



RESEARCH

Open Access

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed Mohammed Khan¹, Gaurav Kumar¹, Shoba Ranganathan^{1, 2, §}

Abstract

Background

HLA-DQ alleles are involved in the pathogenesis of hypersensitivity reactions and autoimmune disorders, with HLA-DQ8 associated with several human autoimmune disorders. Limited success has been achieved using sequence-based computational techniques for predicting HLA-DQ8-restricted T cell epitopes while accuracy and efficiency of recently developed structure-based models need to be improved.

Results

We describe a combined structure-based prediction approach for DQ8-restricted T cell epitope prediction using a recently developed fast and accurate docking protocol, pDOCK, and molecular surface electrostatic potential (MSEP)-based clustering of pMHC binding interfaces. The prediction model was rigorously trained, tested and validated using experimentally verified DQ8 binding and non-binding peptides. High MHC-binding prediction accuracy is validated against independent experimental data (average area under the ROC curve or average $A_{ROC} > 0.94$). Our model also predicts all binding registers correctly and known T cell activators with a positive predictive value (PPV) of 0.91 or 91%. We also studied the binding patterns of DQ8-binding peptides and confirm the existence of peptide epitopes that do not conform to DQ8-specific consensus peptide-binding motifs.

Conclusions

We have developed a model that can be successfully applied as a generic protocol for easy *in silico* identification of HLA-DQ8 binding peptides and thereby potential DQ8-specific T cell epitopes. The current model is therefore applicable for screening potential vaccine candidates from DQ8 binding peptides irrespective of consensus peptide-binding or sequence motifs. We have also illustrated efficient discrimination of different categories of binders from non-binders as well as different categories of pMHC agonists from non-agonists, while accurately predicting the binding registers of DQ8-restricted peptides. This combined approach provides a set of sensitive and specific computational tools to facilitate high-throughput screening of peptides for immunotherapeutic applications such as controlling allergic and autoimmune responses.

Background:

Among many important proteins that take part in adaptive immune responses, major histocompatibility complex (MHC) and T cell receptor (TR) proteins arguably play the most crucial role. The MHC proteins bind and present short antigenic peptides on the cell surface, as peptide-MHC (pMHC) complexes, for recognition by the TR proteins to form T cell receptor-peptide-MHC (TR/pMHC) complexes which subsequently activate the T cells to carry out the immune response [1, 2]. Both of these steps trigger a series of immunological events essential for initiation and regulation of immune responses [3-5].

Broadly classified into two types, MHC class I (MHC-I) proteins bind and present endogenous (processed within the cell) peptides for recognition by the CD8⁺ cytotoxic T cells whereas MHC class II (MHC-II) proteins prefer exogenous (processed outside the cell) peptides for

recognition by the CD4⁺ helper T cells [4]. While MHC-I alleles have been extensively studied [1, 6, 7], investigations pertaining to MHC-II alleles have been hindered, especially in the context of MHC-II restricted T cell epitope prediction, primarily due to the greater promiscuity of MHC-II peptide binding compared to MHC-I peptide binding that makes T cell epitope prediction a challenging task and the lack of MHC-II related biochemical, functional and crystallographic data [3]. However, recent growth in both experimental and structural data for MHC-II alleles has facilitated their analysis for the development of T cell epitope prediction methods applicable to MHC-II proteins [8-10].

Currently available computational protocols for the identification of MHC-II restricted T cell epitopes can be categorized into sequence and structure-based methods. Sequence-based methods are relatively advanced in predicting T cell epitopes for MHC-II alleles, such as HLA-DR [11-13], with abundant biochemical peptide binding data. Nonetheless, for MHC-II alleles with limited peptide data, such as HLA-DQ, these approaches have been used with varying degree of success [14-17], owing to their dependence on experimental data for training purposes. Recently though, Wang *et al.* [18] have reported

¹ Department of Chemistry and Biomolecular Sciences and ARC Center of Excellence in Bioinformatics, Macquarie University, NSW 2109, Australia.

² Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

§Corresponding author

Email addresses:

JMK: javed.khan@mq.edu.au

GK: gaurav.kumar@mq.edu.au

SR: shoba.ranganathan@mq.edu.au



an average area under the ROC curve (A_{ROC}) of 0.88 for HLA-DR, DP and DQ alleles, marking a promising improvement in the applicability of sequence-based prediction approaches for MHC-II alleles with limited peptide binding data. On the other hand, structure-based procedures such as docking [9, 19] have been successfully applied to predict T cell epitopes even for MHC-II alleles with very limited peptide data while addressing the dual issues of docking and scoring for MHC-II binding peptides [8-10]. However, much like all other methods, even the docking and scoring-based approach reported previously [8-10] utilizes only pMHC interaction data such as the docked pMHC binding energy for T cell epitope prediction while not accounting for the TR entity and hence the TR/pMHC interactions, which affects its prediction accuracy. Therefore, in view of the fact that only 50% of strong MHC-binding peptides are known to activate T cells [20] and given the significance of TR/pMHC binding in T cell mediated immunity, it becomes extremely important to factor in TR/pMHC interaction knowledge in conjunction with pMHC binding data for improved prediction of immunogenic T cell epitopes. Also, the speed and efficiency of the docking protocol need to be improved for high-throughput screening of MHC-binding peptides to identify high-binders.

Many HLA-DQ alleles are known to be involved in the pathogenesis of hypersensitivity reactions [21, 22] and autoimmune disorders [23]. Among these, an allele of particular interest is HLA-DQ8 (made up of DQA1*0301 and DQB1*0302, and also known as HLA-DQ3.2 β) due to its association with various human autoimmune disorders such as insulin-dependent diabetes mellitus (IDDM) [24, 25], autoimmune encephalomyelitis [26], autoimmune polyendocrine syndrome type II (APS-II) [27], IDDM-associated periodontal disease [28] and celiac disease [29] and hypersensitivity disorders including house dust mite allergy [21, 22]. DQ8 is found in approximately 20-30% of the human population [30] and is prevalent in about 86% of IDDM patients [31]. Hence, in order to elucidate the role of DQ8 both in autoimmunity and allergenicity, enhanced understanding of DQ8-restricted pMHC and TR/pMHC binding is essential.

Recently, we have developed pDOCK [4] which is a robust new protocol for rapid and accurate fully-flexible docking of peptides to MHC-I and MHC-II alleles. Benchmarking pDOCK with the previous docking technique [9, 19] revealed a 2.5 fold and ~60% increase in its accuracy and speed, respectively. Upon validation against previously published studies, a seven-fold increase was recorded in pDOCK accuracy. pDOCK also accurately determined the binding registers of all MHC-I and MHC-II binding peptides used in that study. We have also analyzed 61 (50 TR/pMHC-I and 11 TR/pMHC-II) available TR/pMHC crystal structures [5] collated from the MPID-T2 [32] database and identified certain structural interaction characteristics such as molecular surface electrostatic potential (MSEP) that can be used as parameters governing TR/pMHC complex formation for T cell epitope prediction. Therefore, we have now combined the power of pDOCK to successfully

identify strong MHC-binding peptides using a previously developed complementary scoring function [9] and the efficient MSEP-based clustering of pMHC binding interfaces [5] to predict DQ8-restricted immunogenic T cell epitopes with high accuracy and correct binding registers. We also investigated the binding patterns of DQ8-restricted peptides and confirm the existence of peptide epitopes that do not conform to previously reported DQ-8 binding motifs [9].

Results and discussion

Evaluation of the accuracy of the DQ8 prediction model was carried out in two steps: (i) assessment of efficiency of the optimized scoring function using test set 1; and (ii) verification of the overall prediction accuracy of the model using test set 2. The accuracy of our model partially relies on the scoring function used. Using default Internal Coordinate Mechanics (ICM) coefficients ($\alpha=\beta=\gamma=1$; $C=0$) in the scoring function for the training set resulted in a correlation with an $r^2=0.79$ and $s=2.05$ kJ/mol between the predicted binding energy values (from docking) and the experimental binding free energy values (computed using IC_{50} values). Better correlation ($r^2=0.82$, $s=1.95$ kJ/mol) was achieved after recalibration of the scoring function by fitting to the training data using multiple linear regression thereby significantly improving the discriminative power of the scoring function.

Following 10-fold cross-validation ($N=51$, $q^2=0.80$, $s_{press}=2.20$ kJ/mol), the optimal scoring function is:

$$\Delta G_{bind} = 0.015 \Delta G_{EL} - 0.859 \Delta G_H + 0.827 \Delta G_{EN} - 1.91 \quad (1)$$

The entropic (ΔG_{EN}) and the electrostatic (ΔG_{EL}) terms are positive, while the overall computed binding energy (ΔG_{bind}) and the hydrophobic (ΔG_H) term are negative (see Methods section for details of each term and their contributions).

Rognan *et al.* [33] performed a leave-one-out cross-validation on training datasets of five and 37 pMHC complexes. However, the current training set of 57 complexes is comparatively larger for such analyses, while it is too small for extensive cross validation as reported by Wang *et al.* [34] and Bock and Gough [35] for training sets of 200 and 2617 complexes, respectively. Therefore, we have chosen to use k -fold cross validation, which is more suited to a dataset of this size. Moreover, such a cross validation approach also helps us compare our results with those reported in the earlier study of Tong *et al.* [9]. It is worth noting that the standard error in the training set after recalibration of the scoring function ($s=1.95$ kJ/mol=0.46 kcal/mol) is less than the standard error after 10-fold cross-validation ($s_{press}=2.20$ kJ/mol=0.52 kcal/mol) as expected and unlike the higher standard error after recalibration ($s=4.77$ kJ/mol=1.13 kcal/mol) than the standard error after 10-fold cross-validation ($s_{press}=2.20$ kJ/mol=0.52 kcal/mol) reported by Tong *et al.* [9], highlighting the ability of pDOCK [4] to handle noise in data and showcasing its robustness. Also,



our standard error values both before and after recalibration ($s=2.05$ kJ/mol and $s=1.95$ kJ/mol, respectively) are significantly lower to the ones ($s=2.91$ kJ/mol and $s=4.77$ kJ/mol, respectively) documented by Tong *et al.* [9].

We have also performed evolutionary regression analysis, similar to the one carried out previously [9], to estimate the robustness of the scoring function for 5-fold ($N=46$, $q^2=0.79$, $s_{press}=2.09$ kJ/mol), 4-fold ($N=43$, $q^2=0.77$, $s_{press}=2.07$ kJ/mol), 3-fold ($N=38$, $q^2=0.78$, $s_{press}=2.05$ kJ/mol) and 2-fold ($N=29$, $q^2=0.74$, $s_{press}=2.03$ kJ/mol) cross-validations and once again the standard error values for our training set were comparatively lower. Importantly, the cross-validation coefficient q^2 and the standard error of prediction s_{press} are stable all through, with mean values of $q^2=0.78$ and $s_{press}=2.09$ kJ/mol, and the respective standard deviation values of 0.02 and 0.07 kJ/mol. These results do not indicate any unusual increase in the standard error values for any of the folds contrasting to the reports of an uncharacteristic increase in the error value for 2-fold cross-validation by Tong *et al.* [9]. The internal consistency of the optimized scoring function used in this prediction model is therefore validated by this iterative regression procedure, rendering it suitable for the identification of MHC-binders within the test datasets and hence for use in our prediction model.

