

How Different are Self and Nonself?

Andreas Tiffeau-Mayer^{1,2,*}, Jonathan A. Levine^{3,4,5,*}, Christopher J. Russo^{2,6}

Quentin Marcou^{7,8}, William Bialek^{2,9}, and Benjamin D. Greenbaum^{5,10,11,‡}

¹*Division of Infection and Immunity and Institute for the Physics of Living Systems, [University College London](#), London WC1E 6JF, United Kingdom*

²*Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, [Princeton University](#), Princeton, New Jersey 08544, USA*

³*Tri-Institutional Program in Computational Biology and Medicine, [Weill Cornell Medicine](#), New York, New York 10021, USA*

⁴*Laboratory of Lymphocyte Dynamics, [Rockefeller University](#), New York, New York 10065, USA*

⁵*The Halvorsen Center for Computational Oncology, Department of Epidemiology and Biostatistics, [Memorial Sloan Kettering Cancer Center](#), New York, New York 10065, USA*

⁶*Program in Biophysical Sciences, [The University of Chicago](#), Chicago, Illinois 60637, USA*

⁷*COMPUTational pharmacology and clinical Oncology, [Centre Inria d'Université Côte d'Azur](#), 06560 Valbonne, France*

⁸*Centre de Recherche en Cancérologie de Marseille (CRCM), Institut Paoli-Calmettes, Inserm UMR1068, CNRS UMR7258, [Aix Marseille University](#), UM105 Marseille, France*

⁹*Center for Studies in Physics and Biology, [Rockefeller University](#), New York, New York 10065, USA*

¹⁰*The Olayan Center for Cancer Vaccines, [Memorial Sloan Kettering Cancer Center](#), New York, New York 10065, USA*

¹¹*Physiology, Biophysics & Systems Biology, [Weill Cornell Medicine](#), Weill Cornell Medical College, New York, New York 10021, USA*



(Received 25 April 2025; accepted 21 January 2026; published 4 March 2026)

Biological and artificial networks routinely make reliable distinctions between similar inputs, and the rules for making these distinctions are learned. In some ways, self/nonself discrimination in the immune system is similar, being both reliable and (partly) learned through thymic selection. In contrast to other examples, we show that the distributions of self- and nonself-peptides are nearly identical but strongly inhomogeneous. Reliable discrimination is possible only because self-peptides are a particular finite sample drawn from this distribution, and T cells can target the “spaces” between these samples. In conventional learning problems, this would constitute memorization or overfitting and lead to disaster. Here, the strong inhomogeneities imply instead that the immune system gains by targeting peptides that are similar to self, with maximum sensitivity for sequences just one or two substitutions away. This model of the structure of the underlying distribution in sequence space predicts, for example, the observed ability of the immune system to respond to mutation-derived cancer neoantigens.

DOI: [10.1103/fbct-vzwm](https://doi.org/10.1103/fbct-vzwm)

I. INTRODUCTION

A basic task of the immune system is to distinguish self from nonself. More specifically, cytotoxic T cells should respond to relatively short peptides from foreign antigens and not overly respond to peptides synthesized by the organism itself. One might expect this task is easier when the differences between peptides are larger, and the targeted peptides are therefore less similar to self. Yet, recent work in cancer immunology has shown that cancer neoantigens that differ from self by just a single amino acid substitution can still generate strong immune responses [1]. More generally, T cells are

often capable of recognizing peptides that are close to self (for instance, see Ref. [2]). The strength of such immune responses can exert a selective pressure on tumors, to which they can adapt by activating immune checkpoints and other resistance mechanisms, thereby attenuating immune cell function [3]. Inhibiting checkpoint blockade proteins has been the basis of transformative immunotherapies [1,3–8].

T cells learn to distinguish between self and nonself, at least to some extent, through the process of thymic selection. Like other learning problems [9,10], this process requires some degree of generalization [2,11–14], because it is not possible for each candidate T cell to be tested against all possible self-peptides. In a usual learning problem, examples are drawn from a distribution, and it is essential to capture features of this underlying distribution and generalize beyond the particular random samples that one has seen. Here we show that for the self/nonself distinction question in peptides, the opposite is the case, and memorization is both allowed and necessary. Our approach predicts that peptides that are close to self are natural targets for the immune system, that is, they are well within the distribution of peptides T cells can recognize.

*These authors contributed equally to this work.

†Contact author: andreas.mayer@ucl.ac.uk

‡Contact author: greenbab@mskcc.org

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

To address these issues, we need a model for the relevant distributions. We use a maximum entropy approach to analyze the occurrence of amino acid motifs that characterize peptides across all domains of life. Maximum entropy models have origins in statistical physics [15] but have also been used in machine learning, particularly in natural language processing [16]. These “small language models” are ideal for capturing the distribution of short peptides presented by the class I major histocompatibility complex (MHC-I) in an interpretable fashion. As generative models, they easily generalize to large language models (LLMs) based on unsupervised generative transformer architectures [17] that are beginning to be applied to more complex computational proteomic tasks [18,19].

Using maximum entropy models, we mathematically characterize the statistical structure of the self/nonself discrimination problem facing cytotoxic T cells. This structure is defined by the distribution of sequences found in the approximately 9-mer length peptides presented to the immune system by MHC-I. The key idea is to build models that match specific features of the data exactly, but otherwise have as little structure as possible. This strategy has produced very accurate models for, e.g., the joint patterns of activity in networks of neurons [20–23], correlated variations of amino acids in protein families [24–28], and fluctuating flight velocities in flocks of birds [29]. Importantly, the maximum entropy method allows an incremental incorporation of constraints, resulting in a more precise description as more features of the data are added.

Here we show that, across eukaryotes, such language models are unable to distinguish between hosts and pathogens. In other words, host and pathogen proteomes are described by the same underlying language model, reflecting their amino acid statistics. This is true even when specific MHC haplotypes are considered, focusing only on the filtered subsets of peptides likely to be presented. We characterize this landscape and find that encountering a pathogen peptide that is only one mutation away from self, which is often the case for cancer neoantigens, is in no way an unusual problem for the immune system to solve when fighting pathogens. On the contrary, the immune recognition of such “close to self” peptides might be evolutionarily favorable.

II. RESULTS

A. Proteome biases reduce diversity of peptides beyond MHC presentation

The 20 amino acids are used with unequal probabilities [30] in vertebrate proteomes and the proteomes of large groups of infectious viruses, bacteria, or parasites (for details of the databases we use see Appendix A). We observe that the correlations between neighboring amino acids are weak, but they extend over a very long distance, far beyond the length of the relevant peptides (Appendix A). For k -mers with $k < 4$, the size of the human proteome allows for sampling of all possible sequences, and we can estimate the probability distribution with reasonable control over errors. However, there are too many possible 9-mers, and a statistical model is required to define the underlying distribution.

Mathematically, we choose a probability distribution of 9-mer sequences, σ , that maximizes the Shannon

entropy,

$$S[P(\sigma)] = - \sum_{\sigma} P(\sigma) \ln P(\sigma), \quad (1)$$

subject to the normalization constraint $\sum_{\sigma} P(\sigma) = 1$, and constraints that enforce the equality of modeled and empirical observables.

We begin by matching global statistical features of observed sequences in the human proteome: first the mean number of each type of amino acid that appears in the peptides (first moment), then the covariances in these numbers (second moment), and then the asymmetry of their probability distribution (third moment). Finally, we include the fact that correlations between pairs of amino acids depend on the distance between them within the 9-mer (two-point). Each of these additional constraints lead to model improvements because each contributes to lowering the entropy [Fig. 1(a)].

Formally, we can define $s_i^{\alpha} = 1$ if the amino acid at site i is of type α , and $s_i^{\alpha} = 0$ otherwise; $i = 1, 2, \dots, 9$ runs along the length of the peptide and $\alpha = 1, 2, \dots, 20$ over the amino acids. We will refer to the entire sequence as $\sigma = \{s_i^{\alpha}\}$.

