Towards guided mutagenesis: Gaussian process regression predicts MHC class II antigen mutant binding

David R. Bell[a*] and Serena H. Chen[b*]

[a] Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21701, USA

[b] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

* To whom correspondence should be addressed. E-mail: david.bell@nih.gov (D.R.B.), chens@ornl.gov (S.H.C.)

Notice: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access- plan).

Abstract

Antigen-specific immunotherapies (ASI) require successful loading and presentation of antigen peptide into the major histocompatibility complex (MHC) binding cleft. One route of ASI design is to mutate native antigens for either stronger or weaker binding interaction to MHC. Exploring all possible mutations is costly both experimentally and computationally. To reduce experimental and computational expense, here we investigate the minimal amount of prior data required to accurately predict the relative binding affinity of point mutations for peptide-MHC class II (pMHCII) binding. Using data from different residue subsets, we interpolate pMHCII mutant binding affinities by Gaussian process (GP) regression of residue volume and hydrophobicity. We apply GP regression to an experimental dataset from the Immune Epitope Database, and theoretical datasets from NetMHCIIpan and Free Energy Perturbation calculations. We find that GP regression can predict binding affinities of 9 neutral residues from a 6-residue subset with an average R^2 coefficient of determination value of 0.62 ± 0.04 (±95% CI), average error of 0.09 ± 0.01 kcal/mol (±95% CI), and with an ROC AUC value of 0.92 for binary classification of enhanced or diminished binding affinity. Similarly, metrics increase to an R^2 value of 0.69 ± 0.04, average error of 0.07 ± 0.01 kcal/mol, and an ROC AUC value of 0.94 for predicting 7 neutral residues from an 8-residue subset. Our work finds that prediction is most accurate for neutral residues at anchor residue sites without register shift. This work holds relevance to predicting pMHCII binding and accelerating ASI design.

Keywords— Antigen-specific immunotherapies, Peptide-MHCII binding, Gaussian process regression, In-silico mutagenesis

Introduction

Major histocompatibility complex class II (MHCII) antigen binding and presentation to the T-cell receptor (TCR) of CD4$^+$ T-cells represents a critical immunological interaction with dysfunction being implicated in autoimmune diseases such as Type-1 Diabetes[1-3], Celiac Disease[4-5], and Multiple Sclerosis[6-7]. Antigen specific immunotherapies (ASI) such as vaccines and HLA blockers work by modulating this MHCII-antigen-TCR interaction[4, 8-9]. Oftentimes, a known antigen is used as a template for ASI design, which is then mutated for enhanced or reduced binding interaction with MHCII and/or TCR[10-11]. Computational techniques, such as sequence-based[12-13], or structure-based[14-18] methods, aim to predict native antigen-MHC binding and mutations to accelerate ASI design and limit costly mutagenesis experiments.

Gaussian Process (GP) regression is a powerful Bayesian supervised machine learning method, which innately provides uncertainty estimates of predictions[19]. GP regression has been used widely in diverse fields such as geostatistics[20] and imaging microscopy[21]. For biological applications, GP regression has been used to predict protein stability[22-23] and turnover[24], protein structure prediction[25] and optimization[26-27], as well as protein-protein binding[22, 28] and quantitative structure activity relationship (QSAR) models for therapeutic design[29-31]. GP regression has also been used to study MHC class I antigen binding from sequence[32-33]. These models, however, have not been extended to the arguably more challenging task of predicting MHCII antigen peptide (pMHCII) binding, with ambiguous registers and variable sized epitopes, nor of predicting the mutational landscape at specific sites. We note that many sequence-based MHC-binding prediction models make use of complex machine learning architectures[13, 34], and that neural networks, in the limit of infinite network width, are equivalent to GP models[35]. Hence, the extension of GP models for pMHCII-binding prediction and characterization of mutational landscapes is a logical outgrowth of current affinity prediction methods.

In this work, we explore the minimal data required to predict relative binding affinities of pMHCII antigen mutants using GP regression across residue volume and hydrophobicity. We study both experimental and theoretical binding affinity datasets from the Immune Epitope Database and Analysis Resource (IEDB), NetMHCIIpan-4.0 server, and Free Energy Perturbation (FEP) calculations. We find that GP regression can accurately predict pMHCII mutant binding affinities for neutral residues at anchor residue sites. More specifically, we find that GP regression across 6- and 8- residue subsets can accurately capture the binding affinities of remaining neutral residues. The corresponding R^2 coefficient of determination values are 0.61 ± 0.04 and 0.69 ± 0.04 (± 95%CI), respectively, with AUC values of 0.92 and 0.94 for binary classification of residues with either enhanced or diminished binding affinity. Finally, we discuss how GP regression can be used to direct mutagenesis experiments and FEP calculations for increased efficiency, which provides opportunities to accelerate MHCII ASI design.

Results

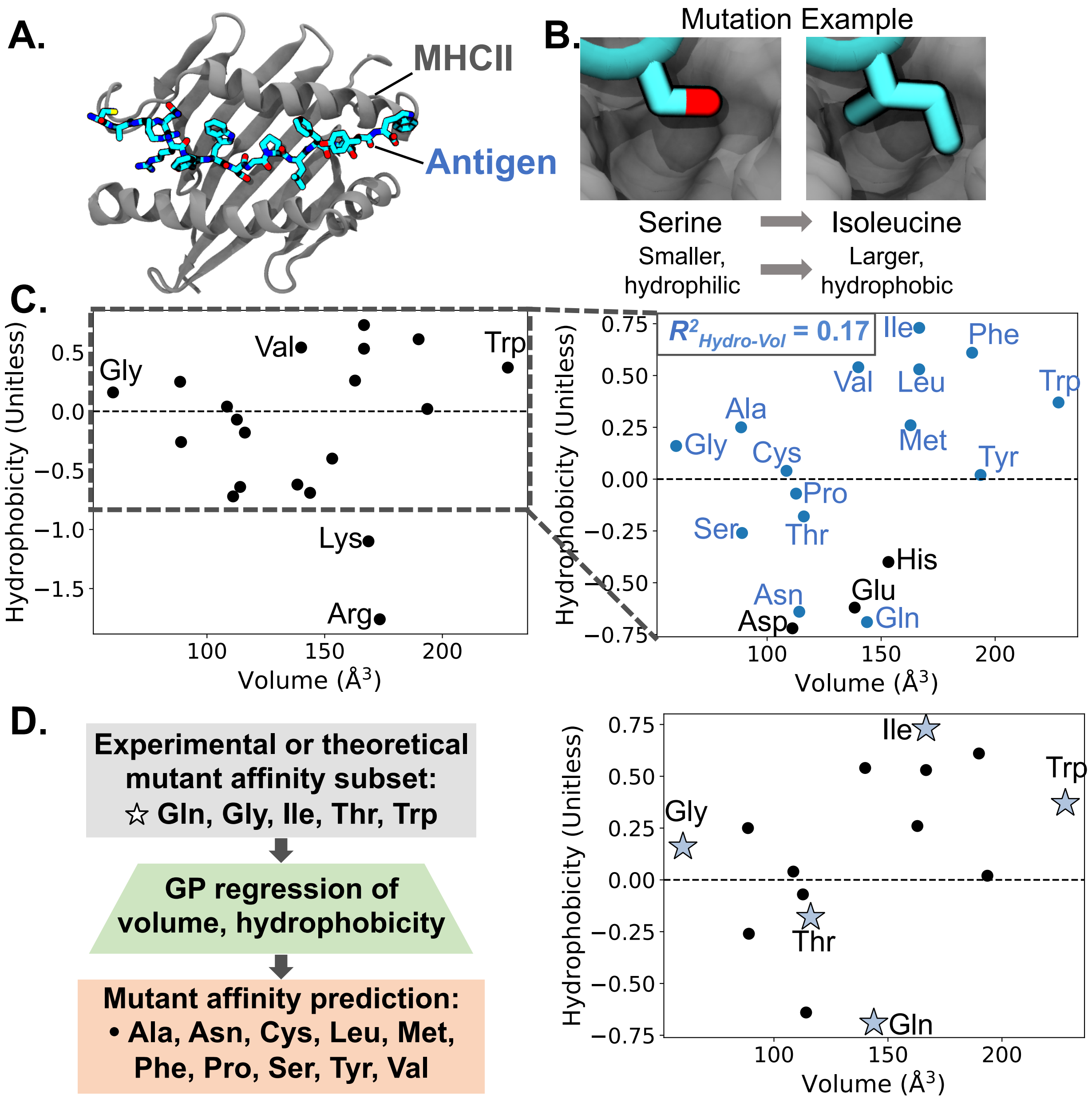

Figure 1. MHCII-antigen structure and Gaussian process (GP) prediction framework. **A.** MHCII in complex with antigen CARQ**RFWSGPLFD**YW from ref[2] with epitope in bold underline. Only the α1 and β1 HLA domains are shown for clarity. **B.** Example antigen mutation of anchor residue Serine to Isoleucine. MHC shown in gray, antigen in cyan. **C.** Twenty standard amino acids presented by Eisenberg hydrophobicity values and residue volume in Å^3. (Inset, right) Close-up view of amino acid plot. Pearson R^2 correlation between hydrophobicity and volume for neutral residues (in blue) is shown in the top left corner. Pearson R^2 correlation for all 20 standard amino acids is 0.01. **D.** GP workflow: use an experimental or theoretical mutant binding affinity residue subset (★,right) to train GP

regression model across residue volume and residue hydrophobicity and then predict mutational binding affinities for the remaining residues (•,right).

Figure 1 presents the MHCII-antigen system together with the Gaussian process (GP) affinity prediction workflow. MHCII is a heterodimer consisting of HLA-α and HLA-β protein chains. These chains construct an antigen binding cleft from two α-helices and one β-sheet, with the antigen binding between the helices. Unlike MHCI, MHCII antigens range in length from 9 to 25 residues, with most around 15 residues[36]. A 9-residue binding core epitope of the antigen binds deeply into the MHCII binding cleft, while the flanking domains contact exterior portions of MHCII. From this 9-residue epitope, 4 anchor residues bury deeply into the MHCII (p1, p4, p6, and p9) while the remaining 5 residues point toward the TCR. For ASI design, antigen mutations, such as the Serine (Ser) to Isoleucine (Ile) mutation of one residue illustrated in Fig. 1B, are tested to identify peptide sequences with desired change in binding affinity.

For this work, we distinguish amino acids by 3 properties: residue volume measured in Å^3 taken from ref[37], residue hydrophobicity according to consensus Eisenberg hydrophobicity values taken from ref[38], and residue charge. The two-dimensional landscape of residue volume and residue hydrophobicity as shown in Fig. 1C is sufficient to distinguish the 20 standard amino acids. In addition, there is low linear correlation between residue volume and residue hydrophobicity. The Pearson's correlation R^2 value between the two features for all 20 residues and 15 neutral residues are 0.01 and 0.17, respectively.

The GP framework we employ for this work is shown in Fig. 1D. We start with a subset of previously determined mutant residue binding affinities, either from experiment or computation, for a residue of a particular MHCII antigen. We then generate a GP model across residue volume and hydrophobicity, fitting to the known affinity values. Lastly, we use the fitted GP model to predict binding affinities of the remaining residues. For example, we start with a 5-residue subset of Gln, Gly, Ile, Thr, and Trp binding affinities for residue site 4 in a particular antigen. We next fit a GP model across residue volume and hydrophobicity to the five known binding affinities. We then use the GP model to predict the binding affinities for the remaining 15 standard residues: Ala, Arg, Asn, Asp, Cys, Glu, His, Leu, Lys, Met, Phe, Pro, Ser, Tyr, and Val. After initial testing of all 20 standard amino acids, we found that excluding charged residues improved predictive results (see Fig. S1). Charged residues are distinguishable on the 2-D plane of residue volume-hydrophobicity, but long-range electrostatic interactions between the antigen and MHCII may affect the binding site conformation and shift registers[10, 39-40], lowering GP prediction accuracy. Hence, we excluded charged residues and only focused on the 15 standard neutral residues for GP prediction. Although our GP models are built only across two dimensions (volume and hydrophobicity), there is an implicit third dimension of charge, which we are accounting for by only targeting neutral residues.

