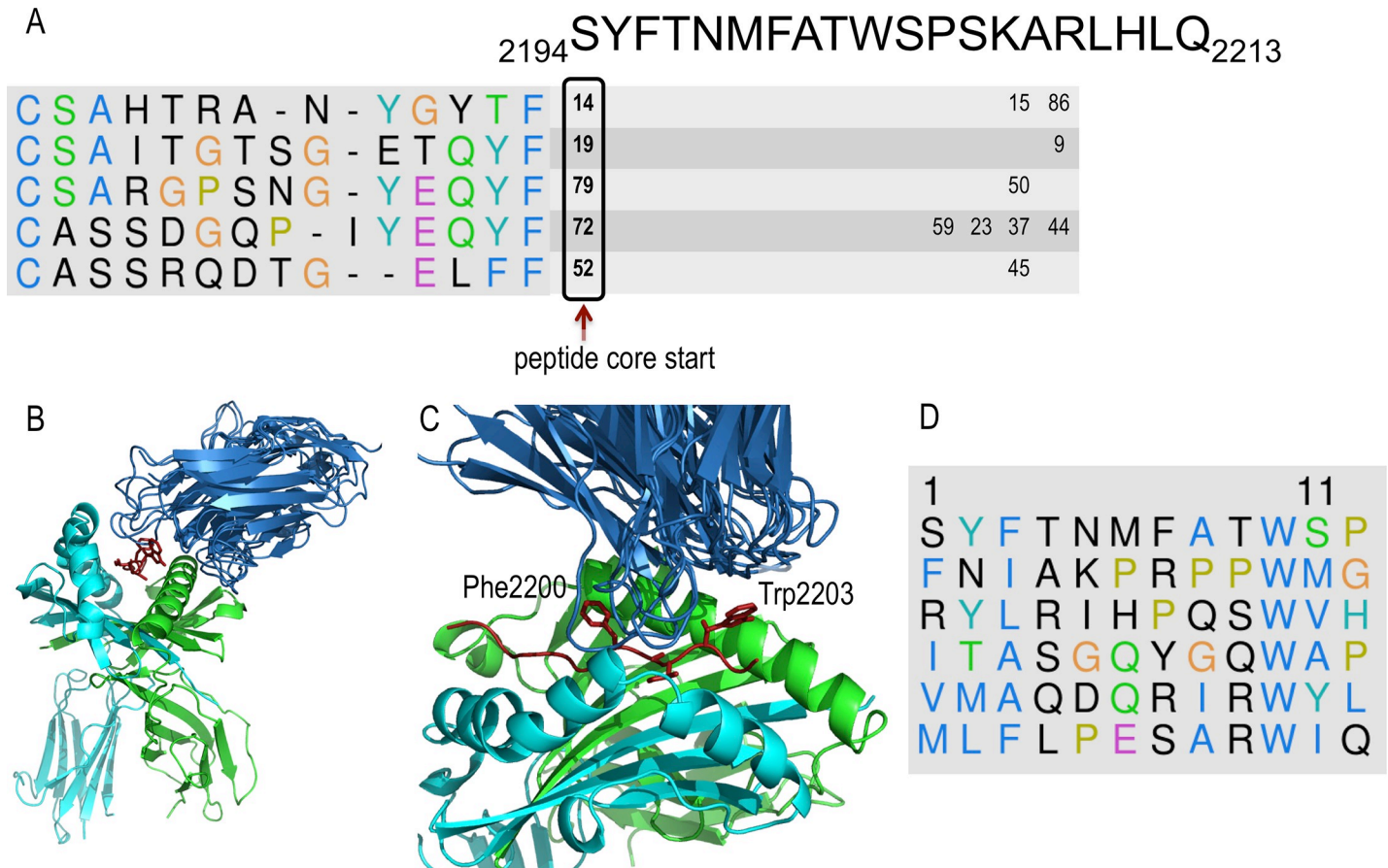
Fig 4. Prediction of FVIII epitopes. A) Consensus based determination of the epitope and its register in the MHCII cavity. CDR3 sequences of the five TCRs and their ranking with respect to different peptide registers in the epitope sequence. Only ranks in the top 100 scores are shown. Peptide FVIII-2194-2205 ranked in the top 100 scoring peptides predicted to bind both HLA-DR1 and each of the 5 indicated TCR-beta variable CDR3 sequences. The predicted consensus core of the epitope for the five TCRs is indicated by an arrow. B) Structural models for the five pMHCII-TCR complexes with the SYFTNMFATWSP 12-mer peptide. The identical MHCII (HLA-DR1) structures within each modeled complex were superimposed, allowing visualization of the different peptide and TCR structural variations. C) The TCR models contact the pMHCII by placing their CDR3 loops in the cavity between FVIII residues Phe2200 and Trp2203. D) The six highest scoring peptides with respect to all five TCRs all have a Trp residue at position 10.

https://doi.org/10.1371/journal.pone.0206654.g004

only six peptides were among the top 100 best scoring for all five TCR sequences; all of them had a Trp residue at position 10 (Fig 4D).

Therefore, using five patient-derived TCR sequences reduced the number of possible 12-residue epitope cores from 2340 to only six FVIII peptides, including the correct epitope core. If we consider consensus peptides only among four of five TCR sequences, we add 12 additional peptides to the prediction list. For comparison, NetMHCIIpan 3.0, which does not rely on TCR sequences, ranks the correct epitope as a weak binder at position 182 only (there is only one NetMHCIIpan prediction per antigen because its predictions do not depend on the TCR sequence).