It is useful to count the number of amino acids of each type that appear in the peptide, $n^{\alpha} = \sum_i s_i^{\alpha}$. With these definitions, the maximum entropy model described above takes the form

$$P(\sigma) = \frac{1}{Z} \exp[-E(\sigma)], \quad (2)$$

where the “energy,” E , depends on the variables s_i^{α} and n^{α} and their moments:

$$E(\sigma) = \sum_{\alpha=1}^{20} \lambda_1^{\alpha} n^{\alpha} + \sum_{\alpha, \beta=1}^{20} \lambda_2^{\alpha\beta} n^{\alpha} n^{\beta} + \sum_{\alpha, \beta, \gamma=1}^{20} \lambda_3^{\alpha\beta\gamma} n^{\alpha} n^{\beta} n^{\gamma} + \frac{1}{2} \sum_{i,j=1}^9 \sum_{\alpha, \beta=1}^{20} J_{\Delta}^{\alpha\beta} s_i^{\alpha} s_j^{\beta}. \quad (3)$$

All coefficients must be adjusted so that the predictions of the model match observed features of the data [31]: λ_1^{α} is fixed by matching the mean number of amino acids of type α , $\lambda_2^{\alpha\beta}$ by matching the covariance in numbers of amino acids of types α and β , $\lambda_3^{\alpha\beta\gamma}$ by matching the third moments of these numbers, and the matrix $J_{\Delta}^{\alpha\beta}$, where $\Delta = |i - j|$, must be adjusted to match the distance dependence of the pairwise correlations. The constant (partition function) Z serves to normalize the distribution, so that $Z = \sum_{\sigma} \exp[-E(\sigma)]$ and can be evaluated numerically by thermodynamic integration [32] (Appendix B). We emphasize that once we match these coefficients to measured features of the proteome, there are no free parameters that can be adjusted. To test the generalizability of the inferred models, we split the proteomes into equal sized train and test partitions at the protein level.

This family of models makes accurate predictions for higher-order statistical properties of the distribution of human 9-mers. As an example, we can compute the correlations among triplets of amino acids at particular sites:

$$C_{ijk}^{\alpha\beta\gamma} = \langle \delta s_i^{\alpha} \delta s_j^{\beta} \delta s_k^{\gamma} \rangle, \quad (4)$$

with $\delta s_i^{\alpha} = s_i^{\alpha} - \langle s_i^{\alpha} \rangle$. While small, the predicted correlations largely agree with those in the observed test data [Fig. 1(b)].

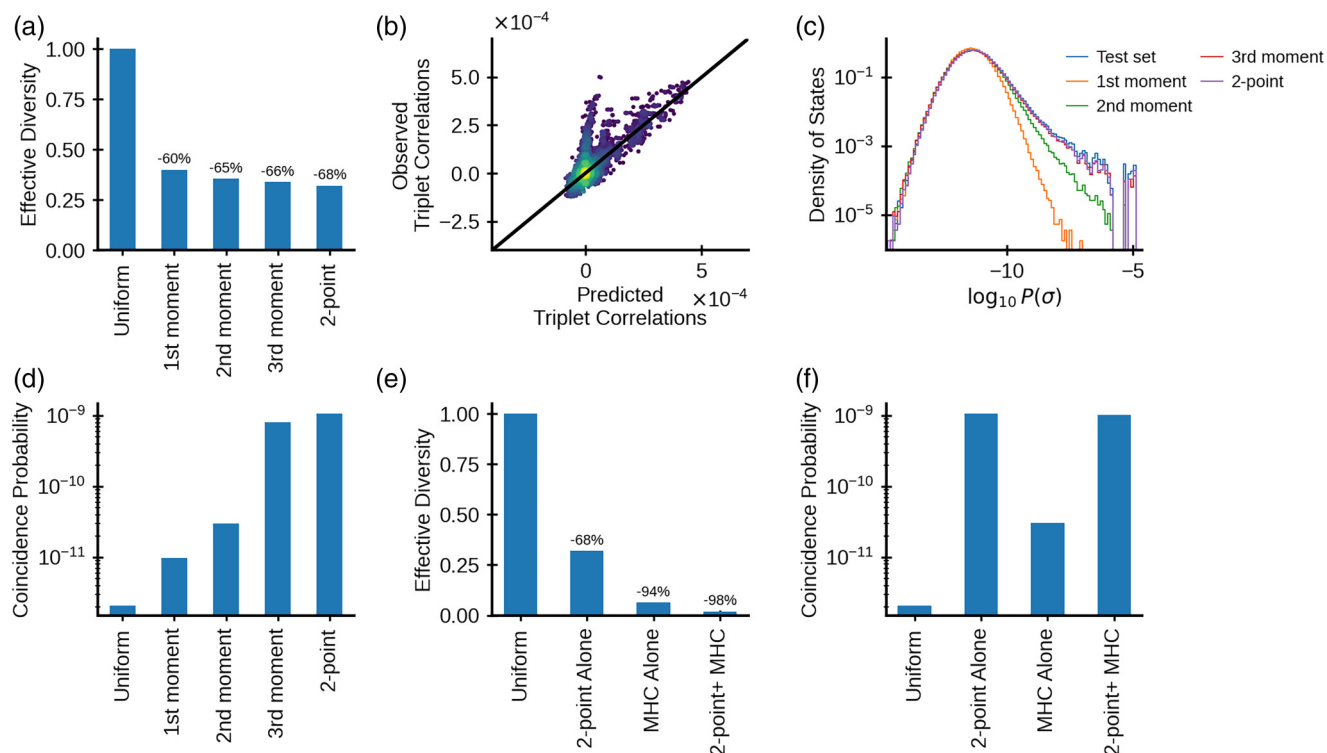


FIG. 1. Maximum entropy models predict peptide statistics. (a) Entropy decreases as we add more constraints on average statistics across sites (moment terms) and correlations between specific sites (two-point term) (see Appendix B). Each model is constructed by adding additional terms to the energy in Eq. (3): The first-moment model contains just the first term, the second-moment model contains the first two terms, the third-moment model contains the first three terms, and the two-point model contains all four constraint terms. (b) Predicted (two-point model) vs observed triplet correlations among amino acids at different sites. (c) Probability density of (log) probability in the full (two-point) model. We compare ensembles of sequences drawn from the series of models, and from the test set data. (d) Probability that two randomly chosen sequences are identical, computed across ensembles drawn from the series of models. (e) Entropy reduction from the full (two-point) model alone, filtering by MHC-I presentation alone, and a joint distribution of the maximum entropy model and filtering by MHC-I presentation together. (f) Probability that two randomly chosen sequences are identical, computed across ensembles drawn from the series of models in panel (e). Effective Diversity is calculated as $D = e^S/20^9$, where S is the Shannon entropy. D can be interpreted as the Shannon diversity (also known as perplexity) relative to that of a uniform distribution of 9-mers. Effective diversity and coincidence for “MHC alone” and “2-point + MHC” (e), (f) are calculated as the average across independent models for the top 500 most common haplotypes (Appendix C).

More subtly, the model assigns a probability of occurrence to every possible 9-mer, and we can calculate the distribution of this (log) probability [Fig. 1(c)]. As we constrain more features, this distribution grows a “heavy tail” of relatively high probability sequences that are also in real data [Fig. 1(c)]. To calibrate this effect, we note that there are nearly 10^{12} possible 9-mers, so the prediction that any sequence occurs with probability above 10^{-7} reflects a startling 10 000-fold enrichment (or more) over random sequences, and many of these predicted sequences are indeed found in real data. A corollary of this prediction is that two 9-mer sequences chosen at random are thousands of times more likely to be the same than if their individual amino acids were chosen at random [Fig. 1(d)]. Large effects on the coincidence probability coexist with relatively small effects on the entropy precisely because the distribution is strongly nonuniform, which has important implications for the immune response. A comparison of the density of states across the model family shows the importance of third-order moment constraints for the close match between model-sampled and empirical distributions and for the existence of this heavy upper tail (Appendix B).

To understand the potential biological importance of the evolutionary constraints uncovered by our approach, we compared their impact to the well-known constraints imposed by MHC-I presentation, which restricts peptides to those with specific amino acids and amino acid motifs that favor MHC binding, such as at anchor sites. Using a previously derived distribution of human HLA haplotypes which accounts for linkage [33], we created a combined model for each haplotype describing the filtered distribution of peptides in the proteome that are predicted to be presented by any allele in that haplotype while still obeying the maximum entropy model constraints. Using these filtered distributions for each of the top 500 most common haplotypes (Appendix C), we compared how the maximum entropy model and MHC-I restriction each contribute to reducing the effective diversity of peptides presented to the immune system [Fig. 1(e)]. On average, filtering by MHC-I presentation alone reduced the effective diversity by 94% relative to the uniform distribution, compared to 68% with the full two-point maximum entropy model. Combining the constraints captured by the maximum entropy model with the restriction of MHC-I presentation acted additively to reduce diversity by a total of 98%

relative to uniform, implying both models restrict nonredundant information.

