Energy landscapes of peptide-MHC binding

Laura Collesano[1], Marta Łuksza[2], and Michael Lässig[1]

[1]Institute for Biological Physics, University of Cologne, Zülpicherstr. 77, 50937 Köln, Germany

[2]Tisch Cancer Institute, Departments of Oncological Sciences and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Abstract

Molecules of the Major Histocompatibility Complex (MHC) present short protein fragments peptides on the cell surface, an important step in T cell immune recognition. MHC-I molecules process peptides from intracellular proteins; MHC-II molecules act in antigen-presenting cells and present peptides derived from extracellular proteins. Here we show that the sequence-dependent energy landscapes of MHC-peptide binding encode class-specific nonlinearities (epistasis). MHC-I has a smooth landscape with global epistasis; the binding energy is a simple deformation of an underlying linear trait. This form of epistasis enhances the discrimination between strong-binding peptides. In contrast, MHC-II has a rugged landscape with idiosyncratic epistasis: binding depends on detailed amino acid combinations at multiple positions of the peptide sequence. The form of epistasis affects the learning of energy landscapes from training data. For MHC-I, a low-complexity problem, we derive a simple matrix model of binding energies that outperforms current models trained by machine learning. For MHC-II, higher complexity prevents learning by simple regression methods. Epistasis also affects the energy and fitness effects of mutations in antigen-derived peptides (epitopes). In MHC-I, large-effect mutations occur predominantly in anchor positions of strong-binding epitopes. In MHC-II, large effects depend on the background epitope sequence but are broadly distributed over the epitope, generating a bigger target for escape mutations from T cell immunity than for MHC-I.

Author Summary

T cell immunity involves the binding of short peptides to the intracellular MHC recognition machinery. Understanding how the binding energy depends on the peptide sequence is key to computationally predict immune recognition and immune escape evolution, for example, of pathogens and cancer cells. We find nonlinear energy landscapes that depend on the recognition pathway: smooth and easy to learn for MHC class I, rugged and difficult to learn for class II. Together, this work establishes links between biophysical origin, nonlinear structure, learnability from data, and biological implications for protein interaction landscapes.

The antigen presentation pathway is pivotal for immune surveillance [1, 2]. Class I and class II Major Histocompatibility Complex (MHC) enable this process by binding and presenting small peptides to CD8+ and CD4+ T cells, respectively. Although there are several steps between the processing of an antigen and its presentation [3], protein-MHC complex formation is considered the most discriminating stage [4]. Therefore, a peptide with high binding affinity has also higher chance of being presented and recognized by the immune system [5]. Quantitatively understanding protein-MHC binding is therefore a critical step for vaccine design and cancer immunology [6, 7, 8].

In the context of cytotoxic T cell recognition, the antigen is presented on the cell surface in a complex with a MHC-I molecule. The peptides able to bind MHC-I are short chains, on average 8 to 11 amino acids long, resulting from the degradation of intracellular proteins carried out by the proteasome [9, 10]. MHC-II acts in specific antigen-presenting cells, presenting peptides transported from the extracellular space for interaction with helper T cells [11]. The amino acid chains that bind to MHCII are more variable in length, on average 15-20 amino acids, and the relevant binding occurs in a subset of about 9 amino acids [12, 13]. MHC-I and MHC-II have different binding mechanisms. The MHC-I binding pocket is closed at both ends, causing peptides to be anchored at their extremities, usually in positions 2 and 9, while a weaker interaction acts in the center [14, 15]. In contrast, MHC-II has an open binding pocket, allowing the peptide to be anchored at multiple positions [12, 16].

The complexity of interactions between binding positions represents the key challenge for affinity models [17, 18]. Ini-

tial studies on MHC-I include pairwise interactions among specific peptide positions in otherwise additive models [17, 19]. In contrast, state-of the-art computational tools based on neural networks trained on experimental data contain pervasive nonlinearities in the binding energy function, albeit implicit in the model output. These models are highly effective for MHC-I alleles [20, 21, 22, 23, 24, 25, 26], they are also increasingly successful in addressing the higher complexity of the binding mechanism for MHC-II alleles [13, 18, 27]. However, given their high number of parameters and their complexity, training becomes difficult for rare alleles insufficiently represented in empirical datasets. Moreover, the neural network architecture impedes a mechanistic interpretation of the model output [28].

This paper presents a systematic study of the energy landscapes of protein-MHC binding, which map a given peptide sequence onto the binding affinity to a given MHC allele. These landscapes have nonlinearities, or epistasis [29, 30, 31, 32, 33, 34], caused by interactions of multiple peptide and MHC amino acids – a common feature of peptide-MHC binding [17]. The form of epistasis, however, depends on the MHC class. For MHC-II, we find rugged landscapes where the binding energy depends on specific amino acid combinations of a peptide in a network of several binding positions. The MHC-I landscapes are much simpler: the binding energy is a nonlinear function of an intermediate trait that is additive in the peptide sequence [31], a less complex alternative to the pairwise interactions approach, already well studied in previous works [19]. In the literature on evolution, these two types of epistasis, characterising MHC-II and MHC-I energy landscapes, are referred to as idiosyncratic and global, respectively [35, 36, 37, 38].

In the first part of the paper, we infer global and idiosyncratic epistasis in energy landscapes from empirical data of several MHC-I and MHC-II alleles. We show that data and predictive models for MHC-I and MHC-II have low and high complexity, respectively. In the second part, we discuss two important consequences of the form of epistasis. First, we infer a matrix energy model for MHC-I that outperforms state-of-the art machine-learning models (NetMHC 4.0), providing predictions of very similar quality with a drastically reduced number of model parameters. This model can be learned by simple regression, demonstrating that the complexity of epistasis sets the optimal learning algorithm. Second, we infer spectra of mutational effects in energy landscapes for MHC-I and MHC-II, and we discuss the differences between these spectra induced by the form of epistasis.

Results

Datasets of peptide-MHC binding affinities. For training and testing, we analyze peptide sequences and MHC binding data from the Immune Epitope Database (IEDB) [39]. The datasets are curated before training (Material and Methods). For MHC-I, we include the most frequent alleles of HLA-A and HLA-B: HLA-A*02:01, HLA-A*11:01, and HLA-B*07:02 with 6953, 1981, and 1024 curated peptides of length 9 aa. The resulting dataset is similar but not identical to the data used for benchmarking and retraining of NetMHC 4.0 [21, 40]. For MHC-II, our dataset contains 7434 peptides from allele HLA-DRB1*01:01 of length 15 aa with a binding core of length 9 aa. In these data, binding affinities are reported as half-inhibitory concentration (IC50), other peptides are annotated qualitatively as binding or non-binding. Throughout this paper, we describe binding affinities by the free energy $\Delta G = \log(\mathrm{IC50}/\rho_0)$, which is measured in units of $k_B T$ and has its zero point at the physiological binding threshold, $\rho_0 = 500$nM (MHC-I) and $\rho_0 = 1000$nM (MHC-II) [41, 42].

Inference of a linear energy model. Fig. 1A shows position-weight matrices, or binding motifs, for two representative alleles HLA-A*02:01 for MHC-I and HLA-DRB1*01:01 for MHC-II. These matrices record the frequencies $p_i(a)$ of amino acid a at a given position in the binding core ($i = 1, \ldots, \ell$), evaluated in the set of binding peptides ($\Delta G \leq 0$). The resulting position-specific sequence information, $\Delta S_i = \sum_a p_i(a) \log(p_i(a)/p_0(a))$, is defined as the Kulback-Leibler distance between the observed amino acid frequency distribution and a null distribution (here $p_0(a) = 1/20$). For MHC-I, the sequence information is concentrated in positions 2 and 9, the known anchor positions of peptide-MHC binding [43] (Fig. 1B). In contrast, MHC-II has a more homogeneously distributed sequence information, i.e., less bias towards the anchor positions 1, 4, 6, and 9. The total additive information, $\Delta S = \sum_{i=1}^{\ell} \Delta S_i$, which measures the power of linear sequence models to discriminate peptides presented on the cell surface, is lower for MHC-II than for MHC-I.

As a starting point for the inference of energy models, we construct a linear energy matrix,

$$L(\mathbf{a}) = \sum_{i=1}^{\ell} \epsilon_i(a_i), \tag{1}$$

where $\mathbf{a} = (a_1, \ldots, a_\ell)$ denotes the peptide sequence and $\epsilon_i(a)$ is the energy contribution of amino acid a at position i. Energy models of this kind are widely used for transcription factor binding and for protein-protein binding, including early-stage predictions of peptide-MHC-I interactions [44, 19]. Here we obtain an optimized 20×9 energy matrix, which minimises the mean squared error (MSE), via 10-fold cross-validation for each of the selected alleles (Material and Methods). At each position, the energy ranking of amino acids reproduces the frequency ranking in the position weight matrix (Table S2 and Table S8).

The heat maps of Fig. 1C show the optimized linear

energy L plotted against the validation data for the two representative alleles. For MHC-I, the linear model systematically underestimates the affinity of strong binders ($L > \Delta G$). A similar effect is observed for the weakest binders, albeit supported by fewer data points. For MHC-II, the linear model has a more drastic mismatch with the data and only explains about half of the variation in ΔG. Likely factors contributing to this mismatch include the variation of binding cores and additional energy contributions of peptide flanking regions not included in the model [45, 46]. In what follows, we will link these deficiencies of the linear model to epistasis, the form of which depends on the MHC class.

Energy landscapes with global and idiosyncratic epistasis. Nonlinear energy landscapes provide a better representation of peptide-MHC binding. Here we refer to such nonlinearities as energy epistasis and parametrize them by interaction terms involving pairs of amino acids in the peptide binding core. For MHC-II, we use a generic landscape of the form

$$\Delta G(\mathbf{a}) = \sum_i \epsilon_i(a_i) + \sum_{i>j} \beta_{ij}(a_i, a_j); \qquad \text{(MHC-II)} \quad (2)$$

for each pair of positions, the 20×20 matrix $\beta_{ij}(a, b)$ characterizes the interaction energy between amino acid a at position i and b at position j. For MHC-I, we use a simpler landscape,

$$\Delta G(\mathbf{a}) = \sum_i \epsilon_i(a_i) + \lambda \sum_{i,j} \epsilon_i(a_i)\epsilon_j(a_j); \qquad \text{(MHC-I)} \quad (3)$$

where the interactions are given by the linear energy terms, $\beta_{i,j} = \lambda\epsilon_i(\alpha_i)\epsilon_j(\alpha_j)$. Following the literature on fitness landscapes, we refer to these two types of nonlinearities as idiosyncratic and global epistasis, respectively [38, 35]. Under idiosyncratic epistasis, the energy effect of a mutation at a given position, $\Delta\Delta G$, depends on the detailed sequence content at other positions. Under global epistasis, it depends only on the energy of the original sequence. In other words, the full energy can be written as a nonlinear function of an intermediate linear trait, here $\Delta G = L + \lambda L^2$ [31, 48].