Discussion

Epitope prediction is a challenging problem in protein engineering, vaccine design, cancer immunotherapy, and autoimmune disease studies. Even though the immune response pathway involves several steps, most current computational approaches focus on MHCII presentation only. Here, we present an integrative structure-based approach that incorporates

additional information by explicitly modeling three steps in the adaptive immune response pathway: antigen cleavage by proteases, MHCII presentation, and TCR recognition. Our approach is *ab initio* as it does not rely on training using known data. The cleavage predictor was built using cleavage sites determined using a random peptide library [24], the peptide-MHC statistical potential was developed using protein-peptide complexes [38], and the pMHCII-TCR statistical potential was developed using a protein-protein docking benchmark [25]. We show that the integrative approach is more accurate compared to considering MHCII presentation only (Table 1, Fig 3). For example, the consideration of cleavage prior to MHCII binding can reduce the number of potential peptide candidates by ~30% (Fig 2C, S2 Table).

As for every optimization problem, the accuracy of the method can be limited by the accuracy of sampling and/or scoring. Here, the sampling exhaustively enumerates all possible epitopes in the input sequence(s) and models the structure of the pMHCII-TCR complex. In turn, the scoring function assesses the possible epitopes using the modeled structures. Therefore, the accuracy of the scoring function and the thoroughness of sampling pMHCII-TCR complex structures determine the accuracy of the entire approach.

We addressed the sampling problem by relying on available structures of pMHCII and pMHCII-TCR complexes as structural templates, thereby significantly improving accuracy compared to current *ab initio* prediction of protein-protein complex structures. While the pMHCII complexes do not show significant structural variability (Fig 5A, Cα RMSD <1Å), the pMHCII-TCR complexes show more variability around the canonical orientation (Fig 6A, Cα RMSD < 24Å). To model this variability, we enriched the available structures with additional pMHCII-TCR template structures obtained by protein-protein docking (Fig 5B), resulting in a more uniform distribution of RMSD values (Fig 6C and 6D, Cα RMSD < 30.0Å, S4 Table). To account for TCR loop variability we generated 10 TCR models.

Our template-based structure modeling approach is not foolproof. For example, recently published structures of an MHCII protein in complex with an insulin peptide and two TCRs revealed a 180° rotation of the TCR compared to the canonical orientation [39]. While currently our approach cannot predict this orientation, we could add such novel structures to our template list to include the rotated TCR orientation if it turns out to be common. To maximize scoring accuracy, we used distance and orientation-dependent potentials for peptide-MHC complexes and a distance-dependent potential for pMHC-TCR interfaces. In particular, the *soap_pMHCII* potential was trained for the current problem based on peptides with known binding affinity (Materials and Methods). Statistical potentials have been shown to be useful in a range of protein structure prediction problems [40].

The accuracy of our approach could be further improved in several ways. First, to increase the accuracy of cleavage sites prediction, the current predictor that relies only on experimental cleavage profiles could be upgraded to also consider structural information [41, 42]. Moreover, greater coverage of the substrate profiles of the enzymes could be obtained with libraries designed to provide greater representation of amino acid space. Second, pMHCII-TCR recognition accuracy could be improved by modeling the TCR loops in the context of the ternary complex, for example by MODELLER [43]. Finally, adding score normalization with respect to peptide length could allow us to predict not only the peptide core, but also the length and conformation of the peptide flanking regions.