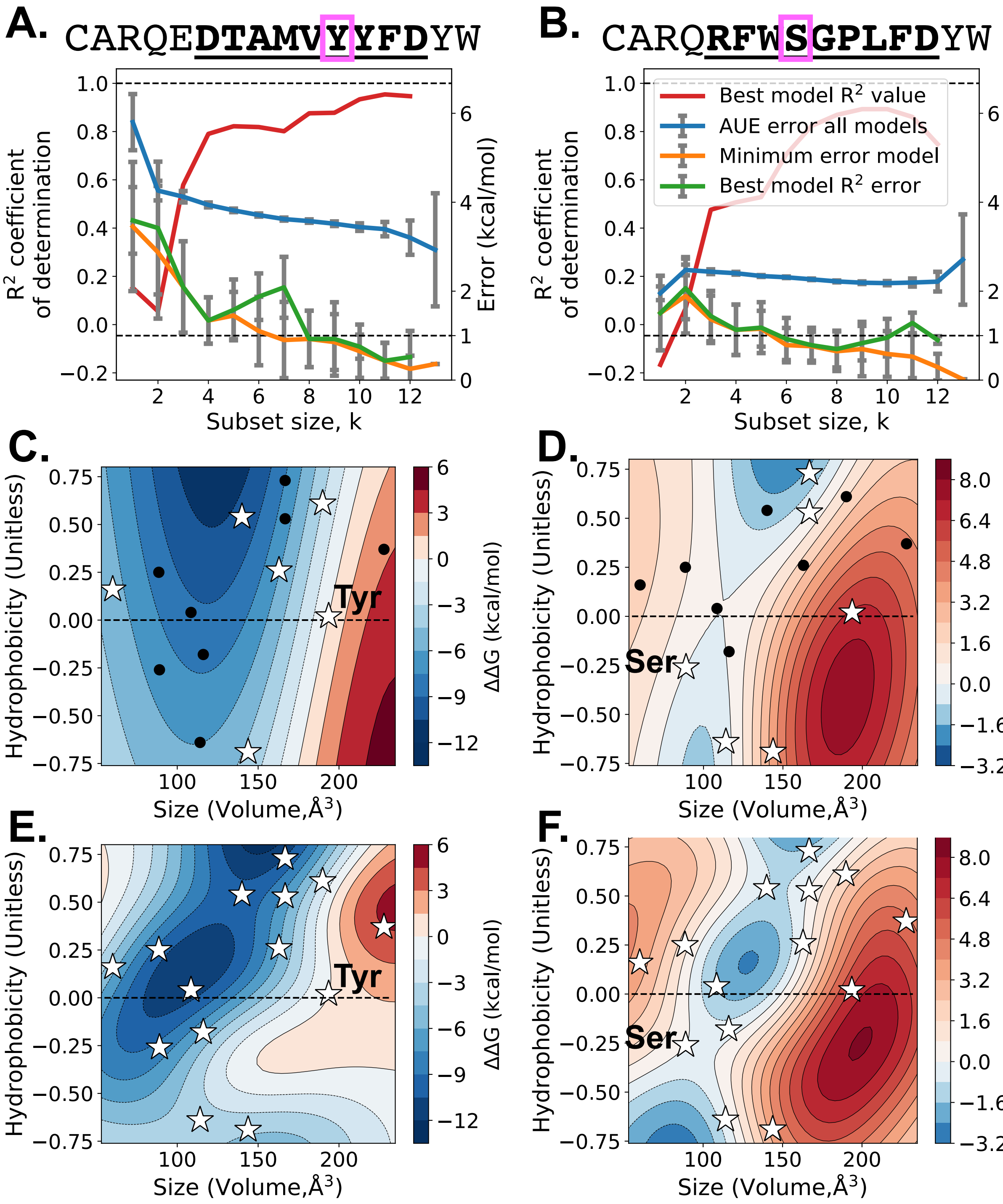

Figure 2. Gaussian process regression for Free Energy Perturbation-calculated binding affinities. **A-B.** Gaussian process regression of mutant binding affinities for residues boxed in magenta. Epitope sequences are in bold underline and are taken from ref[2]. (Plots) Gaussian process regression error and R^2 coefficient of determination for all neutral residues, excluding Proline. For k=13 subset size, only 1 residue ΔG value is predicted, leaving R^2 undefined. Error bars are 95% CI. **C-D.** Interpolated Gaussian process free energy surfaces of neutral residues constructed from a 6-residue subset (shown as white

stars, with native residue labeled). Surfaces represent maximum R^2 models. **E-F.** Gaussian process free energy surfaces of neutral residues constructed from all neutral residues (shown as stars, excluding Pro). **A,C,E.** Tyr6 mutations of CARQE**DTAMVYYFD**YW. **B,D,F.** Ser8 mutations of CARQ**RFWSGPLFD**YW.

Figure 2 presents GP regression results for Free Energy Perturbation (FEP)-predicted binding affinities. The pMHCII systems are the X-idiotype and healthy control antigens taken from Ahmed et al.[2] with sequences shown in Fig. 2A-B. We focused on mutations of the anchor residues Tyr6 of the X-idiotype and Ser4 of the healthy control antigens, as highlighted in magenta boxes in Fig. 2A-B. We first computed all neutral residue relative binding affinities using FEP as described previously[2, 15]. Proline was excluded due to its combined sidechain-backbone structure creating an ambiguity in sidechain mutation for the FEP method. We next studied the accuracy of all possible residue combination subsets and prediction sets for GP regression models. For instance, for a subset size of k=5 residues, we tested a total of *C(n=14,r=5) =* 2002 combinations, including a subset of Ala, Cys, Phe, Gly, Ile residues to predict Leu, Met, Asn, Gln, Ser, Thr, Val, Trp, Tyr residue affinities, a subset of Cys, Phe, Gly, Ile, Leu residues to predict Ala, Met, Asn, Gln, Ser, Thr, Val, Trp, Tyr residues affinities, etc. We then looked at the distribution of GP model accuracy compared to FEP-predicted affinities as measured by R^2 coefficient of determination values, and error, as shown in Fig. 2A-B. We found that R^2 of the highest scoring models reached 0.82 and 0.71 with error below 1 kcal/mol by subset size k=6 residues for both antigen systems. An important caveat of the FEP data is that from previous studies[41], FEP-predicted affinities only agree to within 1 kcal/mol of experimental values, so GP-predicted values to within 1 kcal/mol error agrees to the highest-accuracy of the FEP method. To further illustrate the highest performing GP models, we present the GP-interpolated free energy landscapes from k=6 residues in Fig. 2C-D as well as the GP landscapes using the complete set of 14 neutral residues (absent Proline) in Fig. 2E-F. Although the complete landscapes hold more detail, the GP-interpolated landscapes from residue subset data (Fig. 2C-D) largely capture overall free energy trends.

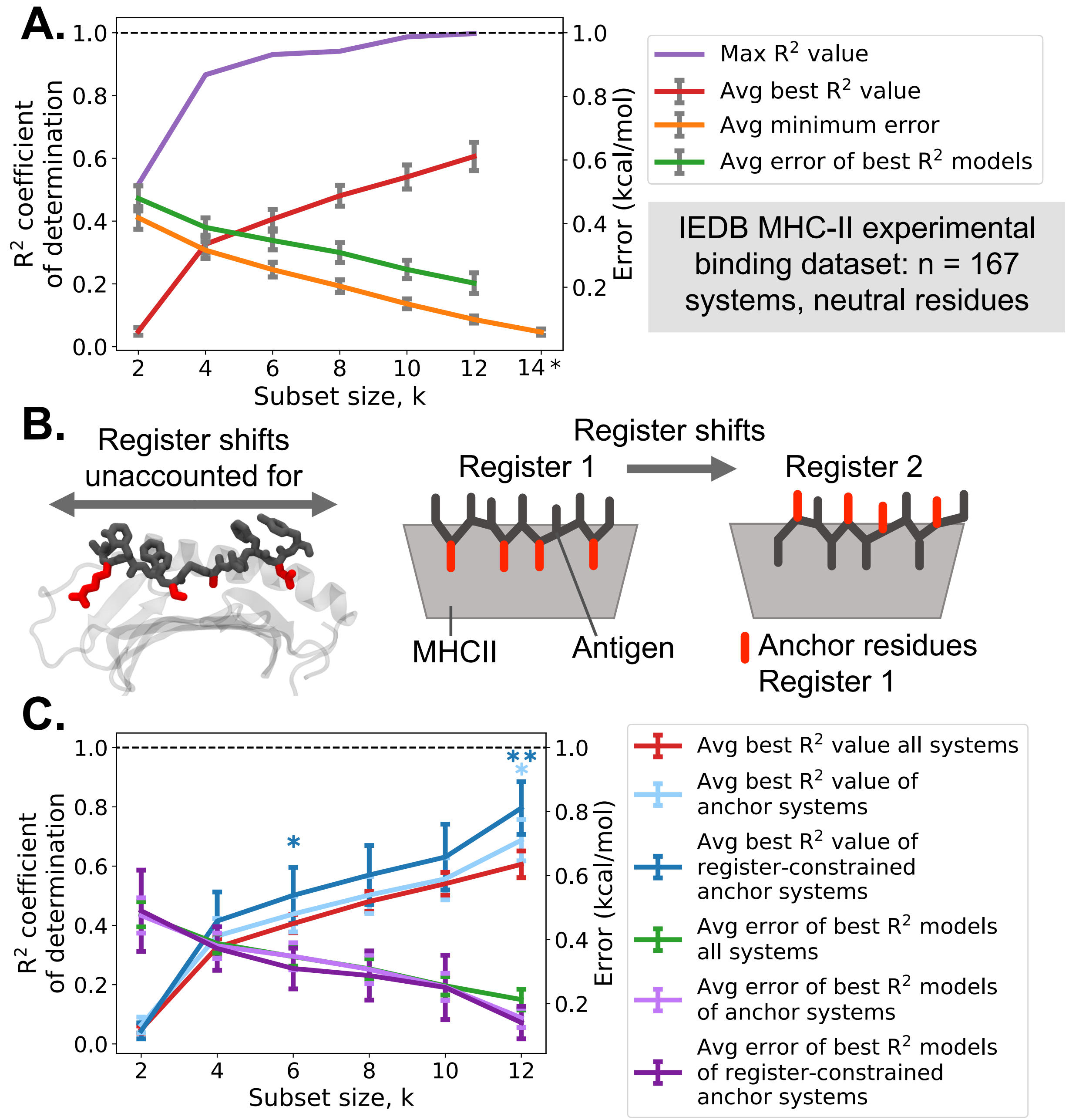

Figure 3. Gaussian process regression for experimentally determined binding affinities from the Immune Epitope Database (IEDB). **A.** Gaussian process regression R^2 coefficient of determination and error for 15 neutral residues. *k* is the subset size used for prediction, e.g., for k=4, 4 residue ΔG values were used to predict the remaining 11 residue ΔG values. Averages are taken across the n=167 systems. *Subset size k=14 is excluded as only 1 residue ΔG value is predicted, leaving R^2 undefined. **B.** IEDB binding affinity values do not account for register shifts, where the antigen moves in the MHCII binding cleft, resulting in a different binding conformation and different anchor residues (red) binding to MHCII. **C.** Gaussian process regression R^2 coefficient of determination and error for all systems and select subsets of anchor residues and register-constrained anchor residues. $*$ and $**$ indicate statistical significance with $p<0.05$ and $p<0.005$ for a one-sided t-test with null

hypothesis: the average best R^2 value of the anchor (or register-constrained anchor) set is not greater than the average best R^2 of all sets. Accounting for register shifts as well as focusing on the anchor residues improved the accuracy of the models. All error bars are 95% CI.