These two types of energy landscapes are strongly different in model complexity. The idiosyncratic interaction term in Equation (2) has 20×20 energy parameters for the amino acid combinations at each each pair of interacting peptide positions. The global epistasis model of Equation (3) requires only $20 \times \ell$ parameters for the linear term and a single additional parameter λ measuring the strength of epistasis.

A common approach to infer epistasis from data is to compare the effect of a mutation on different sequence backgrounds, for example, by analyzing pairs of mutations at two positions in two different orders ($ab \to a'b \to a'b'$ and $ab \to ab' \to a'b'$). Here, given the limited availability of such pairs in our datasets, we map instead the deviations of partial data from the previously inferred linear model [49].

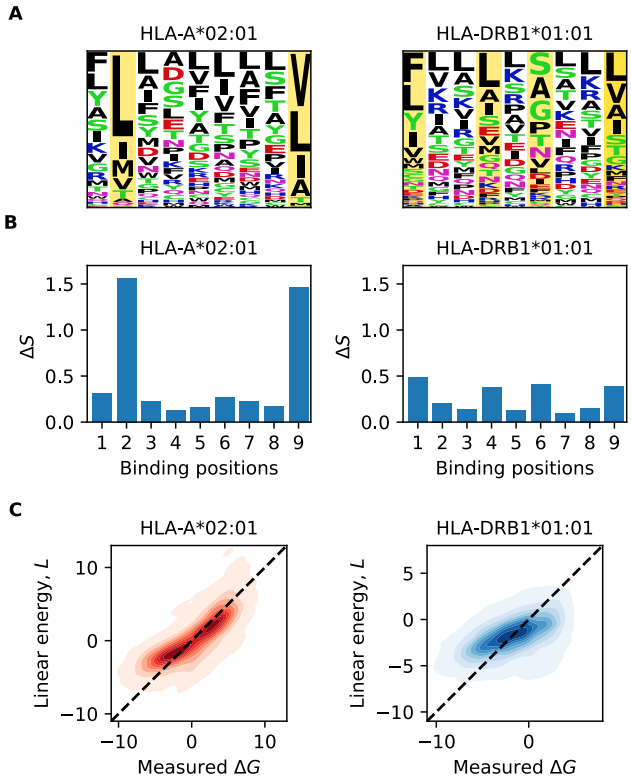

Figure 1. Binding motifs, sequence information, and linear model. (A) Position weight matrix for HLA-A*02:01 (left) and HLA-DRB1*01:01 (right) inferred from training data. Letter size indicates the frequency of a given amino acid at a given position, $p_i(a)$, in the sample of MHC-binding peptides ($\Delta G \leq 0$). Yellow bars highlight anchor positions, the color of letters shows the chemical properties of amino acid (black: hydrophobic, red: acidic, blue: basic, green: polar, purple: neutral) [47]. (B) Position-specific sequence information, ΔS_i, defined as the Kulback-Leibler distance of the observed distribution $p_i(a)$ from an equidistribution. (C) Error plot comparing observed energies, ΔG, and predictions of the linear model, L (contours give densities above 0.01).

For MHC-I, we form data subsamples containing sequences with particular amino acid pairings in positions 2 and 9. For each subsample, we perform a linear regression, the slope and offset of which compare validation data with the linear model L (dashed line). Fig. 2A shows these regressions for a few subsamples of peptides binding HLA-A*02:01; all subsamples are shown in Fig. S2. We observe a systematic modulation of subsample slopes with linear energy L: deviations from the linear model increase from

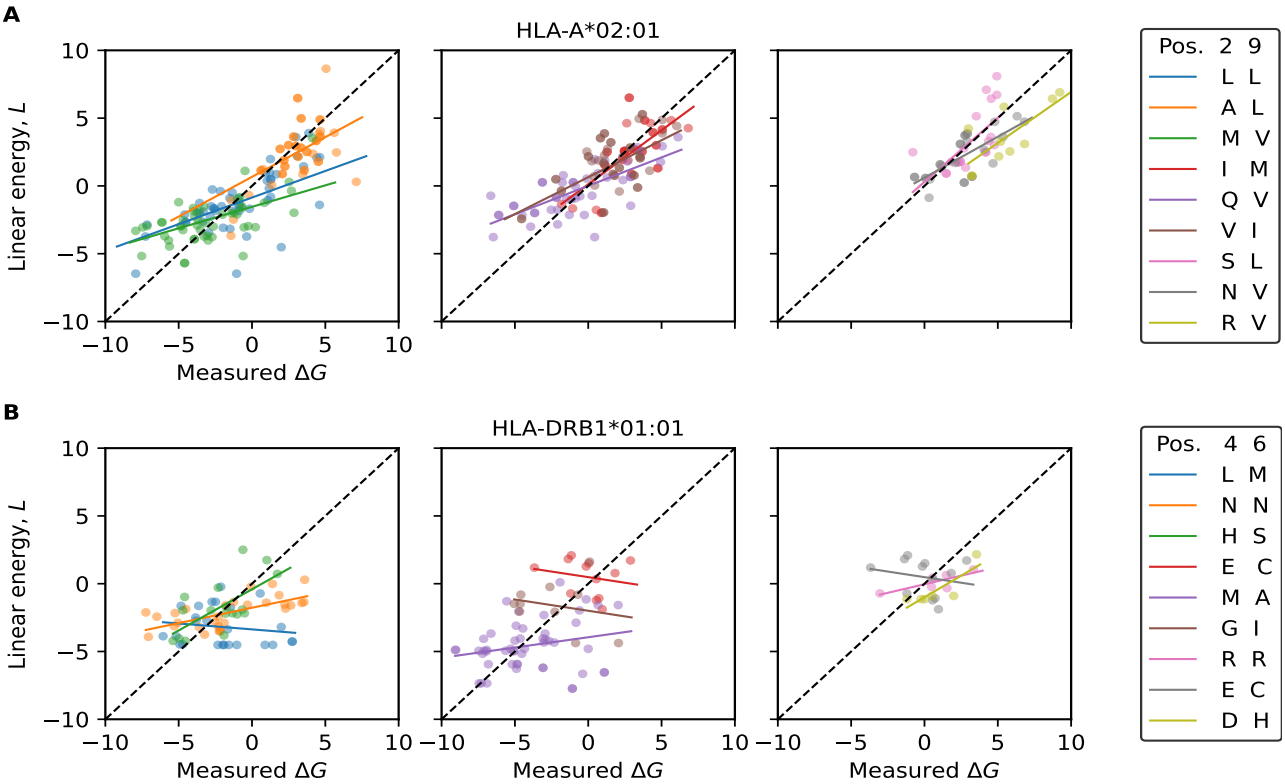

Figure 2. Inference of epistasis. Scatter plots of experimental ΔG values and the linear energy model L for subsamples with given amino acid pairs in selected positions (max 50 random points are displayed). Top: HLA-A*02:01, selected positions 2 and 9; bottom: HLA-DRB1*01:01, selected positions 4 and 6. From left to right: subsamples containing stronger binders, marginal binders, and non-binders, respectively. Lines give linear regressions for each subsample. See Fig. S2 for regression lines of all subsamples.

weak to strong binders (from left to right panel), but are similar for different subsamples in the same range of L. This pattern is consistent with global epistasis of the form $\Delta G(L) = L + \lambda L^2$, as described by Equation (3).

For MHC-II, here allele HLA-DRB1*01:01, we select subsamples with specific amino acid pairings in positions 4 and 6. In this case, the subsample slopes scatter strongly in the same range of L and there is no global trend from weak to strong binders (Fig. 2B, Fig. S2). For example, the pairing Hystidine-Serine (H-S) shows a positive correlation between observed ΔG and L in peptides with marginal binding, the pairing Leucine-Methionine (L-M) shows a negative correlation in the same range of L. Such disorder is seen ubiquitously, pointing to idiosyncratic epistasis depending on individual amino acid pairings, as described by Equation (2).

Complexity of epistatic landscapes. The above analysis can be summarised using the difference between two errors: the MSE produced by a global function $\Delta G(L)$, fitted in the same L region, compared to the MSE produced by a family of functions $\Delta G(L)$ fitted on subsamples of L characterised by different amino-acid pairing in anchor positions (Fig. 2 and S2, Materials and Methods). For the MHC-I allele HLA-A*02:01, we find an error difference $\Delta\text{MSE} = 0.6 \pm 0.4$, indicating that a global model is sufficient to describe the data and the marginal decrease in MSE of the subsample regressions does not justify the increase in model complexity. For the MHC-II allele HLA-DRB1*01:01, we obtain $\Delta\text{MSE} = 2.0 \pm 0.9$, suggesting a more complex epistatic model is appropriate.