Immune responses to protein replacement therapy are a major limitation of protein drugs. Identification and "de-immunization" by removing immunodominant epitopes can prevent immune response and allow wider application of replacement therapy for many conditions, including hemophilia A and B [44], von Willebrand disease (VWD), factor VII deficiency, and Pompe disease [45].
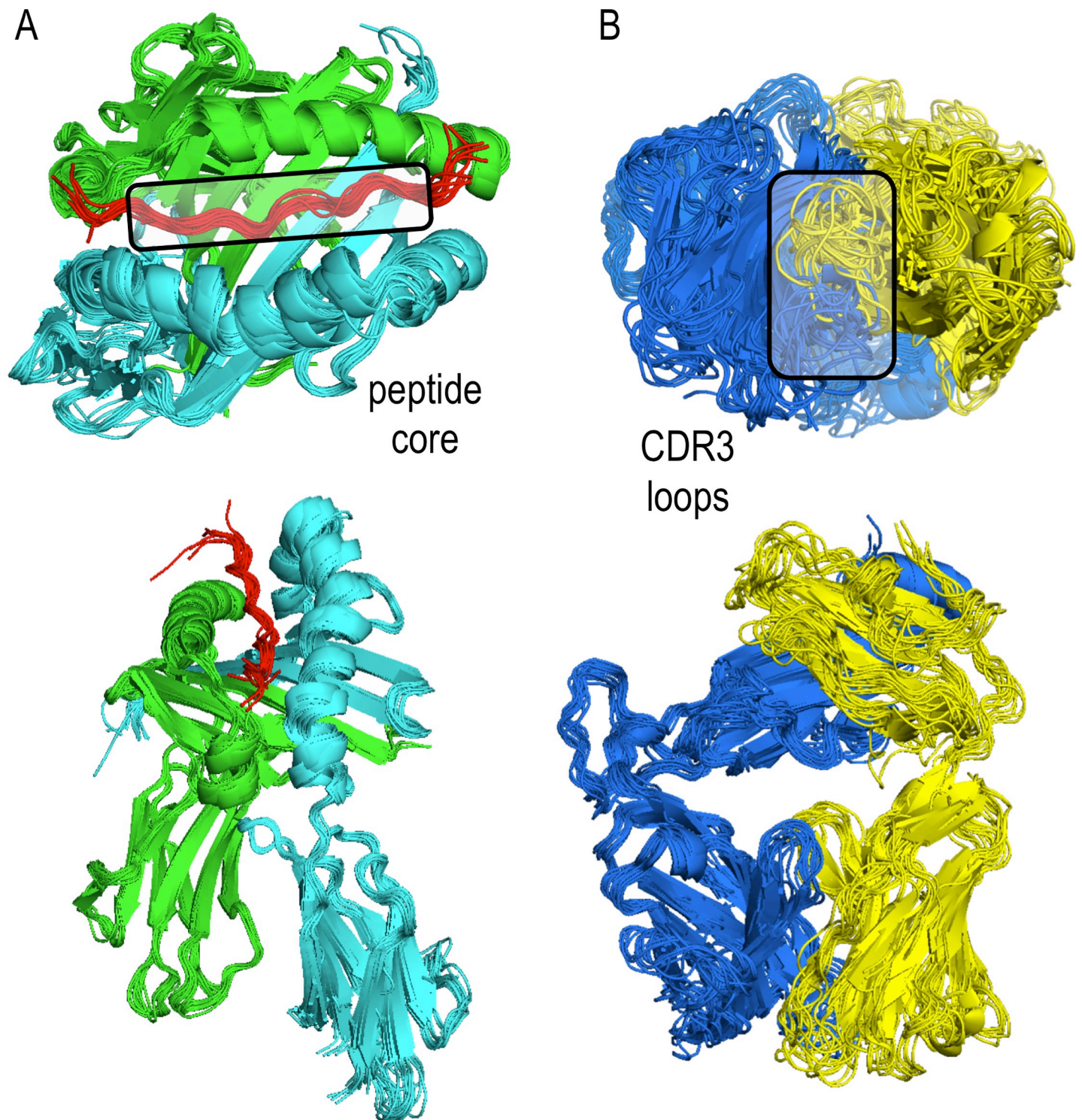
Fig 5. A) Structural alignment of pMHCII complexes from the benchmark set. Alpha and beta chains are green and cyan, respectively. The peptide is shown in red with its core region in the box. B) Structural alignment of TCRs from the benchmark set. TCR alpha and beta chains are shown in yellow and blue, respectively. The variable TCR CDR3 loops are shown in the box. To address TCR loop variability, 10 models are used for each TCR sequence.
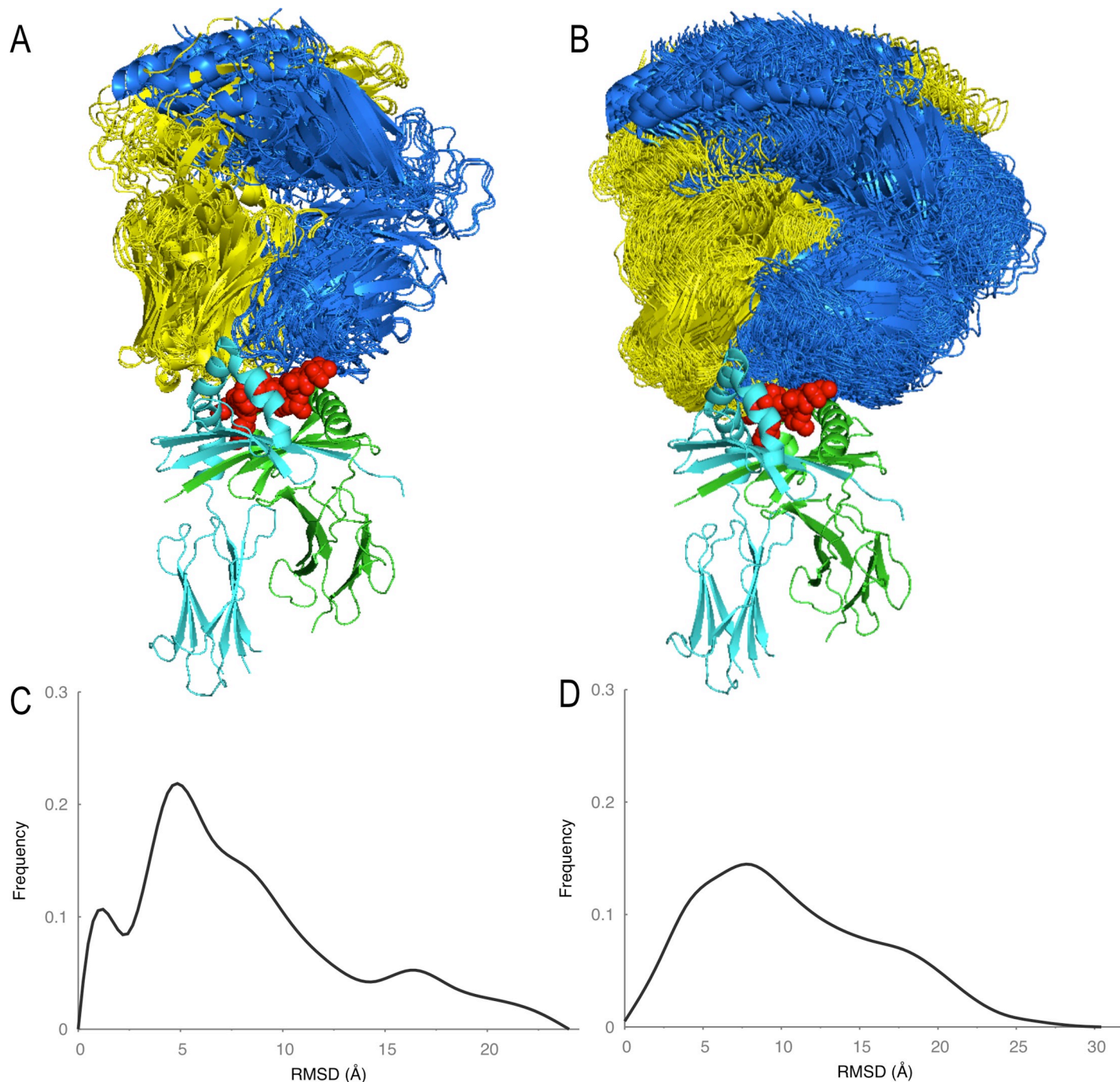
Fig 6. A) Orientations of TCRs with respect to MHCII in the 20 benchmark complexes. B) Orientations of TCRs with respect to MHCII in the 500 template structures generated by docking. C) Distribution of RMSD values between the 20 benchmark pMHCII-TCR structures. D) Distribution of RMSD values between the 500 template structures. Addition of docking generated templates resulted in better coverage of orientation space as indicated by the smoother RMSD distribution on the right.

https://doi.org/10.1371/journal.pone.0206654.g006

A potential use of our method in the development of biotherapeutics was illustrated by its application to FVIII. Specifically, we predicted a FVIII epitope recognized by five TCRs whose TCR β CDR3 sequences were obtained by multiplex PCR [46] (Adaptive Biotechnologies, Inc., Seattle, WA) of cDNA from FVIII-specific T-cell clones that were isolated from hemophilia A

patients who developed an immune response against FVIII replacement therapy [47] (Fig 4). We successfully identified the epitope peptide core among the top 100 choices, among the 2340 possible 12-mer peptides. Consensus scoring for these five TCRs further narrowed the search to six candidate peptides. Moreover, our method has predicted the core of the epitope and its orientation with respect to MHCII and TCR. This strategy could be applied to de-immunizing approaches to identify and eliminate epitopes through strategic amino acid substitutions.