Furthermore, we found that the existence of the “heavy tail” of highly probable peptides in our model increased the probability of coincidence by 10^3 compared to an increase by 10^1 with MHC-I presentation alone [Fig. 1(f)]. Additionally, the percentage of peptides that were classified as binders did not differ significantly between the family of models and did not vary significantly between haplotypes (Appendix C). Thus, while MHC-I presentation determines which peptides can be presented, our maximum entropy model captures independent information on the most common sequence motifs, which are more biologically relevant for comparing two peptides and therefore immune discrimination. Put another way, MHC presentation largely contributes to the reduction in the Shannon entropy of the resulting distribution [Fig. 1(e)], while our model highlights the reduction in Renyi entropy, as reflected in the coincidence statistics [Fig. 1(f)]. Therefore the biophysical constraints on peptide distributions are likely an important additional consideration in understanding the evolution of immune recognition. Indeed, we found that these constraints also have an impact on the observed overlap between self and nonself proteomes.

B. Comparisons between self and nonself proteomes

Having found a class of models that describes the distribution of human 9-mer sequences quantitatively, we can ask whether this distribution differs between humans and other organisms, or between humans and pathogens. We build maximum entropy models for 9-mers coming from a variety of proteomes and compare the resulting distributions. The natural measure of difference between distributions is the Kullback-Leibler divergence (D_{KL}) [34], which measures the average evidence (or log-likelihood ratio) that a single peptide belongs to one distribution rather than the other. For each proteome, we show the KL divergences relative to both the human peptide distribution and relative to a uniform distribution over all peptides [Fig. 2(a)].

The distributions of 9-mers in these proteomes are very different from the uniform distribution, but quite similar to one another and to the statistics of the human proteome [Fig. 2(a)]. As we add successive constraints, we find that not just amino acid frequencies, but also covariances and higher-order statistics are largely shared across proteomes. For instance, the proteomes of bacteria and their eukaryote hosts all differ more from the uniform distribution ($D_{KL} > 1.0$ bit), than from each other ($D_{KL} \lesssim 1.0$ bit), and the maximal $D_{KL} \approx 1.0$ bit still makes individual 9-mers only weakly predictive. The largest divergence is found for the parasite *Plasmodium falciparum*, an organism known to have an unusually AT-rich genome [35]. However, the *P. falciparum* proteome is still closer in statistical distance to the human proteome than it is to the uniform distribution. Most strikingly, if we compare a collection of human viral proteomes with the human proteome, we find $D_{KL} \approx 0.2$ bits, indicating that individual viral peptides are on average almost indistinguishable from self-peptides.

For an alternative visualization, we applied multidimensional scaling (MDS) to embed the symmetrized divergence matrix in two dimensions [Fig. 2(b), Appendix B]. In this

lower dimensional space, vertebrate and viral proteomes cluster close together, with bacterial proteomes only slightly farther away. Again, larger divergences are found for *P. falciparum* relative to the other proteomes, and all proteomes lie far from the uniform distribution.

To understand the biophysical mechanisms shaping the similarity in amino acid usage across host and pathogens, we compared them to those expected under neutral evolution, that we simulated by generating amino acid frequencies for various GC fractions (Appendix D). We found that codon usage explains some of the similarity between host and pathogen proteins (Appendix D), but also observed shared deviations in amino acid frequencies from neutral evolutionary expectations. For example, the human and viral amino acid frequencies correlate more closely with each other than either to the codon null model (Appendix D). Interestingly, viral proteins are closer to neutral expectations than human proteins, compatible with the hypothesis that higher viral mutation rates counteract nonneutral selective forces.

A consequence of the small D_{KL} between the human proteome and the proteome of the human virome is that an immune system trained on general features of these distributions can recognize individual peptides from viruses only if it is willing to accept a nearly comparable false positive rate, corresponding to autoimmune recognition, the ability to recognize self-peptides [Fig. 2(c)].

The trade-off captured by the receiver operating characteristic is independent of the overall probability with which self- or nonself-peptides are encountered. If, for example, there is a 10-fold excess of self-peptides, we can calculate the probability that any one peptide will be identified correctly (precision), but we can trade this against the fraction of nonself-peptides that will be recognized (recall). As shown by the precision-recall curve [Fig. 2(d)], even if such an immune system catches only one in a thousand nonself challenges (those which “stick out” as being most unlikely to occur in the self-distribution), it still reaches only $\sim 50\%$ chance of being correct about a random peptide. Moreover, while skewing MHC presentation to more easily discriminate peptides might be advantageous in principle, empirical analysis of the MHC filtered viral peptides shows that while their distributions are also skewed by presentation, the resulting filtered distributions remain highly similar between viral and self-peptides (Appendix C). In other words, both host and viral peptides are skewed in the same manner: the degree of similarity between the two distributions remains when considering only MHC-I presented peptides. Consequently, a similar classification analysis between self and viral peptides limited to those predicted to be presented on specific MHC-I haplotypes shows no performance gain (Appendix C). We conclude that the adaptive immune system does not distinguish between self and nonself by learning general features of the underlying distributions.

C. Density of nonself-peptides skews close to self

Due to the discrepancy between the $\sim 10^{12}$ possible 9-mers versus the $\sim 10^7$ realized 9-mers in the human proteome, the immune system can recognize pathogen sequences that fall into the spaces between the learned self-sequences even