For comparison, we also analyze epistatic landscapes with random parameters. For idiosyncratic epistasis, we use energy landscapes of the form (2), where the parameters $\epsilon_i(a)$ and $\beta_{ij}(a,b)$ are Gaussian random variables with mean 0 and variance σ_ϵ^2 and σ_β^2, respectively. For a given landscape, we evaluate the energy $\Delta G(\mathbf{a})$ for a sample of random peptides, subject to an additional measurement error of variance $\sigma_m^2 = 1.5$. The strength of epistasis is given by $\tilde{\lambda} = \sigma_\beta^2/\sigma_\epsilon^2$ (details of the simulations are given in Materials and Methods). With an appropriate strength ($\tilde{\lambda} = 2$), the simulated data generate a regression pattern strikingly similar to the empirical data of MHC-II (Fig. S2, Fig. S3B),

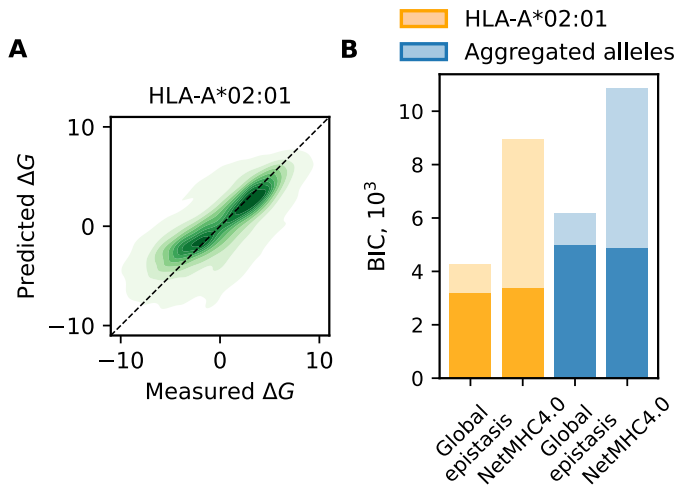

Figure 3. Global-epistasis model for MHC-I. (A) Error plot comparing observed energies, ΔG, and predictions of the optimal global-epistasis model ($\lambda = -0.04$) for HLA-A*02:01. To be compared with the linear model of Figure 1; see Fig. S1 for other alleles. (B) Model comparison of the global-epistasis model and NetMHC-4.0: error function (dark shading) and total BIC score accounting for model complexity (full bars) for allele HLA-A*02*01 (orange) and aggregated over all 3 MHC-I alleles (blue), evaluated by blind testing (see text for details).

a similar error of the linear model (MSE $= 6.2 \pm 0.2$), and a similar error difference between ΔMSE $= 2.3 \pm 1.0$, computed over 10 realizations. For global-epistasis landscapes of the form (3), the energy parameters $\epsilon_i(a)$ are again Gaussian random variables with variance σ_ϵ^2, which is chosen according to the empirical values inferred for MHC-I alleles (Table S2). The simulated data are drawn with a measurement error of variance $\sigma_m^2 = 1$. For intermediate strength of epistasis ($\lambda = -0.04$), the regression pattern is again similar to the empirical data (Fig. S3A) and produces a similar error difference (ΔMSE $= 0.20 \pm 0.1$).

Energy landscapes with global epistasis for MHC-I. The type of epistasis has immediate consequences for the learnability of landscape models from training data. For MHC-I, we can infer a global-epistasis model by regression: we learn the linear model $L(\mathbf{a})$ together with the parameter λ of the nonlinear function $\Delta G(L)$. For the allele HLA-A*02:01, we infer an optimal quadratic model of the form (3) with $\lambda = -0.04$. This curvature is associated with moderate *increasing-return* epistasis: the global nonlinearity reduces the systematic error of the linear model in the strong-binding regime (Fig. 3A, Fig. 1C), leading to a reduced mean square error (MSE = 4.3) compared to the linear model (MSE = 4.6) in the 10 fold cross-validation procedure. Adding a cubic nonlinearity to the function does not further reduce the MSE in a significant way. As shown below, the nonlinearity substantially increases the effects of mutations in strong binders. Very similar global energy landscapes are inferred for the other MHC-I alleles included in our analysis (Fig. S1).

Next, we compare the performance of the global-epistasis model with the most widely used neural network model, NetMHC 4.0, in both binding energy predictions and classification. As validation data for binding energy predictions, we use the set of all IEDB peptides that are not contained in the training data for NetMHC 4.0 (obtained from IEDB Analysis Resources [39, 40]). The set contains 465, 69 and 183 sequences for HLA*A-02:01, HLA*A-11:01 and HLA*B-07:02, respectively. These peptides are also excluded from the training of our model. The global-epistasis model for HLA*A-02:01 reaches a smaller error than NetMHC 4.0 (MSE = 6.8 compared to MSE = 7.2). The aggregated error computed on all MHC-I alleles included in the analysis reaches a slightly larger value (MSE = 6.9 compared to MSE = 6.8 for NetMHC 4.0), reflecting the reduced availability of training sequences (Fig. S4). To assess the classification performance, we collect sets of positive and negative binders from IEDB. We exclude sequences used for training and present in the referenced dataset [39, 40]. The performance metric is reported as Area Under the Curve (AUC, see Material and Methods). For HLA*A-02:01, we obtain a score of 0.837 for the global epistasis model and 0.843 using NetMHC 4.0. Scores for other alleles are detailed in Table S1.

This level of performance is achieved with a drastically lower model complexity: the global-epistasis model has $20 \times 9 + 1$ parameters per allele, significantly less than the number of parameters required to account for all the pairwise interaction terms between each positions [17, 19]. In comparison, the network model employs > 900 parameters [21].

Factoring in the difference in model complexity by a Bayesian Information Criterion (Materials and Methods), the global-epistasis model strongly outperforms NetMHC 4.0 (Fig. 3B). Hence, for the smooth energy landscapes of MHC-I alleles, neural network models are an overkill: they apply a complex multi-level learning procedure to an inference problem that can be solved by simple regression. In contrast, neural network models are an appropriate tool for the intrinsically more complex idiosyncratic energy landscapes of MHC-II. For HLA-DRB1*01:01, low-complexity models of the form (3) fail to map the energy variation (within-dataset MSE = 5.7), similar to the linear model shown in Fig. 1C (cross-validated MSE = 5.9), while NetMHCII-2.3 still provides a faithful representation of the data (within-dataset MSE = 1.6).

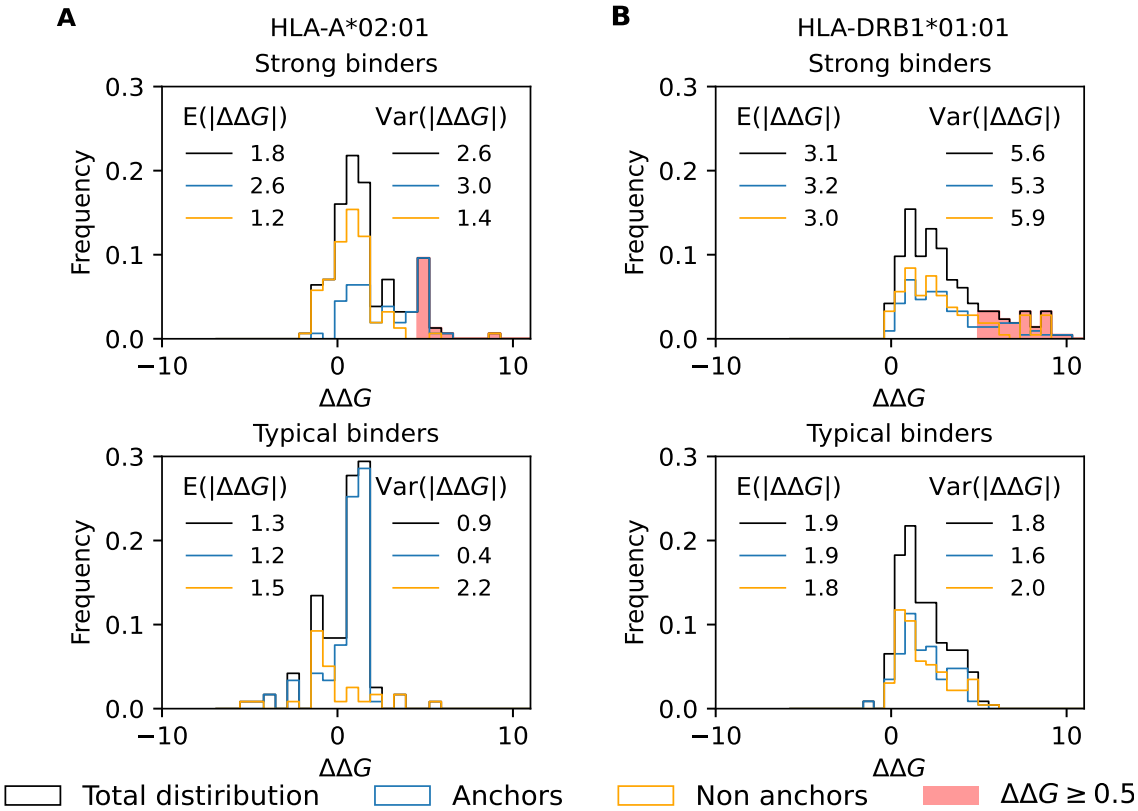

Figure 4. Mutational effect spectra. Distribution of energy changes, $\Delta\Delta G$, for mutations from ancestral peptides in different energy windows: strong binders ($\Delta G \approx -5$, top) and typical binders ($\Delta G \approx -1$). The total distribution (black) is shown together with the components from anchor positions (blue) and non-anchor positions (orange). For strong binders, red shading marks escape mutations leading to loss of presentation ($\Delta\Delta G \geq 5$). Inserts give mean and variance of magnitude effects $|\Delta\Delta G|$. (A) MHC-I, allele HLA-A*02:01 (anchor positions 2 and 9). (B) MHC-II, allele HLA-DRB1*01:01 (anchor positions 1, 4, 6 and 9).

Mutational effect spectra in epistatic landscapes. We can use the energy data of MHC-I and MHC-II for a comparative analysis of mutational effects. In Fig. 4, we show the distribution of energy changes, $\Delta\Delta G$, plotted separately for mutations at anchor and non-anchor positions and for ancestral sequences in two affinity windows: typical binders ($\Delta G_{\text{anc}} \approx -1$) and strong binders ($\Delta G_{\text{anc}} \approx -5$). For MHC-I, we observe homogeneously upscaled effects for mutations of strong binders, reflecting the larger slope of the global epistasis function $\Delta G(L)$ (Fig. 4A). This upscaling increases mean and variance of magnitude effects, $E(|\Delta\Delta G|)$ and $\text{Var}(|\Delta\Delta G|)$, in non-anchor and, at larger amplitudes, in anchor positions (inserts in Fig. 4). For MHC-II, the mean effect $E(|\Delta\Delta G|)$ is also enhanced in strong binders (Fig. 4B). This is the known global pattern of idiosyncratic epistasis [38]: strong binders have a higher fraction of affinity-increasing bonds (i.e., amino acid pairings for which $\beta_{ij}(a_i, a_j) < 0$), generating an additional energy increase for mutations breaking these bonds. However, the variance of magnitude effects is largely independent of ΔG_{anc}, and the effect distributions are similar in anchor and non-anchor positions. This is again consistent with strong idiosyncratic epistasis: large mutational effects at non-anchor positions can be generated by energy bonds to anchor positions.

To quantify idiosyncratic epistasis directly, we record mutational effect spectra generated by background sequence variation from calibrated and random energy models for MHC-II (these spectra are undersampled in the data). For a given mutation $a \to a'$ of a peptide with given energy ΔG_{anc}, we resample the ancestral background sequence at the other positions subject to the constraint of maintaining ΔG_{anc}; the resulting effect spectrum is then shifted to mean 0. In Fig. S5, we plot these spectra for a sample of mutations in typical binders ($\Delta G_{\text{anc}} \approx -1$) and strong binders ($\Delta G_{\text{anc}} \approx -5$), together with the corresponding sample-averaged distributions. The rms effect given by these distributions measures idiosyncratic deviations from global epistasis. This effect is sizeable for typical and for strong binders and similar to the global shift in $E(|\Delta\Delta G|)$ between the two binding regimes (Fig. 4B, Fig. S5). Hence, in MHC-II, the energy effect of a given mutation depends

strongly on the specific amino acid pairings of the ancestral peptide.