While our method can be applied to any MHC allele without reliance on training datasets like allele-specific machine-learning methods (Introduction), it requires knowledge of the TCR sequences. The gain is the additional information provided by ITCell that associates a TCR with its cognate pMHC. This information could also be used in the context of designing T cells with specific high-affinity TCRs to target known antigens for T cell therapy of cancer or autoimmunity. The method could also aid in identifying the epitopes for lists of TCR sequences that are correlated with a specific condition (such as TCRs of tumor infiltrating T cells, or TCRs found in patients with an autoimmune disease). These TCRs may be identified by high-throughput TCR sequencing [26–28] and determination of the corresponding MHC alleles of the blood donors is straightforward. Moreover, the ITCell method can also be applied to the MHCI pathway (Table 2), combined with existing methods for prediction of proteasomal cleavage and substrate specificity of the transporter associated with antigen processing (TAP). In addition to pMHCI-TCR and pMHCII-TCR, our approach may also be applicable to other immune recognition complexes, such as the pMHCI-KIR3D complexes [48].

Materials and methods

The T-cell epitope prediction problem is defined as follows. Given the sequences of an antigen, MHCII allele(s), and antigen-specific TCR(s) that result in the immune response, predict the peptide(s) in the antigen sequence that are both cleaved in the endosome and form a complex with a given MHCII and TCR. The method proceeds in three steps, corresponding to modeling of antigen cleavage, MHCII presentation, and TCR recognition. First, the antigen cleavage sites are identified based on the cleavage profiles of cathepsins S, B, and H. Second, we build a structural model of the peptide-MHCII complex and assess it with a statistical potential. Third, we assess whether or not any of the top-scoring peptide-MHCII complexes can bind to a given T-cell receptor (TCR), based on a score of the best structural model of the ternary peptide-MHCII-TCR complex and the locations of predicted cleavage sites.

1. Antigen cleavage

Peptide specificity matrices. Peptide specificity datasets for cathepsins S, B, and H were obtained using Multiplex Substrate Profiling by Mass Spectrometry (MSP-MS) [24, 49]. We note that in the MSP-MS peptide library, norleucine is used in place of methionine. Recombinant human cathepsin H (catalog #: 7516-CY-010), cathepsin S (catalog #: 1183-CY-010), and cathepsin B (catalog #: 953-CY-010) were purchased from R&D Systems and activated according to the supplier's instructions. Cathepsin H was activated using bacterial thermolysin (catalog # 3097-ZN-020), which was inhibited with 2 mM phosphoramidon (catalog # EI006) prior to cathepsin H profiling. MSP-MS assays were carried out as described previously [24]. Briefly, 0.2 μg/mL cathepsin, a matched no-enzyme control, and a phosphoramidon-inhibited thermolysin control (cathepsin H only) were assayed against a diverse library of 228 tetradecapeptides pooled at 500 nM in D-PBS (pH 6.5) containing 1 mM TCEP. After 15, 60, and 240 min, 30 μL of assay mixture was removed, quenched with 7.5 μL 20% formic acid, and flash-frozen in liquid N_2. Prior to mass spectrometry acquisition, peptide samples were desalted using

Millipore C_{18} ZipTips and rehydrated in 0.2% formic acid. LC-MS/MS data were acquired using a Thermo Scientific LTQ-Orbitrap-XL or LTQ-FT-ICR mass spectrometer that were equipped with a Thermo Scientific EASY-Spray Ion Source, EASY-Spray PepMap C_{18} Column (3 μM, 100 Å), and Waters nanoACQUITY UPLC System. The LC was operated at a 600 nL/min flow rate during sample loading for 14 min and then at a 300 nL/min flow rate for peptide separation over 65 min using a linear gradient from 2% to 50% (vol/vol) acetonitrile in 0.1% formic acid. For MS/MS analysis using the LTQ-Orbitrap-XL, survey scans were recorded over a mass range of 325–1500 *m/z*. Peptide fragmentation was performed using collision-induced dissociation (CID) on the six most intense precursor ions, with a minimum of 1,000 counts, using an isolation width of 2.0 Th, and a minimum normalized collision energy of 25. Instrument parameters for the LTQ-FT-ICR were used as described previously [24]. Mass spectrometry peak lists were generated using MSConvert from the ProteoWizard Toolkit [50], and data were searched against the 228-member peptide library using Protein Prospector software (v.5.12.4) [51]. All cleavages were allowed in the search by designating no enzyme specificity. The following variable modifications were used: amino acid oxidation (proline, tryptophan, and tyrosine) and N-terminal pyroglutamate conversion from glutamine. Protein Prospector score thresholds were selected with a minimum protein score of 15 and a minimum peptide score of 10. Maximum expectation values of 0.01 and 0.05 were used for protein and peptide matches, respectively. For cathepsins B and S, octapeptides corresponding to P4-P4' were used as the positive dataset. Octapeptides corresponding to all possible cleavages in the MSP-MS library (n = 2,964) were used as the negative data set. For cathepsin H, selected cleavages were restricted to primary mono-aminopeptidase cleavage products (P1-P4'), due to background thermolysin cleavages, and corresponding pentapeptides were used as the negative dataset (n = 228). iceLogo representations and Z-scores were calculated using iceLogo software (v.1.2) [52]. All cleavages identified are available in S1 File. All raw spectrum (.RAW) files are available at the ProteoSAFe resource (ftp://MSV000079617@massive.ucsd.edu; username: MSV000079617; password: MHCII).