Figure 3 presents GP prediction results using mutant affinity data from the Immune Epitope Database and Analysis Resource (IEDB)[42]. For details on the IEDB dataset, see Methods. From Fig. 3A, the overall R^2 coefficient of determination values are lower than the FEP dataset; however, the prediction error remains low. Also seen in Fig. 3A is the large difference between the maximum R^2 models and the average of the highest scoring models, indicating that high accuracy can be achieved, but is limited over the average. We explored this difference in prediction accuracies by generating scatterplots of the highest scoring models at k=12, as presented in Fig. S2. We found that neither the inner-quartile range of affinity values, ΔG IQR, nor the residue position show a marked trend with high scoring R^2 models.

We suspected that structural properties inherent in the pMHCII system but not accounted for in the IEDB dataset might explain the difference in high- and low-accuracy models. For instance, though MHCII antigens are ~15 residues long, the 9-residue core epitope, termed the antigen register, has the greatest contact with the MHCII core binding cleft. Furthermore, in this 9-residue core epitope, residue positions 1, 4, 6, and 9, termed anchor residues, maintain large interactions with MHCII, because these residues sidechains are buried into deep MHCII binding pockets. Importantly, the IEDB does not provide information on which register the antigen is binding, nor which residues are anchor residues. If the register is shifted by one position, the anchor residues as well as the TCR-facing residues will completely change, as shown in Fig. 3B. For the GP regression we calculate here, we implicitly assume that the binding register will not change upon mutation: mutations capture the same MHCII residue environment, and the remainder of the binding surface remains unchanged. The IEDB data is not register constrained, therefore the mutations we are interpolating may, in fact, be entirely different binding surfaces. To investigate further into this hypothesis, we compared average GP model results for all systems to GP model results from anchor residues and register-constrained anchor residues, as shown in Fig. 3C. We used the NetMHCIIpan 4.0[12, 34] webserver to predict binding registers for all 167 IEDB systems, and then selected wild-type anchor residue systems (n=42) based on residue position and register-constrained anchor residue systems (n=13) based on systems where the binding register did not change for all 15 neutral mutations. We found that GP models of register-constrained anchor residue systems were the most accurate, followed in accuracy by anchor residue systems, and lastly all systems combined. Given the favorable results of building GP models for anchor and register-constrained systems, we next sought to further explore GP model accuracy for these systems using a larger NetMHCIIpan 4.0 dataset.

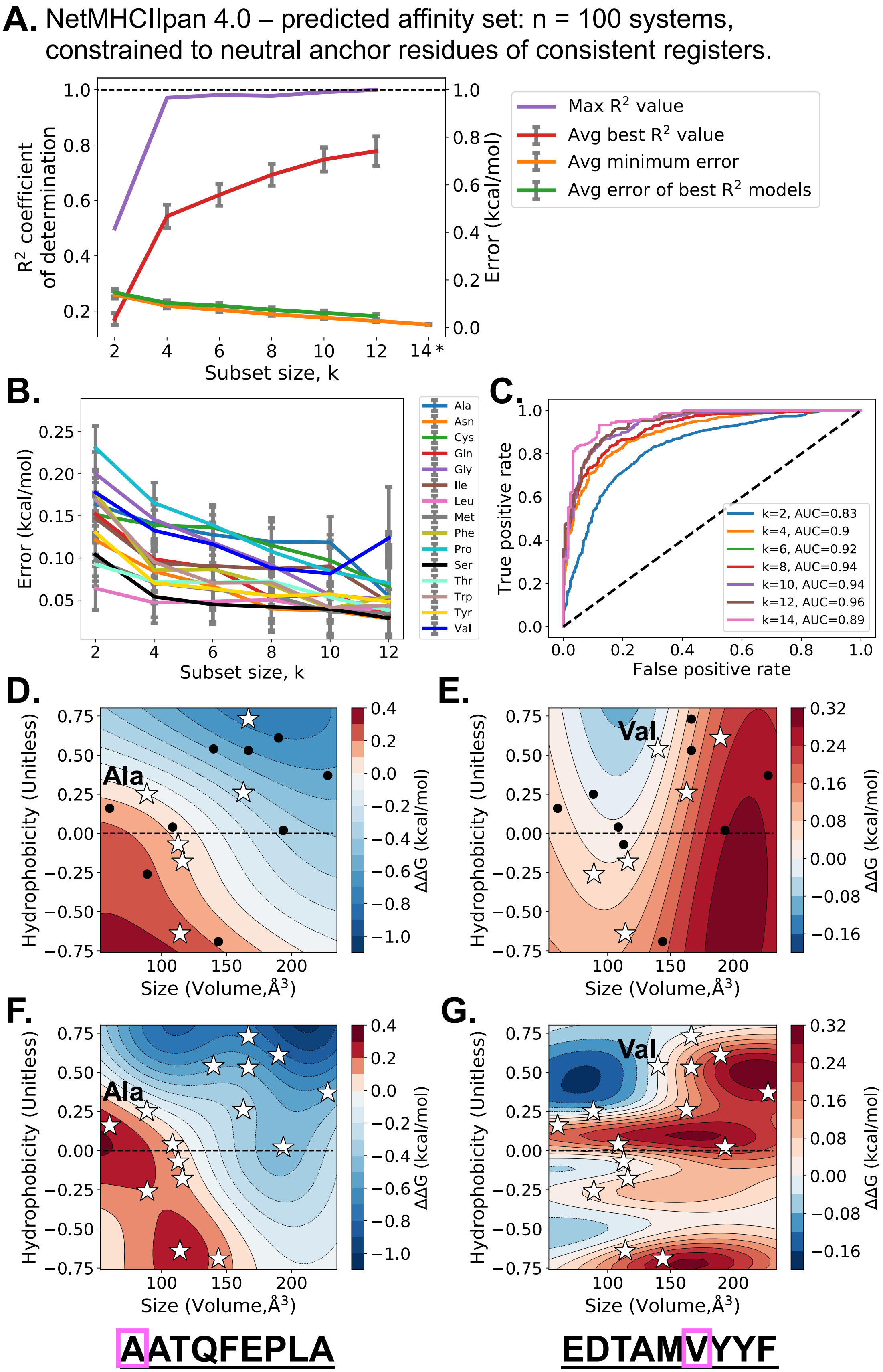
A. NetMHCIIpan 4.0 – predicted affinity set: n = 100 systems, constrained to neutral anchor residues of consistent registers.
R² coefficient of determination
Error (kcal/mol)
Subset size, k
14 *
Max R² value
Avg best R² value
Avg minimum error
Avg error of best R² models
B.
Error (kcal/mol)
Subset size, k
Ala
Asn
Cys
Gln
Gly
Ile
Leu
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
C.
True positive rate
False positive rate
k=2, AUC=0.83
k=4, AUC=0.9
k=6, AUC=0.92
k=8, AUC=0.94
k=10, AUC=0.94
k=12, AUC=0.96
k=14, AUC=0.89
D.
Ala
Hydrophobicity (Unitless)
Size (Volume,Å³)
ΔΔG (kcal/mol)
E.
Val
Hydrophobicity (Unitless)
Size (Volume,Å³)
ΔΔG (kcal/mol)
F.
Ala
Hydrophobicity (Unitless)
Size (Volume,Å³)
ΔΔG (kcal/mol)
G.
Val
Hydrophobicity (Unitless)
Size (Volume,Å³)
ΔΔG (kcal/mol)
AATQFEPLA
EDTAMVYYF

Figure 4. Gaussian process regression for register-constrained binding affinities from the NetMHCIIpan 4.0 server. **A.** Gaussian process regression R^2 coefficient of determination and error for neutral residues at anchor residue positions. k is the subset size used for prediction, e.g., for k=4, 4 residue ΔG values were used to predict the remaining 11 residue ΔG values. Averages are taken across the n=100 systems. * For k=14 subset size, only 1 residue ΔG value is predicted, leaving R^2 undefined. Error bars are 95% CI. **B.** Average error per residue for each k subset size for the top n=100 R^2 models. Error bars are 95% CI. k=14 was excluded from B. as explained in A. **C.** ROC curve for multiple subset sizes, k with AUC values for classifying either enhanced or diminished binding affinity. **D-G.** Gaussian process regression binding affinity predicted surfaces for k=6 residue subsets (**D-E.**) and the full neutral residue sets (k=15, **F-G.**). **D.-E.** Surfaces represent maximum R^2 models for k=6 residue subset shown as white stars with native residue labeled. Two examples are shown, (**D,F.**) predicting A1 mutants of **AATQFEPLA** binding to HLA-DQA10501-DQB10201 favoring larger, hydrophobic residues, and (**E,G.**) predicting V6 mutants of **EDTAMVYYF** binding to HLA-DQA10103-DQB10601 favoring smaller residues. Note that the full antigen sequence in D,F was included in the NETMHCIIpan training set, while the sequence in E,G was not.

In addition to FEP-calculated and experimentally determined binding affinity datasets, we further compared the prediction accuracy of GP regression with sequence-based computing methods. Figure 4 presents results from GP regression prediction of register-constrained binding affinities predicted by the NetMHCIIpan 4.0 webserver. In this analysis, we assume NetMHCIIpan predictions to be 'ground truth' values in order to compare the two methods; however, we note that NetMHCIIpan is itself a predictive method. To exclude the effect of possible register shifts on prediction results, for the NetMHCIIpan dataset, we only predicted mutant binding affinities of the 9-residue core epitope without the flanking domains, thereby implicitly constraining the binding register. Further, to ensure we are capturing the binding interaction between the MHCII and the antigen, we limited our dataset to anchor residue positions. The antigens selected for NetMHC prediction included antigen/MHCII systems from the IEDB dataset as well as systems implicated in Type-1 Diabetes[2, 43]. For more details on the NetMHCIIpan dataset, see Methods. Following our IEDB analysis, we focused on neutral mutations of anchor residues. We found that average best R^2 coefficient of determination values were greater than 0.54 as $k \geq 4$, as shown in Fig. 4A. The minimum error for the top performing R^2 models remains low, around 0.2 kcal/mol, with the average maximum error for the same models around 0.3 kcal/mol with the maximum error value observed for all systems falling below 1.0 kcal/mol at k=4 (Fig. S3). When we decompose the error between GP prediction and NetMHCIIpan prediction in Fig. 4B, certain residues have higher error, particularly residues Ala, Cys, Gly, and Pro. Meanwhile, residues Leu and Ser consistently have lower error and are predicted with higher accuracy than other residues. We further decomposed GP models by residue occurrence, calculating if certain residues occurred in top-scoring GP models more frequently than others, but we did not find large differences (Fig. S4). Similarly, scatterplot charts analogous to Fig. S2 do not reveal marked trends between R^2, ΔG IQR, or residue position (Fig. S5). To evaluate whether GP regression can be generalized to perform binary classification of either enhanced or reduced binding affinity, we computed ROC curves and area under the curve (AUC) values, shown in Fig. 4C. Positive binding affinity values relative to the native antigen are considered reduced binding affinity while negative relative binding affinity values are considered enhanced. As shown, from the ROC curve of the raw data (Fig. 4C), GP regression has high accuracy for predicting the net binding affinity effects of mutations, with AUC values greater than 0.83 for all subsets with the maximum AUC of 0.96 for k=12. Other antigen binding prediction models including NetMHCIIpan 4.0, which use the entire antigen sequence, report AUC values ranging from 0.7 to 0.9 compared to experiment[12, 44-46]. Our method differs from these models by focusing on individual residue sites for guiding in-vivo or in-situ mutagenesis, incorporating predetermined information to increase the prediction accuracy in real time.