Immune escape evolution. The energy statistics of peptide-MHC binding affects the escape evolution of antigens from T cell immune recognition. In both MHC classes, even strong-binding antigenic epitopes can effectively suppress presentation by large-effect mutations ($\Delta\Delta G \gtrsim 5$, shaded areas in Fig. 4). For MHC-I, large-effect mutations occur in anchor positions, while most mutations in non-anchor positions maintain presentation. This reflects the modularity of presentation and T cell recognition encoded in peptide sequences. For MHC-II, escape mutations acting on presentation are broadly distributed across anchor and non-anchor positions. The overall target for this type of escape evolution is larger for MHC-II: 21 % of all mutations have $\Delta\Delta G \geq 5$, compared to 12 % for MHC-I. This suggests the idiosyncrasy of peptide-MHC-II binding facilitates the escape evolution of antigens by generating large epistatic mutation steps.

Discussion

Peptide-MHC binding interactions have strong, class-specific nonlinearities. For MHC-I, these interactions can be described by an energy landscape with global epistasis, $\Delta G(L)$, where the intermediate trait L is additive in the peptide sequence. The curvature of the landscape increases in slope at strong binding. This enhances the energy discrimination of presented peptides and increases selection on affinity-decreasing escape mutations in pathogen evolution. Moreover, the recognition chain of class I is modular: strong MHC binding energy effects governing peptide presentation are predominantly determined by the anchor positions 2 and 9, while non-anchor positions encode the extracellular interactions of presented peptides with T cell receptors. For MHC-II, we find rugged energy landscapes with idiosyncratic epistasis, describing cooperative binding that depends on the amino acid content at multiple peptide binding positions. Mutational effects at a given binding position depend strongly on the peptide sequence background, and there is no modularity: weak and strong effects are found at anchor and non-anchor positions.

These differences in epistasis between energy landscapes are in tune with distinct biophysical binding mechanisms. MHC-I molecules bind short peptides in a fixed contact map; global epistasis of the binding energy is consistent with mutual reinforcement of binding between the anchor positions that is mediated by the intermediate chain. In contrast, MHC-II binds longer peptides with variable contact maps [16], suggesting that idiosyncratic epistasis can partly be attributed to changes in the core contact points and in the peptide configuration in between these points.

The type of epistasis also governs the efficient learning of energy landscapes from data. For MHC-I, we derive a simple energy matrix model that outperforms state-of-the art machine-learning models [21] in BIC score. This model can be inferred by a joint regression procedure for the linear trait $L(\mathbf{a})$ and for the nonlinear function $\Delta G(L)$. Such regression methods fail for MHC-II, while neural networks still produce a faithful energy landscape [13]. The difference in learnabilty is not surprising, given the strongly different complexity of global and idiosyncratic energy models. This compression can be understood as a dimensionality reduction in phenotype space. In the idiosyncratic case, sequence variation affects not only the focal trait, here the binding energy ΔG, but also a number of co-varying traits $A, B, \dots,$ including the spatial configuration of the interaction surface and allosteric interactions. A given mutation $a \to a'$ on two different backgrounds b, b' has effects $\Delta\Delta G_b = \Delta\Delta G(\Delta G_{ab}, A_{ab}, B_{ab}, \dots)$ and $\Delta\Delta G_{b'} = \Delta\Delta G(\Delta G_{ab'}, A_{ab'}, B_{ab'}, \dots)$. These effects differ because they probe the slope of the energy landscape at different values of the co-varying traits, even if they start from similar energies, $\Delta G_{ab} \approx \Delta G_{ab'}$. Under global epistasis, the ensemble of traits collapses to a single relevant trait ΔG. This implies $\Delta\Delta G_b = \Delta\Delta G(\Delta G_{ab})$ and $\Delta\Delta G_{b'} = \Delta\Delta G(\Delta G_{ab'})$; the only remaining source of epistasis for the mutation $a \to a'$ is the energy effect of the background mutation $b \to b'$. The (lack of) collapse from idiosyncratic to global epistasis can be evaluated in a given training dataset and serve as a diagnostic for the appropriate learning procedure.

In summary, we have highlighted the roles of epistasis in the peptide-MHC energy landscapes for T cell immunity. The type of epistasis is linked to the underlying biophysical mechanism, governs the learnabilty of landscape models from data, and shapes the evolution of binding agents (here, antigenic peptides). We expect this broad picture of trait-based epistasis to be applicable to other protein-protein interaction landscapes.

Acknowledgement

We thank Leon Seeger and Denny Trimcev for discussions and comments on the manuscript. This work was supported by Deutsche Forschungsgemeinschaft CRC 1310 (to ML). M. Łuksza was supported as a Pew Biomedical Scholar.

Material and Methods

Dataset curation. The datasets for both MHC-I and MHC-II were sourced from the IEDB database (version April 2023 for HLA-A*02:01, May 2023 for all other alleles), reporting measured IC50 values in nM. The experimental methods involve only in vitro binding assays, no mass spectrometry measurements were selected. These dataset contains quantitative and qualitative measurements. We average ΔG values over multiple measurements relative to the same sequence from training and test data. We first discard peptides containing non-canonical amino acids. Several sequences contain multiple measurements, which are curated in the following way. For a set of quantitative ΔG, we accept values within one standard deviation from the mean and discard others. For sets of qualitative measurements ($\Delta G > \Delta G_{\text{thr}}$), we select the minimum ΔG_{thr}. We generate cross-validation training and test sets, taking care not to include the same sequence in both sets at the same time.

Extraction of MHC-II cores. From IEDB, we collect sequences of 15-mers along with their binding affinities. However, our binding model focuses their core of 9 aa that directly binds to the complex. To extract this core, we use to NetMHCII-2.3, which provides an estimation of the core for longer input sequences. The extracted cores are paired with the experimental measures provided by IEDB and used for the inference of the linear model.

Model optimisation. The energy matrix and coefficients were trained and tested using a 10-fold cross-validation approach: we randomly divide the data were randomly in 10 subsets, 9 of which are used for training and one for testing. We optimize the error function

$$n \times \text{MSE} = \sum_{i=1}^{n} (\Delta G_{i,\text{model}} - \Delta G_{i,\text{data}})^2, \tag{4}$$

where n is the number of sequences within the test set.

The function in Equation (4) is adapted for inequalities present in the datasets. For qualitative data with $\Delta G > \Delta G_{\text{thr}}$, we use the error 0 if $\Delta G_{\text{model}} > \Delta G_{\text{thr}}$ and $(\Delta G_{\text{model}} - \Delta G_{\text{thr}})^2$ otherwise. For the linear model, we start from a matrix of randomly generated ϵ, each drawn by a Gaussian distribution with $\sigma^2 = 1/180$. In order to speed up the minimization, we add another degree of freedom, ϵ_0, which can be inserted back into a 20×9 energy matrix without changes in the outcome of the minimisation. The results of the linear optimization are used as starting point to optimize the global-epistasis model. For this reason, we use the same training partitions, in order to avoid biased results. The minimised global model is of the form $\Delta G = k + aL' + bL'^2$, where 2 degrees of freedom are added in order to speed up the computation. Via the transformation for $L = k + aL'$, we can express ΔG as $L + \lambda L^2$, where λ has units of $1/k_bT$. With this transformation, we bring the essential parameters to 181. The linear and epistatic energy matrices trained on the full datasets of all alleles studied are given in Table S2–S8.

Complexity estimates. For the analysis of Fig. 2, we divide linear predictions into bins of the linear trait L. For each bin, we fit the data ΔG to an average trait $L'_{\text{ave}} = aL + b$ and to subset-specific forms $L'_p = a_pL + b_p$ for specific combinations of alleles at anchor positions. The difference between the corresponding errors, $\Delta\text{MSE} = \text{MSE}_{\text{ave}} - \text{MSE}_p$, serves as an estimate for the added value of models with idiosyncratic epistasis, as given by Eq. (2).

Simulations of random energy models. The idiosyncratic model used for simulations incorporates 4 pairwise interactions among position 2 and 6, 2 and 9, 4 and 6, 4 and 9. We draw randomly the energy and interaction matrices from a Gaussian distribution centered at 0. The distribution of experimental data exhibits a total variance $\sigma^2 = \sigma_m^2 + \sigma_\epsilon^2 + \sigma_\beta^2 = 8.7$. From data relative to HLA-DRB1*01:01, we estimate $\sigma_m^2 = 1.5$ as a measure of the experimental noise. We can then choose σ_β^2 and σ_ϵ^2 by fixing their ratio, defined as $\tilde{\lambda}$. We simulate 3 different regimes of epistatic strength: strong ($\tilde{\lambda} = 2$), moderate ($\tilde{\lambda} = 1$) and weak ($\tilde{\lambda} = 0.3$).

In the simulations of the global-epistasis model, each column of the energy matrix is inferred from a Gaussian distribution with a zero mean and a standard deviation corresponding to the respective column of the linear energy matrix (Table S2). We consider 3 different regimes of epistatic strength: weak ($\lambda = -0.01$), moderate ($\lambda = -0.04$) and strong ($\lambda = -0.1$). The systematic noise, denoted as σ_m, is introduced from a Gaussian distribution with a variance of 1, computed from data for HLA-A*02:01.

For both classes, we perform 10 separate realizations for each $\tilde{\lambda}$ and λ, characterized by different random sets of training sequences and initial conditions for ϵ. A linear model is optimised and tested on random test sets. The partial regressions, performed on similar L ranges, are used to estimate the complexity of each landscape.

Complexity estimates. For estimating the complexity of the given MHC class, we compute ΔMSE in bins of L with size 2, spanning L values from -11 to 15 for both MHC classes (other bin sizes give broadly consistent results). In each bin, we evaluate $\Delta\text{MSE} = \text{MSE}_{\text{ave}} - \text{MSE}_p$ over a set of p subsamples with specific allele combinations, including only subsamples with more than 2 data points.

The ΔMSE values from simulations are derived from randomly generated sets of 20,000 peptides. The values reported in the main text are obtained by averaging over 10 realizations, and the standard deviation is computed across bins over these realizations. The range of L spanned in the simulated data is determined based on the minimum and maximum values observed in the specific realization under consideration.