Prediction of cleavage sites. The cleavage site predictor was built for each cathepsin individually. First, the MSP-MS results were summarized in a matrix that contained amino acid residue counts from the cleaved peptides for each of the eight positions (P4 to P4'). This matrix was then used to compute a Hidden Markov Model (HMM) for cleavage prediction [53]: The residue counts were converted to probabilities, normalized by the natural occurrence of amino acid residues; and the score, *cleavage_score*, is the sum of the log of normalized probabilities. All octapeptides in the antigen sequence are considered for proteolytic cleavage at the P1-P1' position. For a confident cleavage prediction, we require *cleavage_score* > 3.0, corresponding to at least three times higher likelihood of cleavage than by chance. We first consider cleavage by endopeptidases cathepsins B and S, followed by the aminopeptidase cathepsin H for additional trimming at the N-terminus; we do not use the cathepsin H profile for endoprotease cleavage, because our specificity matrix was constructed using only aminopeptidase cleavage sites.

2. MHCII presentation

Given the antigen peptides predicted in the previous step, we build a comparative model of the pMHCII complex for each possible 12-mer peptide. We can rely on comparative modeling due to low C_α RMSD (<1Å) between pMHCII complexes in the PDB (Fig 5A).

Human pMHCII template structure dataset. Structures of common human MHCII alleles along with their peptides were extracted from the PDB or modeled using other MHCII structures as templates (S2 Table). Because the sequence identity for all MHCII alleles to

existing templates in the PDB is high (87–100%), only non-identical side chains were re-packed on a fixed backbone using SCWRL4 [54]. Additional alleles can be easily modeled and added to the dataset using the same protocol.

pMHCII presentation. Modeling of the pMHCII complex optimizes side chain orientations [54] on a fixed backbone, which is obtained from the template in the human pMHCII template structure dataset. The peptide-MHCII interface in the pMHCII model is scored with an atomic orientation-dependent statistical potential, either *soap_peptide* or *soap_pMHCII*, resulting in the *pMHC_score* for each peptide. These potentials were computed using a previously described procedure [25], as follows: *soap_peptide* was trained using the FlexPepDock dataset of protein-peptide structures and decoys [38]; *soap_peptide* was used for modeling pMHCII complexes of known peptide affinity [55]; *soap_pMHCII* was trained using the resulting models.

3. TCR recognition

Generation of FVIII-specific T cells and TCRBV sequencing. Two hemophilia A subjects provided written informed consent for use of their samples in accordance with the Principles of Helsinki, under protocols approved by the University of Washington and Seattle Children's Hospital IRBs. Clones were isolated by single-cell sorting of CD4[+] T cells stained with phycoerythrin (PE)-labeled HLA-DR1/FVIII2194-2213 MHCII tetramers (Benaroya Research Institute Tetramer Core, Seattle, WA). Clones were expanded and their specificity to HLA-DR1 complexed with FVIII-2194-2213 was confirmed by tetramer staining, proliferation, and cytokine production assays [36]. CDNA libraries were constructed and the TCRβ CDR3 regions expressed by the clones were sequenced by Adaptive Biotechnologies (Seattle, WA) as described [47, 56].

TCR template structure dataset. Human TCR structures were extracted from the PDB to serve as templates for comparative modeling (S5 Table). For each target, 10 models were constructed based on the top 10 templates as ranked by Blast [57], using MODELLER v9.8 protocol for multiple templates [43]. To account for the variability of the CDR3 TCR loops (Fig 5B), 10 models were computed. In cases were the TCRα chain sequence was not available, models for TCRβ chain only were constructed and used.

pMHCII-TCR template structure dataset. The complex between pMHCII and TCR tends to adopt a canonical orientation [58], but variations are possible. We aligned the 20 complexes in our benchmark dataset based on the MHCII structure and calculated the Cα RMSD between TCRs (Fig 6, S4 Table). While most of the structures had RMSD values within 5Å of each other, RMS deviations of up to 24 Å could also be seen (Fig 6C). Therefore, we relied on protein-protein docking algorithm to generate template structures for reliable coverage of possible orientations. We separated the pMHCII and TCR structures from five pMHCII-TCR complexes in the benchmark dataset (PDB codes: 1j8h, 2ian, 2wbj, 4may, and 4gg6) and re-docked them using PatchDock [59, 60]. Finally, for each of these 5 complexes we selected the 100 docking models with the lowest RMSD to the crystal structure to serve as templates. The resulting 500 template structures provide good coverage of possible orientations with an average RMSD of 1.75Å (standard deviation 0.7) between the closest template and the crystal structure for the 20 benchmark cases (S4 Table).

TCR recognition. For each pMHCII model from the previous step, we built a comparative model of the ternary pMHCII-TCR complex, based on each individual ternary complex among the 500 complex templates in the pMHCII-TCR template dataset. The pMHCII and TCR structures were aligned on the template using structural alignment [61]. Similarly to the *pMHC_score*, the interface between TCR and pMHCII was then scored with the soap-PP

atomic distance-dependent statistical potential derived for protein-protein docking [25], resulting in a *TCR_pMHC_score*. We then selected the best scoring pMHCII-TCR model among the 5,000 computed models (ten TCR models times 500 pMHC-TCR templates).

The final peptide ranking was based on first removing the peptides with a *cleavage_score* >3, followed by ranking the surviving peptides by the sum of *pMHC_score* and *TCR_pMHC_score* (*ie*, the final score). These final scores were converted into normalized Z-scores using the mean and standard deviation of the final scores for all antigen sequence 12-mer peptides.