In order to evaluate the efficiency of the optimized scoring function, three decision threshold binding energy values (Table 1), which define levels of specificities suitable for practical applications [36], were used to determine the corresponding sensitivity values on different subsets – H (high-affinity binders only; $A_{ROC}=0.89$); MH (medium- and high-affinity binders; $A_{ROC}=0.96$); and LMH (low-, medium-, high-affinity binders; $A_{ROC}=0.98$) from test set 1. The suitable use of structural data for discriminating MHC-II binding peptides from the background with almost excellent accuracy ($A_{ROC}\geq 0.89$) is advocated by these outcomes. In general, very few false positives and a large number of true positives were observed at $SP=0.95$ contrasting to the previous report of fewer true positives at $SP=0.95$ [9], shedding light on pDOCK's efficiency even at higher levels of specificity. High-sensitivity predictions are commonly expected at

$SP=0.80$ [9]. Our MHC-binding prediction results for test set 1 (Table 1) fit almost perfectly with the expected binding patterns of DQ8-binding peptides, providing a sensitivity of 99% (at $SP=0.80$ for MH and H).

With higher levels of specificity, a gradual decrease in sensitivity values (at $SP=0.90$, $SE=0.97$ for LMH and MH and $SE=0.96$ for H; at $SP=0.95$, $SE=0.97$, 0.94 and 0.89 for LMH, MH and H, respectively) is observed. On an average however, the sensitivity values are above 96%, with $SE=0.89$ (89% of high-affinity binders are correctly identified) being the worst case scenario. The efficacy of pDOCK [4] in accurately detecting binding registers was also evaluated with experimentally determined registers. Our findings reconfirm our earlier observation [4] that pDOCK accurately determines binding registers for all the peptides docked from the training set and test set 1 (Additional File 1 – Table S1 and Additional File 2 – Table S2, respectively). All 22 experimentally determined registers from test set 2 (Additional File 3 – Table S3) were also correctly predicted by pDOCK.

Finally, using the binding energy decision threshold (-37.91 kJ/mol) defined above for high-affinity binders at the specificity of 95% ($SE=0.89$), the predictive performance and the accuracy of the scoring function was tested on a functional dataset of 36 peptides (test set 2) known to bind DQ8, out of which, 31 were T cell activators and 5 were non-activators. However, upon application of the binding energy threshold of -37.91 kJ/mol five peptides that were known T cell activators were eliminated and the top 31 predictions included the 5 non-activators. Through structural analysis of all available TR/pMHC crystal structures, we have recently shown that MSEF at both pMHC and TR binding interfaces play a major role in TR/pMHC complex formation and thus in T cell activation [5]. Therefore, we applied this knowledge and performed MSEF-based clustering of the above identified top 31 peptide docked pMHC interfaces from test set 2 along with pMHC interfaces from all six available TR/pMHC-II structures using the webPIPSA [37] server, to identify electrostatic similarities between them. The electrostatic distances for all pMHC interfaces clustered, varied from 0.077 to 1.187 (colour key and density plot-inset in Additional File 4 – Fig. S1).

Clustering identified 8 pMHC interfaces (group B in Additional File 4 – Fig. S1) as non-agonists which included three known non-activators (pMHC 2, 3 and 9 in group B from Additional File 4 – Fig. S1; peptides 32, 33 and 36 in Additional File 3 – Table S3). Hence, combining the results from the application of the optimized scoring function and from clustering, peptides that were predicted to be T cell activators in both instances were selected, resulting in 23 peptides (Additional File 5 – Table S4). The predicted binding energy values for these 23 peptides range from -38.05 to -43.67 kJ/mol and 21 of these are known T cell activators (true positives), thereby, resulting in a positive predictive value ($PPV = \text{True Positives} / (\text{True Positives} + \text{False Positives})$; a measure of accuracy) of 0.91 or 91%. This is in accordance with existing reports that high-affinity binders have a greater

Table 1 – Identification of MHC-binders to DQ8.

Sensitivity values and binding energy thresholds for specificity levels of 0.80, 0.90 and 0.95.

Specificity (SP) Level	Group	Sensitivity (SE)	Binding Energy Threshold (kJ/mol)
SP = 0.80	LMH	0.98	-29.55
	MH	0.99	-34.00
	H	0.99	-35.20
SP = 0.90	LMH	0.97	-29.63
	MH	0.97	-34.90
	H	0.96	-36.50
SP = 0.95	LMH	0.97	-29.70
	MH	0.94	-35.25
	H	0.89	-37.91



chance of stimulating T cells [38-40] and that they are critical for peptide vaccine design. Nonetheless, TR agonistic properties of pMHC interfaces obtained from MSEP-based clustering are used as primary criterion and predicted pMHC binding energy is used as secondary parameter in ranking the peptides in Additional File 5 – Table S4. This is primarily because although strong pMHC binding is a prerequisite for TR recognition [40], it does not necessarily mean T cell activation as there is only a 50% chance of immunogenicity [20] even among strong MHC-binding peptides.

Our combined prediction model has successfully eliminated the *Dermatophagoides pteromyssinus* (*Der p 2*) allergenic peptides 11-30, 21-40 and 81-100 (pMHC 2, 3 and 9 in group B from Additional File 4 – Fig. S1), which are all known non-activators, from the list of our top 23 predictions as opposed to their inclusion and eight, tenth and ninth ranking, respectively, in the list of top 12 predictions for T cell activators by Tong *et al.* [9]. Similarly, the *Der p 2* peptide 41-60 (pMHC 5 in group B from Additional File 4 – Fig. S1; peptide 23 in Additional File 3 – Table S3) is categorized as a non-agonist by our MSEP-based clustering which is in complete agreement with an experimental study [21]. Moreover, our top 23 predictions include several peptides for which our model was able to correctly predict the known experimental binding registers (Additional File 5 – Table S4). For example, the *Der p 2* peptide 1-20 (ranked 15th in Additional File 5 – Table S4), where our results are in absolute accordance with both the experimental study [22] and previous reports by Tong *et al.* [9]. The above examples shed light on the efficiency of our prediction model. This combined approach has correctly identified 21 of the 31 T cell activators with a PPV of 91%, representing high prediction accuracy for successful prediction of immunogenic T cell epitopes. Therefore, the current prediction model is suitable to screen for DQ8 high-affinity binders at SP=0.95 and then, to identify potential T cell epitopes among the DQ8 high-binding peptide set.

Epitopes not conforming to binding motifs exist

Identification of potential immunodominant epitopes within autoantigenic proteins has been done for many MHC-II alleles by developing allele-specific consensus peptide-binding motifs. Nevertheless, prior reports have suggested that the existence of these motifs in a given peptide does not necessarily render allele-specificity to the peptide [16]. This study has revealed that considering all relevant residue positions (P1, P4, P6, P7, P9) for peptides, all 70 peptide sequences (Additional File 6 – Table S5) have one or more amino acid residues and 56 peptides (Additional File 6 – Table S5) have two or more residues that do not conform to available DQ8 binding motifs [14, 17], in test set 1 alone. Using existing DQ8 binding motifs, the peptides A-gliadin 49-63 (#1), VP16 (#23) and MHC Ia 46-63 (#24) are generally considered negatives, however, from Additional File 6 – Table S5 it is evident that these T cell epitopes are easily identified

just by using our scoring function, thereby, reaffirming earlier observations by Tong *et al.* [9]. This yet again proves that many other factors such as peptide and MHC binding groove physicochemical composition have to be considered in T cell epitope prediction systems and binding motifs by themselves are inadequate for identifying T cell epitopes.

Conclusions

Despite rapid advances in peptide vaccine development, identifying allele-specific T cell epitopes, especially MHC-II restricted epitopes, suitable for designing vaccines and immunotherapies remains a challenging prospect. Various excellent approaches [41-43] have been adopted by researchers to address this issue. However, training of predictive models using peptides preselected based on existing binding motifs renders them unable to predict epitopes that do not conform to binding motifs.

Our overall significant outcomes along with increasing evidence for inadequacy of binding motifs in defining T cell epitopes, suggest that we have developed a model that can be successfully applied as a generic protocol for easy *in silico* identification of HLA-DQ8 binding peptides and thereby potential DQ8-specific T cell epitopes. The current model is therefore applicable for screening vaccine candidates irrespective of sequence motifs. Also, pDOCK [4] accurately predicts all binding registers, eliminating the use of the nine-residue sliding window approach used by Tong *et al.* [9] that resulted in multiple registers within candidate DQ8-binding peptides. The key to the success of our prediction technique lies in the fact that we have employed a vital combined two-step methodology, where we have utilized pMHC binding energy as the first parameter to segregate DQ8 high-affinity binding peptides and then a TR/pMHC interaction characteristic (MSEP) as the second criterion to filter out potential DQ8-specific T cell epitopes amongst the DQ8 high-affinity binding peptides. It could also be inferred that this two-step prediction model theoretically or computationally mimics the two-steps, namely pMHC binding and TR recognition of pMHC complex, that lead to the formation of TR/pMHC complex and subsequently cause T cell activation, thereby making it more suitable for the identification of potentially immunogenic T cell epitopes. We have also illustrated efficient discrimination of different categories of peptide binders from non-binders, using the scoring function, as well as different categories of pMHC agonists from non-agonists, using MSEP, while accurately predicting the binding registers of DQ8-restricted peptides. Our overall PPV (0.91 or 91%), which denotes the accuracy of our model in identifying immunogenic T cell epitopes, represents a significant improvement over that of Tong *et al.* [9] (0.58 or 58%), across a larger dataset. Therefore, this combined approach provides a set of sensitive and specific computational tools to facilitate high-throughput screening of peptides for immunotherapeutic applications such as controlling allergic and autoimmune diseases.

Methods

Data

Structural data

The crystal structure of Insulin B9-23-DQ8 pMHC complex [44], with the Protein Data Bank or PDB [45] code 1jk8, was used to extrapolate the structural coordinates of the DQ8 allele. ICM package version 3.6-1 [46] was then used to relax the extracted structure by conjugate gradient minimization.