Classification score. The classification score is computed on set of positive and negative binders, sourced from IEDB on the March 4th 2024 for all the considered alleles. After curations, the validation sets consist of 2488 positive and 1497 negative binders for HLA*A-02:01, 391 positive and 184 negative binders for HLA*A-11:01 and 382 positive and 505 negative binders for HLA*B-07:02.

We use the curated sets to construct the Receiver Operating Characteristic (ROC) curves, illustrating the relationship between *sensitivity*, representing the ratio of true positives to the total number of positives, and $1-$*specificity*, which denotes the ratio of false positives to the total negatives [17, 50]. We define positives and negatives based on the binding probability defined below:

$$p = \frac{1}{1 + e^{\Delta G}}. \tag{5}$$

We calculate these metrics by varying the thresholds that distinguish positives from negatives and computing the corresponding variables. The area under this curve provides the AUC score, which is presented for the relevant alleles in Table S1.

Model comparison. We use the BIC score [51],

$$\text{BIC} = n \times \text{MSE} + k \log(n), \tag{6}$$

where n is the number of sequences in the test set and k the number of parameters, respectively. For NetMHC 4.0, we compute $k = k_{\text{h}} + k_{\text{out}}$, in which $k_h = (n_{\text{in}} + 1) \times n_h$, where n_{in} and n_h are the number of inputs and hidden neurons. Similarly, $k_{\text{out}} = (n_h + 1) \times n_{\text{out}}$, where n_{out} is the number of output neurons [52]. With the architecture information provided in reference [21], $n_{\text{in}} = 180$, $n_h = 5$ and $n_{\text{out}} = 1$, we obtain k is 911. The total BIC is reported in Fig. 3B.

Mutational effect spectra. The mutational effect in Fig. 4 is computed collecting all pairs of single mutants available in data. Specifically, we collect 156 pairs in the strong binders regime and 119 for typical binders regime for HLA-A*02:01. For HLA-DRB1*01:01, we collect 214 and 230 sequence pairs for strong and typical binders regimes, respectively. For each energy window, both quantitative and qualitative measurements are collected. For mutations with qualitative data, $\Delta G_{\text{wt}} > \Delta G_{\text{thr}}$, we define $\Delta\Delta G = 0$ if $\Delta G_{\text{mut}} > \Delta G_{\text{thr}}$ and $\Delta\Delta G = \Delta G_{\text{mut}} - \Delta G_{\text{thr}}$ otherwise; mutations with a qualitative ΔG_{mut} are treated in an analogous way. The spectra of background mutations in Fig. S5 are computed based on 50 single mutants presenting the same mutation at the chosen position. The ancestral peptides are selected in the same energy window for strong and typical binders, as above. For both the idiosyncratic model and NetMHCII 2.3, positions 4 and 6 are chosen as main anchors.

Data and Code availability. All curated data, raw data sources, and materials used in this study are available in Supplementary Tables, at: https://github.com/lcollesano/MHC-gem and in the IEDB Analysis Resources at http://tools.iedb.org/mhci/download.

References

1. Bosselut R. T cell antigen recognition: Evolution-driven affinities. Proceedings of the National Academy of Sciences 2019; 116:21969–71
2. Batalia MA and Collins EJ. Peptide binding by class I and class II MHC molecules. Peptide Science 1997; 43:281–302
3. Yewdell JW, Reits E, and Neefjes J. Making sense of mass destruction: quantitating MHC class I antigen presentation. Nature Reviews Immunology 2003; 3:952–61
4. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, Sette A, Peters B, and Nielsen M. Automated benchmarking of peptide-MHC class I binding predictions. Bioinformatics 2015; 31:2174–81
5. Yewdell JW and Bennink JR. Immunodominance in major histocompatibility complex class I–restricted T lymphocyte responses. Annual review of immunology 1999; 17:51–88
6. Sahin U and Türeci Ö. Personalized vaccines for cancer immunotherapy. Science 2018; 359:1355–60
7. Blass E and Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nature reviews Clinical oncology 2021; 18:215–29
8. Lin MJ, Svensson-Arvelund J, Lubitz GS, Marabelle A, Melero I, Brown BD, and Brody JD. Cancer vaccines: the next immunotherapy frontier. Nature cancer 2022; 3:911–26
9. Madden DR. The three-dimensional structure of peptide-MHC complexes. Annual review of immunology 1995; 13:587–622
10. Pamer E and Cresswell P. Mechanisms of MHC class I–restricted antigen processing. Annual review of immunology 1998; 16:323–58
11. Castellino F, Zhong G, and Germain RN. Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. Human immunology 1997; 54:159–69
12. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, and Wiley DC. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 1993; 364:33–9
13. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, and Nielsen M. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 2018; 154:394–406
14. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, and Sette A. Prominent role of secondary anchor residues in peptide binding to HLA-A2. 1 molecules. Cell 1993; 74:929–37
15. Falk K, Rötzschke O, Stevanović S, Jung G, and Rammensee HG. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 1991; 351:290–6
16. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, and Wiley DC. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 1994; 368:215–21
17. Peters B, Tong W, Sidney J, Sette A, and Weng Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. Bioinformatics 2003; 19:1765–72
18. Nielsen M, Lund O, Buus S, and Lundegaard C. MHC class II epitope predictive algorithms. Immunology 2010; 130:319–28
19. Doytchinova IA, Blythe MJ, and Flower DR. Additive method for the prediction of protein- peptide binding affinity. Application to the MHC class I molecule HLA-A* 0201. Journal of proteome research 2002; 1:263–72
20. Hattotuwagama CK, Guan P, Doytchinova IA, Zygouri C, and Flower DR. Quantitative online prediction of peptide binding to the major histocompatibility complex. Journal of Molecular Graphics and Modelling 2004; 22:195–207
21. Andreatta M and Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics 2016; 32:511–7
22. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, and Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 2009; 61:1–13
23. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, and Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. Nucleic acids research 2008; 36:W509–W512

24. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. PloS one 2007; 2:e796

25. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, and Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. Cell systems 2018; 7:129–32

26. Nielsen M and Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome medicine 2016; 8:1–9

27. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, and Lund O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS computational biology 2008; 4:e1000107

28. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Annals of translational medicine 2018; 6

29. Johnson MS, Martsul A, Kryazhimskiy S, and Desai MM. Higher-fitness yeast genotypes are less robust to deleterious mutations. Science 2019; 366:490–3

30. Chou HH, Chiu HC, Delaney NF, Segrè D, and Marx CJ. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 2011; 332:1190–2

31. Nourmohammad A, Held T, and Lässig M. Universality and predictability in molecular quantitative genetics. Current opinion in genetics & development 2013; 23:684–93

32. Malmberg RL, Held S, Waits A, and Mauricio R. Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse. Genetics 2005; 171:2013–27

33. Moulana A, Dupic T, Phillips AM, Chang J, Nieves S, Roffler AA, Greaney AJ, Starr TN, Bloom JD, and Desai MM. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA. 1. Nature Communications 2022; 13:7011

34. Moulana A, Dupic T, Phillips AM, Chang J, Roffler AA, Greaney AJ, Starr TN, Bloom JD, and Desai MM. The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA. 1 evolution. Elife 2023; 12:e83442

35. Johnson MS, Reddy G, and Desai MM. Epistasis and evolution: recent advances and an outlook for prediction. BMC biology 2023; 21:120

36. Lyons DM, Zou Z, Xu H, and Zhang J. Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. Nature Ecology & Evolution 2020; 4:1685–93

37. Reddy G and Desai MM. Global epistasis emerges from a generic model of a complex trait. Elife 2021; 10:e64740

38. Bakerlee CW, Nguyen Ba AN, Shulgina Y, Rojas Echenique JI, and Desai MM. Idiosyncratic epistasis leads to global fitness–correlated trends. Science 2022; 376:630–5

39. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, and Peters B. The immune epitope database (IEDB): 2018 update. Nucleic acids research 2019; 47:D339–D343

40. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, and Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC bioinformatics 2014; 15:1–9

41. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief C, Oseroff C, Yuan L, Ruppert J, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. The Journal of Immunology 1994; 153:5586–92

42. Zhao W and Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLoS computational biology 2018; 14:e1006457

43. Gfeller D and Bassani-Sternberg M. Predicting antigen presentation—what could we learn from a million peptides? Frontiers in immunology 2018; 9:1716

44. Mustonen V, Kinney J, Callan Jr CG, and Lässig M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences 2008; 105:12376–81

45. Arnold PY, La Gruta NL, Miller T, Vignali KM, Adams PS, Woodland DL, and Vignali DA. The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. The Journal of Immunology 2002; 169:739–49

46. Holland CJ, Cole DK, and Godkin A. Re-directing CD4+ T cell responses with the flanking residues of MHC class II-bound peptides: the core is not enough. Frontiers in immunology 2013; 4:172

47. Tareen A and Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics 2020; 36:2272–4

48. Nourmohammad A, Schiffels S, and Lässig M. Evolution of molecular phenotypes under stabilizing selection. Journal of Statistical Mechanics: Theory and Experiment 2013; 2013:P01012

49. Otwinowski J, McCandlish DM, and Plotkin JB. Inferring the shape of global epistasis. Proceedings of the National Academy of Sciences 2018; 115:E7550–E7558

50. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 1997; 30:1145–59

51. Kass RE and Raftery AE. Bayes factors. Journal of the american statistical association 1995; 90:773–95

52. Goodfellow I, Bengio Y, and Courville A. Deep learning. MIT press, 2016

Supplementary Figures

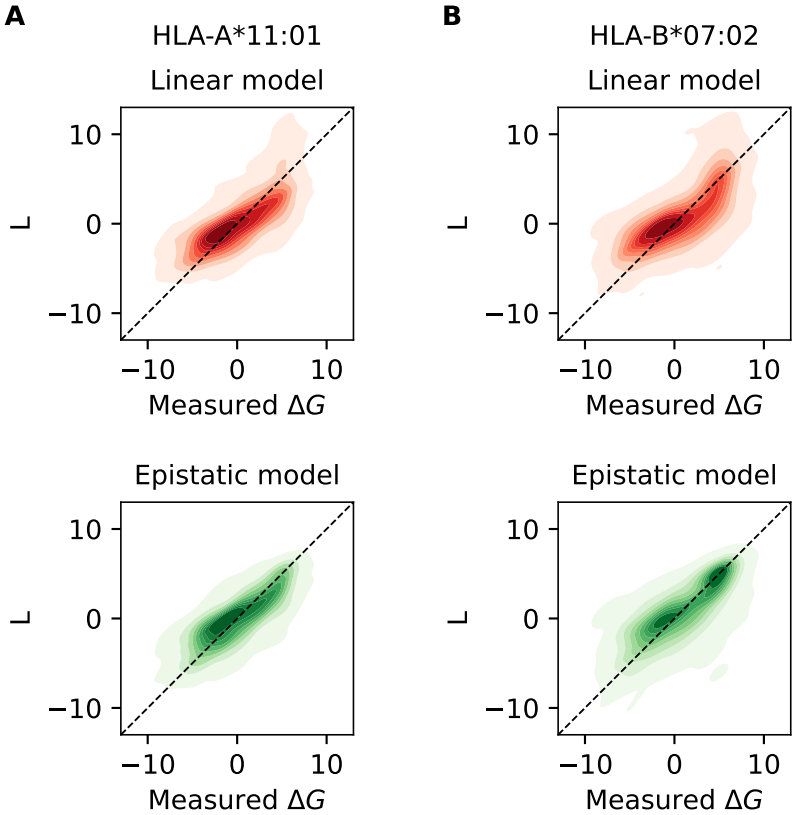

Figure S1. Error plots for HLA-A*11:01 and HLA-B*07:02. Error plot compairs ΔG from validation data and predictions of the linear model (in red), L (dashed line, left). On the right (in green), validation data for ΔG are compared to the global epistasis model (for both alleles, $\lambda = -0.04$). The error is computed across 10 cross-validation test sets (contours give densities above 0.01). A) HLA-A*11:01 (Linear model MSE = 5.0, Global model MSE = 4.4) B) HLA-B*07:02 (Linear model MSE = 6.3, Global model MSE = 5.6).

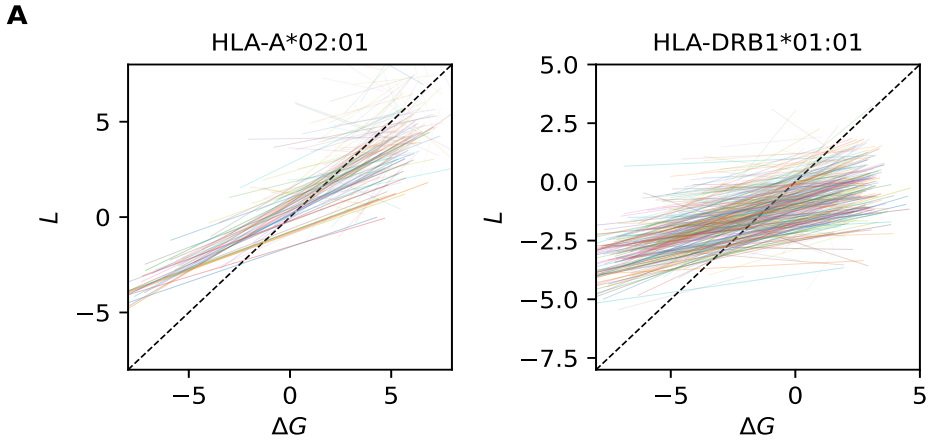

Figure S2. Inference of epistasis on full datasets. Linear regressions summarising errors of L with respect to ΔG, for each subsample available for HLA-A*02:01 (194 linear fits, on the left) and HLA-DRB1*01:01 (351 linear fits, on the right right). Each line is color-coded depending on the start of the L region spanned by the given subsample. The transparency of each regression line is proportional to the number of datapoints in the given subsample.

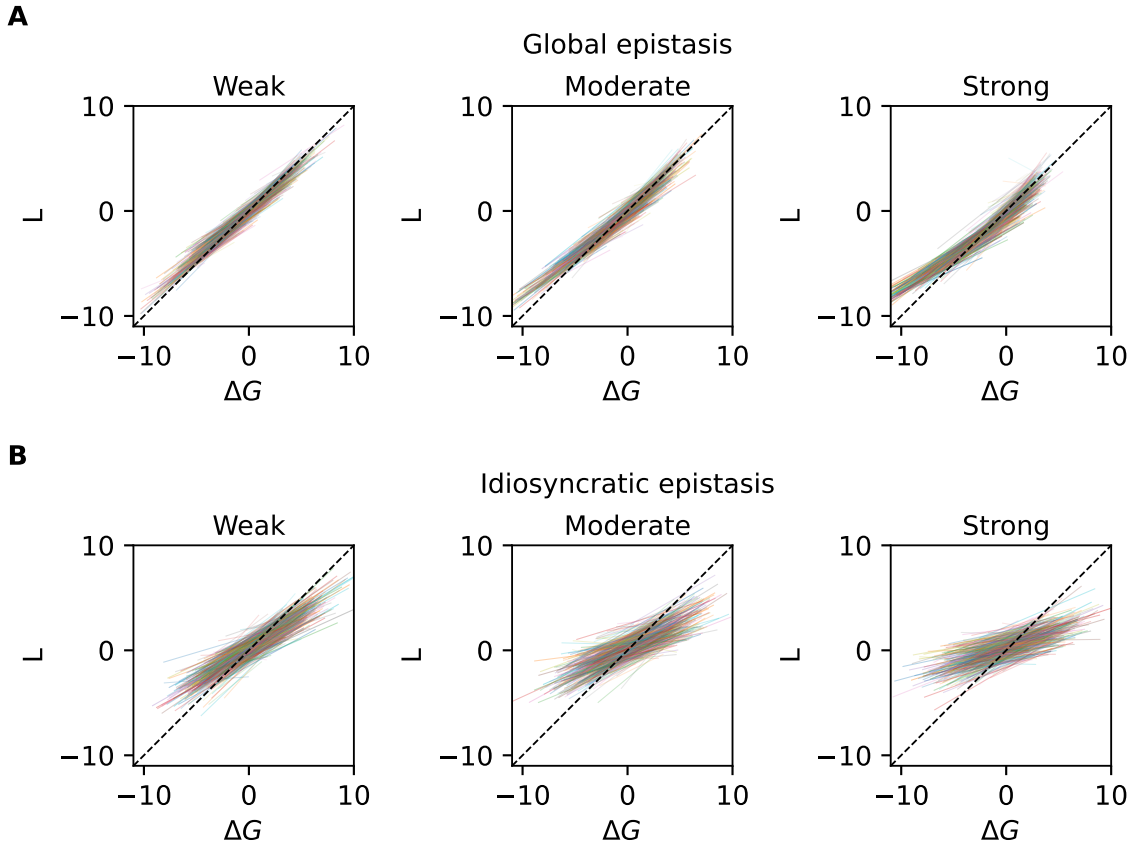

Figure S3. Regression lines and measure of complexity in simulation. Linear regression for subsamples of simulated data (one realization) according to the global epistasis model and the idiosyncratic epistasis model. A) From left to right, Weak global epistasis with $\lambda = -0.01$, moderate with $\lambda = -0.04$ and strong $\lambda = -0.1$. The lines are color coded based on the minimum of the L spanned by the given subsample. B) From left to right, Weak idiosyncratic epistasis with $\tilde{\lambda} = 0.3$, moderate with $\tilde{\lambda} = 1$ and strong $\tilde{\lambda} = 2$. The transparency of each regression line is proportional to the number of datapoints in the given subsample.

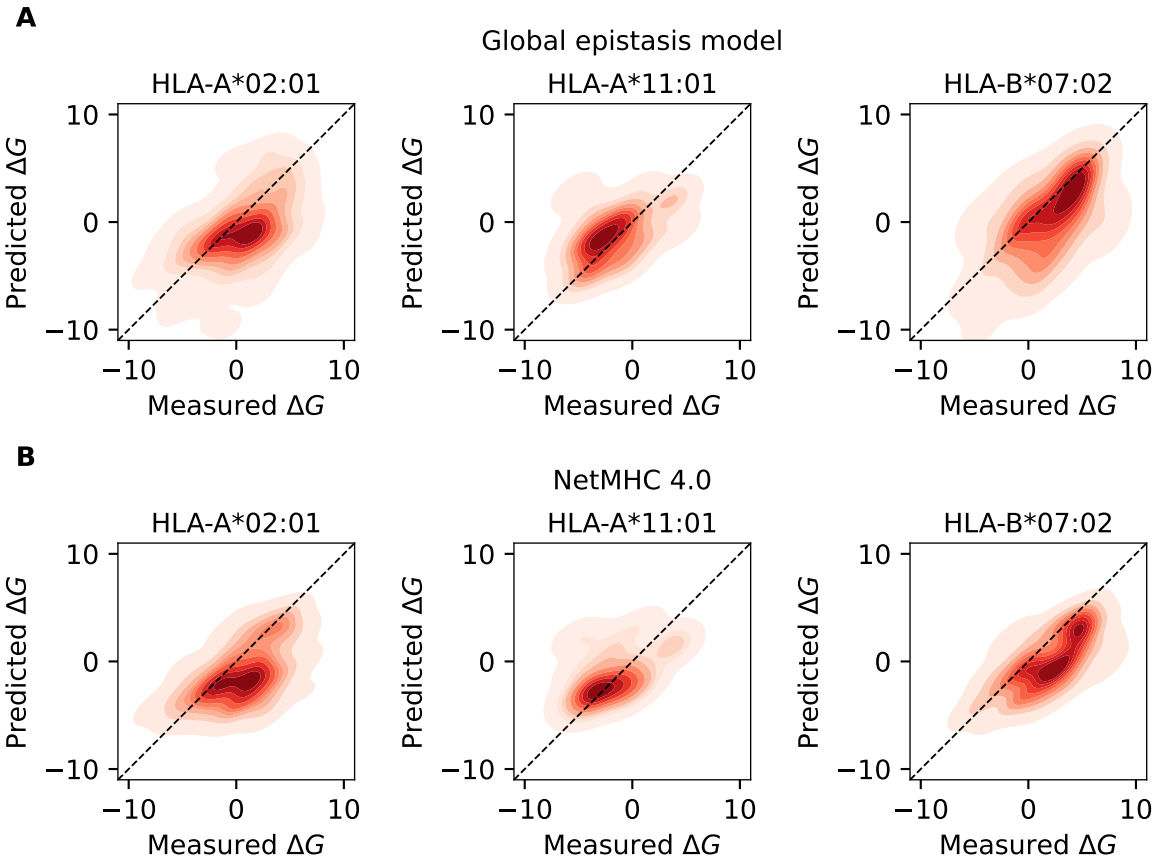

Figure S4. Blind set validation for MHC-I alleles. Error plot compairs observed energies contained in the blind set, ΔG, and predictions of the global epistasis model with $\lambda = -0.04$ for all the tested alleles (A, on top) and NetMHC 4.0 (B, bottom). Contours give densities above 0.01. The blind set contains 465 sequences for HLA-A*02:01, 69 for HLA-A*11:01 and 183 for HLA-B*07:02.