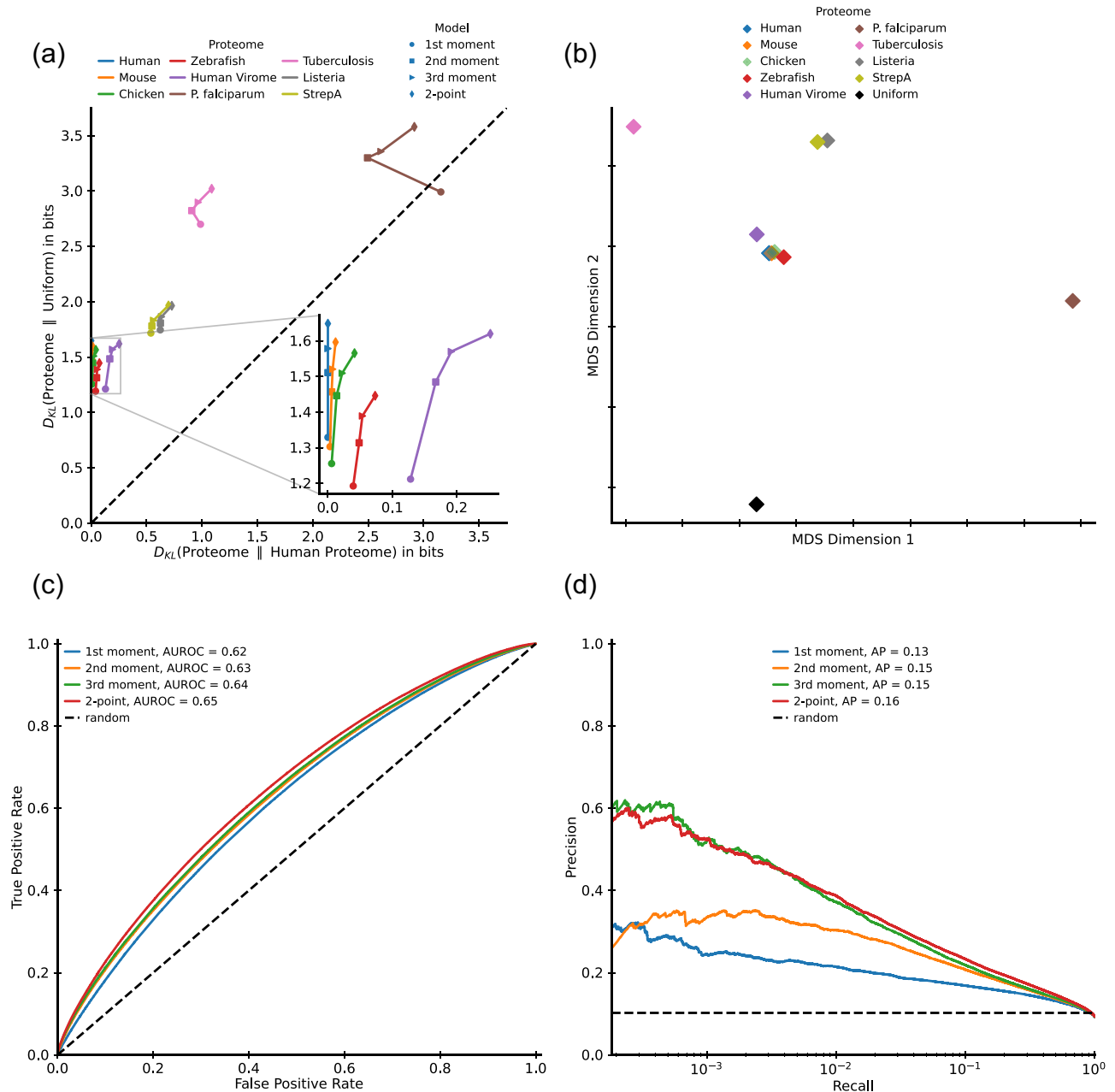


FIG. 2. Divergence between peptides from different proteomes. (a) Kullback-Leibler (KL) divergences between peptide distributions of different pathogen proteomes relative to human host peptides (x axis) and relative to a uniform distribution over all peptides (y axis). For each proteome we show the statistical distance calculated according to a nested set of models including a different number of constraints. The inset shows a zoom on the set of proteomes close to the human statistics. (b) Multidimensional scaling (MDS) dimensionality reduction of the matrix of pairwise, symmetrized (Jeffreys) divergences between proteomes based on the full two-point models. MDS arranges the points on the plot so that the distances among each pair of proteomes correlates as best as possible to the divergence between the peptide distributions. (c), (d) Performance of the models as classifiers, where we assign peptides as self or nonself (viral) based on the likelihood ratio $L = \ln[P_{\text{nonself}}(\sigma)/P_{\text{self}}(\sigma)]$. (c) If we choose a nonself viral peptide at random, then there is some probability that it will be identified correctly (true positive), and if we choose a self-peptide and random there is some probability it will be identified incorrectly as nonself (false positive). We can trade these probabilities against one another by changing the threshold L at which we make the decision. (d) If self-peptides are in 10-fold excess, then there is a probability that some random peptide will be classified correctly (precision), and this will “catch” a certain fraction of the nonself-peptides (recall). Note that if peptides are assigned self/nonself at random, then we will be right 10% of the time since self is in 10-fold excess (dashed line).

if they come from the same overall language model. But these gaps are tight: if sequences were random, typical self-peptides would be separated by only three amino acid substitutions. In fact, the distribution is very inhomoge-

neous, as discussed above, so most of the gaps are even smaller.

We can make this geometric picture explicit by asking, for each 9-mer in a pathogen proteome, what is the dis-

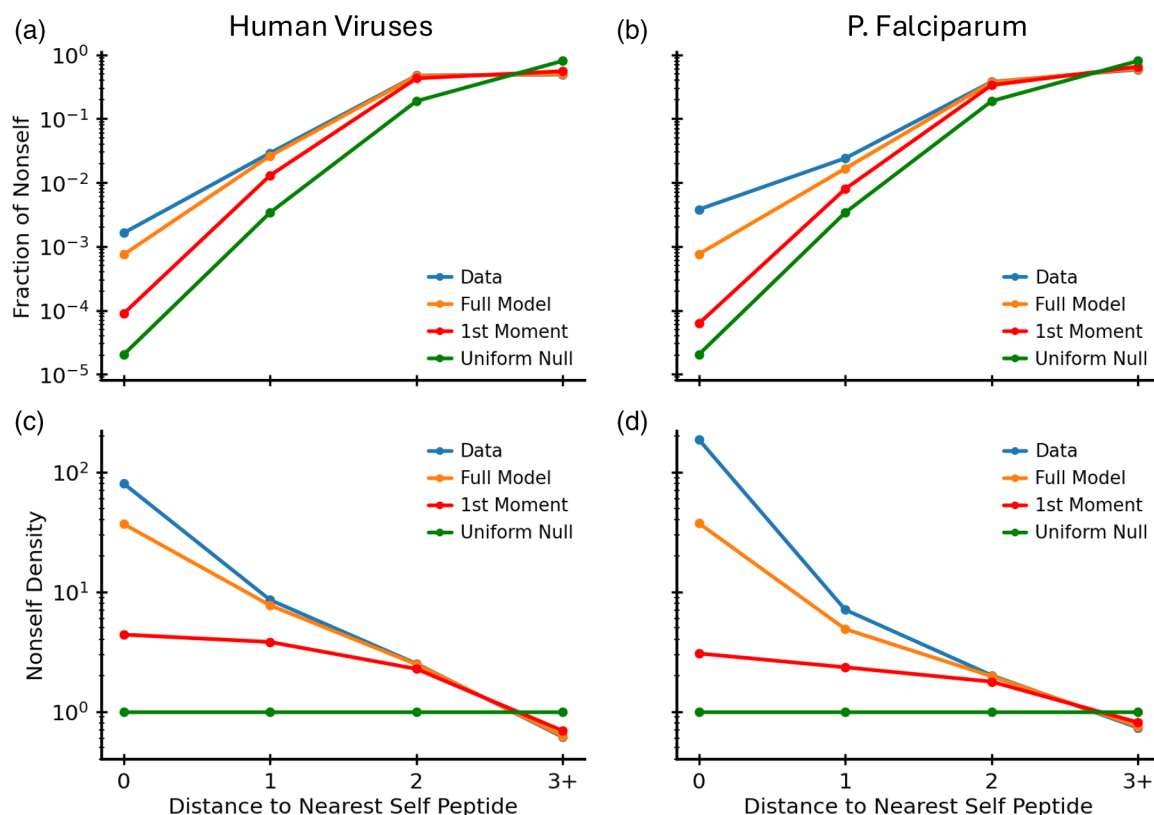


FIG. 3. Many nonself-peptides are close to self. (a), (b) Distribution of hamming distances to the nearest human peptide for peptides from an ensemble of human viruses (a) and *P. falciparum* (b) as found in the pathogen proteome data (blue), in predictions (orange) from the pathogen specific full model, pathogen specific independent amino acid frequency model (red), and in a null model with a uniform distribution over all 20^9 possible 9-mers (green). (c), (d) Relative density of distances to the nearest self-peptide for peptides from human viruses (c) or *P. falciparum* (d). Density is computed by normalizing the distributions by the corresponding value for the uniform (null) distribution. Colors as in panels (a), (b). Owing to the size of these proteomes, statistical error bars on these estimates are too small relative to the data to appear on the plots.

tance to the nearest human sequence. Looking at an ensemble of human viruses, we find that more than $\sim 0.1\%$ of these pathogenic sequences are identical with a human peptide, and more than 1% are just one amino acid away [Fig. 3(a)]. These coincidences and near coincidences are 10–100 times more frequent than expected if sequences were generated by a uniformly random null model, a model that only considers independent amino acid frequencies [Fig. 3(a)], or a model with shuffled amino acid frequencies (Appendix E). This shifted nearest-distance distribution is recapitulated with increasing precision as more constraints are added to the model (Appendix E) and near coincidences are captured perfectly by the full model in Eq. (2) [Fig. 3(a)]. The scale of these effects is biologically meaningful: the empirical coincidence probability of $\sim 10^{-3}$ implies that for a viral proteome with $\sim 10^4$ amino acids, exact coincidences are common, whereas they would be rare in the absence of the biases captured in our model. Last, we repeated the analyses with peptides from a variety of pathogen proteomes and the same pattern emerged (Appendix E). Even for the most statistically distinguishable proteome, *P. falciparum*, we again found a large enrichment of near coincidences [Fig. 3(b)], showing our findings generalize even to comparatively more dissimilar proteomes.

There are more pathogenic sequences within one amino acid of a human sequence than perfect coincidences in part

because there are $9 \times 19 = 171$ ways to be at distance one, but only one way to be at distance zero. Accounting for this effect by normalizing each distribution by the uniform null, we see that the density of nonself sequences declines monotonically with distance from the nearest self-sequence. Again, this is very well described by our models of nonself sequences [Figs. 3(c) and 3(d)], and holds true for parasites, bacteria, and viruses (Appendix E).

The model that we use here does not require that some foreign sequences may have evolved under selection to mimic specific human sequence motifs [36], but rather a similar distribution might simply stem from shared biophysical constraints for protein folding, metabolism, and function. We do see a small enhancement of exact coincidences in the actual sequences above that predicted by the models, which might reflect evolutionary pressure toward stronger mimicry of genes by pathogens in specific cases. But even this phenomenon is made easier to evolve by the fact that many pathogen peptides start out only a few mutations away from self at baseline.