Biochemical and functional peptide binding data

The dataset of 1719 peptides with known binding affinity values for DQ8, used by Wang *et al.* [18], is both not publicly available and/or listed in the published article. Therefore, we have used the available set of peptides known to bind DQ8 and/or elicit T cell proliferation, for this study. The experimental data used for this investigation was primarily divided into two datasets: (i) peptides from biochemical studies with experimental IC_{50} values and (ii) peptides from functional T cell assays that are known to cause T cell activation.

127 peptides with experimentally determined IC_{50} values, obtained from biochemical studies [15, 47-49], comprised dataset I. These peptides with known IC_{50} values were further classified as high-affinity MHC-binders: $IC_{50} \leq 500$ nM, medium-affinity binders: $500 \text{ nM} < IC_{50} \leq 1500$ nM, low-affinity binders: $1500 < IC_{50} \leq 5000$ nM and non-binders: $5000 \text{ nM} < IC_{50}$. Therefore, dataset I was made up of 70 high-affinity, 14 medium-affinity, 29 low-affinity binders and 14 non-binders. Although 14 peptides were considered non-binding based on their IC_{50} values, some of them have reported binding registers. Similarly, some of the peptides that are regarded as binders (high-, medium- and low-affinity binders; 113 peptides) did not have any known binding registers. Therefore, 87 (84 binding and three non-binding) peptides in this dataset had experimentally determined binding registers and 40 (29 binding and 11 non-binding) peptides had

no known binding registers. This dataset was further divided into the training set (Additional File 1 – Table S1; 57 peptides with 43 high-, five medium-, five low- and four non-binders) and test set 1 (Additional File 2 – Table S2; remaining 70 peptides with 27 high-, nine medium-, 24 low- and 10 non-binders).

36 DQ8-specific peptides, out of which, 31 were known to cause DQ8-restricted T cell proliferation and five were known to not activate T cells, formed dataset II. These peptides were listed in previously published functional studies [21, 22, 49, 50] which utilized *in vitro* immunoassays to detect T cell activity. Therefore, this peptide set was collated and was subsequently used as test set 2 (Additional File 3 – Table S3) in this study.

Prediction Model

DQ8-binding and non-binding peptide sequences were docked into the MHC peptide binding groove of the X-ray crystallographic structure for DQ8 using pDOCK [4]. Following this, a previously reported customized free energy scoring function [9] was utilized to improve the predictive performance of the model. Finally, MSEP-based clustering [5] of the peptide docked pMHC binding interfaces from test set 2 was performed to enhance the accuracy of the model and effectively predict DQ8-restricted immunogenic T cell epitopes. Figure 1 illustrates the prediction model developed using the combined approach in this study.

Docking of peptides to DQ8

Docking of all peptides to the extracted template crystal structure of DQ8 was performed using pDOCK [4]. pDOCK utilizes the ICM [46] optimal-bias Monte Carlo minimization procedure [51] which in turn uses the Merck Molecular Force Field or MMFF [52] and Empirical Conformational Energy Program for Peptides 3 (ECEPP/3) force field parameters [53] to perform each docking. In brief, the pDOCK protocol involves: (i) preparatory step 1: MHC receptor modeling and/or position-

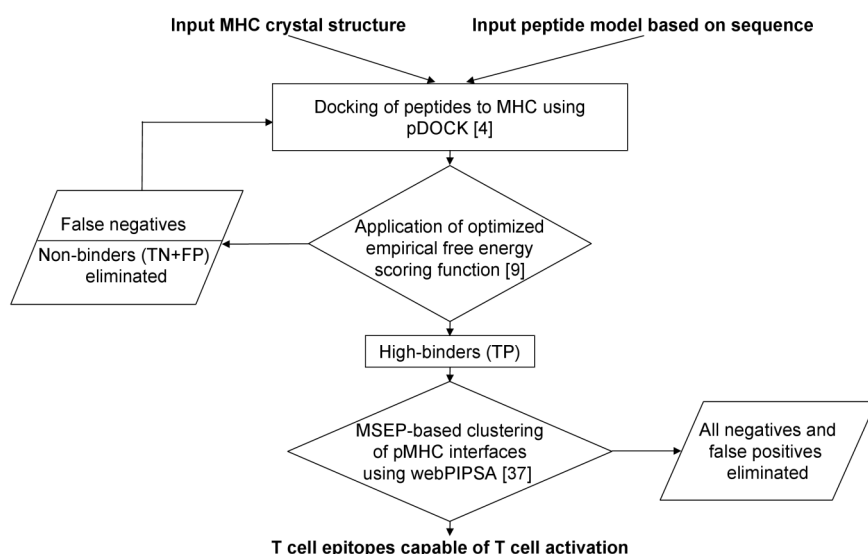


Figure 1 – Flowchart of the prediction model used in this work.

ing using the ICM global optimization algorithm [46]; (ii) preparatory step 2: determining the docking grid by defining the grid dimensions (length x breadth x height) based on standardized values [4] for MHC supertypes (MHC-I and MHC-II; MHC-II in this case) for peptide placement and grid map generation within the vicinity of the MHC peptide binding site and; (iii) a single consolidated final docking and refinement step: peptide positioning within the grid, fully flexible docking of the peptides into the peptide binding groove followed by iterative *ab initio* refinements of all peptide residues along with the backbone and peptide interacting side-chain dihedral angles of the MHC binding site residues to eliminate or minimize atomic clash regions at the pMHC interface, using the ICM global optimization docking algorithm [46] and a biased Monte Carlo procedure [51]. The preparatory steps were together used to generate the MHC receptor maps and the final single docking and refinement step was used to carry out peptide docking, generate the final least energy docked peptide conformation and further refine the product. pDOCK was run on a 2 CPU 3.20 GHz 3 GB RAM workstation.

Empirical free energy scoring function

A previously reported scoring function [9] is employed in this investigation. Originally based on the free energy potential [51] in the ICM 3.6-1 software package [46], this adopted scoring function has its binding free energy calculated as the difference between the energy of the solvated pMHC complex and the sum of the energy of the solvated MHC receptor and the peptide. The fully relaxed conformation of the free peptide in water [54] is chosen as the reference state for a given peptide. The MHC and the peptide are separated after docking and their relaxed energies are computed, following energy minimization in water for all binding energy calculations. Therefore, the binding free energy (ΔG_{bind}) function used here is expressed as follows:

$$\Delta G_{\text{bind}} = \alpha \Delta G_{\text{EL}} + \beta \Delta G_{\text{H}} + \gamma \Delta G_{\text{EN}} + C \quad (2)$$

where, ΔG_{EL} is the electrostatic contributions from the desolvation of partial charges transferred from an aqueous medium to a protein core environment and the pMHC coulombic interactions. Using an implementation of the boundary element algorithm [54, 55], the numeric solution of the Poisson equation determines ΔG_{EL} . The hydrophobic energy (ΔG_{H}) is composed of the product of solvent accessible surface area (determined by rolling a sphere of radius 1.4Å along the surface of the molecule) by the surface tension. The entropic term of the protein side-chains is denoted by ΔG_{EN} and is computed from the maximal burial entropies for each type of amino acid and their relative accessibilities. Entropy change in the system due to the decrease of free molecular concentration and the loss of rotational/translational degrees of freedom upon binding [54], is accounted for by the constant term C or K [33]. Generally, physical parameters that are independent of the dataset used represent C . It has been noted [56] that, among various research groups, there are great

variations in the value used for C . To obtain the best separation of binders and non-binders, the coefficients (α , β , γ) assigned to each energy term in this scoring function were optimized. Many previous studies [9, 54, 57, 58] have successfully used this separation schema consisting of the most significant potentials contributing to protein-protein, protein-ligand and protein-peptide interactions.

Optimizing the scoring function

A similar approach to that employed by Tong *et al.* [9] was again utilized for optimization of the above described scoring function. Initially, the concentration of ligand required to saturate half of the available binding sites of the protein [35], in other words, the reported IC_{50} values (for dataset I), were considered to be similar to the equilibrium dissociation constants (K_d) since the concentration of the ligand in the unbound state is much lower than K_d of the ligand in the binding assay, such that $\Delta G_{\text{bind}} \approx -RT \ln (\text{IC}_{50})$ [33]. This was followed by recalibration of the coefficients for different energy terms by standard least-square multivariate regression analysis, as previously described by Wang *et al.* [34], of the training set to improve the discriminative power of the scoring function.

Subsequently, quality of the scoring function was assessed using 10-fold cross-validation (Fig. 2) [35]. The technique utilized here is called k -fold cross-validation, where a scoring function is trained on $(k-1)$ partitions by constructing k random, (approximately) equal-sized, disjoint partitions of the sample data, and tested on the excluded partition. After k such experiments, the results are averaged and an estimate of the error rate expected upon generalization to new data is given by the observed error rate. Finally, the cross-validation coefficient q^2 and the standard error of prediction s_{press} were used to evaluate the predictive power of the scoring function. Further evaluation using evolutionary regression analysis [34] with different subsets representing 5-fold, 4-fold, 3-fold and 2-fold cross-validation (Fig. 2), was also conducted to assess the robustness of the scoring function.

Clustering of pMHC interfaces

This step was carried out as recently outlined by us [5]. However, in this case the peptide docked pMHC complexes for the top 31 peptides (obtained upon the application of the binding energy decision threshold of -37.91 kJ/mol, defined for high-affinity binders at the specificity of 95% and sensitivity of 89% from Table 1) from test set 2 were used to extract the coordinates for truncated versions of the pMHC complexes, encompassing only the pMHC binding interfaces. These pMHC interfaces were subjected to MSEP-based clustering along with known human T cell activating pMHC-II binding interfaces [5] from all six currently available human TR/pMHC-II crystal structures to identify MSEP similarities between the peptide docked pMHC interfaces and the human pMHC-II interfaces known to activate T cells depending on the electrostatic distances between them. The webPIPSA [37] server was used to calculate MSEP and compare

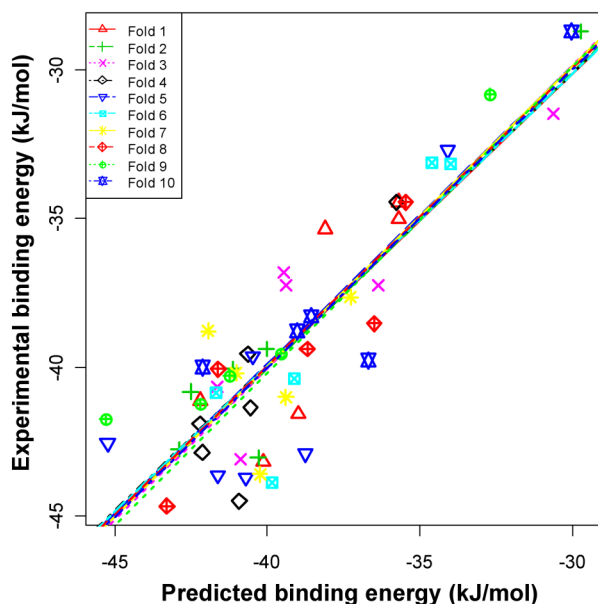


Figure 2 – Relationship between experimental and predicted binding energies from 1-fold to 10-fold cross-validations. Although the figure depicts folds 1-10, cross-validation results only for folds 2, 3, 4, 5 and 10 are discussed.

electrostatic interaction properties of the pMHC interfaces.