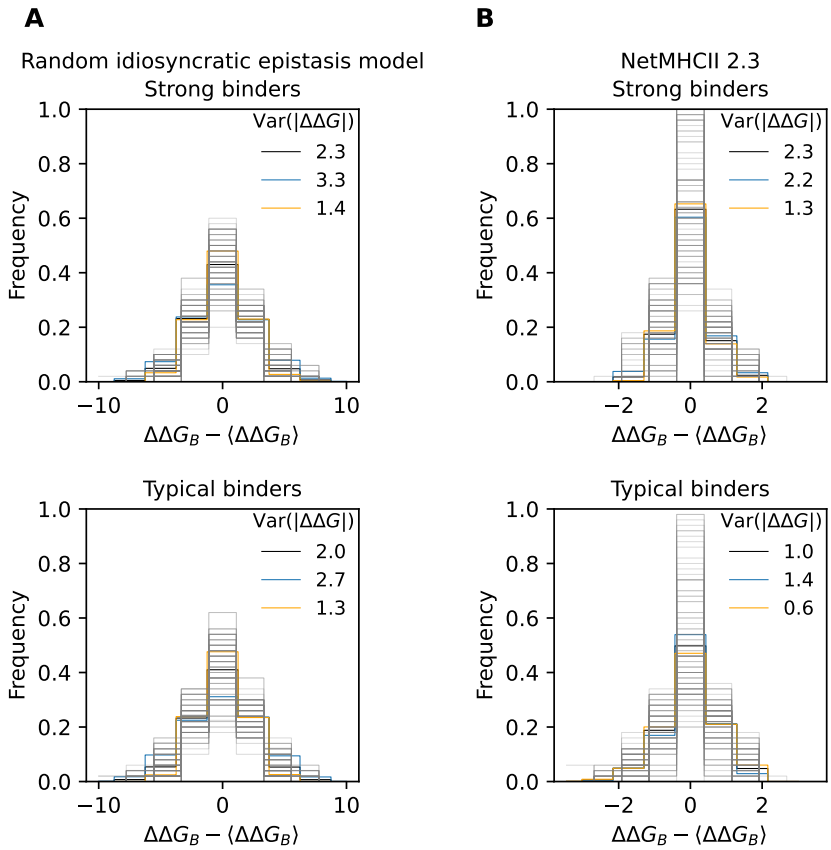

Figure S5. Idiosyncratic deviations from global epistasis. Background mutation spectra, shifted to mean 0, for samples of mutations in allele HLA-DRB1*01:01 for strong binders ($\Delta G \approx -5$, top) and typical binders ($\Delta G \approx -1$, bottom). (A) Spectra computed from the random idiosyncratic model with $\tilde{\lambda} = 2$; (B) spectra computed with NetMHCII-2.3.

Supplementary Tables

	Global epistasis model	**NetMHC 4.0**
HLA-A*02:01	0.837	0.843
HLA-A*11:01	0.952	0.960
HLA-B*07:02	0.929	0.947

Table S1. AUC scores for classification comparison between the global epistasis model and NetMHC 4.0.

AA	1	2	3	4	5	6	7	8	9
R	0.512	3.5876	2.0418	1.3152	1.3159	1.5414	1.7443	0.7501	1.7317
H	1.0834	1.5191	1.4644	0.9598	0.4466	1.2164	0.6584	0.8577	1.9219
K	-0.0303	2.5524	2.2813	0.9938	1.1933	1.8261	2.2849	1.0825	2.0481
D	2.4512	1.9741	0.4724	-0.0515	0.6914	0.8638	1.3873	1.564	1.546
E	2.387	2.2986	1.8864	0.103	1.0557	0.9791	1.1859	0.1864	2.6628
S	0.514	1.0906	0.5856	0.4608	0.9619	0.5208	1.0832	0.4254	0.848
T	1.0105	-0.3311	1.1822	0.6488	0.8729	0.2117	0.5497	0.5928	-0.3231
N	1.2892	2.654	0.9721	0.7493	1.1727	0.4819	1.1439	1.0146	1.516
Q	1.2178	-0.5661	1.0273	1.0448	1.0916	0.7374	1.1037	1.1354	2.1441
C	0.9998	1.8242	0.6929	0.4225	0.2378	0.2871	0.7131	0.6704	0.8281
G	0.8303	0.9945	1.3567	0.4986	1.0557	1.6548	1.8598	0.5328	1.6046
P	2.6186	1.8143	1.0735	0.7375	1.5483	0.579	0.2024	0.0355	1.5665
A	0.3152	-0.0293	0.1495	0.3792	0.9334	1.0265	0.2665	0.49	-1.4211
V	0.4312	-0.8632	0.843	0.9922	0.3784	-0.1686	0.3294	1.1262	-2.9387
I	0.1985	-1.3071	-0.0147	0.9023	0.1983	-0.3386	0.176	0.8522	-1.9094
L	0.2364	-2.9056	-0.2345	0.9099	0.4863	-0.139	0.4006	0.2916	-1.9346
M	-0.4881	-2.9776	-1.0266	0.6223	0.5011	-0.1315	0.0619	1.0294	-0.9927
F	-1.37	0.6495	-0.3426	0.7041	-0.2892	-0.0355	-0.9511	-0.2345	1.2408
Y	-1.2055	0.5	-0.2012	0.5234	-0.4121	0.4477	-0.3922	-0.5803	1.7167
W	0.4283	0.4647	-0.9197	-0.1689	-0.3713	0.7016	-1.1509	-0.0903	1.058

Table S2. Position weight matrix for the linear model, $L(\mathbf{a})$, in HLA-A*02:01

AA	1	2	3	4	5	6	7	8	9
R	0.6948	5.7649	2.9966	1.6864	1.7039	2.1967	2.3107	0.9852	2.5068
H	1.347	2.6977	1.7968	1.2181	0.7187	1.4887	0.7981	1.0581	3.0408
K	0.0329	4.0471	3.0798	1.3289	1.5468	2.2972	3.0024	1.4319	3.2513
D	3.7467	3.5361	0.6945	0.0873	1.0574	1.1151	1.7452	2.0478	2.8989
E	3.2307	3.8652	2.46	0.2342	1.3046	1.3643	1.4697	0.3017	4.4669
S	0.6612	0.9003	0.8059	0.6533	1.2387	0.7936	1.3592	0.6513	0.3907
T	1.1015	-1.1118	1.3843	0.8833	1.2304	0.3914	0.768	0.9199	-1.0131
N	1.5863	3.096	1.2038	0.9546	1.505	0.6942	1.4377	1.1949	2.527
Q	1.4733	-1.3309	1.2226	1.4264	1.3865	0.9565	1.3062	1.4804	3.5803
C	1.1548	2.0952	1.0697	0.6653	0.5617	0.5787	0.8789	1.108	1.2379
G	0.993	1.7958	1.7089	0.7366	1.2664	2.116	2.5162	0.7942	2.3022
P	3.9642	3.2104	1.2747	1.0104	1.9541	0.9162	0.2928	0.1898	2.513
A	0.4574	-0.5237	0.3431	0.6537	1.2289	1.3085	0.5311	0.7288	-2.4389
V	0.6859	-1.6056	1.0941	1.4934	0.7604	0.1097	0.6947	1.5987	-3.9536
I	0.3815	-2.1345	0.1784	1.324	0.4914	-0.0999	0.4333	1.2317	-2.9595
L	0.4211	-3.6896	-0.051	1.4095	0.8615	0.11	0.7199	0.5989	-2.9368
M	-0.2227	-3.8818	-0.7079	1.1653	0.7895	0.0113	0.3366	1.2769	-1.9197
F	-1.1973	1.2164	-0.133	1.1167	-0.0247	0.3423	-0.6264	0.0606	1.4403
Y	-0.9992	0.6149	0.059	0.787	-0.0446	0.5538	-0.1357	-0.3014	2.6981
W	0.599	0.8037	-0.5831	0.23	0.0218	1.0761	-0.9123	0.1489	1.6878

Table S3. Position weight matrix for the global-epistasis model, with $\lambda = -\mathbf{0.04}$ in HLA-A*02:01

AA	1	2	3	4	5	6	7	8	9
R	0.0153	1.7718	2.1049	0.8375	0.4371	0.5981	1.4507	0.7516	-2.0733
H	0.1885	1.8994	0.5781	0.8111	0.2803	0.524	0.2833	1.343	-0.1594
K	0.1579	4.5616	1.666	0.885	1.1914	1.1609	2.1172	1.0039	-4.2919
D	2.7862	0.9144	0.8687	0.5055	1.2762	1.4024	1.0695	1.6338	-0.036
E	1.9253	2.8511	1.7568	1.2569	1.1758	1.5968	1.58	0.6914	0.3566
S	-0.8258	-1.5626	-0.0331	0.5609	0.3246	0.5745	0.5687	0.5178	0.9822
T	-0.416	-1.888	0.8163	0.7525	0.9647	0.282	0.5931	1.0096	1.7411
N	1.4124	2.0531	0.138	0.2826	0.9865	0.6627	0.8594	0.5874	1.7855
Q	0.4105	-0.4249	0.5747	0.8501	1.2233	0.7863	1.6558	0.7178	0.7881
C	0.7005	2.7486	0.4907	0.6457	0.4815	0.8813	0.7227	0.7903	0.4577
G	-0.097	1.3111	1.0918	0.8375	0.9387	1.1583	0.7905	1.0561	1.394
P	2.1405	0.8025	1.8812	0.8124	0.4573	0.9767	0.5242	0.3412	2.4907
A	-0.3549	-0.6372	0.3824	0.3176	0.8024	0.7349	0.6172	1.0093	1.3196
V	0.2709	-1.7777	1.1124	0.6791	-0.2214	0.3898	0.2413	0.9064	1.775
I	0.0829	-1.2795	0.2442	0.7569	0.3684	0.343	0.0442	1.429	1.6631
L	0.7031	-0.5922	0.4031	0.8998	0.6466	0.1325	0.0688	-0.0007	1.5496
M	-0.4238	-1.4057	-1.0453	0.2469	0.2943	-0.1592	-0.3699	0.2316	1.0866
F	0.9879	0.6541	-0.9369	0.2923	-0.1807	-0.5135	-0.591	-0.3434	1.7103
Y	0.7089	0.61	-0.4734	0.7507	0.3129	-0.1526	-0.3359	-0.5313	-2.1077
W	2.022	0.9195	0.3607	-0.3579	0.8358	1.0073	-0.2337	0.3984	2.2169