D. Immunogenicity as a function of distance from self

Our results challenge the notion that easier to recognize T-cell antigens are farther from self [37–39]. In extension of the efficient sensing concept from neuroscience [40,41], we

might expect that the immune system allocates more resources to more commonly encountered stimuli. This predicts that mutation-derived neoantigens, which are close to self, are not anomalous, but rather inhabit the same space as common pathogenic epitopes. We therefore queried the immune epitope database (IEDB) [42] to determine the relationship between a peptide's distance from the nearest self-peptide and its likelihood of being experimentally validated as immunogenic. We analyzed data in IEDB that was measured by enzyme-linked immunosorbent spot (ELISPOT), as it detects functional cytokine release following antigen recognition (see Appendix E and accompanying Fig. 17). Consistent with our predictions, we find class I peptides from IEDB which are one or two amino acids away from self-peptides are more likely to be designated as immunogenic (ELISPOT positive) compared to peptides with three or more amino acid substitutions [Fig. 4(a)]. As some amino acid substitutions may be more biochemically significant than others, we further characterized the set of peptides that were only one mutation away from self by using a predicted T-cell "cross-reactivity distance" between the mutant peptide and the self-peptide, defined in recent work studying tumor neoantigens [7]. This distance metric quantifies the likelihood with which a substitution to an index peptide at a specific position and with a specific amino acid will affect T-cell recognition, and was inferred from activation experiments of T cells against peptide mutant libraries. The cross-reactivity distance between two peptides models how likely a substitution will lead to a differential T-cell response. For the epitopes in IEDB that are one amino acid away from a self-peptide, this cross-reactivity metric was predictive of immunogenicity status. While some of these Hamming-distance-one peptides are biochemically too similar to self and thus presumably tolerized, there are many that are sufficiently far away by cross-reactivity distance and can be recognized by the immune system [Fig. 4(b)].

The likelihood of an epitope under the human maximum entropy model gives another measure of distance to self, which focuses on the similarity to the ensemble rather than to the nearest existing self-peptide. Using this measure, we found a small but significant shift toward higher statistical similarity to self among immunogenic epitopes [Fig. 4(c)], especially at the high probability tail. As a result, epitopes in the highest likelihood bins are more likely to be immunogenic than 9-mers which are statistically unlikely to be found in a human protein [Fig. 4(d)]. Taken together, these two measures both demonstrate that the immune system favors targeting epitopes that are more "self-like" over epitopes that are more foreign.

Next, we wanted to test whether this preference of cytotoxic T cells toward recognition of peptides with similarity to self extends beyond the virome. MHC class I presents peptides derived from intracellular proteins and viral recognition is thought to be the major evolutionary pressure on MHC-I T-cell recognition [43]. As the virome is particularly close to self, we reasoned that this might provide a dominant pressure for recognition to focus on close-to-self-peptides even when peptides are not viral in origin. As a case study we considered MHC class I presented epitopes coming from the AT rich genome of *P. falciparum*. While the average *P. falciparum*

peptide comes from a different distribution than a human peptide [Figs. 2(a) and 2(b)], there are specific peptides that are disproportionately close to human self-peptides [Fig. 3(b)]. Despite the existence of more dissimilar peptides in this pathogen, immunogenic *P. falciparum* epitopes were more likely to be one or two amino acids away from the nearest self-peptide than peptides found not to be immunogenic or than random 9-mers from the pathogen proteome [Fig. 4(e)]. In accordance with the high D_{KL} , peptides from *P. falciparum* have a left-shifted (less likely) probability distribution under Eq. (2) which was derived from human proteome statistics [Fig. 4(f)]. While nonimmunogenic peptides followed this same left-shifted distribution, immunogenic peptides were right-shifted and more closely resemble the human proteome [Fig. 4(f)]. Although epitopes annotated in IEDB for *P. falciparum* tend to come from the most highly abundant proteins, the abundance of source proteins showed no discernible difference between immunogenic and nonimmunogenic epitopes [Fig. 4(g)]. Thus, despite the wide range of epitopes that are highly dissimilar from self that come from *P. falciparum*, the immune system tends to target the subset of epitopes that are closer and more statistically similar to the self-proteome.

III. DISCUSSION

Taken together, our results lead to a "shell model" of immunogenicity (Fig. 5). Because self- and nonself-peptides come from the same distribution, and this distribution is strongly nonuniform, an immune system targeting only peptides that are very dissimilar from self is unlikely to find much to react against. The highest density of nonself targets is in regions of sequence space where there is also the highest density of self-peptides. To maximize the probability of hitting targets and avoiding self, the immune system responds to peptides that are very close to but not exactly self (a shell around self-peptides). Consequently, even when challenged with a range of similar and highly dissimilar peptides that come from truly different distributions, like in parasitic infection by *P. falciparum*, the immune system preferentially targets the subset of class-I epitopes closer to self.

While counterintuitive, the shell model predicts many currently seen phenomena within a uniform framework. The recognition of mutation derived neoantigens in cancer, along with the overrepresentation of close peptides in IEDB are both predictions of this model. Likewise, the view of negative selection in the thymus as a memorization, rather than a statistical learning procedure also implies the need for peripheral tolerance to further censor false positives, since it is not possible for a T cell to be trained on every possible presented antigen perfectly [44]. Self-specific T cells are, therefore, still present at small numbers in the repertoire [2,45]. Negative feedback would be another mechanism to enforce the need to discriminate between similar peptides [46,47]. Our model therefore predicts two phenomena: bias toward recognition of close-to-self-peptides and the need for the "corrective" of peripheral tolerance as a trade-off for an immune system that is sensitive to small peptide differences.