Integration of NetMHCIIpan score. The predicted affinity was converted to a normalized Z-score similarly to *pMHC_score* and *TCR_pMHC_score*. The normalized score was added to other scores.

Supporting information

S1 Fig. Success rate for SOAP_PEP, NetMHCIIpan3.1 and the two scoring functions combined for A) all, B) high, C) intermediate, and D) low affinity peptides.
(TIFF)

S1 Table. Predicted cleavage sites for the benchmark antigens and invariant chain. The predicted cleavage sites are marked with a dot. The confirmed epitope sequences from the ternary complex crystal structures are bolded.
(DOCX)

S2 Table. Human MHCII allele dataset. [1]Alleles frequencies are taken from http://www.allelefrequencies.net. [2]Highest sequence identity to a PDB structure.
(DOCX)

S3 Table. Benchmark of epitope predictions based on 32 crystal structures of peptide-MHCII complexes. The core residues that fit within the 9-residue MHCII groove are bolded.
(DOCX)

S4 Table. RMSD between the complexes, measured as the RMSD between the TCRs after superposition of the MHCII. RMSD to closest structure in the template library in the last column.
(DOCX)

S5 Table. TCR templates dataset (PDB codes).
(DOCX)

S1 File. Identified cathepsin cleavage sites.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Dina Schneidman-Duhovny, Pedro Paz, Andrej Sali.

Data curation: Guang Qiang Dong, Michael B. Winter, Eric Shifrut, Nir Friedman, Charles S. Craik, Kathleen P. Pratt, Pedro Paz, Fred Aswad.

Formal analysis: Dina Schneidman-Duhovny, Michael B. Winter, Nir Friedman, Kathleen P. Pratt, Andrej Sali.

Funding acquisition: Fred Aswad, Andrej Sali.

Investigation: Dina Schneidman-Duhovny, Pedro Paz, Andrej Sali.

Methodology: Dina Schneidman-Duhovny, Guang Qiang Dong, Michael B. Winter, Nir Friedman, Andrej Sali.

Resources: Fred Aswad.

Software: Dina Schneidman-Duhovny, Natalia Khuri, Guang Qiang Dong.

Supervision: Andrej Sali.

Validation: Dina Schneidman-Duhovny, Natalia Khuri, Michael B. Winter, Eric Shifrut, Nir Friedman, Charles S. Craik, Kathleen P. Pratt, Pedro Paz, Fred Aswad.

Visualization: Dina Schneidman-Duhovny, Michael B. Winter, Charles S. Craik.

Writing – original draft: Dina Schneidman-Duhovny, Natalia Khuri, Michael B. Winter, Charles S. Craik, Andrej Sali.

Writing – review & editing: Dina Schneidman-Duhovny, Natalia Khuri, Guang Qiang Dong, Michael B. Winter, Eric Shifrut, Nir Friedman, Charles S. Craik, Kathleen P. Pratt, Pedro Paz, Fred Aswad, Andrej Sali.

References

1. Abbas AK, Lichtman AH, Pillai S. Basic immunology: functions and disorders of the immune system: Elsevier Health Sciences; 2012.

2. Murphy KP, Travers P, Walport M, Janeway C. Janeway's immunobiology. 7th ed. New York: Garland Science; 2008. xxi, 887 p. p.

3. Sadegh-Nasseri S. A step-by-step overview of the dynamic process of epitope selection by major histocompatibility complex class II for presentation to helper T cells. F1000Research. 2016; 5. https://doi.org/10.12688/f1000research.7664.1 PMID: 27347387; PubMed Central PMCID: PMC4902097.

4. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, et al. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nature medicine. 2007; 13 (1):100–6. https://doi.org/10.1038/nm1461 PMID: 17187074.

5. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature. 2014; 515(7528):577–81. https://doi.org/10.1038/nature13988 PMID: 25428507; PubMed Central PMCID: PMC4279952.

6. Tye-Din JA, Stewart JA, Dromey JA, Beissbarth T, van Heel DA, Tatham A, et al. Comprehensive, quantitative mapping of T cell epitopes in gluten in celiac disease. Science translational medicine. 2010; 2(41):41ra51. https://doi.org/10.1126/scitranslmed.3001012 PMID: 20650871.

7. Michels AW, Ostrov DA. New approaches for predicting T cell-mediated drug reactions: A role for inducible and potentially preventable autoimmunity. The Journal of allergy and clinical immunology. 2015; 136(2):252–7. https://doi.org/10.1016/j.jaci.2015.06.024 PMID: 26254052; PubMed Central PMCID: PMC4529959.

8. Nguyen PC, Lewis KB, Ettinger RA, Schuman JT, Lin JC, Healey JF, et al. High-resolution mapping of epitopes on the C2 domain of factor VIII by analysis of point mutants using surface plasmon resonance. Blood. 2014; 123(17):2732–9. https://doi.org/10.1182/blood-2013-09-527275 PMID: 24591205; PubMed Central PMCID: PMC3999758.

9. Newell EW, Sigal N, Nair N, Kidd BA, Greenberg HB, Davis MM. Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. Nature biotechnology. 2013; 31(7):623–9. https://doi.org/10.1038/nbt.2593 PMID: 23748502; PubMed Central PMCID: PMC3796952.

10. Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, et al. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. Cell. 2018; 172(3):549–63 e16. https://doi.org/10.1016/j.cell.2017.11.043 PMID: 29275860; PubMed Central PMCID: PMC5786495.

11. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. Immunology. 2010; 130(3):319–28. https://doi.org/10.1111/j.1365-2567.2010.03268.x PMID: 20408898; PubMed Central PMCID: PMC2913211.

12. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC bioinformatics. 2014; 15:241. https://doi.org/10.1186/1471-2105-15-241 PMID: 25017736; PubMed Central PMCID: PMC4111843.