The web-server begins with calculation of pMHC interface MSEP using the University of Houston Brownian Dynamics (UHBD) program [59] and then compares their electrostatic properties by calculating similarity indices for all pairs of pMHC interfaces based on their electrostatic similarity, using the PIPSA algorithm [60]. These similarity indices are then converted to electrostatic distances which are clustered and displayed as a colour coded matrix called heat map (Additional File 4 – Fig. S1) using the R [61] software package. This clustering output (Additional File 4 – Fig. S1) was divided into five groups with the six known human T cell activating pMHC-II interfaces in group A, the test set 2 pMHC interfaces with the largest electrostatic distance values compared to group A forming group B (regarded as non-agonists), group C comprising the weak-agonists, test set 2 pMHC interfaces with the least electrostatic distance values compared to group A making up group D (considered as strong-agonists) and group E being composed of moderate-agonists.

Training, testing and validation

The bound conformations of binding peptides with experimentally determined registers and the best conformations of non-binding peptides without any preferred registers were sampled to initially train the DQ8 prediction model. Among the 57 peptides in the training set, 55 (53 binding and two non-binding) peptides had known binding registers and 2 non-binding peptides had no known binding conformations. After optimization of the empirical free energy scoring function by statistical analyses performed on the training set, the optimized scoring function was tested on test set 1 to further assess its predictive ability. Test set 1 had 32 (31 binding and one non

-binding) peptides with known binding registers and 38 (29 binding and nine non-binding) peptides with no known binding registers. Following this, the optimized scoring function and the MSEP-based clustering approach were together applied on test set 2 to improve the overall accuracy of the prediction model, thereby, validating it against experimental T cell activation data. Test set 2 had 22 DQ8-binding peptides with known binding registers and 14 DQ8-binding peptides with no known binding registers.

Similar to the method reported by Tong *et al.* [9], we performed sensitivity (SE), specificity (SP) and receiver operating characteristic (ROC) analyses, described previously by Brusic *et al.* [36], on test set 1 to evaluate the efficiency of the optimized scoring function. The percentages of correctly predicted MHC-binders and non-binders are given by $SE = TP / (TP + FN)$ and $SP = TN / (TN + FP)$, respectively. Experimental MHC-binders with at least one predicted binding register and experimental non-binders with no predicted binding register are represented by true positives (TP) and true negatives (TN), respectively. Whereas, experimental MHC-binders predicted as non-binders and experimental non-binders predicted as MHC-binders, are denoted by false negatives (FN) and false positives (FP), respectively. Similarly for test set 2, experimental T cell activators and experimental non-activators are regarded as true positives (TP) and true negatives (TN), respectively. While experimental T cell activators predicted as non-activators and experimental non-activators predicted as activators, are considered false negatives (FN) and false positives (FP), respectively. ROC analysis, where the ROC curve is generated by plotting SE as a function of (1-SP) for various classification thresholds, was used to verify the accuracy of our MHC-binding predictions. A measure of the MHC-binding prediction accuracy is provided by the area under the ROC curve (A_{ROC}), where $A_{ROC} < 70\%$ denotes poor, $A_{ROC} > 80\%$ is for good and $A_{ROC} > 90\%$ represents excellent predictions [36]. The values of $SP \geq 80\%$ are considered useful in practice [9]. Thus, SE values for three values of SP (80%, 90% and 95%) in test set 1, were assessed. Finally, a PPV analysis was also conducted to determine the accuracy of our T cell epitope predictions within test set 2.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JMK developed the methodology, carried out the computational simulation studies and drafted the manuscript. JMK, GK and SR performed the statistical calculations and analyses. JMK and SR participated in the design of the study and interpretation of data. SR conceived the project and finalized the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

JMK gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship. We also thank Mr. H.R. Cheruku, Macquarie University, Sydney for assistance with additional files and Dr. Vladimir Brusic, Dana-Farber Cancer Institute, Boston for help with binding energy threshold determination. Open access publication charges were borne by Macquarie University.

References

- Rammensee HG, Falk K, Rotzschke O: **Peptides naturally presented by MHC class I molecules.** *Annu Rev Immunol* 1993, **11**:213-244.
- Lefranc MP, Lefranc G: *The T cell receptor FactsBook*. San Diego: Academic Press; 2001.
- Khan JM, Tong JC, Ranganathan S: **Structural Immunoinformatics: Understanding MHC-peptide-TR binding.** In *Bioinformatics for Immunomics. Volume 3*. Edited by Davies MN, Ranganathan S, Flower DR. New York: Springer; 2010: 77-94: *Immunomics Reviews Series*].
- Khan JM, Ranganathan S: **pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes.** *Immunome Res* 2010, **6** Suppl 1:S2.
- Khan JM, Ranganathan S: **Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one.** *PLoS One* 2011, **6**:e17194.
- Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs.** *Hum Immunol* 2002, **63**:701-709.
- Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M: **NetMHCpan, a method for MHC class I binding prediction beyond humans.** *Immunogenetics* 2009, **61**:1-13.
- Tong JC, Tan TW, Sinha AA, Ranganathan S: **Prediction of desmoglein-3 peptides reveals multiple shared T-cell epitopes in HLA DR4- and DR6-associated pemphigus vulgaris.** *BMC Bioinformatics* 2006, **7** Suppl 5:S7.
- Tong JC, Zhang GL, Tan TW, August JT, Brusci V, Ranganathan S: **Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides.** *Bioinformatics* 2006, **22**:1232-1238.
- Tong JC, Zhang ZH, August JT, Brusci V, Tan TW, Ranganathan S: **In silico characterization of immunogenic epitopes presented by HLA-Cw*0401.** *Immunome Res* 2007, **3**:7.
- Brusic V, Bajic VB, Petrovsky N: **Computational methods for prediction of T-cell epitopes--a framework for modelling, testing, and applications.** *Methods* 2004, **34**:436-443.
- Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O: **Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan.** *PLoS Comput Biol* 2008, **4**:e1000107.
- Dimitrov I, Garnev P, Flower DR, Doytchinova I: **Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis.** *Eur J Med Chem* 2010, **45**:236-243.
- Godkin A, Friede T, Davenport M, Stevanovic S, Willis A, Jewell D, Hill A, Rammensee HG: **Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2).** *Int Immunol* 1997, **9**:905-911.
- Godkin AJ, Davenport MP, Willis A, Jewell DP, Hill AV: **Use of complete eluted peptide sequence data from HLA-DR and -DQ molecules to predict T cell epitopes, and the influence of the nonbinding terminal regions of ligands in epitope selection.** *J Immunol* 1998, **161**:850-858.
- Harfouch-Hammoud E, Walk T, Otto H, Jung G, Bach JF, van Ender PM, Caillat-Zucman S: **Identification of peptides from autoantigens GAD65 and IA-2 that bind to HLA class II molecules predisposing to or protecting from type 1 diabetes.** *Diabetes* 1999, **48**:1937-1947.
- Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50**:213-219.
- Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B: **Peptide binding predictions for HLA DR, DP and DQ molecules.** *BMC Bioinformatics* 2010, **11**:568.
- Tong JC, Tan TW, Ranganathan S: **Modeling the structure of bound peptide ligands to major histocompatibility complex.** *Protein Sci* 2004, **13**:2523-2532.
- Yewdell JW, Bennink JR: **Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses.** *Annu Rev Immunol* 1999, **17**:51-88.
- Neeno T, Krco CJ, Harders J, Baisch J, Cheng S, David CS: **HLA-DQ8 transgenic mice lacking endogenous class II molecules respond to house dust allergens: identification of antigenic epitopes.** *J Immunol* 1996, **156**:3191-3195.
- Krco CJ, Harders J, Chapoval S, David CS: **Immune response of HLA-DQ transgenic mice to house dust mite allergen p2: identification of HLA-DQ restricted minimal epitopes and critical residues.** *Clin Immunol* 2000, **97**:154-161.
- Klein L, Klugmann M, Nave KA, Tuohy VK, Kyewski B: **Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells.** *Nat Med* 2000, **6**:56-61.
- Erlach HA, Zeidler A, Chang J, Shaw S, Raffel LJ, Klitz W, Beshkov Y, Costin G, Pressman S, Bugawan T, et al.: **HLA class II alleles and susceptibility and resistance to insulin dependent diabetes mellitus in Mexican-American families.** *Nat Genet* 1993, **3**:358-364.
- Nepom GT, Kwok WW: **Molecular basis for HLA-DQ associations with IDDM.** *Diabetes* 1998, **47**:1177-1184.
- Mangalam A, Luckey D, Basal E, Jackson M, Smart M, Rodriguez M, David C: **HLA-DQ8 (DQB1*0302)-restricted Th17 cells exacerbate experimental autoimmune encephalomyelitis in HLA-DR3-transgenic mice.** *J Immunol* 2009, **182**:5131-5139.
- Robles DT, Fain PR, Gottlieb PA, Eisenbarth GS: **The genetics of autoimmune polyendocrine syndrome type II.** *Endocrinol Metab Clin North Am* 2002, **31**:353-368, vi-vii.
- Faustman D, Li XP, Lin HY, Fu YE, Eisenbarth G, Avruch J, Guo J: **Linkage of faulty major histocompatibility complex class I to autoimmune diabetes.** *Science* 1991, **254**:1756-1761.
- Sollid LM, Thorsby E: **HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis.** *Gastroenterology* 1993, **105**:910-922.
- Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR: **Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations.** *Nucleic Acids Res* 2011, **39**:D913-919.
- Graham J, Hagopian WA, Kockum I, Li LS, Sanjeevi CB, Lowe RM, Schaefer JB, Zarghami M, Day HL, Landin-Olsson M, et al: **Genetic effects on age-dependent onset and islet cell auto-antibody markers in type 1 diabetes.** *Diabetes* 2002, **51**:1346-1355.
- Khan JM, Cheruku HR, Tong JC, Ranganathan S: **MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures.** *Bioinformatics* 2011, **27**:1192-1193.
- Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V: **Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins.** *J Med Chem* 1999, **42**:4650-4658.
- Wang R, Lai L, Wang S: **Further development and validation of empirical scoring functions for structure-based binding affinity prediction.** *J Comput Aided Mol Des* 2002, **16**:11-26.
- Bock JR, Gough DA: **A new method to estimate ligand-receptor energetics.** *Mol Cell Proteomics* 2002, **1**:904-910.
- Brusic V, Petrovsky N, Zhang G, Bajic VB: **Prediction of promiscuous peptides that bind HLA class I molecules.** *Immunol Cell Biol* 2002, **80**:280-285.
- Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC: **webPIPSA: a web server for the comparison of protein interaction properties.** *Nucleic Acids Research* 2008, **36**:W276-W280.
- Deng Y, Yewdell JW, Eisenlohr LC, Bennink JR: **MHC affinity, peptide liberation, T cell repertoire, and immunodominance all contribute to the paucity of MHC class I-restricted peptides recognized by antiviral CTL.** *J Immunol* 1997, **158**:1507-1515.
- Keogh E, Fikes J, Southwood S, Celis E, Chesnut R, Sette A: **Identification of new epitopes from four different tumor-associated antigens: recognition of naturally processed epitopes correlates with HLA-A*0201-binding affinity.** *J Immunol* 2001, **167**:787-796.
- Jensen PE: **Recent advances in antigen processing and presentation.** *Nat Immunol* 2007, **8**:1041-1048.
- Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L: **Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network.** *Bioinformatics* 1998, **14**:121-130.
- Mallios RR: **Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 2001, **17**:942-948.
- Doytchinova IA, Flower DR: **Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction.** *Bioinformatics* 2003, **19**:2263-2270.
- Lee KH, Wucherpfennig KW, Wiley DC: **Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes.** *Nat Immunol* 2001, **2**:501-507.
- Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al: **The**



- Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 2002, **58**:899-907.
46. Abagyan RA, Totrov MM, Kuznetsov DA: **ICM – a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation.** *J Comput Chem* 1994, **15**:488-506.
47. Sidney J, del Guercio MF, Southwood S, Sette A: **The HLA molecules DQA1*0501/B1*0201 and DQA1*0301/B1*0302 share an extensive overlap in peptide binding specificity.** *J Immunol* 2002, **169**:5098-5108.
48. Suri A, Walters JJ, Gross ML, Unanue ER: **Natural peptides selected by diabetogenic DQ8 and murine I-A(g7) molecules show common sequence specificity.** *J Clin Invest* 2005, **115**:2268-2276.
49. Chang KY, Unanue ER: **Prediction of HLA-DQ8beta cell peptide using a computational program and its relationship to autoreactive T cells.** *Int Immunol* 2009, **21**:705-713.
50. Paisansinsup T, Deshmukh US, Chowdhary VR, Luthra HS, Fu SM, David CS: **HLA class II influences the immune response and antibody diversification to Ro60/Sjogren's syndrome-A: heightened antibody responses and epitope spreading in mice expressing HLA-DR molecules.** *J Immunol* 2002, **168**:5876-5884.
51. Abagyan RA, Totrov MM: **Ab Initio Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure.** *J Comput Phys* 1999, **151**:402-421.
52. Halgren TA: **Merck Molecular Force Field. I-V.** *J Comp Chem* 1995, **17**:490-641.
53. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA: **Energy parameters in polypeptides. 10: Improved geometric parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides.** *J Phys Chem* 1992, **96**:6472-6484.
54. Schapira M, Totrov M, Abagyan R: **Prediction of the binding energy for small molecules, peptides and proteins.** *J Mol Recognit* 1999, **12**:177-190.
55. Bharadwaj R, Windemuth A, Sridharan S, Honig B, Nicholls A: **The fast multipole boundary element method for molecular electrostatics: An optimal approach for large systems.** *Journal of Computational Chemistry* 1995, **16**:898-913.
56. Janin J: **Protein-protein recognition.** *Prog Biophys Mol Biol* 1995, **64**:145-166.
57. Krystek S, Stouch T, Novotny J: **Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures.** *J Mol Biol* 1993, **234**:661-679.
58. Novotny J, Bruccoleri RE, Davis M, Sharp KA: **Empirical free energy calculations: a blind test and further improvements to the method.** *J Mol Biol* 1997, **268**:401-411.
59. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, McCammon JA: **Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program.** *Computer Physics Communications* 1995, **91**:57-95.
60. Blomberg N, Gabdoulline RR, Nilges M, Wade RC: **Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity.** *Proteins* 1999, **37**:379-387.
61. Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics.** *Journal of Computational and Graphical Statistics* 1996, **5**:299-314.

Additional files

Additional File 1 – Table S1. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used in the training set for this study. (*.pdf)

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#55).

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-82-2-SP.pdf>

Additional File 2 – Table S2. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used as test set 1 for this study. (*.pdf)

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#32).

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-83-2-SP.pdf>

Additional File 3 – Table S3. HLA-DQ8 specific peptides used as test set 2 for this study. (*.pdf)

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#22). Peptides that do not elicit T cell response after binding to the MHC are italicised.

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-84-2-SP.pdf>

Additional File 4 – Figure S1. Heat map showing the clustering output of the 31 pMHC interfaces for the top 31 peptides (obtained upon the application of the binding energy decision threshold of -37.91 kJ/mol, defined for high-affinity binders at the specificity of 95% and sensitivity of 89%) from test set 2 along with all six available human pMHC-II interfaces from TR/pMHC-II crystal structures based on the calculated MSEF values depicted as a colour coded matrix. (*.pdf)

Group A: the six available human pMHC-II interfaces from TR/pMHC-II crystal structures, Group B: non-agonists, Group C: weak agonists (WA), Group D: strong-agonists (SA) and Group E: moderate-agonists (MA).

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-85-2-SP.pdf>

Additional File 5 – Table S4. Top 23 predictions from test set 2 using our prediction model. (*.pdf)

21 out of the 23 predictions are known T cell activators, indicated by the '+' sign. Experimental binding registers are in underlined bold font. IA-2, *Der p* 2, PPI, Ro60, and GAD65 refer to insulinoma-associated (islet) antigen-2, *dermatophagoides ptermyssinus* allergenic peptides, pre-proinsulin, human ribonucleoprotein-60 and human glutamic acid decarboxylase-65, respectively. Similarly, SA, MA and WA denote the predicted strong-agonist, moderate-agonist and weak-agonist, respectively.

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-86-2-SP.pdf>

Additional File 6 – Table S5. Analysis of DQ8 binding motifs for peptides from test set 1. (*.pdf)

The residues concurrent with available binding motifs for DQ8 [15, 17] are underlined in bold font. All peptides have one or more residues that do not conform to binding motifs and 56 peptides have 2 or more such residues. Y and N refer to yes and no, respectively.

<http://immunome-research.net/journal/homeimmunomeojsfiles/journals/1/articles/14/supp/14-87-2-SP.pdf>



Additional File 1

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S1. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used in the training set for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#55).

S.No	Description	Peptide sequence	IC ₅₀ (nM)	Reference
1	Thyroid per 632-645Y	IDV <u>WLGGLAEN</u> FLPY	39.00	[47]
2	Thyroid per 632-645Y analog	IDV <u>WLGGLAENV</u> LPY	22.97	[47]
3	Thyroid per 632-645Y analog	IDV <u>WLGGLAESF</u> LPY	17.94	[47]
4	Thyroid per 632-645Y analog	IDV <u>WLGGLAEDF</u> LPY	33.85	[47]
5	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPD	25.74	[47]
6	Thyroid per 632-645Y analog	IDV <u>LLGGLAENF</u> LPY	119.29	[47]
7	Thyroid per 632-645Y analog	IDV <u>WLGGLAEYF</u> LPY	24.58	[47]
8	Thyroid per 632-645Y analog	IDV <u>WLGGLAENY</u> LPY	30.49	[47]
9	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPL	32.18	[47]
10	Thyroid per 632-645Y analog	IDV <u>WLGGLAEKF</u> LPY	31.47	[47]
11	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> YPY	57.61	[47]
12	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> DPY	25.35	[47]
13	Thyroid per 632-645Y analog	IDV <u>WLGGLAEN</u> DLPY	16.77	[47]
14	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPY	325.00	[47]
15	Thyroid per 632-645Y analog	IDV <u>SLGGLAENF</u> LPY	72.43	[47]
16	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LSY	195.00	[47]
17	Thyroid per 632-645Y analog	IDV <u>WLGGLAEQF</u> LPY	34.50	[47]
18	Thyroid per 632-645Y analog	IDV <u>WLGGVAENF</u> LPY	92.86	[47]
19	Thyroid per 632-645Y analog	IDV <u>WYGGLAENF</u> LPY	62.95	[47]
20	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LLY	139.29	[47]
21	Thyroid per 632-645Y analog	IDV <u>WLLGLAENF</u> LPY	130.00	[47]
22	Thyroid per 632-645Y analog	IDV <u>WSGGLAENF</u> LPY	35.88	[47]
23	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LKY	278.57	[47]
24	Thyroid per 632-645Y analog	IDV <u>WLGGSAENF</u> LPY	216.67	[47]
25	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> VYPY	76.82	[47]
26	Thyroid per 632-645Y analog	IDV <u>WLYGLAENF</u> LPY	125.81	[47]
27	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPK	66.00	[47]
28	Thyroid per 632-645Y analog	IDV <u>WLGGALN</u> FLPY	130.00	[47]
29	Thyroid per 632-645Y analog	IDV <u>WLGGKAENF</u> LPY	177.27	[47]
30	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPF	50.47	[47]
31	Thyroid per 632-645Y analog	IDV <u>WLGGYAENF</u> LPY	139.29	[47]
32	Thyroid per 632-645Y analog	IDV <u>WLKGLAENF</u> LPY	325.00	[47]
33	Thyroid per 632-645Y analog	IDV <u>WLDGLAENF</u> LPY	100.00	[47]

S.No	Description	Peptide	IC ₅₀ (nM)	Reference
34	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LAENFLPS	69.57	[47]
35	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LAENKLPY	1677.00	[47]
36	Thyroid per 632-645Y analog	IDV <u>KLGG</u> LAENFLPY	2028.00	[47]
37	Thyroid per 632-645Y analog	IDV <u>WDGG</u> LAENFLPY	83.22	[47]
38	Thyroid per 632-645Y analog	IDV <u>WKGG</u> LAENFLPY	96.97	[47]
39	Thyroid per 632-645Y analog	IDV <u>WLSG</u> LAENFLPY	105.41	[47]
40	Thyroid per 632-645Y analog	IDV <u>WLGS</u> LAENFLPY	78.00	[47]
41	Thyroid per 632-645Y analog	IDV <u>WLGD</u> LAENFLPY	105.41	[47]
42	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LSENFLPY	390.00	[47]
43	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LDENFLPY	177.27	[47]
44	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LAENFLYY	108.33	[47]
45	Thyroid per 632-645Y analog	IDV <u>WLGG</u> LAENFSPY	53.94	[47]
46	MHC II E8 51-65	FDG <u>DEIF</u> HVDIEKSE	1000.00	[48]
47	MHC II E8 51-65 analog	FDG <u>DEIA</u> HVDIEKSE	3300.00	[48]
48	TRAIL receptor 2 364-380 analog	GRFTYQNA ~~AA~~ QPAATGPG	1000.00	[48]
49	Nicastrin 65-78	ISGDTGVIH <u>VVE</u> KE	1000.00	[48]
50	Nicastrin 65-78 analog	ISGDTGVIH <u>VVA</u> KE	4300.00	[48]
51	E25B protein 112-126	YQTI <u>EENIKIFE</u> EDA	800.00	[48]
52	E25B protein 112-126 analog	YQTI <u>EENIKIFE</u> EKA	1700.00	[48]
53	ZnT8 diabetic autoantigen	LYPDYQIQAGIMIT	700.00	[49]
54	ZnT8 diabetic autoantigen	AVD <u>GVISVHSL</u> HIW	18000.00	[49]
55	ZnT8 diabetic autoantigen	SKRLTFGWYRAEIL	20200.00	[49]
56	B2m 91–104	TPTEKDEYCARVNH	10000.00	[47]
57	Artificial sequence	YARFQSQTTLKQKT	10000.00	[47]

Additional File 2

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S2. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used as test set 1 for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#32).