Fig. 4D-G shows the GP-interpolated free energy surfaces for 2 systems, with both partial k=6 residue GP predictions (Fig. 4D-E), and complete set surfaces (Fig. 4F-G). Similar to GP-interpolated surfaces of the FEP values, the complete set surfaces preserve more detail than the partial subset surfaces, but the partial subsets are still able to capture broad trends and accurately predict mutant affinity values with 0.05 and 0.06 kcal/mol average error. Importantly, unlike other machine learning techniques which commonly favor large, hydrophobic residues for enhanced binding affinity, the GP prediction technique we employ here does not, as seen by the energy surfaces in Fig. 4E,G. Our GP prediction technique implicitly captures the actual MHCII-antigen binding interaction and then interpolates that interaction to other mutant residues. Hence, hydrophilic residues, and likewise more easily deliverable antigens, may also be favorably predicted.

Discussion and future directions

GP regression can accelerate immunotherapy design by guiding mutagenesis and thereby decreasing the experimental and computational cost of determining MHCII binding affinity. We show that GP regression across a two-dimensional surface of neutral residue volume and hydrophobicity is sufficient to classify enhanced and reduced affinity mutations (AUC~0.9+) and capture affinity trends across the mutational landscape (R^2 coefficient of determination greater than 0.6 and low errors of 0.1-1.0 kcal/mol). This prediction method can be used concurrently with experimental and theoretical investigations to direct mutagenesis with real-time data. From a small residue subset, we can determine which residues are predicted to be favorable or have large uncertainty, and then select those particular residues to investigate next via experiment or computation. Furthermore, our GP prediction method benefits from two aspects. First, GP regression offers an estimate of error, so that we have knowledge of prediction confidence. Likewise, our GP regression method across a two-dimensional surface is intuitive; the impact of residue hydrophobicity and residue volume can be easily grasped rather than black box neural network methods or common QSAR techniques with large feature spaces.

For our GP prediction method, we implicitly assume antigen mutations do not affect binding register. Bound antigens are known to be dynamic[47-49], and localized T-cell populations can recognize different register shifts of the same antigen[50]. However, for many ASI design studies, the target antigen, along with a probable binding register is known beforehand. Another assumption of our technique is that the bound antigens will not change binding conformations upon mutation, and/or that any residue size-dependent conformational changes will be captured by the GP model. For some proteins, single point mutations can drastically change structure[51-52]. But assuming the antigens are initially bound in the MHCII binding cleft, large structural changes are not anticipated, and if they do occur, we anticipate the binding effects will be implicitly captured by the GP-predicted free energy landscape.

We selectively chose to score GP regression models with coefficient of determination R^2 values rather than Pearson product moment correlation coefficient R^2 values. Indeed, as presented in Fig. S6, we tested scoring GP models by Pearson R^2 values for the FEP systems of Fig. 2 and found quite high Pearson R^2 values of 0.7 to 0.9 for subset size k=4. However, Pearson R^2 values only quantify the existence of a linear relationship between the predicted affinities and the actual values. Pearson R^2 values do not account for the variance of the data, so even though a GP model might have high Pearson R^2 correlation, it may also have high error between the predicted affinities and actual values. The R^2 coefficient of determination normalizes the error to the variance of the actual values, ensuring that high R^2 coefficient of determination values also have low error between predicted affinities and actual values. Further, scikit-learn[53], the machine learning library used to generate the GP models, implicitly scores models based on the R^2 coefficient of determination values. We therefore conclude that the R^2 coefficient of determination is a more logical metric of model accuracy and use it here for all investigations.

Our GP prediction method also opens doors for incorporating non-standard amino acids[54] into antigen-specific immunotherapies. If a particular region of the interpolated GP landscape is predicted to be beneficial but not contain a standard amino acid, a non-standard amino acid in that region would be recommended for affinity testing. Extending ASI design to multiple mutation predictions simultaneously is also an area where GP regression-prediction methods might be useful. This would allow GP regression to gather information from alanine scan mutations[55] and other multiple mutants with differential MHCII affinity.

Methods

Datasets. *Free Energy Perturbation, FEP*. FEP data were taken from references[2, 15] based on mutagenesis calculations of antigens from a Type-1 Diabetes-implicated cell line. Briefly, two antigens were studied here, the X-idiotype: CARQE**DTAMVYYFD**YW and the healthy control: CARQ**RFWSGPLFD**YW. These antigens were modeled binding to HLA-DQ8, with the core epitope in bold underline. After 500ns of MD simulation to ensure stable pMHCII binding, FEP was conducted following protocols developed in previous works[41, 56] using a custom short-range potential on NAMD2[57]. FEP was conducted over 34λ windows with at least 200ps/window and 5 replicas/mutation. All error bars are 95% CI unless otherwise noted.

Immune Epitope Database and Analysis Resource, IEDB. The IEDB webserver[42] was accessed on Nov. 15, 2020 and the MHC-II binding dataset was downloaded from: http://tools.iedb.org/mhcii/download/. The IEDB dataset contains antigen binding data across multiple HLA types. An internal program was written to collect and collate antigens and their mutants, as well as convert binding affinity values to kcal/mol. Only antigens with binding affinity data for all 20 standard amino acids were selected for analysis. In total, 167 systems across 6 antigens and 11 HLA types met these criteria and were used for GP regression. The systems are presented in Table S1.

NetMHCIIpan 4.0 webserver. The NETMHCIIpan 4.0 webserver[12, 34, 58] was accessed at http://www.cbs.dtu.dk/services/NetMHCIIpan/ on Feb. 1, 2021. Twenty antigens were selected for prediction based on their occurrence in the IEDB as well as implication in Type-1 Diabetes. The systems are presented in Table S2. The epitopes were register-constrained during the NetMHCIIpan prediction by truncating the flanking domains, leaving only the predicted 9-residue epitope core. Only mutants for anchor residues 1, 4, 6 and 9 were computed, leading to a total of 20 antigens x 4 anchors = 100 systems. The binding affinity rather than the eluted ligand metric was used for affinity values.

Gaussian Process (GP) Regression. GP models were generated using scikit-learn's machine learning python library[53]. A constant kernel combined with an RBF kernel was used for model generation. Median kernel parameters are presented in Fig. S7. For each system, all possible residue subset combinations were tested for GP generation, only saving the top 100-best scoring and minimum error models.

Acknowledgements

The authors would like to thank Leili Zhang, Guojing Cong, Giacomo Domeniconi, Chih-Chieh Yang, Ruhong Zhou, Jeffrey K Weber, and Sangyun Lee for insightful discussions. SHC would like to acknowledge Program Development funding from Oak Ridge National Laboratory which helped spur this collaboration.

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Author Contributions

DRB and SHC designed and conceived the study as well as ran experiments. DRB performed data analysis and generated the figures. DRB and SHC wrote the manuscript.

References

1. Unanue, E. R.; Turk, V.; Neefjes, J., Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annual review of immunology* **2016,** *34* (1), 265-297.

2. Ahmed, R.; Omidian, Z.; Giwa, A.; Cornwell, B.; Majety, N.; Bell, D. R.; Lee, S.; Zhang, H.; Michels, A.; Desiderio, S.; Sadegh-Nasseri, S.; Rabb, H.; Gritsch, S.; Suva, M. L.; Cahan, P.; Zhou, R.; Jie, C.; Donner, T.; Hamad, A. R. A., A Public BCR Present in a Unique Dual-Receptor-Expressing Lymphocyte from Type 1 Diabetes Patients Encodes a Potent T Cell Autoantigen. *Cell* **2019,** *177* (6), 1583-1599.e16.

3. Jones, E. Y.; Fugger, L.; Strominger, J. L.; Siebold, C., MHC class II proteins and disease: a structural perspective. *Nature Reviews Immunology* **2006,** *6* (4), 271-282.

4. Alhassan, E.; Yadav, A.; Kelly, C. P.; Mukherjee, R., Novel Nondietary Therapies for Celiac Disease. *Cellular and Molecular Gastroenterology and Hepatology* **2019,** *8* (3), 335-345.

5. Erlichster, M.; Bedo, J.; Skafidas, E.; Kwan, P.; Kowalczyk, A.; Goudey, B., Improved HLA-based prediction of coeliac disease identifies two novel genetic interactions. *European Journal of Human Genetics* **2020,** *28* (12), 1743-1752.

6. Dendrou, C. A.; Fugger, L.; Friese, M. A., Immunopathology of multiple sclerosis. *Nature Reviews Immunology* **2015,** *15* (9), 545-558.

7. Jelcic, I.; Al Nimer, F.; Wang, J.; Lentsch, V.; Planas, R.; Jelcic, I.; Madjovski, A.; Ruhrmann, S.; Faigle, W.; Frauenknecht, K.; Pinilla, C.; Santos, R.; Hammer, C.; Ortiz, Y.; Opitz, L.; Grönlund, H.; Rogler, G.; Boyman, O.; Reynolds, R.; Lutterotti, A.; Khademi, M.; Olsson, T.; Piehl, F.; Sospedra, M.; Martin, R., Memory B Cells Activate Brain-Homing, Autoreactive CD4(+) T Cells in Multiple Sclerosis. *Cell* **2018,** *175* (1), 85-100.e23.

8. Pozsgay, J.; Szekanecz, Z.; Sármay, G., Antigen-specific immunotherapies in rheumatic diseases. *Nature Reviews Rheumatology* **2017,** *13* (9), 525-537.

9. Nicholas, D.; Odumosu, O.; Langridge, W. H., Autoantigen based vaccines for type 1 diabetes. *Discovery medicine* **2011,** *11* (59), 293-301.

10. Wang, Y.; Sosinowski, T.; Novikov, A.; Crawford, F.; White, J.; Jin, N.; Liu, Z.; Zou, J.; Neau, D.; Davidson, H. W.; Nakayama, M.; Kwok, W. W.; Gapin, L.; Marrack, P.; Kappler, J. W.; Dai, S., How C-terminal additions to insulin B-chain fragments create superagonists for T cells in mouse and human type 1 diabetes. *Science Immunology* **2019,** *4* (34), eaav7517.

11. Xia, J.; Siegel, M.; Bergseng, E.; Sollid, L. M.; Khosla, C., Inhibition of HLA-DQ2-mediated antigen presentation by analogues of a high affinity 33-residue peptide from alpha2-gliadin. *J. Am. Chem. Soc.* **2006,** *128* (6), 1859-1867.

12. Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M., NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **2020,** *48* (W1), W449-w454.

13. O'Donnell, T. J.; Rubinsteyn, A.; Bonsack, M.; Riemer, A. B.; Laserson, U.; Hammerbacher, J., MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems* **2018,** *7* (1), 129-132.e4.

14. Antunes, D. A.; Abella, J. R.; Devaurs, D.; Rigo, M. M.; Kavraki, L. E., Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes. *Curr Top Med Chem* **2018,** *18* (26), 2239-2255.