13. Zhang H, Wang P, Papangelopoulos N, Xu Y, Sette A, Bourne PE, et al. Limitations of Ab initio predictions of peptide binding to MHC class II molecules. PloS one. 2010; 5(2):e9272. https://doi.org/10.1371/journal.pone.0009272 PMID: 20174654; PubMed Central PMCID: PMC2822856.

14. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S. NetMHCIIpan-2.0—Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. Immunome research. 2010; 6:9. https://doi.org/10.1186/1745-7580-6-9 PMID: 21073747; PubMed Central PMCID: PMC2994798.

15. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013; 65(10):711–24. https://doi.org/10.1007/s00251-013-0720-y PMID: 23900783; PubMed Central PMCID: PMC3809066.

16. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015; 43(Database issue):D405–12. https://doi.org/10.1093/nar/gku938 PMID: 25300482; PubMed Central PMCID: PMC4384014.

17. Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS computational biology. 2008; 4 (4):e1000048. https://doi.org/10.1371/journal.pcbi.1000048 PMID: 18389056; PubMed Central PMCID: PMC2267221.

18. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics. 2015; 67(11–12):641–50. https://doi.org/10.1007/s00251-015-0873-y PMID: 26416257; PubMed Central PMCID: PMC4637192.

19. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. Cellular and molecular life sciences: CMLS. 2005; 62(9):1025–37. https://doi.org/10.1007/s00018-005-4528-2 PMID: 15868101.

20. Vider-Shalit T, Sarid R, Maman K, Tsaban L, Levi R, Louzoun Y. Viruses selectively mutate their CD8+ T-cell epitopes—a large-scale immunomic analysis. Bioinformatics. 2009; 25(12):i39–44. https://doi.org/10.1093/bioinformatics/btp221 PMID: 19478014; PubMed Central PMCID: PMC2687975.

21. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome research. 2011; 21(5):790–7. https://doi.org/10.1101/gr.115428.110 PMID: 21349924; PubMed Central PMCID: PMC3083096.

22. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. Journal of immunology. 2000; 164(11):5782–7. PMID: 10820256.

23. Hartman IZ, Kim A, Cotter RJ, Walter K, Dalai SK, Boronina T, et al. A reductionist cell-free major histocompatibility complex class II antigen processing system identifies immunodominant epitopes. Nature medicine. 2010; 16(11):1333–40. https://doi.org/10.1038/nm.2248 PMID: 21037588; PubMed Central PMCID: PMC3058316.

24. O'Donoghue AJ, Eroy-Reveles AA, Knudsen GM, Ingram J, Zhou M, Statnekov JB, et al. Global identification of peptidase specificity by multiplex substrate profiling. Nat Methods. 2012; 9(11):1095–100. https://doi.org/10.1038/nmeth.2182 PMID: 23023596; PubMed Central PMCID: PMC3707110.

25. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics. 2013; 29(24):3158–66. https://doi.org/10.1093/bioinformatics/btt560 PMID: 24078704; PubMed Central PMCID: PMC3842762.

26. Shifrut E, Baruch K, Gal H, Ndifon W, Deczkowska A, Schwartz M, et al. CD4(+) T Cell-Receptor Repertoire Diversity is Compromised in the Spleen but Not in the Bone Marrow of Aged Mice Due to Private and Sporadic Clonal Expansions. Frontiers in immunology. 2013; 4:379. https://doi.org/10.3389/fimmu.2013.00379 PMID: 24312094; PubMed Central PMCID: PMC3832891.

27. Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. Genome medicine. 2013; 5(10):98. https://doi.org/10.1186/gm502 PMID: 24172704; PubMed Central PMCID: PMC3979016.

28. Newell EW, Davis MM. Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. Nature biotechnology. 2014; 32(2):149–57. https://doi.org/10.1038/nbt.2783 PMID: 24441473; PubMed Central PMCID: PMC4001742.

29. Kim A, Hartman IZ, Poore B, Boronina T, Cole RN, Song N, et al. Divergent paths for the selection of immunodominant epitopes from distinct antigenic sources. Nature communications. 2014; 5:5369. https://doi.org/10.1038/ncomms6369 PMID: 25413013; PubMed Central PMCID: PMC4241505.

30. Choe Y, Leonetti F, Greenbaum DC, Lecaille F, Bogyo M, Bromme D, et al. Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. The Journal of biological chemistry. 2006; 281(18):12824–32. https://doi.org/10.1074/jbc.M513331200 PMID: 16520377.

31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000; 28(1):235–42. PMID: 10592235; PubMed Central PMCID: PMC102472.

32. Stoeckle C, Tolosa E. Antigen processing and presentation in multiple sclerosis. Results and problems in cell differentiation. 2010; 51:149–72. https://doi.org/10.1007/400_2009_22 PMID: 19582405.

33. Shao W, Pedrioli PG, Wolski W, Scurtescu C, Schmid E, Vizcaíno JA, et al. The SysteMHC Atlas project. Nucleic Acids Research. 2017.

34. Yin Y, Li Y, Kerzic MC, Martin R, Mariuzza RA. Structure of a TCR with high affinity for self-antigen reveals basis for escape from negative selection. The EMBO journal. 2011; 30(6):1137–48. https://doi.org/10.1038/emboj.2011.21 PMID: 21297580; PubMed Central PMCID: PMC3061028.

35. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome research. 2014; 24(10):1603–12. https://doi.org/10.1101/gr.170753.113 PMID: 25024161; PubMed Central PMCID: PMC4199372.

36. Ettinger RA, James EA, Kwok WW, Thompson AR, Pratt KP. Lineages of human T-cell clones, including T helper 17/T helper 1 cells, isolated at different stages of anti-factor VIII immune responses. Blood. 2009; 114(7):1423–8. https://doi.org/10.1182/blood-2009-01-200725 PMID: 19549986; PubMed Central PMCID: PMC2727410.