S.No	Description	Peptide sequence	IC ₅₀ (nM)	Reference
1	Thyroid per 632-645Y analog	IDV WLGGLAEL FLPY	22.71	[47]
2	Thyroid per 632-645Y analog	IDV WLGGLAENS LPY	34.94	[47]
3	Thyroid per 632-645Y analog	IDV DLGGLAENF LPY	20.28	[47]
4	Thyroid per 632-645Y analog	IDV WLGGLYENF LPY	195.00	[47]
5	Thyroid per 632-645Y analog	IDV WLGGLAYNF LPY	100.00	[47]
6	Thyroid per 632-645Y analog	IDV YLGGLAENF LPY	52.00	[47]
7	Thyroid per 632-645Y analog	IDV WVGGLAENF LPY	44.30	[47]
8	Thyroid per 632-645Y analog	IDV WLGGLAENF LDY	41.94	[47]
9	Thyroid per 632-645Y analog	IDV WLGGLASNF LPY	354.55	[47]
10	Thyroid per 632-645Y analog	IDV WLGLLAENF LPY	50.65	[47]
11	Thyroid per 632-645Y analog	IDV WLGGLAENF KPY	74.56	[47]
12	E25B protein 112-126 analog	YQTI EENIKIF KEDA	1000.00	[48]
13	E25B protein 112-126 analog	YQTI EENIKIFE ADA	1800.00	[48]
14	E25B protein 112-126 analog	YQTI EENIKIFE AAA	2500.00	[48]
15	E25B protein 112-126 analog	YQTI EENIKIFE AAA	1700.00	[48]
16	E25B protein 112-126 analog	YQTI KENIKIFE EDA	3800.00	[48]
17	TRAIL receptor 2 364-380 analog	GRFT KQNAAAQPE TGPG	3700.00	[48]
18	TRAIL receptor 2 364-380 analog	GRFT AQNAAAQPA TGPG	3100.00	[48]
19	TRAIL receptor 2 364-380 analog	GRFT AQNAAAQPE TGPG	1700.00	[48]
20	TRAIL receptor 2 364-380 analog	GRFT YQNAAAQPE TGPG	1700.00	[48]
21	Nicestrin 65-78 analog	ISG DTGVIHV AKE	4300.00	[48]
22	Nicestrin 65-78 analog	ISG ATGVIHV VEKE	2300.00	[48]
23	Superoxide dimutase 1 90-103	AGK DGVANVS IEDR	2000.00	[48]
24	Superoxide dimutase 1 90-103 analog	AGK AGVANVS IEDR	1800.00	[48]
25	Superoxide dimutase 1 90-103 analog	AGK DGVANAS IEDR	2800.00	[48]
26	MHC II E8 51-65 analog	FDG KEIFHVD IEKSE	2800.00	[48]
27	MHC II E8 51-65 analog	FDG AEIFHVD IEKSE	2000.00	[48]
28	MHC II E8 51-65 analog	FDG DEIFHAD IEKSE	3100.00	[48]
29	ZnT8 diabetic autoantigen	LTI QIESAAD QDPS	2500.00	[49]
30	ZnT8 diabetic autoantigen	RTG IAQALSS FDLH	2500.00	[49]
31	ZnT8 diabetic autoantigen	ILS VHVATAA SQDS	4900.00	[49]
32	ZnT8 diabetic autoantigen	AIL TDAAHLL IDL	7200.00	[49]
33	A-gliadin 30-44	FPGQQQQFPQPYP	600.00	[15]

S.No	Description	Peptide	IC ₅₀ (nM)	Reference
34	A-gliadin 34-48	QQQFPPQQPYPQPQP	10000.00	[15]
35	A-gliadin 41-55	QPYPQPQPFPSQQPY	1120.00	[15]
36	A-gliadin 49-63	FPSQQPYLQLQFPFQ	20.00	[15]
37	A-gliadin 56-70	LQLQFPFQPQPFPL	20.00	[15]
38	A-gliadin 77-91	SFPPQQPYPQPQPQY	370.00	[15]
39	A-gliadin 196-210	PSSQFQQPLQQYPLG	10000.00	[15]
40	A-gliadin 201-215	QQPLQQYPLGQGSFR	2180.00	[15]
41	A-gliadin 207-221	YPLGQGSFRPSQQNP	100.00	[15]
42	A-gliadin 227-241	VQPQQQLPQFEIRNL	73.00	[15]
43	34P3A	IARAKMFPAVAEK	541.00	[47]
44	Artificial sequence	AAAAVAAEAY	48.00	[47]
45	Artificial sequence	YARFQRQTTLKAAA	10000.00	[47]
46	Artificial sequence (ROIV)	YAHAAHAHAHAHAHAA	2942.00	[47]
47	CD20 249–262 analog	EEDIEIPIQEEY	21.00	[47]
48	CLIP 95-102	KPVSKMRMATPLLMQALP	650.00	[47]
49	CLIP 96-114	KLKPKPKPVSKMRMATPLL	10000.00	[47]
50	DQa1 0501 16-30	YQSYGPSGQYTHEFD	10000.00	[47]
51	FceR 104–122	SQDLELSWNLNGLQADLSS	123.00	[47]
52	FceR 104–122 analog	SQDLELSWNLNGLQAY	118.00	[47]
53	GAD 101–115	CDGERPTLAFLQDVM	69.00	[47]
54	MHC Ia 46-63	EPRAPWIEQEGPEYW	519.00	[47]
55	GAD65 253–265	IARFKMFPEVKEK	3712.00	[47]
56	HA 255–271Y	FESTGNLIAPEYGFKISY	62.00	[47]
57	HSV	DMTPADALDDFDL	173.00	[47]
58	IA-2 499-509	GVAGLLVALAV	95.00	[47]
59	IA-2 499–509	MSSGSFINISV	2470.00	[47]
60	Insulin B 5–15	FVNQHLCGSHLVEAL	10000.00	[47]
61	Lamba repressor 12–24	LEDARRLKAIYEK	717.00	[47]
62	Lol p1 101–120	APYHFDLSGHAFGSMAKKGE	3602.00	[47]
63	MHC Ia 51–63 analog	YPFIEQEGPEFFDQE	1156.00	[47]
64	ML LSR2 5–17	GVTYEIDLTNKN	10000.00	[47]
65	OVA 267-276 Y	LTEWTSSNVMEERY	62.00	[47]
66	p21 51–66; C out	LLDILDTAGLEEYSAMRD	202.00	[47]
67	Pf ABRA 487–506	DSNIMNSINNVMDIDFFEK	171.00	[47]
68	Pf cp 379–396 truncated analog	IEKKIAKMEKASY	10000.00	[47]
69	Pf MSP-1 250-271	FGYRKPLDNIKDNVGKMEDYIKK	10000.00	[47]
70	VP16	PPLYATGRLSQAQLMPSPPM	538.00	[47]

Additional File 3

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S3. HLA-DQ8 specific peptides used as test set 2 for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#22). Peptides that do not elicit T cell response after binding to the MHC are italicised.

S.No	Description	Amino acid residue numbers	Peptide sequence	Reference
1	<i>Dermatophagoides pteromyssinus</i> (Der p 2) allergenic peptide	1-20	DQV <u>DVKDCANHE</u> IKKVLVPG	[21] and [22]
2	Human GAD65	126-140	F <u>DRSTKVIDFH</u> YPNE	[49]
3	Human GAD65	206-220	TY <u>EIAPVFVLL</u> EYVT	[49]
4	Human GAD65	231-250	PGGSGDGIF <u>SPGGAISNM</u> YA	[49]
5	Human GAD65	431-445	KH <u>YDLSYDTGD</u> KALQ	[49]
6	Human GAD65	461-475	AKG <u>TTGFEAHVD</u> KCL	[49]
7	Human GAD65	471-490	VDKCLELAEY <u>LYNIIKN</u> REG	[49]
8	Human GAD65	536-550	RM <u>MEYGT</u> TMVSYQPL	[49]
9	Islet Antigen-2	601-618	R <u>QHARQQDK</u> ERLAALGPE	[49]
10	Islet Antigen-2	616-633	GPE <u>GAHGDTTFE</u> YQDLCR	[49]
11	Islet Antigen-2	646-663	EGPPEP <u>SRVSSVSSQ</u> FSD	[49]
12	Islet Antigen-2	661-678	FSD <u>AAQASPSH</u> SSTPSW	[49]
13	Islet Antigen-2	721-738	AEP <u>NTCATAQGE</u> GNIKKN	[49]
14	Islet Antigen-2	826-843	DE <u>GASLYHVYE</u> VNLVSEH	[49]
15	Islet Antigen-2	931-948	KGVK <u>EIDIAATLE</u> HVRDQ	[49]
16	Islet Antigen-2	961-979	F <u>ALTAVAE</u> EVNAILKALP	[49]
17	Preproinsulin	1-24	MALWMRLLPLL <u>ALLALWGPD</u> PAAA	[49]
18	Preproinsulin	14-33	LALWGPD <u>PAAAFVNQH</u> LCGSHLVE	[49]