15. Bell, D.; Domeniconi, G.; Yang, C.-C.; Zhang, L.; Cong, G., Dynamics-based peptide-MHC binding optimization by a convolutional variational autoencoder: a use-case model for CASTELO. *arXiv preprint arXiv:2012.00672* **2020**.
16. Riley, T. P.; Keller, G. L. J.; Smith, A. R.; Davancaze, L. M.; Arbuiso, A. G.; Devlin, J. R.; Baker, B. M., Structure Based Prediction of Neoantigen Immunogenicity. *Frontiers in Immunology* **2019,** *10* (2047).
17. Abella, J. R.; Antunes, D. A.; Clementi, C.; Kavraki, L. E., Large-Scale Structure-Based Prediction of Stable Peptide Binding to Class I HLAs Using Random Forests. *Frontiers in Immunology* **2020,** *11* (1583).
18. Weber, J.; Chowell, D.; Krishna, C.; Chan, T.; Zhou, R., Predicting HLA-I peptide immunogenicity with deep learning and molecular dynamics. Research Square: 2020.
19. Rasmussen, C. E. W., Christopher K I, *Gaussian Processes for Machine Learning*. MIT Press: 2006.
20. Murakami, D.; Yamagata, Y.; Hirano, T., Chapter Four - Geostatistics and Gaussian process models. In *Spatial Analysis Using Big Data*, Yamagata, Y.; Seya, H., Eds. Academic Press: 2020; pp 57-112.
21. Ziatdinov, M.; Kim, D.; Neumayer, S.; Vasudevan, R. K.; Collins, L.; Jesse, S.; Ahmadi, M.; Kalinin, S. V., Imaging mechanism for hyperspectral scanning probe microscopy via Gaussian process modelling. *npj Computational Materials* **2020,** *6* (1), 21.
22. Pires, D. E. V.; Ascher, D. B.; Blundell, T. L., mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014,** *30* (3), 335-342.
23. Jokinen, E.; Heinonen, M.; Lähdesmäki, H., mGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion. *Bioinformatics* **2018,** *34* (13), i274-i283.
24. Rahman, M.; Previs, S. F.; Kasumov, T.; Sadygov, R. G., Gaussian Process Modeling of Protein Turnover. *Journal of proteome research* **2016,** *15* (7), 2115-2122.
25. Huang, Y.-F.; Golding, G. B., Phylogenetic Gaussian Process Model for the Inference of Functionally Important Regions in Protein Tertiary Structures. *PLOS Computational Biology* **2014,** *10* (1), e1003429.
26. Bedbrook, C. N.; Yang, K. K.; Rice, A. J.; Gradinaru, V.; Arnold, F. H., Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Computational Biology* **2017,** *13* (10), e1005786.
27. Romero, P. A.; Krause, A.; Arnold, F. H., Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences* **2013,** *110* (3), E193-E201.
28. Zhou, P.; Tian, F.; Chen, X.; Shang, Z., Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands using genetic algorithm-Gaussian processes. *Peptide Science* **2008,** *90* (6), 792-802.
29. Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K. R., Predicting lipophilicity of drug-discovery molecules using Gaussian process models. *ChemMedChem* **2007,** *2* (9), 1265-7.
30. Tiño, P.; Nabney, I. T.; Williams, B. S.; Lösel, J.; Sun, Y., Nonlinear prediction of quantitative structure-activity relationships. *Journal of chemical information and computer sciences* **2004,** *44* (5), 1647-53.
31. Burden, F. R., Quantitative structure-activity relationship studies using Gaussian processes. *Journal of chemical information and computer sciences* **2001,** *41* (3), 830-5.

32. You, L.; Brusic, V.; Gallagher, M.; Bodén, M., Using Gaussian process with test rejection to detect T-cell epitopes in pathogen genomes. *IEEE/ACM transactions on computational biology and bioinformatics* **2010,** *7* (4), 741-51.
33. Ren, Y.; Chen, X.; Feng, M.; Wang, Q.; Zhou, P., Gaussian process: a promising approach for the modeling and prediction of Peptide binding affinity to MHC proteins. *Protein and peptide letters* **2011,** *18* (7), 670-8.
34. Reynisson, B.; Barra, C.; Kaabinejadian, S.; Hildebrand, W. H.; Peters, B.; Nielsen, M., Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *Journal of proteome research* **2020,** *19* (6), 2304-2315.
35. Lee, J.; Bahri, Y.; Novak, R.; Schoenholz, S. S.; Pennington, J.; Sohl-Dickstein, J., Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165* **2017**.
36. Wieczorek, M.; Abualrous, E. T.; Sticht, J.; Álvaro-Benito, M.; Stolzenberg, S.; Noé, F.; Freund, C., Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in Immunology* **2017,** *8* (292).
37. Zamyatnin, A. A., Protein volume in solution. *Progress in Biophysics and Molecular Biology* **1972,** *24*, 107-123.
38. Zhou, R.; Silverman, B. D.; Royyuru, A. K.; Athma, P., Spatial profiling of protein hydrophobicity: native vs. decoy structures. *Proteins* **2003,** *52* (4), 561-72.
39. Wucherpfennig, K. W.; Yu, B.; Bhol, K.; Monos, D. S.; Argyris, E.; Karr, R. W.; Ahmed, A. R.; Strominger, J. L., Structural basis for major histocompatibility complex (MHC)-linked susceptibility to autoimmunity: charged residues of a single MHC binding pocket confer selective presentation of self-peptides in pemphigus vulgaris. *Proceedings of the National Academy of Sciences of the United States of America* **1995,** *92* (25), 11935-11939.
40. Wang, Y.; Sosinowski, T.; Novikov, A.; Crawford, F.; Neau, D. B.; Yang, J.; Kwok, W. W.; Marrack, P.; Kappler, J. W.; Dai, S., C-terminal modification of the insulin B:11–23 peptide creates superagonists in mouse and human type 1 diabetes. *Proceedings of the National Academy of Sciences* **2018,** *115* (1), 162-167.
41. Xia, Z.; Chen, H.; Kang, S.-g.; Huynh, T.; Fang, J. W.; Lamothe, P. A.; Walker, B. D.; Zhou, R., The complex and specific pMHC interactions with diverse HIV-1 TCR clonotypes reveal a structural basis for alterations in CTL function. *Sci Rep* **2014,** *4*, 4087.
42. Vita, R.; Mahajan, S.; Overton, J. A.; Dhanda, S. K.; Martini, S.; Cantrell, J. R.; Wheeler, D. K.; Sette, A.; Peters, B., The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **2019,** *47* (D1), D339-d343.
43. Patel, S. D.; Cope, A. P.; Congia, M.; Chen, T. T.; Kim, E.; Fugger, L.; Wherrett, D.; Sonderstrup-McDevitt, G., Identification of immunodominant T cell epitopes of human glutamic acid decarboxylase 65 by using HLA-DR(α1*0101,β1*0401) transgenic mice. *Proceedings of the National Academy of Sciences* **1997,** *94* (15), 8082-8087.
44. Zhang, L.; Chen, Y.; Wong, H.-S.; Zhou, S.; Mamitsuka, H.; Zhu, S., TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. *PLoS One* **2012,** *7* (2), e30483.
45. Laimer, J.; Lackner, P., MHCII3D-Robust Structure Based Prediction of MHC II Binding Peptides. *International journal of molecular sciences* **2020,** *22* (1), 12.
46. Bordner, A. J.; Mittelmann, H. D., MultiRTA: a simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes. *BMC bioinformatics* **2010,** *11*, 482-482.

47. Abella, J. R.; Antunes, D. A.; Clementi, C.; Kavraki, L. E., APE-Gen: A Fast Method for Generating Ensembles of Bound Peptide-MHC Conformations. *Molecules (Basel, Switzerland)* **2019,** *24* (5), 881.

48. Knapp, B.; Demharter, S.; Esmaielbeiki, R.; Deane, C. M., Current status and future challenges in T-cell receptor/peptide/MHC molecular dynamics simulations. *Briefings in Bioinformatics* **2015,** *16* (6), 1035-1044.

49. Petrone, P. M.; Garcia, A. E., MHC-peptide binding is assisted by bound water molecules. *J Mol Biol* **2004,** *338* (2), 419-35.

50. Mohan, J. F.; Petzold, S. J.; Unanue, E. R., Register shifting of an insulin peptide–MHC complex allows diabetogenic T cells to escape thymic deletion. *Journal of Experimental Medicine* **2011,** *208* (12), 2375-2383.

51. Chen, S.-H.; Elber, R., The energy landscape of a protein switch. *Physical Chemistry Chemical Physics* **2014,** *16* (14), 6407-6421.

52. Chen, S.-H.; Meller, J.; Elber, R., Comprehensive analysis of sequences of a protein switch. *Protein Science* **2016,** *25* (1), 135-146.

53. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011,** *12*, 2825-2830.

54. Jin, X.; Park, O.-J.; Hong, S. H., Incorporation of non-standard amino acids into proteins: challenges, recent achievements, and emerging applications. *Applied Microbiology and Biotechnology* **2019,** *103* (7), 2947-2958.

55. Chen, S. H.; Perez-Aguilar, J. M.; Zhou, R., Graphene-extracted membrane lipids facilitate the activation of integrin αvβ8. *Nanoscale* **2020,** *12* (14), 7939-7949.

56. Chen, S. H.; Kang, S.-g.; Luo, J.; Zhou, R., Charging nanoparticles: increased binding of Gd@C82(OH)22 derivatives to human MMP-9. *Nanoscale* **2018,** *10* (12), 5667-5677.

57. Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E., Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **2020,** *153* (4), 044130.

58. Jurtz, V.; Paul, S.; Andreatta, M.; Marcatili, P.; Peters, B.; Nielsen, M., NetMHC pan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *bioRxiv* **2017**, 149518.

Supplementary Information

Towards guided mutagenesis: Gaussian process regression predicts MHC class II antigen mutant binding

David R. Bell[a*] and Serena H. Chen[b*]

[a] Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21701, USA

[b] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

* To whom correspondence should be addressed. E-mail: david.bell@nih.gov (D.R.B.), chens@ornl.gov (S.H.C.)

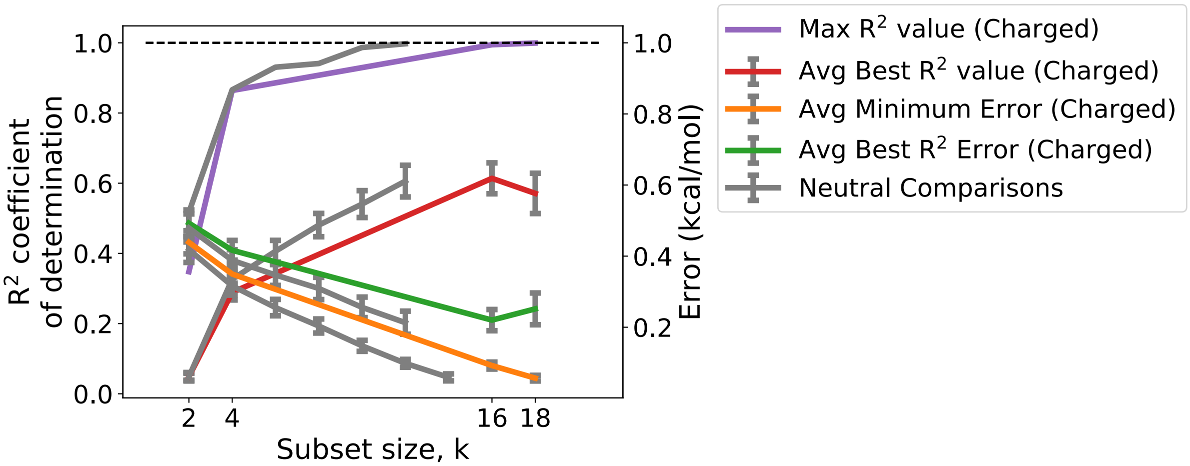

Figure S1. Gaussian process regression results using all 20 standard amino acids in color, compared to using 15 neutral amino acids in gray. Data is experimentally determined binding affinities from the Immune Epitope Database (IEDB). R^2 is the coefficient of determination. For the 20 standard amino acids, GP regression was evaluated for k=2,4,16, and 18 subset sizes, while neutral residues were evaluated at k=2,4,6,8,10,12, and 14. Averages are taken across the n=167 systems. The neutral comparison plots in gray are the same as presented in Figure 3 in main text. Error bars are 95% confidence intervals.