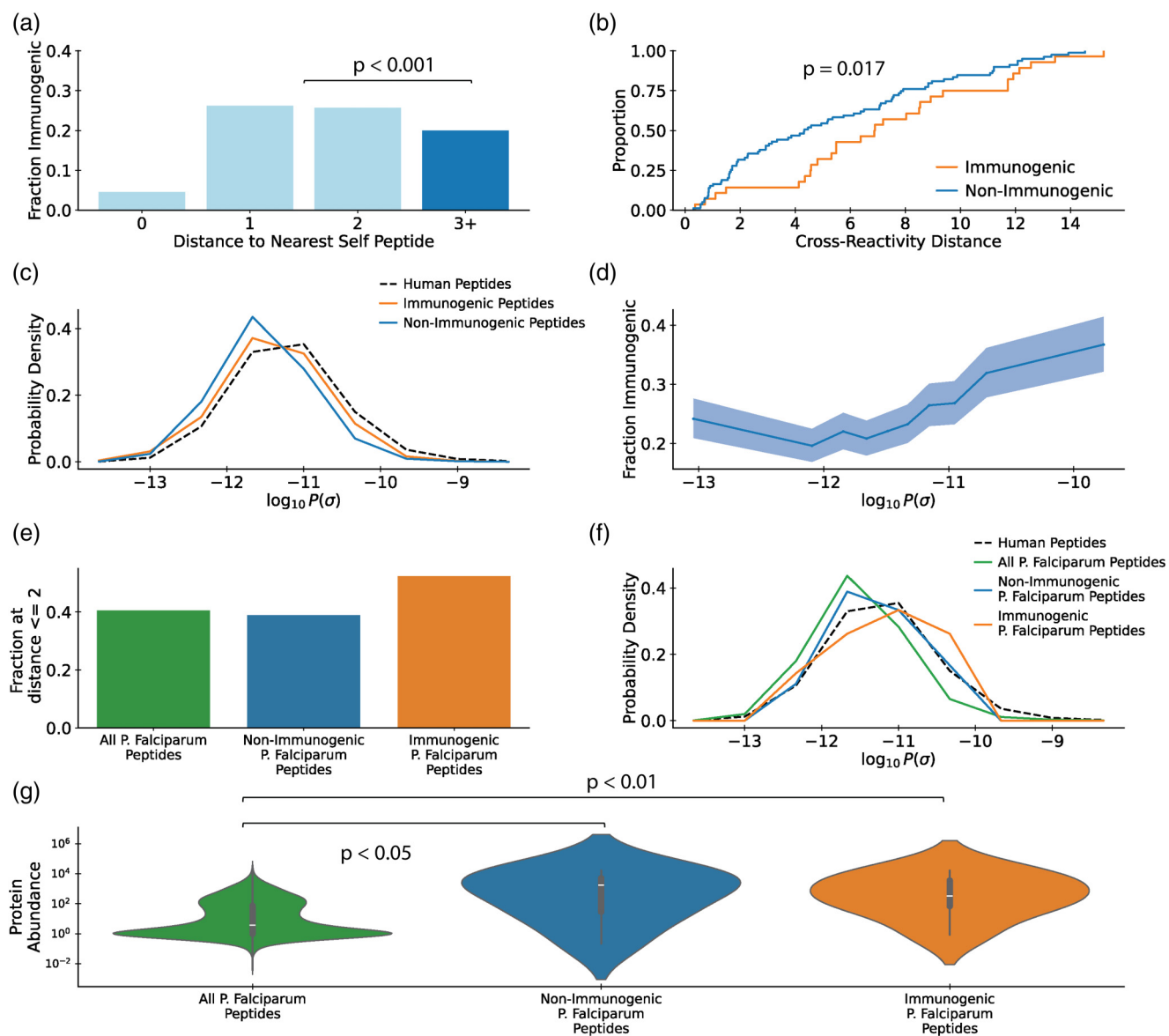


FIG. 4. Immunogenicity as a function of distance from self. (a) Immunogenicity (by ELISPOT status) varies as a function of Hamming distance to the nearest self-peptide from the human reference proteome. Peptides within Hamming distance 2 of the nearest self-peptide in the human reference proteome were significantly more likely to be immunogenic than peptides at distance 3 or more (χ^2 test). (b) Empirical cumulative distribution functions (ECDF) of T-cell cross-reactivity distances [7] for immunogenic (orange) and nonimmunogenic (blue) epitopes differing from the nearest self-peptide by a single amino acid. For these distance one epitopes, immunogenic epitopes tend toward higher T-cell cross-reactivity distance than nonimmunogenic epitopes (Kolmogorov–Smirnov test). (c) Distributions of model (log) probabilities for immunogenic (orange) and nonimmunogenic (blue) epitopes calculated using the two-point maximum entropy model for the human proteome. The distribution of probabilities for the human proteome is shown for comparison (black). Immunogenic epitopes are right-shifted relative to nonimmunogenic epitopes (Mann–Whitney U -test $p < 1 \times 10^{-10}$). (d) Immunogenicity varies as a function of model probability. The fraction of epitopes that are immunogenic is shown (solid line) for binned values of model probability using the two-point maximum entropy model for the human proteome. Shaded region represents the 95% confidence bounds of the binomial proportion using a Clopper–Pearson interval. (e) Fraction of different sets of peptides from the *P. falciparum* proteome that are within Hamming distance 2 of the nearest self-peptide. Immunogenic MHC class I presented 9-mers from *P. falciparum* (orange) are not depleted in peptides with similarity to self, compared to the overall proteome (green) and MHC class I presented peptides from *P. falciparum* found experimentally to be nonimmunogenic (blue). In fact, the proportion of peptides with similarity to self was highest among the immunogenic peptides, although the difference did not reach statistical significance (green vs orange, $p = 0.058$ and blue vs orange, $p = 0.12$, via χ^2 test). (f) Distributions of (log) probabilities of sets of peptides from the *P. falciparum* proteome calculated using the two-point maximum entropy model for the human proteome. (g) Average protein abundance of the source protein of different sets of peptides from the *P. falciparum* proteome. Abundance significantly differed between the source proteins for IEDB epitopes (blue and orange) and the background distribution of all peptides (green). Abundance did not differ between immunogenic (orange) and nonimmunogenic peptides (blue) (Mann–Whitney U -test). Data: Foreign epitopes from IEDB [42] presented by MHC-I tested by ELISPOT assay. Protein abundance data is from Pax-DB [48] as described in Appendix E.

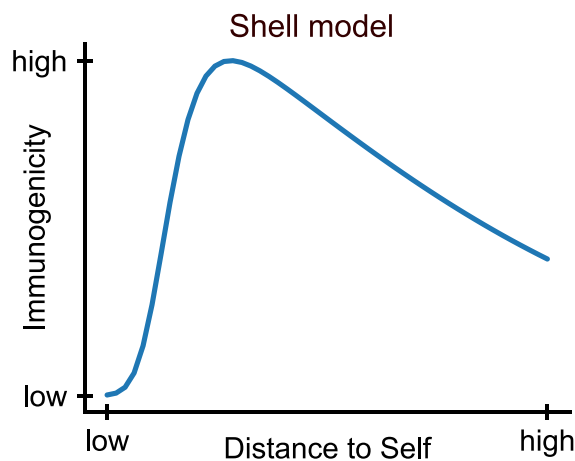


FIG. 5. Schematic of the predicted biases in immunogenicity. A shell model of nonself-peptide recognition where the immune system biases its search based on the more likely “close to self” peptides. Only (nearly) exact matches to self are nonimmunogenic due to tolerance. The most immunogenic peptides are not those that are most foreign, but those in a “shell” around self-peptides.

Self/nonself discrimination thus is not a typical classification problem. The usual regime for making distinctions is one in which examples are drawn from very different distributions, as with filtering out spam emails from normal emails. It might be complicated to find the proper axes along which to measure these differences, but they are essential for reliable discrimination. This is not so in self/nonself discrimination (Fig. 2); indeed, a good first approximation is that the distributions of peptides found in humans and pathogens are the same. Discrimination is still possible, but only because self-peptides come from a fixed and finite sample, $\sim 10^{12}$ 9-mers are possible, but only $\sim 10^7$ are realized.

Exploiting the finiteness of the sample usually amounts to “overfitting” or memorization [9,10], but for the immune system this is essential. For the self/nonself distinction in peptides, memorization is both allowed and necessary [49]. Training on peptides from the self-proteome can guide immune targeting to dense regions of nonself space around self-peptides. This provides a new perspective on the role of positive selection [50,51]: positively selecting T cells trained on self-peptides would create a set of T cells which can more easily target nonself-peptides, especially given T-cell cross reactivity. When faced by highly nonuniform peptide distributions, the immune system would find it more efficient to allocate resources to the recognition of more common peptides. This general prediction is the immunological application of the principles of efficient coding, which is long-standing principle in sensory neuroscience. For instance, work in neuroscience has linked the excess of retinal ganglion cells devoted to detecting dark contrasts to the excess of dark contrasts in natural visual scenes [52]. Similarly, we propose the immune system might be tuned to more readily respond to more common peptides. Moreover, by overindexing on the relatively small set of “close to self” peptides, the immune system can leverage a small set of peptides to recognize a pathogen, providing a novel perspective on immunodom-

inance, where indeed a relatively small peptide set drives immune responses, as a further consequence of this model.

The IEDB measurements and our theory both relate to immunogenicity as an emergent property of the immune system to respond at the collective level and do not explicitly predict the behavior of individual TCRs. Both experimental and theoretical studies [11,53] have pointed to the role of collective signal integration among T cells in self/nonself discrimination. In light of the current study, an important benefit of such collective decision making is the enabling of discrimination between closely related antigens. The integration of signals from polyclonal T-cell pools with a spectrum of avidities to related self- and nonself-peptides can allow discrimination despite cross reactivity of individual T cells [54].

One may also wonder why viruses do not just evolve to have farther-from-self-peptides and avoid immune detection? Viral transmission often occurs before adaptive immunity has taken over from innate detection, which is the first line of defense against viruses and focuses on evolutionarily conserved features which can be classified using maximum entropy [55,56]. A virus undetectable to the adaptive immune system would therefore be both transmissible and highly lethal, eventually impairing the existence of a new host. Moreover, the shell model assumes a certain homology between viral and host proteins exists at baseline, and viruses can only move so far away without impaired protein function given their higher error rate. Our model therefore represents an equilibrium between these considerations.

The adaptive immune system has long been viewed as a system for learning the pathogenic environment [57]. T cells are quite capable of, and arguably more likely to, recognize peptides that are close to self. This has important implications for understanding the success of cancer immunotherapies targeting neoantigens, and for guiding epitope selection for vaccines and cellular therapies. If we make assumptions about costs and benefits, then we can turn the schematic of Fig. 5 into quantitative predictions about the optimal parts of the pathogenic environment for the immune system to target [58,59]. Importantly, the qualitative conclusions of our work are independent of any specific quantitative assumptions: self and nonself are not very different, the distributions of these peptides are strongly inhomogeneous, and the combination of these results means that the immune system benefits by targeting antigens close to those represented in the organism’s own proteome.

ACKNOWLEDGMENTS

We thank V. Balachandran, C. Brown, C. Callan, W. James, A. Levine, M. Łuksza, S. Martis, T. Merghoub, A. Schietinger, Z. Sethna, and G. Victora for insightful discussions. We thank D. Hoyos for help with the MHC analysis. This work was supported in part by the National Science Foundation through the Center for the Physics of Biological Function (PHY-1734030); by a Lewis-Sigler fellowship (A.T.M.); by a Médecine Sciences fellowship of the Fondation pour la Recherche Médicale (Q.M.); by fellowships from the Simons Foundation and the John Simon Guggenheim Memorial Foundation (W.B.); and by the Lustgarten Foundation, the Mark Foundation (ASPIRE Award), the NCI (P30CA008748,

U01CA228963), the Pershing Square Sohn Cancer Research Alliance (Pershing Square Sohn Prize–Mark Foundation Fellowship), and Stand Up to Cancer (B.D.G.). B.D.G. would like to thank the Halvorsen Center for Computational Oncology, the Olayan Center for Cancer Vaccines, and the Tow Foundation.

B.D.G. has received honoraria for speaking engagements from Merck, Bristol Meyers Squibb, and Chugai Pharmaceuticals; has received research funding from Bristol Meyers Squibb and Merck; and has been a compensated consultant for Darwin Health, Merck, PMV Pharma and Rome Therapeutics of which he is a co-founder. All other authors declare no conflict of interests.