S.No	Description	Amino acid residue numbers	Peptide sequence	Reference
19	Preproinsulin	34-53	HLV <u>EA</u> LYLV <u>CG</u> ERGFFYTPKTRRE	[49]
20	Preproinsulin	44-63	GERGFFY <u>TPKTR</u> REAE <u>DL</u> QVGQVE	[49]
21	Preproinsulin	74-93	<u>AGSLQPLA</u> LEGSQKRGIVEQCCT	[49]
22	Preproinsulin	94-110	QCCT <u>SIC</u> SLY <u>Q</u> LENYCN	[49]
23	<i>Der p 2</i> allergenic peptide	41-60	FEAVQNTKTAKIEIKASIDG	[21] and [22]
24	<i>Der p 2</i> allergenic peptide	51-70	KIEIKASIDGLEVDVPGIDP	[21] and [22]
25	<i>Der p 2</i> allergenic peptide	61-80	LEVDPGIDPNACHYMKCPL	[21] and [22]
26	<i>Der p 2</i> allergenic peptide	91-110	TWNVPKIAPKSENVVTVKV	[21] and [22]
27	<i>Der p 2</i> allergenic peptide	101-120	SENVVTVKVMGDDGVLACA	[21] and [22]
28	<i>Der p 2</i> allergenic peptide	111-129	MGDDGVLACAIATHAKIRD	[21] and [22]
29	Human Ro60	401-425	MVVTRTEKDSYVVASFDEMVPCT	[50]
30	Human Ro60	466-485	VFTDNETFAGGVHPAIALRE	[50]
31	Human Ro60	501-525	MTSNGFTIADPDDRGMLDMCGFDTG	[50]
32	<i>Der p 2</i> allergenic peptide	11-30	HEIKKVLVPGCHGSEPCIIN	[21] and [22]
33	<i>Der p 2</i> allergenic peptide	21-40	CHGSEPCIHHRGKPFQLEAV	[21] and [22]
34	<i>Der p 2</i> allergenic peptide	31-50	RGKPFQLEAVFEAVQNTKTA	[21] and [22]
35	<i>Der p 2</i> allergenic peptide	71-90	HACHYMKCPLVKGQYDIDKY	[21] and [22]
36	<i>Der p 2</i> allergenic peptide	81-100	VKGQQYDIKYTWNVPIAPK	[21] and [22]

Additional File 4

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

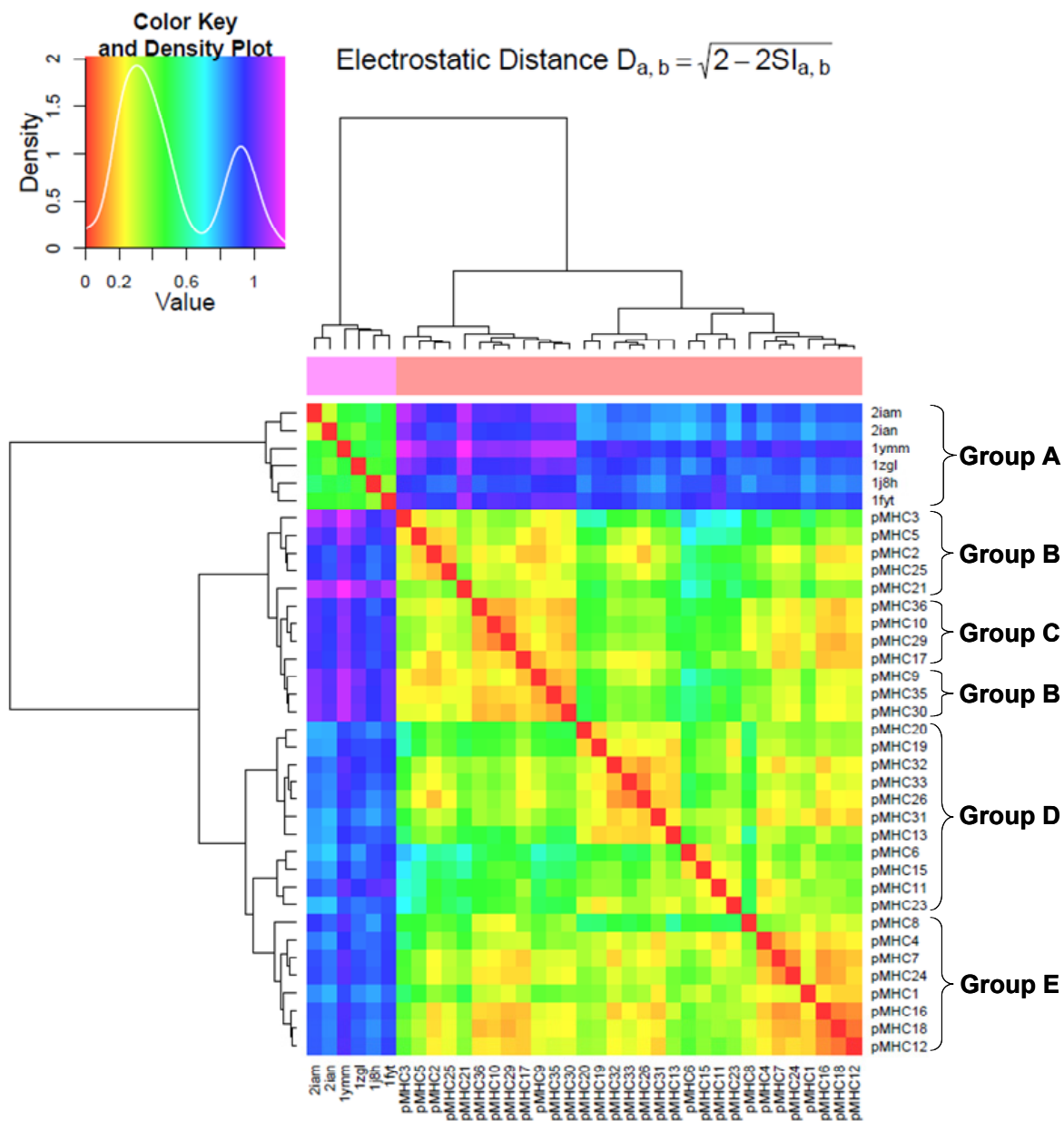


Figure S1. Heat map showing the clustering output of the 31 pMHC interfaces for the top 31 peptides (obtained upon the application of the binding energy decision threshold of -37.91 kJ/mol, defined for high-affinity binders at the specificity of 95% and sensitivity of 89%) from test set 2 along with all six available human pMHC-II interfaces from TR/pMHC-II crystal structures based on the calculated MSEF values depicted as a colour coded matrix. Group A: the six available human pMHC-II interfaces from TR/pMHC-II crystal structures, Group B: non-agonists, Group C: weak-agonists (WA), Group D: strong-agonists (SA) and Group E: moderate-agonists (MA).

Additional File 5

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S4. Top 23 predictions from test set 2 using our prediction model.

21 out of the 23 predictions are known T cell activators, indicated by the '+' sign. Experimental binding registers are in underlined bold font. IA-2, *Der p 2*, PPI, Ro60, and GAD65 refer to insulinoma-associated (islet) antigen-2, *dermatophagoides pternyssinus* allergenic peptides, pre-proinsulin, human ribonucleoprotein-60 and human glutamic acid decarboxylase-65, respectively. Similarly, SA, MA and WA denote the predicted strong-agonist, moderate-agonist and weak-agonist, respectively.

Description	Peptide sequence	Reference	T cell response	Predicted binding energy (kJ/mol)	pMHC clustering result	Rank
<i>Der p 2</i> 101-120	SENVVTVKVMGDDGVLACA	[22]	+	-41.83	SA	1
GAD65 461-475	AKG <u>TTGFEAHVD</u> KCL	[49]	+	-41.59	SA	2
IA-2 931-948	KGVK <u>EIDIAATLE</u> HVRDQ	[49]	+	-41.53	SA	3
PPI 1-24	MALWMRLLPLL <u>ALLALWGPD</u> PAAA	[49]	+	-41.04	SA	4
Ro60 401-425	MVVTRTEKDSYVVASFDEMVPDVT	[50]	+	-40.76	SA	5
PPI 94-110	QCCT <u>SICSLYQLE</u> NYCN	[49]	+	-40.37	SA	6
GAD65 431-445	KH <u>YDLSYDTGD</u> KALQ	[49]	+	-40.01	SA	7
PPI 34-53	HLV <u>EALYLVCGE</u> RGFFYTPKTRRE	[49]	+	-39.82	SA	8
Ro60 501-525	MTSNGFTIADPDDRGMLDMCGFDTG	[50]	+	-38.88	SA	9

Description	Peptide sequence	Reference	T cell response	Predicted binding energy (kJ/mol)	pMHC clustering result	Rank
IA-2 826-843	DE <u>GASLYHVE</u> VNLVSEH	[49]	+	-38.54	SA	10
<i>Der p 2</i> 51-70	KIEIKASIDGLEVDVPGIDP	[22]	+	-38.52	SA	11
GAD65 536-550	RMM <u>EYGTMMVSYQ</u> PL	[49]	+	-43.67	MA	12
<i>Der p 2</i> 61-80	LEVDPGIDPNACHYMKCPL	[21] and [22]	+	-43.14	MA	13
GAD65 206-220	TY <u>EIAPVFVLE</u> YVT	[49]	+	-42.49	MA	14
<i>Der p 2</i> 1-20	DQV <u>DKDCANHE</u> IKKVLVPG	[22]	+	-41.72	MA	15
<i>Der p 2</i> 71-90	HACHYMKCPLVKGQYDIDKY	[21] and [22]	-	-41.63	MA	16
<i>Der p 2</i> 31-50	RGKPFQLEAVFEAVQNTKTA	[21] and [22]	-	-41.28	MA	17
<i>Der p 2</i> 111-129	MGDDGVLACAIATHAKIRD	[21]	+	-40.58	MA	18
IA-2 721-738	AEP <u>NTCATAQGE</u> GNIKKN	[49]	+	-38.13	MA	19
GAD65 126-140	F <u>DRSTKVIDFH</u> YPNE	[49]	+	-40.70	WA	20
<i>Der p 2</i> 91-110	TWNVPKIAPKSENVVTVKV	[21] and [22]	+	-40.01	WA	21
PPI 14-33	LALWGPD <u>PAAAFVNQH</u> LCGSHLVE	[49]	+	-39.59	WA	22
IA-2 661-678	FSD <u>AAQASPSH</u> SSTPSW	[49]	+	-38.05	WA	23

Additional File 6

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S5. Analysis of DQ8 binding motifs for peptides from test set 1.

The residues concurrent with available binding motifs for DQ8 [15, 17] are underlined in bold font. All peptides have one or more residues that do not conform to binding motifs and 56 peptides have 2 or more such residues. Y and N refer to yes and no, respectively.