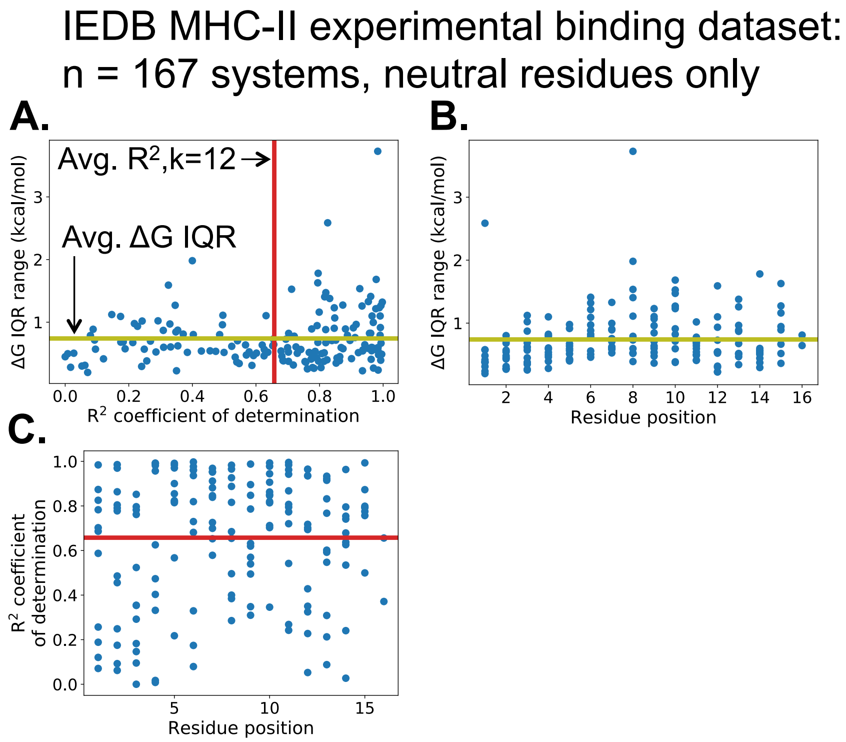

Figure S2. ΔG interquartile range (IQR) and Gaussian process regression performance for the n=167 experimentally determined systems from the Immune Epitope Database (IEDB). **A.** Binding affinity ΔG IQR for n=167 systems as a function of Gaussian process regression R^2 coefficient of determination values for the k=12 subset size. **B.** Binding affinity ΔG IQR for n=167 systems as a function of residue position in the MHCII antigen. **C.** Gaussian process regression R^2 values for the k=12 subset size across residue position in the MHCII antigen. **A-C.** Avg ΔG IQR from expt. and Avg R^2 values for k=12 subset size is shown in olive and red lines respectively.

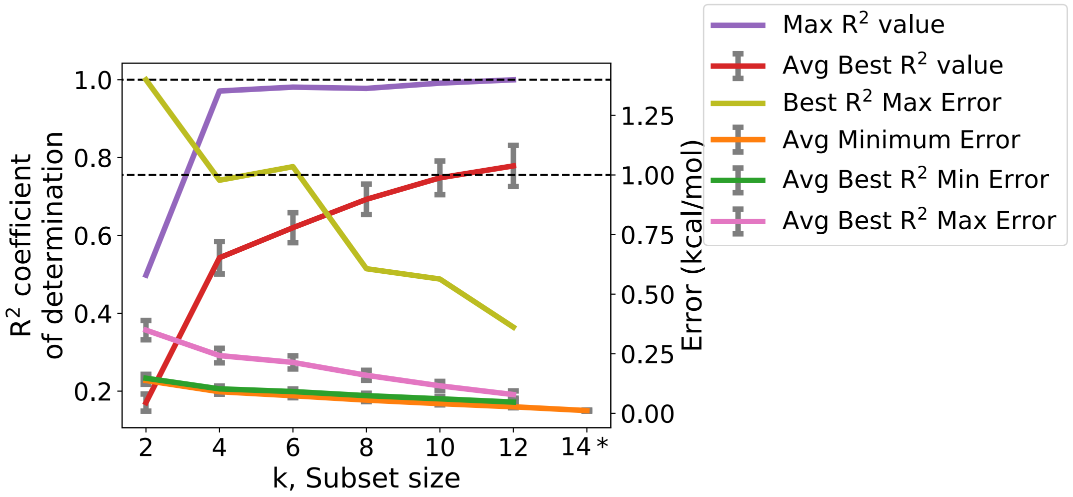

Figure S3. Gaussian process regression R^2 coefficient of determination and error for only neutral residues (15 total neutral residues) using the NETMHCIIpan dataset. k is the subset size used for prediction, e.g. for k=4, 4 residue ΔG values were used to predict the remaining 11 residue ΔG values. Averages are taken across the n=100 register-constrained systems predicted from NetMHCIIpan 4.0. Error bars are 95% CI. "Best R^2 Max Error" means the highest residue error of the top scoring models for the n=100 systems at k subset size.

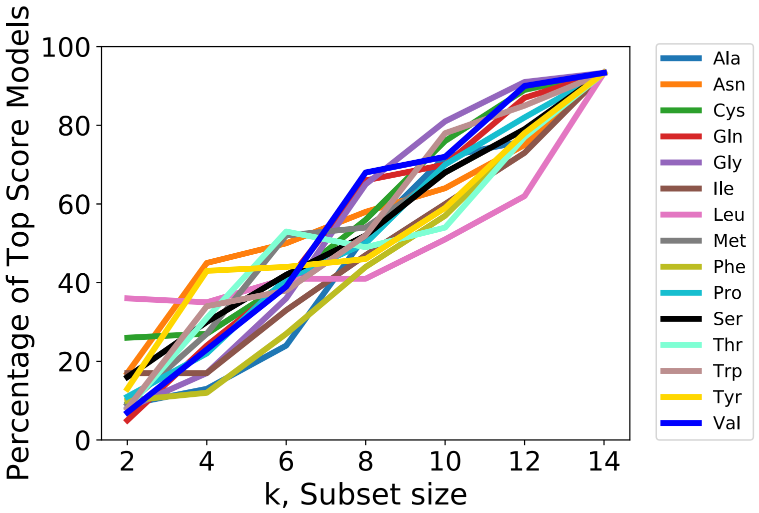

Figure S4. Residue occurrence in Gaussian process regression models across subset size, k for the NetMHCIIpan dataset. Note: only neutral residues are shown, charged residues were excluded from our analysis.

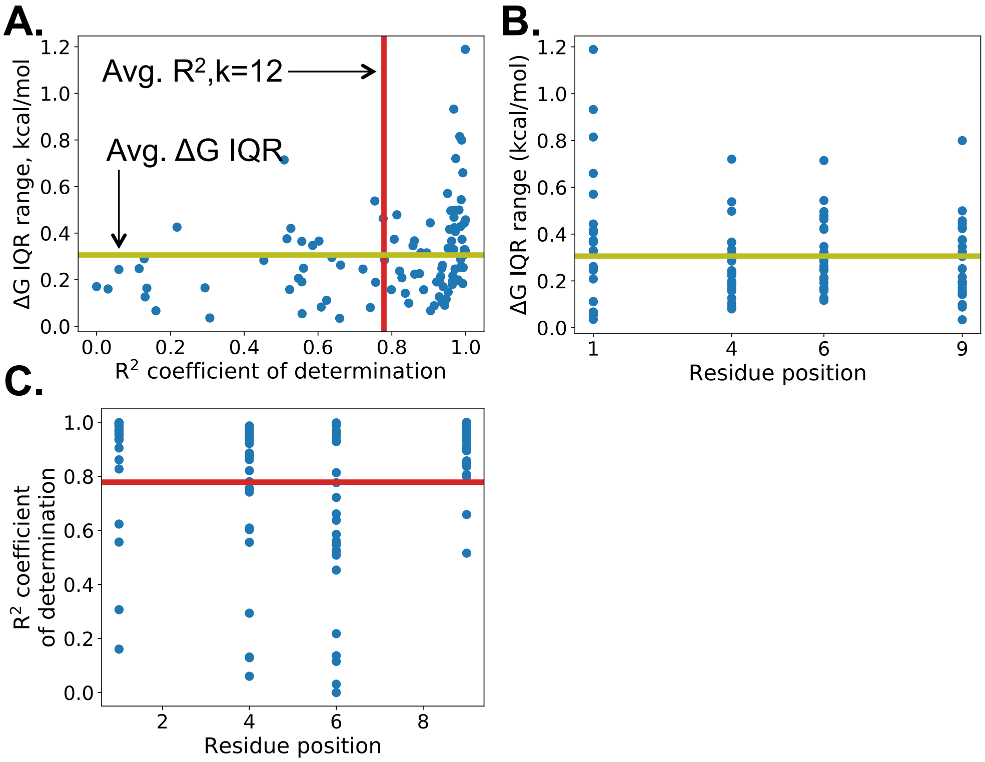

Figure S5. ΔG interquartile range (IQR) and Gaussian process regression performance for the n=100 register-constrained NetMHCIIpan 4.0-predicted set. **A.** Binding affinity ΔG IQR for n=100 systems as a function of Gaussian process regression R^2 values for the k=12 subset size. R^2 is the coefficient of determination. **B.** Binding affinity ΔG IQR for n=100 systems as a function of residue position in the MHCII antigen. **C.** Gaussian process regression R^2 values for the k=12 subset size across residue position in the MHCII antigen. **A-C.** Avg ΔG IQR from expt. and Avg R^2 values for k=12 subset size is shown in olive and red lines respectively.

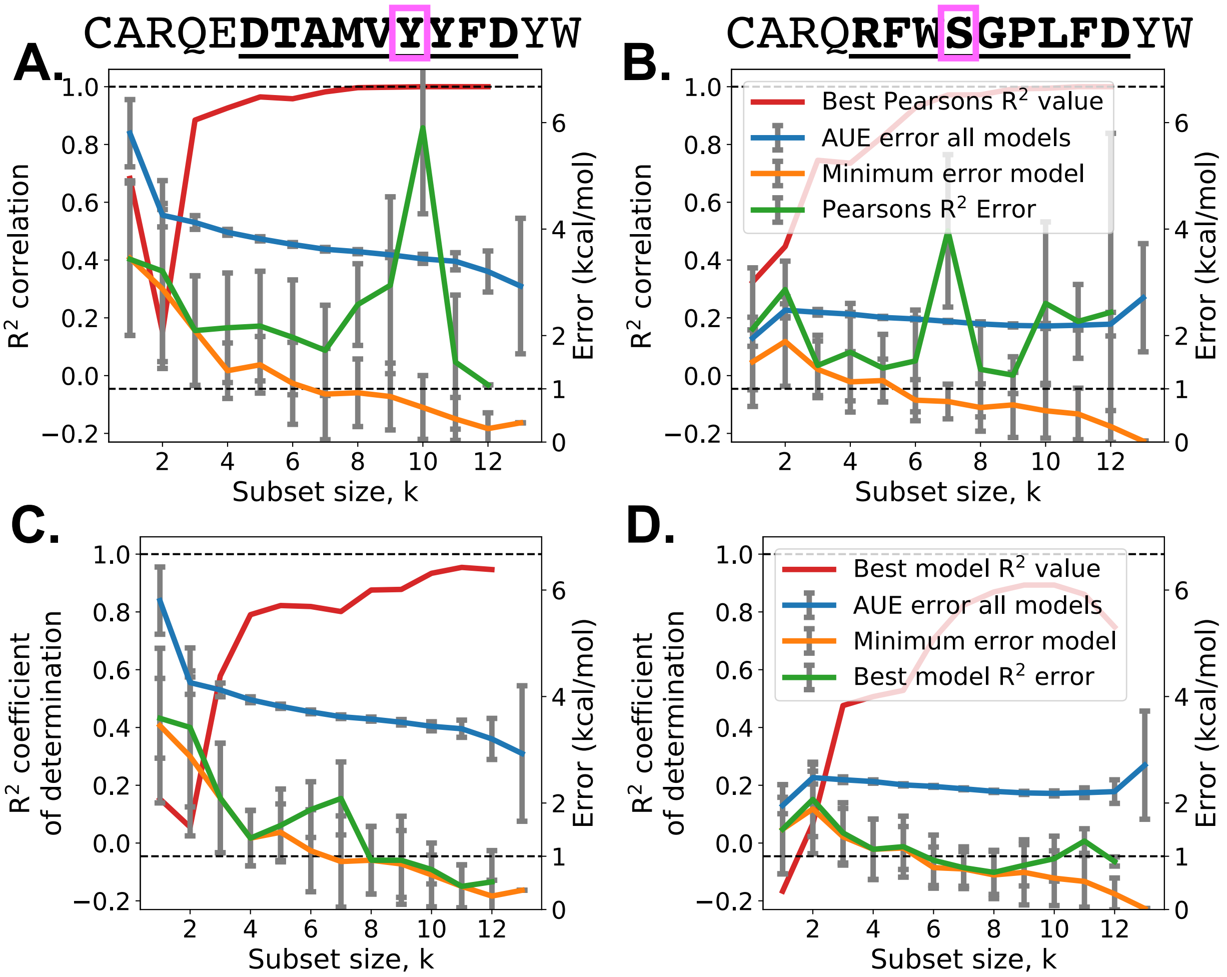

Figure S6. Gaussian process results comparing scoring by Pearson's product moment correlation coefficient (**A-B**) and the coefficient of determination (**C-D**). Note that systems are the same as shown in Fig. 2. Error bars represent 95% confidence intervals. Although the Pearson correlation coefficient is high, the variance of the data is not accounted for, hence the overall error does not decrease with increasing correlation coefficient values. In contrast, the R^2 coefficient of determination does account for data variance or spread, and thus higher coefficient of determination values generally results in lower errors.