DATA AVAILABILITY

All code is available on github. The code to generate figures and an example tutorial on inferring maximum entropy models can be found at Ref. [61], and the full code for all analyses can be found at Ref. [62]. The data supporting this study’s findings are available within the article.

APPENDIX A: HUMAN AND PATHOGEN SEQUENCES

We compared proteome statistics from model organisms across different branches of jawed vertebrates: human and mouse as two examples of mammals, chicken as an example of a bird, and zebrafish as an example of a jawed fish. We also compared them against a set of pathogen proteomes: a collection of human viruses (individual viral proteomes are too small for reliable statistical analyses), several pathogenic bacterial species (*Mycobacterium tuberculosis*, *Listeria monocytogenes*, and *Streptococcus pyogenes*), as well as a parasite (*P. falciparum*). These examples were chosen because of their representation IEDB (see Appendix E).

For each species we downloaded its reference proteome from Uniprot [63] using the proteome identifiers shown in Table I. The table also displays shortened names used in main text figure legends. For the pan-viral proteome we downloaded all viral sequences annotated with human as a host species (taxon ID No. 9606) from Uniref90, a protein database clustered at 90. For each proteome we then generated a list of all possible 9-mers by iterating over all possible starting positions within each protein from the proteome. We note that the total number of peptides is roughly equal to the total length of the proteome (except for forbidden start

positions at the edges of the protein). Our collection of 9-mer peptides used for training and testing the models for each organism can be downloaded from [67]. The 20 amino acids are used with unequal probabilities and the correlations between neighboring amino acids are weak, but they extend over a very long distance, far beyond the length of the relevant peptides (Fig. 6). For each proteome, we split the data into equal portions train and test sets.

APPENDIX B: MAXIMUM ENTROPY MODELS

1. Model definition

We use the maximum entropy framework [15] as a principled way to include increasingly detailed statistical structure into a series of nested models for peptide statistics. Using this approach we constrain average features of the sequence $\langle f_\mu(\sigma) \rangle$ to equal their empirical values $\bar{f}_\mu$, while otherwise keeping the probability distribution as random as possible. Mathematically, this means we choose a probability distribution that maximizes the Shannon entropy

$$S[P(\sigma)] = - \sum_{\sigma} P(\sigma) \ln P(\sigma), \quad (B1)$$

subject to the normalization constraint $\sum_{\sigma} P(\sigma) = 1$, and constraints that enforce the equality of modeled and empirical observables

$$\langle f_\mu(\sigma) \rangle = \sum_{\sigma} P(\sigma) f_\mu(\sigma) = \bar{f}_\mu. \quad (B2)$$

Optimizing with respect to the normalization constraint yields a Boltzmann distribution of the form

$$P(\sigma) = \frac{1}{Z} \exp[-E(\sigma)], \quad (B3)$$

where

$$E(\sigma) = \sum_{\mu=1}^K \lambda_\mu f_\mu(\sigma) \quad (B4)$$

is a sum of terms involving each constraint, and

$$Z = \sum_{\sigma} \exp[-E(\sigma)] \quad (B5)$$

is a normalization factor, called the partition function. This is mathematically equivalent to the statistical mechanics of a system in thermal equilibrium, where $E(\sigma)$ is the energy as a function of its configuration.

TABLE I. Reference proteomes used in this study.

Short name	Full name	Proteome ID	Approximate length
Human	<i>Homo sapiens</i>	UP000005640	1×10^7
Mouse	<i>Mus musculus</i>	UP000000589	1×10^7
Chicken	<i>Gallus gallus</i>	UP000000539	1×10^7
Zebrafish	<i>Danio rerio</i>	UP000000437	1×10^7
Malaria	<i>Plasmodium falciparum</i>	UP000001450	4×10^6
Tuberculosis	<i>Mycobacterium tuberculosis</i>	UP000001584	1×10^6
Listeria	<i>Listeria monocytogenes</i>	UP000000817	8×10^5
StrepA	<i>Streptococcus pyogenes</i>	UP000000750	5×10^5
Human Virome	Human viruses	Uniref90-Filtered	1×10^6

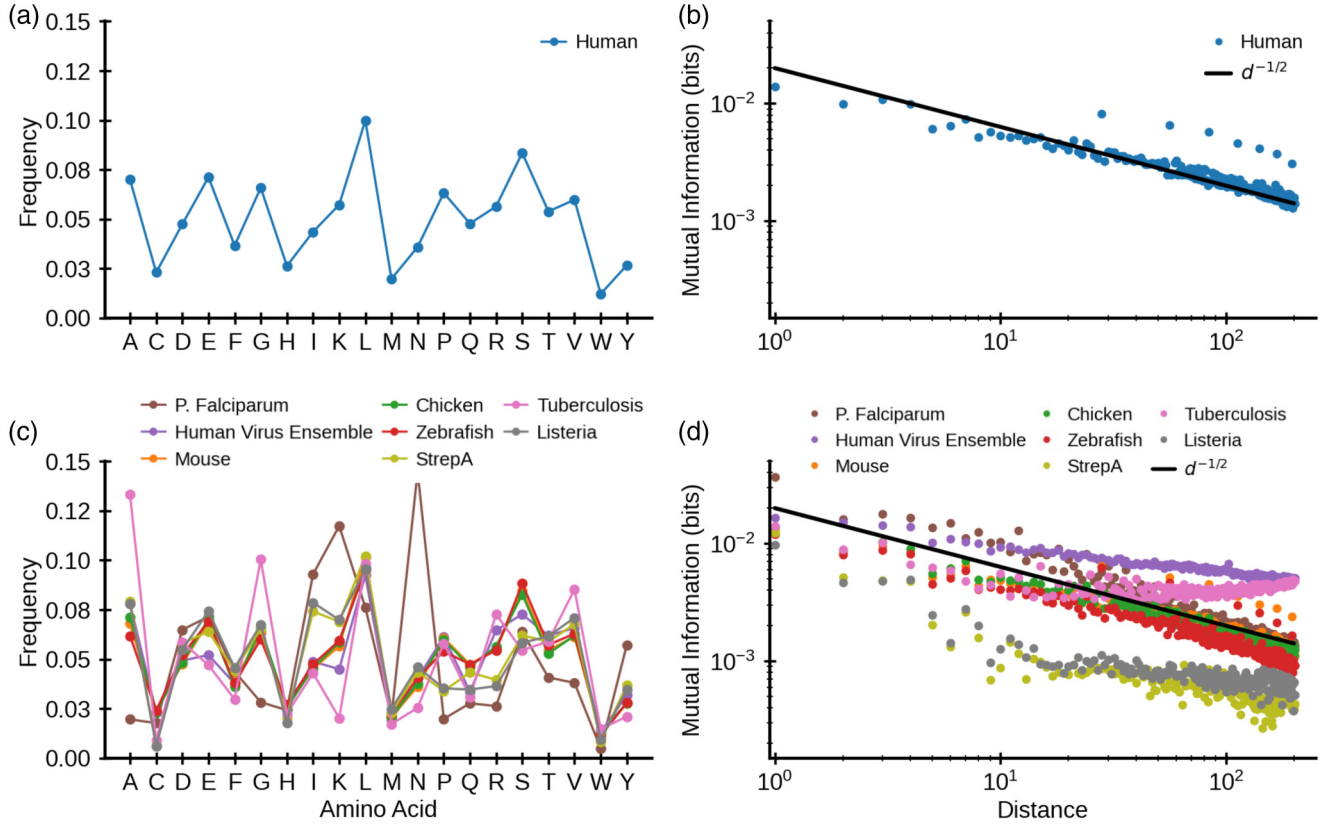


FIG. 6. Proteome statistics. Basic statistics for the human reference proteome (a), (b) and various other proteomes (c), (d) analyzed in this paper (Table I). (a) Amino acids are used with unequal probabilities in the human proteome. (b) Correlations between amino acids extend across a very long distance. Decay of mutual information with separation for the human proteome. Here the mutual information in bits per symbol is shown as a function of separation distance $d(X, Y) = |i - j|$, where the symbols X and Y are located at positions i and j in the protein sequence. This decay is modeled (black) as a power law [60] with exponent $-1/2$. (c) Same as panel (a) but for several example nonhuman vertebrate proteomes and various pathogen proteomes. (d) Same as panel (b) but for several example nonhuman vertebrate proteomes and various pathogen proteomes.

To fix the parameters λ_μ , we follow a standard method: At current values of the parameters we estimate $\langle f_\mu(\sigma) \rangle$ using Monte Carlo sampling, then change parameters to reduce the discrepancy between estimated and empirical observables, and iterate until these discrepancies are nearly zero [31].