37. Ettinger RA, Liberman JA, Gunasekera D, Puranik K, James EA, Thompson AR, et al. FVIII proteins with a modified immunodominant T-cell epitope exhibit reduced immunogenicity and normal FVIII activity. Blood advances. 2018; 2(4):309–22. https://doi.org/10.1182/bloodadvances.2017013482 PMID: 29444872; PubMed Central PMCID: PMC5858479.

38. Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins. 2010; 78(9):2029–40. Epub 2010/05/11. https://doi.org/10.1002/prot.22716 PMID: 20455260.

39. Beringer DX, Kleijwegt FS, Wiede F, van der Slik AR, Loh KL, Petersen J, et al. T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. Nature immunology. 2015; 16(11):1153–61. https://doi.org/10.1038/ni.3271 PMID: 26437244.

40. Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. Advances in protein chemistry and structural biology. 2014; 94:77–120. https://doi.org/10.1016/B978-0-12-800168-4.00004-4 PMID: 24629186.

41. Barkan DT, Hostetter DR, Mahrus S, Pieper U, Wells JA, Craik CS, et al. Prediction of protease substrates using sequence and structure features. Bioinformatics. 2010; 26(14):1714–22. https://doi.org/10.1093/bioinformatics/btq267 PMID: 20505003; PubMed Central PMCID: PMC2894511.

42. Song J, Tan H, Boyd SE, Shen H, Mahmood K, Webb GI, et al. Bioinformatic approaches for predicting substrates of proteases. Journal of bioinformatics and computational biology. 2011; 9(1):149–78. PMID: 21328711.

43. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234(3):779–815. Epub 1993/12/05. S0022-2836(83)71626-8 [pii] https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673.

44. Sharma R, Anguela XM, Doyon Y, Wechsler T, DeKelver RC, Sproul S, et al. In vivo genome editing of the albumin locus as a platform for protein replacement therapy. Blood. 2015; 126(15):1777–84. https://doi.org/10.1182/blood-2014-12-615492 PMID: 26297739; PubMed Central PMCID: PMC4600017.

45. Elder ME, Nayak S, Collins SW, Lawson LA, Kelley JS, Herzog RW, et al. B-Cell depletion and immuno-modulation before initiation of enzyme replacement therapy blocks the immune response to acid alpha-glucosidase in infantile-onset Pompe disease. The Journal of pediatrics. 2013; 163(3):847–54 e1. https://doi.org/10.1016/j.jpeds.2013.03.002 PMID: 23601496; PubMed Central PMCID: PMC3981605.

46. Akatsuka Y, Martin E, Madonik A, Barsoukov A, Hansen JA. Rapid screening of T-cell receptor (TCR) variable gene usage by multiplex PCR: Application for assessment of clonal composition. HLA. 1999; 53(2):122–34.

47. Ettinger RA, Paz P, James EA, Gunasekera D, Aswad F, Thompson AR, et al. T cells from three Hemophilia A subjects recognized the same HLA-restricted FVIII epitope with a narrow TCR repertoire. Blood. 2016:blood-2015-11-682468.

48. Vivian JP, Duncan RC, Berry R, O'Connor GM, Reid HH, Beddoe T, et al. Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. Nature. 2011; 479(7373):401–5. https://doi.org/10.1038/nature10517 PMID: 22020283; PubMed Central PMCID: PMC3723390.

49. O'Donoghue AJ, Knudsen GM, Beekman C, Perry JA, Johnson AD, DeRisi JL, et al. Destructin-1 is a collagen-degrading endopeptidase secreted by Pseudogymnoascus destructans, the causative agent of white-nose syndrome. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112(24):7478–83. https://doi.org/10.1073/pnas.1507082112 PMID: 25944934.

50. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology. 2012; 30(10):918–20. https://doi.org/10.1038/nbt.2377 PMID: 23051804; PubMed Central PMCID: PMC3471674.

51. Chalkley RJ, Baker PR, Medzihradszky KF, Lynn AJ, Burlingame AL. In-depth analysis of tandem mass spectrometry data from disparate instrument types. Molecular & cellular proteomics: MCP. 2008; 7 (12):2386–98. https://doi.org/10.1074/mcp.M800021-MCP200 PMID: 18653769.

52. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. Nature methods. 2009; 6(11):786–7. https://doi.org/10.1038/nmeth1109-786 PMID: 19876014.

53. Rabiner L, Juang B. An introduction to hidden Markov models. ASSP Magazine, IEEE. 1986; 3(1):4–16.

54. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009; 77(4):778–95. Epub 2009/07/16. https://doi.org/10.1002/prot.22488 PMID: 19603484; PubMed Central PMCID: PMC2885146.

55. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC bioinformatics. 2007; 8:238. https://doi.org/10.1186/1471-2105-8-238 PMID: 17608956; PubMed Central PMCID: PMC1939856.

56. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009; 114(19):4099–107. https://doi.org/10.1182/blood-2009-04-217604 PMID: 19706884; PubMed Central PMCID: PMC2774550.

57. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic acids research. 2008; 36(Web Server issue):W5–9. https://doi.org/10.1093/nar/gkn201 PMID: 18440982; PubMed Central PMCID: PMC2447716.

58. Khan JM, Ranganathan S. Understanding TR binding to pMHC complexes: how does a TR scan many pMHC complexes yet preferentially bind to one. PLoS ONE. 2011; 6(2):e17194. Epub 2011/03/03. https://doi.org/10.1371/journal.pone.0017194 PMID: 21364947; PubMed Central PMCID: PMC3043089.

59. Duhovny D, Nussinov R, Wolfson HJ, editors. Efficient Unbound Docking of Rigid Molecules. Second International Workshop, WABI 2002; 2002; Rome, Italy: Springer Berlin / Heidelberg.

60. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005; 33(Web Server issue):W363–7. Epub 2005/06/28. 33/suppl_2/W363 [pii] https://doi.org/10.1093/nar/gki481 PMID: 15980490; PubMed Central PMCID: PMC1160241.

61. Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins. 2004; 56(1):143–56. Epub 2004/05/27. https://doi.org/10.1002/prot.10628 PMID: 15162494.