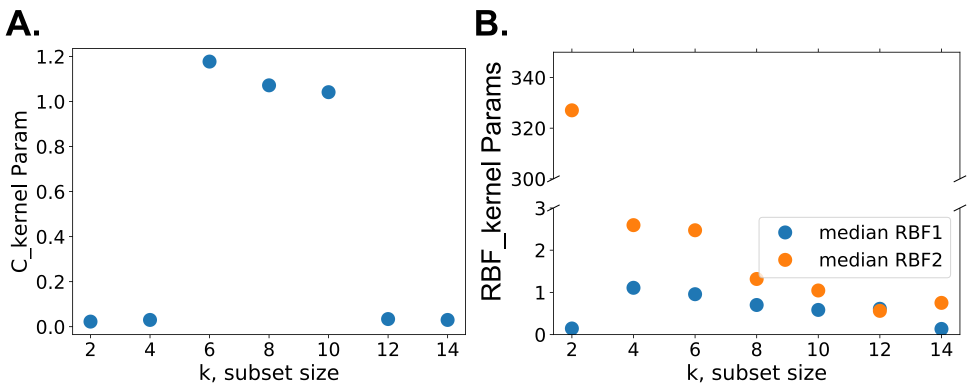

Figure S7. Gaussian process regression kernel parameters found from the NetMHCIIpan dataset. A constant kernel and a radial basis function kernel were used from scikit-learn's framework. **(A)** The C kernel parameter median value and **(B)** the RBF kernel parameters median values across k, subset size: RBF1=RBF parameter volume, RBF2=RBF parameter hydrophobicity.

Table S1. IEDB dataset

#	MHCII	Antigen Sequence, with mutation "."
1	HLA-DQA1*0102-DQB1*0602	A.ATAGTTVYGAFAA
2	HLA-DQA1*0102-DQB1*0602	AA.TAGTTVYGAFAA
3	HLA-DQA1*0102-DQB1*0602	AAATA.TTVYGAFAA
4	HLA-DQA1*0102-DQB1*0602	AAATAGT.VYGAFAA
5	HLA-DQA1*0102-DQB1*0602	AAATAGTTV.GAFAA
6	HLA-DQA1*0102-DQB1*0602	AAATAGTTVY.AFAA
7	HLA-DQA1*0102-DQB1*0602	AAATAGTTVYG.FAA
8	HLA-DQA1*0501-DQB1*0201	.FILDGDNLFPKV
9	HLA-DQA1*0501-DQB1*0201	A.ILDGDNLFPKV
10	HLA-DQA1*0501-DQB1*0201	AF.LDGDNLFPKV
11	HLA-DQA1*0501-DQB1*0201	AFI.DGDNLFPKV
12	HLA-DQA1*0501-DQB1*0201	AFIL.GDNLFPKV
13	HLA-DQA1*0501-DQB1*0201	AFILD.DNLFPKV
14	HLA-DQA1*0501-DQB1*0201	AFILDG.NLFPKV
15	HLA-DQA1*0501-DQB1*0201	AFILDGD.LFPKV
16	HLA-DQA1*0501-DQB1*0201	AFILDGDN.FPKV
17	HLA-DQA1*0501-DQB1*0201	AFILDGDNL.PKV
18	HLA-DQA1*0501-DQB1*0201	AFILDGDNLF.KV
19	HLA-DQA1*0501-DQB1*0201	AFILDGDNLFP.V
20	HLA-DQA1*0501-DQB1*0201	AFILDGDNLFPK.
21	HLA-DQA1*0501-DQB1*0301	.VKFPGGGQIVGGVY
22	HLA-DQA1*0501-DQB1*0301	D.KFPGGGQIVGGVY
23	HLA-DQA1*0501-DQB1*0301	DV.FPGGGQIVGGVY
24	HLA-DQA1*0501-DQB1*0301	DVK.PGGGQIVGGVY

25	HLA-DQA1*0501-DQB1*0301	DVKF.GGGQIVGGVY
26	HLA-DQA1*0501-DQB1*0301	DVKFP.GGQIVGGVY
27	HLA-DQA1*0501-DQB1*0301	DVKFPG.GQIVGGVY
28	HLA-DQA1*0501-DQB1*0301	DVKFPGG.QIVGGVY
29	HLA-DQA1*0501-DQB1*0301	DVKFPGGG.IVGGVY
30	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQ.VGGVY
31	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQI.GGVY
32	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQIV.GVY
33	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQIVG.VY
34	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQIVGG.Y
35	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQIVGGV.
36	HLA-DQA1*0101-DQB1*0501	.KKYFAATQFEPLAA
37	HLA-DQA1*0101-DQB1*0501	E.KYFAATQFEPLAA
38	HLA-DQA1*0101-DQB1*0501	EK.YFAATQFEPLAA
39	HLA-DQA1*0101-DQB1*0501	EKK.FAATQFEPLAA
40	HLA-DQA1*0101-DQB1*0501	EKKY.AATQFEPLAA
41	HLA-DQA1*0101-DQB1*0501	EKKYF.ATQFEPLAA
42	HLA-DQA1*0101-DQB1*0501	EKKYFA.TQFEPLAA
43	HLA-DQA1*0101-DQB1*0501	EKKYFAA.QFEPLAA
44	HLA-DQA1*0101-DQB1*0501	EKKYFAAT.FEPLAA
45	HLA-DQA1*0101-DQB1*0501	EKKYFAATQ.EPLAA
46	HLA-DQA1*0101-DQB1*0501	EKKYFAATQF.PLAA
47	HLA-DQA1*0101-DQB1*0501	EKKYFAATQFE.LAA
48	HLA-DQA1*0101-DQB1*0501	EKKYFAATQFEP.AA
49	HLA-DQA1*0101-DQB1*0501	EKKYFAATQFEPL.A
50	HLA-DQA1*0101-DQB1*0501	EKKYFAATQFEPLA.
51	HLA-DQA1*0301-DQB1*0302	.KKYFAATQFEPLAA
52	HLA-DQA1*0301-DQB1*0302	E.KYFAATQFEPLAA
53	HLA-DQA1*0301-DQB1*0302	EK.YFAATQFEPLAA
54	HLA-DQA1*0301-DQB1*0302	EKK.FAATQFEPLAA
55	HLA-DQA1*0301-DQB1*0302	EKKY.AATQFEPLAA
56	HLA-DQA1*0301-DQB1*0302	EKKYF.ATQFEPLAA
57	HLA-DQA1*0301-DQB1*0302	EKKYFA.TQFEPLAA
58	HLA-DQA1*0301-DQB1*0302	EKKYFAA.QFEPLAA
59	HLA-DQA1*0301-DQB1*0302	EKKYFAAT.FEPLAA
60	HLA-DQA1*0301-DQB1*0302	EKKYFAATQ.EPLAA
61	HLA-DQA1*0301-DQB1*0302	EKKYFAATQF.PLAA
62	HLA-DQA1*0301-DQB1*0302	EKKYFAATQFE.LAA
63	HLA-DQA1*0301-DQB1*0302	EKKYFAATQFEP.AA
64	HLA-DQA1*0301-DQB1*0302	EKKYFAATQFEPL.A
65	HLA-DQA1*0301-DQB1*0302	EKKYFAATQFEPLA.
66	HLA-DQA1*0401-DQB1*0402	.KKYFAATQFEPLAA
67	HLA-DQA1*0401-DQB1*0402	E.KYFAATQFEPLAA

68	HLA-DQA1*0401-DQB1*0402	EK.YFAATQFEPLAA
69	HLA-DQA1*0401-DQB1*0402	EKK.FAATQFEPLAA
70	HLA-DQA1*0401-DQB1*0402	EKKY.AATQFEPLAA
71	HLA-DQA1*0401-DQB1*0402	EKKYF.ATQFEPLAA
72	HLA-DQA1*0401-DQB1*0402	EKKYFA.TQFEPLAA
73	HLA-DQA1*0401-DQB1*0402	EKKYFAA.QFEPLAA
74	HLA-DQA1*0401-DQB1*0402	EKKYFAAT.FEPLAA
75	HLA-DQA1*0401-DQB1*0402	EKKYFAATQ.EPLAA
76	HLA-DQA1*0401-DQB1*0402	EKKYFAATQF.PLAA
77	HLA-DQA1*0401-DQB1*0402	EKKYFAATQFE.LAA
78	HLA-DQA1*0401-DQB1*0402	EKKYFAATQFEP.AA
79	HLA-DQA1*0401-DQB1*0402	EKKYFAATQFEPL.A
80	HLA-DQA1*0401-DQB1*0402	EKKYFAATQFEPLA.
81	HLA-DQA1*0501-DQB1*0201	.KKYFAATQFEPLAA
82	HLA-DQA1*0501-DQB1*0201	E.KYFAATQFEPLAA
83	HLA-DQA1*0501-DQB1*0201	EK.YFAATQFEPLAA
84	HLA-DQA1*0501-DQB1*0201	EKK.FAATQFEPLAA
85	HLA-DQA1*0501-DQB1*0201	EKKY.AATQFEPLAA
86	HLA-DQA1*0501-DQB1*0201	EKKYF.ATQFEPLAA
87	HLA-DQA1*0501-DQB1*0201	EKKYFA.TQFEPLAA
88	HLA-DQA1*0501-DQB1*0201	EKKYFAA.QFEPLAA
89	HLA-DQA1*0501-DQB1*0201	EKKYFAAT.FEPLAA
90	HLA-DQA1*0501-DQB1*0201	EKKYFAATQ.EPLAA
91	HLA-DQA1*0501-DQB1*0201	EKKYFAATQF.PLAA
92	HLA-DQA1*0501-DQB1*0201	EKKYFAATQFE.LAA
93	HLA-DQA1*0501-DQB1*0201	EKKYFAATQFEP.AA
94	HLA-DQA1*0501-DQB1*0201	EKKYFAATQFEPL.A
95	HLA-DQA1*0501-DQB1*0201	EKKYFAATQFEPLA.
96	HLA-DQA1*0501-DQB1*0301	.KKYFAATQFEPLAA
97	HLA-DQA1*0501-DQB1*0301	E.KYFAATQFEPLAA
98	HLA-DQA1*0501-DQB1*0301	EK.YFAATQFEPLAA
99	HLA-DQA1*0501-DQB1*0301	EKK.FAATQFEPLAA
100	HLA-DQA1*0501-DQB1*0301	EKKY.AATQFEPLAA
101	HLA-DQA1*0501-DQB1*0301	EKKYFA.TQFEPLAA
102	HLA-DQA1*0501-DQB1*0301	EKKYFAA.QFEPLAA
103	HLA-DQA1*0501-DQB1*0301	EKKYFAAT.FEPLAA
104	HLA-DQA1*0501-DQB1*0301	EKKYFAATQ.EPLAA
105	HLA-DQA1*0501-DQB1*0301	EKKYFAATQF.PLAA
106	HLA-DQA1*0501-DQB1*0301	EKKYFAATQFE.LAA
107	HLA-DQA1*0501-DQB1*0301	EKKYFAATQFEP.AA
108	HLA-DQA1*0501-DQB1*0301	EKKYFAATQFEPL.A
109	HLA-DQA1*0102-DQB1*0602	.NEPTAAAIAYGLDR
110	HLA-DQA1*0102-DQB1*0602	I.EPTAAAIAYGLDR