S.No.	Position				2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1	4	6	7 9				
Binding Motif	T	V	RR	R				
	S	D	VV	E				
	W	M	DD	N				
	K	A	AA	G				
	E	I	II	D				
	D		YY	P				
	Q			Q				
	F							
	L							
	M							
Epitopes	1	FPSQ Q PYLQLQPFPQ			Y	A-gliadin 49-63	-49.00	[15]

S.No.	Position	2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1 4 6 7 9				
2	IDV <u>W</u> LGGL <u>A</u> ENFKPY	Y	Thyroid per 632-645Y analog	-49.12	[47]
3	IDV <u>W</u> LGLL <u>A</u> ENFLPY	Y	Thyroid per 632-645Y analog	-48.74	[47]
4	IDV <u>W</u> LGGL <u>A</u> ENFLDY	Y	Thyroid per 632-645Y analog	-48.03	[47]
5	IDV <u>W</u> VGGL <u>A</u> ENFLPY	Y	Thyroid per 632-645Y analog	-46.79	[47]
6	IDVYLGGGL <u>A</u> ENFLPY	Y	Thyroid per 632-645Y analog	-45.00	[47]
7	LQLQP <u>E</u> PQPQPFP <u>P</u> L	Y	A-gliadin 56-70	-44.55	[15]
8	IDV <u>W</u> LGGL <u>A</u> <u>Y</u> NFLPY	Y	Thyroid per 632-645Y analog	-44.35	[47]
9	LTE <u>W</u> TSSN <u>V</u> ME <u>E</u> RY	Y	OVA 267-276 Y	-43.73	[47]
10	V <u>Q</u> PQQQLPQFEIRNL	Y	A-gliadin 227-241	-43.44	[15]
11	YPL <u>G</u> QGSF <u>R</u> PSQQNP	Y	A-gliadin 207-221	-43.12	[15]
12	EED <u>I</u> EE <u>I</u> IP <u>I</u> QE <u>E</u> EEY	N	CD20 249-262 analog	-42.25	[47]
13	IDV <u>W</u> LGGL <u>A</u> SNFLPY	Y	Thyroid per 632-645Y analog	-42.25	[47]

S.No	Position	2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1 4 6 7 9				
14	IDV <u>W</u> LGGL <u>Y</u> ENFLPY	Y	Thyroid per 632-645Y analog	-41.81	[47]
15	AAAA <u>A</u> V <u>A</u> A EAY	Y	Artificial sequence	-41.75	[47]
16	GVAGLL <u>V</u> ALAV	Y	IA-2 499-509	-41.34	[47]
17	CDGERPTL <u>A</u> FLQDVM	Y	GAD 101-115	-41.25	[47]
18	FEST <u>T</u> GNLI <u>A</u> PEYGFKISY	Y	HA 255-271Y	-41.20	[47]
19	SQDLELS <u>W</u> NLNGLQAD <u>L</u> SS	Y	FceR 104-122	-41.19	[47]
20	DSNI <u>M</u> NS <u>I</u> NN <u>V</u> M <u>D</u> EIDFFEK	N	Pf ABRA 487-506	-40.53	[47]
21	SQDLEL <u>S</u> WNLNGLQAY	Y	FceR 104-122 analog	-40.12	[47]
22	IDV <u>D</u> LGGL <u>A</u> ENFLPY	Y	Thyroid per 632-645Y analog	-40.11	[47]
23	PPL <u>L</u> YATG <u>R</u> LS <u>Q</u> AQLMPSPPM	Y	VP16	-38.87	[47]
24	<u>E</u> PR <u>A</u> PW <u>I</u> E <u>Q</u> EGPEYW	N	MHC Ia 46-63	-38.74	[47]
25	S <u>E</u> PPQQP <u>Y</u> P <u>Q</u> PQPQY	Y	A-gliadin 77-91	-38.26	[15]
26	IDV <u>W</u> LGGL <u>A</u> ELFLPY	Y	Thyroid per 632-645Y analog	-37.93	[47]
27	KPV <u>S</u> KMRM <u>A</u> TPLLMQALP	Y	CLIP 95-102	-37.56	[47]

S.No.	Position	2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1 4 6 7 9				
28	IDV <u>W</u> LGGL <u>A</u> ENSLPY	Y	Thyroid per 632-645Y analog	-36.95	[47]
29	YQTIE <u>E</u> EN <u>I</u> K <u>I</u> FK <u>E</u> DA	N	E25B protein 112-126 analog	-36.76	[48]
30	LL <u>D</u> IL <u>D</u> T <u>A</u> GL <u>E</u> EYSAMRD	N	p21 51–66; C out	-36.50	[47]
31	YPFIE <u>E</u> QEGPEFF <u>D</u> QE	Y	MHC Ia 51–63 analog	-36.41	[47]
32	<u>L</u> ED <u>A</u> R <u>R</u> LKAIYEK	Y	Lamba repressor 12– 24	-36.10	[47]
33	IARAKMFP <u>A</u> <u>V</u> A <u>E</u> K	Y	34P3A	-35.68	[47]
34	D <u>M</u> TP <u>A</u> D <u>A</u> L <u>D</u> D <u>F</u> DL	N	HSV	-35.67	[47]
35	LTIE <u>Q</u> IESA <u>A</u> <u>D</u> <u>Q</u> <u>D</u> PS	N	ZnT8 diabetic autoantigen	-35.46	[49]
36	RTGIAQ <u>A</u> LSSF <u>D</u> LG	Y	ZnT8 diabetic autoantigen	-35.01	[49]
37	FPG <u>Q</u> QQQFPP <u>Q</u> QPY	Y	A-gliadin 30-44	-34.98	[15]
38	YQTIE <u>E</u> EN <u>I</u> K <u>I</u> FAAAA	Y	E25B protein 112-126 analog	-34.92	[48]
39	GRFTAQN <u>A</u> <u>A</u> Q <u>P</u> <u>E</u> TGPG	Y	TRAIL receptor 2 364- 380 analog	-34.87	[48]

S.No.	Position	2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1 4 6 7 9				
40	Q P Y P <u>Q</u> P Q P F P S <u>Q</u> P Y	Y	A-gliadin 41-55	-34.78	[15]
41	I L S V H V <u>A</u> T <u>A</u> A S <u>Q</u> D S	N	ZnT8 diabetic autoantigen	-34.21	[49]
42	M <u>S</u> S G S F <u>I</u> N I S V	Y	IA-2 499–509	-34.09	[47]
43	G R F T A Q N <u>A</u> A <u>A</u> Q P A T G P G	Y	TRAIL receptor 2 364- 380 analog	-33.94	[48]
44	G R F T Y Q N <u>A</u> A <u>A</u> Q P <u>E</u> T G P G	Y	TRAIL receptor 2 364- 380 analog	-33.84	[48]
45	F D G A E I F H <u>V</u> <u>D</u> <u>I</u> <u>E</u> K S E	Y	MHC II E8 51-65 analog	-33.63	[48]
46	A P Y H <u>F</u> D L S G H <u>A</u> <u>F</u> <u>G</u> S M A K K G E	Y	Lol p1 101–120	-33.62	[47]
47	Y Q T I <u>E</u> E N <u>I</u> <u>K</u> <u>I</u> F E A D A	Y	E25B protein 112-126 analog	-33.48	[48]
48	Y Q T I <u>E</u> E N <u>I</u> <u>K</u> <u>I</u> F E A A A	Y	E25B protein 112-126 analog	-33.23	[48]
49	A G K A G V <u>A</u> N <u>V</u> S I <u>E</u> D R	Y	Superoxide dimutase 1 90-103 analog	-33.19	[48]
50	I A R <u>F</u> K M F P E <u>V</u> <u>K</u> <u>E</u> K	Y	GAD65 253–265	-32.91	[47]
51	Y Q T I <u>K</u> E N <u>I</u> <u>K</u> <u>I</u> F E <u>E</u> D A	Y	E25B protein 112-126 analog	-32.81	[48]

S.No.	Position	2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1 4 6 7 9				
52	FDG <u>K</u> E I F H <u>V</u> <u>D</u> I <u>E</u> K S E	N	MHC II E8 51-65 analog	-32.76	[48]
53	I S G A T G <u>V</u> I H <u>V</u> <u>E</u> K E	Y	Nicastrin 65-78 analog	-32.43	[48]
54	FDG <u>D</u> E I F H <u>A</u> <u>D</u> I <u>E</u> K S E	N	MHC II E8 51-65 analog	-31.71	[48]
55	A G K <u>D</u> G V <u>A</u> N <u>A</u> S I <u>E</u> D R	N	Superoxide dimutase 1 90-103 analog	-31.67	[48]
56	A G K <u>D</u> G V <u>A</u> N <u>V</u> S I <u>E</u> D R	N	Superoxide dimutase 1 90-103	-30.94	[48]
57	Y A H A A H A <u>A</u> H <u>A</u> A H A A H A A	Y	Artificial sequence (ROIV)	-30.92	[47]
58	G R F T <u>K</u> Q N <u>A</u> A <u>A</u> Q P <u>E</u> T G P G	N	TRAIL receptor 2 364- 380 analog	-29.86	[48]
59	Y A R F <u>Q</u> R Q T T L K A A A	Y	Artificial sequence	-29.67	[47]
60	Q Q P <u>L</u> Q Q Y P L G <u>Q</u> <u>G</u> S F R	Y	A-gliadin 201-215	-29.57	[15]
61	F G Y R K P <u>L</u> D N <u>I</u> K <u>D</u> N V <u>G</u> K M E D Y I K K	N	Pf MSP-1 250-271	-29.55	[47]
62	F V N Q H <u>L</u> C G S H L <u>V</u> E A L	Y	Insulin B 5-15	-29.26	[47]
63	I S G <u>D</u> T G <u>V</u> I H <u>V</u> V A K E	Y	Nicastrin 65-78 analog	-29.11	[48]
64	P S <u>S</u> Q F Q Q P L Q <u>Q</u> Y P L G	Y	A-gliadin 196-210	-28.82	[15]

S.No.	Position					2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference	
	1	4	6	7	9					
65	G	V	<u>I</u>	Y	<u>I</u>	DLTNKN	Y	ML LSR2 5–17	-28.46	[47]
66	A	I	L	<u>T</u>	D	<u>A</u> HLLI <u>D</u> LT	Y	ZnT8 diabetic autoantigen	-26.68	[49]
67	Q	Q	Q	<u>F</u>	P	PQQP <u>Y</u> P <u>Q</u> PQP	Y	A-gliadin 34-48	-26.49	[15]
68	Y	Q	S	Y	G	P <u>S</u> GQYTHE <u>F</u> <u>D</u>	Y	DQa1 0501 16-30	-24.86	[47]
69	K	L	P	K	P	<u>K</u> PVSKM <u>R</u> MATPLL	Y	CLIP 96-114	-24.64	[47]
70	I	E	<u>K</u>	K	I	<u>A</u> KMEKASY	Y	Pf cp 379–396 truncated analog	-23.89	[47]