Table S4. Position weight matrix for the linear model, $L(\mathbf{a})$, in HLA-A*11:01

AA	1	2	3	4	5	6	7	8	9
R	0.2873	3.2905	3.1277	1.1578	0.8721	0.9283	2.308	1.2854	-3.9503
H	0.736	2.49	1.1459	1.1056	0.7273	1.184	0.5747	1.7339	-1.9004
K	0.3977	5.9311	2.1508	1.4872	1.6992	1.6742	2.5515	1.4454	-6.0666
D	3.6218	4.1692	1.2906	0.9911	1.7081	1.9373	1.5904	2.1274	1.0745
E	2.3462	6.4217	2.3883	1.5685	1.6127	1.9821	2.1174	1.2594	0.0597
S	-0.4569	-2.3357	0.302	0.9787	0.7329	0.9394	0.9377	0.8009	1.3512
T	-0.0706	-2.6369	1.343	1.2048	1.5747	0.5532	0.9708	1.3924	2.7654
N	1.85	2.0979	0.3995	0.6201	1.4108	1.1044	1.1563	0.8982	3.6572
Q	1.0053	-1.2238	0.8692	1.1797	1.6206	1.2948	1.9602	1.3046	4.6549
C	1.1287	2.5418	0.7063	0.8111	1.0907	1.5058	1.1969	0.9194	0.4229
G	0.0437	0.7261	1.344	1.2386	1.5321	1.6961	1.4311	1.4743	3.517
P	2.9756	4.5926	2.3418	1.2231	0.8049	1.3444	0.5855	0.6107	5.1412
A	0.0455	-1.3713	0.771	0.7193	1.2624	1.0082	0.8316	1.5898	0.6628
V	0.5319	-2.5293	1.2856	1.11	0.1726	0.7268	0.6112	1.4575	1.0585
I	0.4261	-2.1318	0.5096	1.2545	0.8339	0.722	0.3174	1.9115	2.6248
L	1.0307	-1.4892	0.7577	1.2008	0.9009	0.6118	0.4054	0.4524	1.9005
M	-0.0558	-2.3259	-0.5882	0.8111	0.6491	0.1536	0.0123	0.8683	0.8978
F	1.3267	0.2164	-0.5871	0.7513	-0.081	-0.2457	-0.458	0.0117	3.1903
Y	1.0076	1.0119	-0.2327	1.1416	0.6199	0.1149	-0.0016	-0.1105	-3.9399
W	2.2248	1.4726	0.3675	0.2375	1.0012	1.1511	0.0363	0.9369	3.7153

Table S5. Position weight matrix for the global-epistasis model, $L(\mathbf{a})$, in HLA-A*11:01

AA	1	2	3	4	5	6	7	8	9
R	-0.7472	2.5149	-0.8649	0.0078	-0.1694	-0.4574	0.4109	0.6835	1.5848
H	-0.0497	0.3124	-0.1217	0.6072	0.8995	0.5873	0.8527	1.2805	1.3938
K	0.3489	1.2694	1.1996	0.1647	0.9191	0.6588	1.6138	1.8781	1.1285
D	3.686	0.8628	1.5075	1.941	0.3064	1.7863	1.4489	1.2118	1.4394
E	2.6568	3.7181	1.2493	1.7357	2.061	1.909	1.4145	0.56	1.9324
S	-0.006	1.3973	0.5139	0.7409	0.6369	0.9829	0.048	0.1481	3.3317
T	1.1719	0.2993	0.8472	1.4164	0.525	-0.3996	0.7006	0.1799	-0.6353
N	1.1593	1.0697	1.156	0.7284	1.1593	1.1428	0.4272	1.9377	3.3289
Q	1.4024	1.5726	1.4684	0.6193	1.6326	1.6055	0.6423	1.1121	2.287
C	0.9039	1.5807	2.7015	1.372	0.2252	1.8005	0.3046	0.8571	2.8135
G	1.1443	0.0088	0.6617	0.4316	0.7196	0.1153	0.7552	1.5764	3.3917
P	2.7886	-3.9686	2.2499	1.5345	0.9175	1.5952	1.4691	-0.7156	2.0513
A	-0.562	0.771	0.0026	1.2082	0.7292	0.6597	0.2852	-0.834	-0.4665
V	0.6407	0.0563	0.7974	1.117	0.8532	1.0375	0.8203	0.0074	-1.373
I	-0.3428	1.306	0.2625	1.0785	0.7321	1.3838	0.6627	0.7237	-1.2565
L	0.2591	0.7765	1.1604	1.0559	1.0027	1.1167	0.2378	0.3945	-2.52
M	0.0516	1.8698	-1.6157	1.0535	0.5755	-0.2553	0.4141	0.5741	-2.2916
F	0.1725	0.7346	1.9978	0.7511	1.4821	-0.2606	0.9648	0.9533	-1.2434
Y	0.2029	-1.3787	-0.3557	0.2385	0.2684	0.856	0.9424	1.4336	1.2086
W	0.4346	0.1854	0.1848	-1.1528	-0.0375	1.1518	0.9748	1.687	-1.0925

Table S6. Position weight matrix for the linear model, $L(\mathbf{a})$, in HLA-B*07:02

AA	1	2	3	4	5	6	7	8	9
R	-0.7438	5.4551	-1.1562	0.8095	-0.0919	-0.435	0.9935	1.4793	5.4362
H	0.2527	0.6591	-0.1146	1.2764	1.066	0.9586	0.7829	1.7441	1.4375
K	0.2349	1.8678	1.4734	0.4226	1.3669	0.9608	2.1084	2.5537	3.5764
D	4.842	3.9628	2.1711	2.7831	1.495	2.3494	2.3169	1.8041	4.2515
E	3.4884	6.0129	1.6032	2.6555	2.4262	2.6026	1.9036	0.4651	3.0646
S	0.669	3.0596	0.8295	1.1298	1.4948	1.2939	0.3293	0.2365	3.1488
T	1.4657	0.0683	1.5285	1.8739	0.9396	-0.2025	0.8447	0.6604	-1.1579
N	1.8755	1.0617	1.4065	0.9042	1.9044	2.1505	1.1159	2.1717	4.06
Q	2.8283	2.0341	1.9135	0.8596	2.3191	2.1911	0.924	2.0219	2.5208
C	1.5073	3.114	3.4583	1.6	0.5841	2.2619	0.6244	2.0076	4.6272
G	1.7048	1.1519	1.3181	1.4429	1.0681	0.452	0.7684	2.569	6.3681
P	3.5888	-5.9444	3.1471	1.9217	1.1578	2.1534	1.6914	-0.9774	3.0625
A	-0.2909	-0.52	0.5683	1.7801	1.2504	1.4701	0.6466	-0.8723	-1.3376
V	0.8419	-0.9495	1.3482	1.8483	1.482	1.5593	1.3742	0.4323	-2.3351
I	0.2157	1.7469	0.5241	1.6947	1.3467	1.9728	1.4428	0.8553	-2.4149
L	0.6921	0.2909	1.7651	1.1998	1.4562	1.9813	0.5064	1.551	-3.7431
M	0.2475	2.0183	-1.1838	1.2619	0.5027	0.6529	0.8527	0.7214	-3.5015
F	0.5093	1.3576	3.0037	1.2983	1.9662	0.2627	1.367	1.5037	-2.2904
Y	0.0705	-2.2589	0.0186	0.6481	1.1639	1.7181	1.4895	2.1098	1.4657
W	0.3938	-0.4662	0.1822	-0.5105	-0.2749	1.2355	2.45	1.9827	-2.415

Table S7. Position weight matrix for the global-epistasis model, with $\lambda = -\mathbf{0.04}$ in HLA-B*07:02

AA	1	2	3	4	5	6	7	8	9
R	0.9684	-0.0074	-0.0556	0.9759	-0.3467	1.1456	-0.3978	-0.1749	0.1091
H	0.2164	-0.0965	-0.0261	0.093	0.3911	0.2522	-0.4412	0.2263	0.1907
K	0.8845	-0.0057	0.2543	0.5794	-0.2852	1.0588	0.0217	-0.0162	0.3186
D	0.6316	0.8446	1.8352	1.1589	0.9278	0.932	0.8845	1.0643	0.9019
E	1.0245	0.7678	0.884	0.2729	0.5909	0.8066	0.4836	1.0849	0.8747
S	0.4655	0.6664	0.1651	-0.1095	-0.0932	-1.039	-0.3357	0.2574	-0.7554
T	0.8732	0.4818	0.2786	0.223	0.305	-0.2797	-0.2592	0.5081	-0.7262
N	0.5954	0.3452	0.7892	0.192	0.3909	-0.8373	-0.4327	0.6737	0.0697
Q	0.8875	0.1857	-0.003	-0.0367	-0.128	0.5547	-0.5076	0.6435	0.3701
C	-0.338	0.6082	0.6147	0.7931	0.701	0.6665	0.3448	0.4076	0.6968
G	0.7053	0.9131	0.3328	0.1494	0.5095	-0.9354	0.5128	0.4591	-0.2676
P	1.1664	1.9504	0.4273	1.085	-0.2327	-0.7068	0.0617	0.013	0.6297
A	1.0773	0.138	-0.1705	-0.6654	0.0106	-1.4323	-0.6304	-0.2045	-1.2025
V	-0.3327	-0.3592	0.0835	-0.3197	0.1411	-0.2737	-0.5926	0.2097	-1.5044
I	-0.9493	-0.6115	-0.1923	-0.7682	0.116	0.3758	-0.3975	0.3738	-1.1503
L	-0.8375	-0.3966	-0.444	-1.2106	-0.2272	0.134	-0.3368	-0.1929	-1.197
M	-0.4534	-0.1414	-0.5604	-1.578	0.0727	0.1957	-0.7342	-0.1129	-0.8431
F	-1.8811	0.1216	-0.1468	0.1863	0.3481	0.3403	-0.4217	0.4706	-0.1125
Y	-1.8356	-0.662	-0.4354	1.4381	0.2723	0.4374	-0.3967	0.2418	0.2501
W	-1.3229	0.2423	0.4556	1.0435	-0.3376	0.9612	-0.0294	1.2394	0.4372

Table S8. Position weight matrix for the linear model, $L(\mathbf{a})$, in HLA-DRB1*01:01