111	HLA-DQA1*0102-DQB1*0602	IN.PTAAAIAYGLDR
112	HLA-DQA1*0102-DQB1*0602	INE.TAAAIAYGLDR
113	HLA-DQA1*0102-DQB1*0602	INEP.AAAIAYGLDR
114	HLA-DQA1*0102-DQB1*0602	INEPT.AAIAYGLDR
115	HLA-DQA1*0102-DQB1*0602	INEPTA.AIAYGLDR
116	HLA-DQA1*0102-DQB1*0602	INEPTAA.IAYGLDR
117	HLA-DQA1*0102-DQB1*0602	INEPTAAA.AYGLDR
118	HLA-DQA1*0102-DQB1*0602	INEPTAAAI.YGLDR
119	HLA-DQA1*0102-DQB1*0602	INEPTAAAIA.GLDR
120	HLA-DQA1*0102-DQB1*0602	INEPTAAAIAY.LDR
121	HLA-DQA1*0102-DQB1*0602	INEPTAAAIAYG.DR
122	HLA-DQA1*0102-DQB1*0602	INEPTAAAIAYGL.R
123	HLA-DQA1*0102-DQB1*0602	INEPTAAAIAYGLD.
124	HLA-DQA1*0401-DQB1*0402	.NEPTAAAIAYGLDR
125	HLA-DQA1*0401-DQB1*0402	I.EPTAAAIAYGLDR
126	HLA-DQA1*0401-DQB1*0402	IN.PTAAAIAYGLDR
127	HLA-DQA1*0401-DQB1*0402	INE.TAAAIAYGLDR
128	HLA-DQA1*0401-DQB1*0402	INEP.AAAIAYGLDR
129	HLA-DQA1*0401-DQB1*0402	INEPT.AAIAYGLDR
130	HLA-DQA1*0401-DQB1*0402	INEPTA.AIAYGLDR
131	HLA-DQA1*0401-DQB1*0402	INEPTAA.IAYGLDR
132	HLA-DQA1*0401-DQB1*0402	INEPTAAA.AYGLDR
133	HLA-DQA1*0401-DQB1*0402	INEPTAAAI.YGLDR
134	HLA-DQA1*0401-DQB1*0402	INEPTAAAIA.GLDR
135	HLA-DQA1*0401-DQB1*0402	INEPTAAAIAY.LDR
136	HLA-DQA1*0401-DQB1*0402	INEPTAAAIAYG.DR
137	HLA-DQA1*0401-DQB1*0402	INEPTAAAIAYGL.R
138	HLA-DQA1*0401-DQB1*0402	INEPTAAAIAYGLD.
139	HLA-DQA1*0101-DQB1*0501	.QDLELSWNLNGLQAY
140	HLA-DQA1*0101-DQB1*0501	S.DLELSWNLNGLQAY
141	HLA-DQA1*0101-DQB1*0501	SQ.LELSWNLNGLQAY
142	HLA-DQA1*0101-DQB1*0501	SQD.ELSWNLNGLQAY
143	HLA-DQA1*0101-DQB1*0501	SQDLE.SWNLNGLQAY
144	HLA-DQA1*0101-DQB1*0501	SQDLEL.WNLNGLQAY
145	HLA-DQA1*0101-DQB1*0501	SQDLELSW.LNGLQAY
146	HLA-DQA1*0101-DQB1*0501	SQDLELSWN.NGLQAY
147	HLA-DQA1*0101-DQB1*0501	SQDLELSWNL.GLQAY
148	HLA-DQA1*0101-DQB1*0501	SQDLELSWNLN.LQAY
149	HLA-DQA1*0101-DQB1*0501	SQDLELSWNLNG.QAY
150	HLA-DQA1*0101-DQB1*0501	SQDLELSWNLNGL.AY
151	HLA-DQA1*0101-DQB1*0501	SQDLELSWNLNGLQA.
152	HLA-DQA1*0301-DQB1*0302	.QDLELSWNLNGLQAY
153	HLA-DQA1*0301-DQB1*0302	S.DLELSWNLNGLQAY

154	HLA-DQA1*0301-DQB1*0302	SQ.LELSWNLNGLQAY
155	HLA-DQA1*0301-DQB1*0302	SQD.ELSWNLNGLQAY
156	HLA-DQA1*0301-DQB1*0302	SQDL.LSWNLNGLQAY
157	HLA-DQA1*0301-DQB1*0302	SQDLE.SWNLNGLQAY
158	HLA-DQA1*0301-DQB1*0302	SQDLEL.WNLNGLQAY
159	HLA-DQA1*0301-DQB1*0302	SQDLELS.NLNGLQAY
160	HLA-DQA1*0301-DQB1*0302	SQDLELSW.LNGLQAY
161	HLA-DQA1*0301-DQB1*0302	SQDLELSWN.NGLQAY
162	HLA-DQA1*0301-DQB1*0302	SQDLELSWNL.GLQAY
163	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLN.LQAY
164	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLNG.QAY
165	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLNGL.AY
166	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLNGLQ.Y
167	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLNGLQA.

Table S2. NetMHCIIpan 4.0 sequence dataset.

#	Source/ Protein antigen	HLA	Full Antigen Sequence	NETMHCII-predicted epitope core	*Anchor residues applied GP regression
1	IEDB	HLA-DQA1*0102-DQB1*0602	AAATAGTTVYGAFAA	TAGTTVYGA	*AGTTVYGA
2					TAG*TVYGA
3					TAGTT*YGA
4					TAGTTVYG*
5	IEDB	HLA-DQA1*0501-DQB1*0201	AFILDGDNLFPKV	FILDGDNLF	*ILDGDNLF
6					FIL*GDNLF
7					FILDG*NLF
8					FILDGDNL*
9	IEDB	HLA-DQA1*0501-DQB1*0301	DVKFPGGGQIVGGVY	PGGGQIVGG	*GGGQIVGG
10					PGGGQ*VGG
11					PGG*QIVGG
12					PGGGQIVG*
13	IEDB	HLA-DQA1*0102-DQB1*0602	INEPTAAAIAYGLDR	TAAAIAYGL	*AAAIAYGL
14					TAA*IAYGL
15					TAAAI*YGL
16					TAAAIAYG*
17	IEDB	HLA-DQA1*0401-DQB1*0402	INEPTAAAIAYGLDR	EPTAAAIAY	*PTAAAIAY
18					EPT*AAIAY

19					EPTAA*IAY
20					EPTAAAIA*
21	IEDB	HLA-DQA1*0101-DQB1*0501	SQDLELSWNLNGLQAY	LELSWNLNG	*ELSWNLNG
22					LEL*WNLNG
23					LELSW*LNG
24					LELSWNLN*
25	IEDB	HLA-DQA1*0301-DQB1*0302	SQDLELSWNLNGLQAY	WNLNGLQAY	*NLNGLQAY
26					WNL*GLQAY
27					WNLNG*QAY
28					WNLNGLQA*
29	IEDB	HLA-DQA1*0101-DQB1*0501	EKKYFAATQFEPLAA	FAATQFEPL	*AATQFEPL
30					FAA*QFEPL
31					FAATQ*EPL
32					FAATQFEP*
33	IEDB	HLA-DQA1*0301-DQB1*0302	EKKYFAATQFEPLAA	FAATQFEPL	*AATQFEPL
34					FAA*QFEPL
35					FAATQ*EPL
36					FAATQFEP*
37	IEDB	HLA-DQA1*0401-DQB1*0402	EKKYFAATQFEPLAA	FAATQFEPL	*AATQFEPL
38					FAA*QFEPL
39					FAATQ*EPL
40					FAATQFEP*
41	IEDB	HLA-DQA1*0501-DQB1*0201	EKKYFAATQFEPLAA	AATQFEPLA	*ATQFEPLA
42					AAT*FEPLA
43					AATQF*PLA
44					AATQFEPL*
45	IEDB	HLA-DQA1*0501-DQB1*0301	EKKYFAATQFEPLAA	AATQFEPLA	*ATQFEPLA
46					AAT*FEPLA
47					AATQF*PLA
48					AATQFEPL*
49	Insulin + DQ8	HLA-DQA1*0301-DQB1*0302	SHLVEALYLVCGERG	EALYLVCGE	*ALYLVCGE
50					EAL*LVCGE
51					EALYL*CGE
52					EALYLVCG*
53	Insulin	HLA-DQA1*0501-	SHLVEALYLVCGERG	LYLVCGERG	*YLVCGERG

	+ DQ2	DQB1*0201			
54					LYL*CGERG
55					LYLVC*ERG
56					LYLVCGER*
57	Insulin + DQ7	HLA-DQA1*0301-DQB1*0301	SHLVEALYLVCGERG	VEALYLVCG	*EALYLVCG
58					VEA*YLVCG
59					VEALY*VCG
60					VEALYLVC*
61	X-id + DQ8	HLA-DQA1*0301-DQB1*0302	CARQEDTAMVYYFDYW	DTAMVYYFD	*TAMVYYFD
62					DTA*VYYFD
63					DTAMV*YFD
64					DTAMVYYF*
65	X-id+ DQ2	HLA-DQA1*0501-DQB1*0201	CARQEDTAMVYYFDYW	EDTAMVYYF	*DTAMVYYF
66					EDT*MVYYF
67					EDTAM*YYF
68					EDTAMVYY*
69	X-id + DQ7	HLA-DQA1*0301-DQB1*0301	CARQEDTAMVYYFDYW	EDTAMVYYF	*DTAMVYYF
70					EDT*MVYYF
71					EDTAM*YYF
72					EDTAMVYY*
73	X-id + DQ6	HLA-DQA1*0103-DQB1*0601	CARQEDTAMVYYFDYW	EDTAMVYYF	*DTAMVYYF
74					EDT*MVYYF
75					EDTAM*YYF
76					EDTAMVYY*
77	GAD65	HLA-DR0401	MNILLQYVV	MNILLQYVV	*NILLQYVV
78					MNI*LQYVV
79					MNILL*YVV
80					MNILLQYV*
81	GAD65	HLA-DR0401	LIAFTSEHS	LIAFTSEHS	*IAFTSEHS
82					LIA*TSEHS
83					LIAFT*EHS
84					LIAFTSEH*
85	GAD65	HLA-DR0401	FTSEHSHFS	FTSEHSHFS	*TSEHSHFS
86					FTS*HSHFS
87					FTSEH*HFS
88					FTSEHSHF*
89	GAD65	HLA-DR0405	LYNIIKNREG	YNIIKNREG	*NIIKNREG
90					YNI*KNREG

91					YNIIK*REG
92					YNIIKNRE*
93	GAD65	HLA-DR0405	SLRTLEDNEE	LRTLEDNEE	*RTLEDNEE
94					LRT*EDNEE
95					LRTLE*NEE
96					LRTLEDNE*
97	GAD65	HLA-DR0405	FFRMVISNPAA	FRMVISNPA	*RMVISNPA
98					FRM*ISNPA
99					FRMVI*NPA
100					FRMVISNP*