Correlations extend beyond the scale of peptides of interest (Fig. 6), so we start by considering compositional constraints on the covariation of the total count of amino acids of different types. That is, we count the number of amino acids of each type in the peptide,

$$n^\alpha(\sigma) = \sum_i s_i^\alpha, \quad (\text{B6})$$

and constrain its expectation value, and then do the same with the second and third moments:

$$n^{\alpha\beta}(\sigma) = n^\alpha(\sigma)n^\beta(\sigma) = \left(\sum_{i=1}^L s_i^\alpha \right) \left(\sum_{j=1}^L s_j^\beta \right), \quad (\text{B7})$$

$$\begin{aligned} n^{\alpha\beta\gamma}(\sigma) &= n^\alpha(\sigma)n^\beta(\sigma)n^\gamma(\sigma) \\ &= \left(\sum_{i=1}^L s_i^\alpha \right) \left(\sum_{j=1}^L s_j^\beta \right) \left(\sum_{k=1}^L s_k^\gamma \right). \end{aligned} \quad (\text{B8})$$

This leads to a maximum entropy probability distribution with an effective energy:

$$E(\sigma) = \sum_{\alpha=1}^{20} \lambda_1^\alpha n^\alpha + \sum_{\alpha,\beta=1}^{20} \lambda_2^{\alpha\beta} n^\alpha n^\beta + \sum_{\alpha,\beta,\gamma=1}^{20} \lambda_3^{\alpha\beta\gamma} n^\alpha n^\beta n^\gamma. \quad (\text{B9})$$

This model only involves global couplings between amino acids independent of their distance.

Another common constraint involves the two-point frequencies $f_{ij}^{\alpha\beta}(\sigma) = s_i^\alpha s_j^\beta$; the expectation value $\langle f_{ij}^{\alpha\beta} \rangle$ is the probability of finding amino acid of type α at site i and amino acid of type β at site j . By construction the data are translation invariant, except for edge effects arising from the finite length of proteins, so this should depend only on the distance $\Delta = |i - j|$ between the two amino acids. We explicitly enforce this invariance in the model by instead constraining the expectation value of

$$f_\Delta^{\alpha\beta}(\sigma) = \frac{1}{2} \sum_{\substack{i,j=1 \\ |i-j|=\Delta}}^9 s_i^\alpha s_j^\beta. \quad (\text{B10})$$

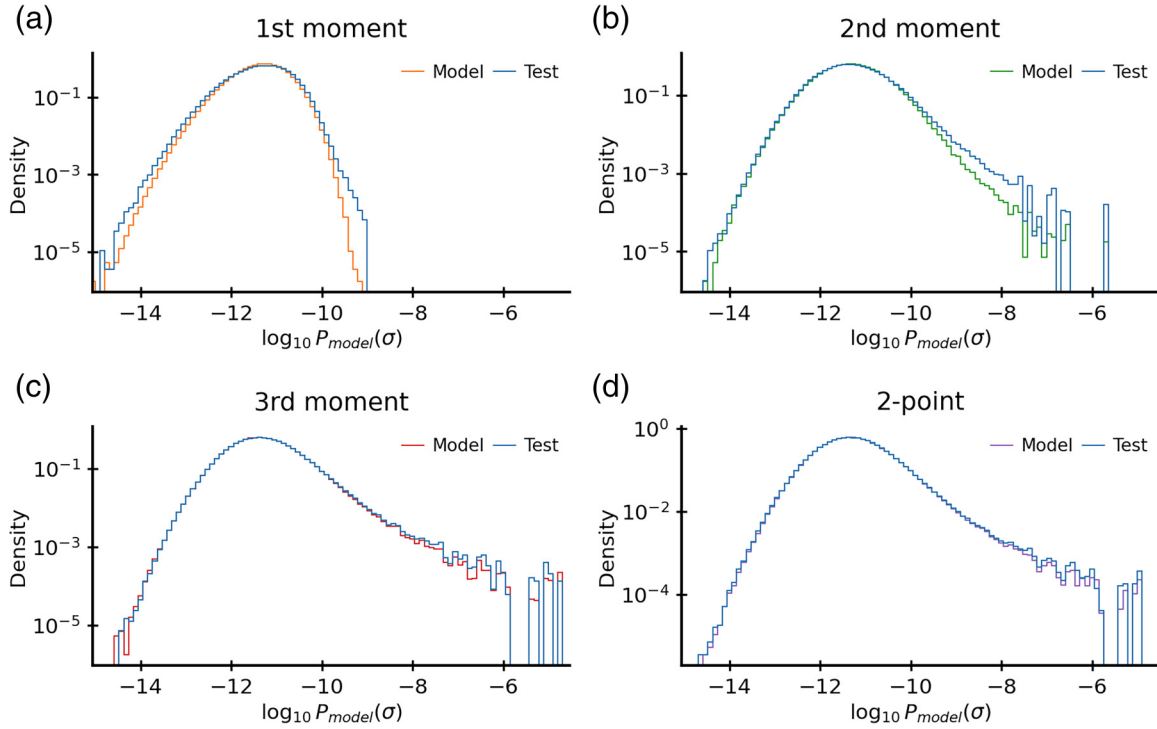


FIG. 7. Assessing the self-consistency of the density of states predicted by the different models. Probability density of (log) probability in the first-moment model (a), second-moment model (b), third-moment model (c), and full two-point model (d). The distributions of predicted probabilities of ensembles of sequences drawn from the series of models is compared to sequences from the test set.

Taking all terms together, we obtain the energy of the full model (with $J_0^{\alpha\beta} = 0$)

$$E(\sigma) = \sum_{\alpha=1}^{20} \lambda_1^{\alpha} n^{\alpha} + \sum_{\alpha,\beta=1}^{20} \lambda_2^{\alpha\beta} n^{\alpha} n^{\beta} + \sum_{\alpha,\beta,\gamma=1}^{20} \lambda_3^{\alpha\beta\gamma} n^{\alpha} n^{\beta} n^{\gamma} + \frac{1}{2} \sum_{i,j=1}^9 \sum_{\alpha,\beta=1}^{20} J_{\Delta}^{\alpha\beta} s_i^{\alpha} s_j^{\beta}. \quad (\text{B11})$$

All of the coefficients must be adjusted so that the predictions of the model match observed features of the data: λ_1^{α} is fixed by matching the mean number of amino acids of type α ; $\lambda_2^{\alpha\beta}$ by matching the covariance in numbers of amino acids of types α and β ; $\lambda_3^{\alpha\beta\gamma}$ by matching the third moments of these numbers; the matrix $J_{\Delta}^{\alpha\beta}$, where $\Delta = |i - j|$, must be adjusted to match the distance dependence of the pairwise correlations; and the constant (partition function) Z serves to normalize the distribution, so that

$$Z = \sum_{\sigma} \exp[-E(\sigma)]. \quad (\text{B12})$$

To test the predictive power of the maximum entropy model, we compare the density of states between model and data (Fig. 7), and we calculate triplet correlations according to

$$C_{ijk}^{\alpha\beta\gamma} = \langle \delta s_i^{\alpha} \delta s_j^{\beta} \delta s_k^{\gamma} \rangle, \quad (\text{B13})$$

with $\delta s_i^{\alpha} = s_i^{\alpha} - \langle s_i^{\alpha} \rangle$.

2. Model complexity and Bayesian information criterion

Each additional constraint term in the energy function in Eq. (B11) adds additional free parameters to be inferred during model training. The independent-site (first-moment) model has 20 parameters, one for the frequency of each amino acid, corresponding to the λ_1 in Eq. (B11). Each additional set of λ values and the coefficients in the matrix J increase the number of parameters of the higher complexity models [Fig. 8(a)]. With increased complexity, we gain an increase in predictive power, as shown in the boost in log-likelihood on the training data [Fig. 8(b)]. The tradeoff between the model complexity (and potential overfitting) and the predictive power of each model can be quantified using information criteria [64,65]. Specifically, we computed the Bayesian information criterion (BIC) of each model, relative to the first-moment model [Fig. 8(c)]:

$$\Delta \text{BIC}_{\text{model}} = \text{BIC}_{\text{model}} - \text{BIC}_{\text{independent}},$$

where for each model,

$$\begin{aligned} \text{BIC} &= -2 \ln \hat{\mathcal{L}} + k \ln n \\ &= 2 \sum_{j=1}^n E(\sigma_j) - 2nF + k \ln n, \end{aligned} \quad (\text{B14})$$

where k is the number of parameters, n is the size of the training data, and the likelihood $\hat{\mathcal{L}}$ of the training data is computed using the model energy function and the model partition function, $F = -\ln Z$ (see below for details on estimating F).

For the number of parameters, k , we use the full sizes of the parameter matrices as a conservative upper bound on the

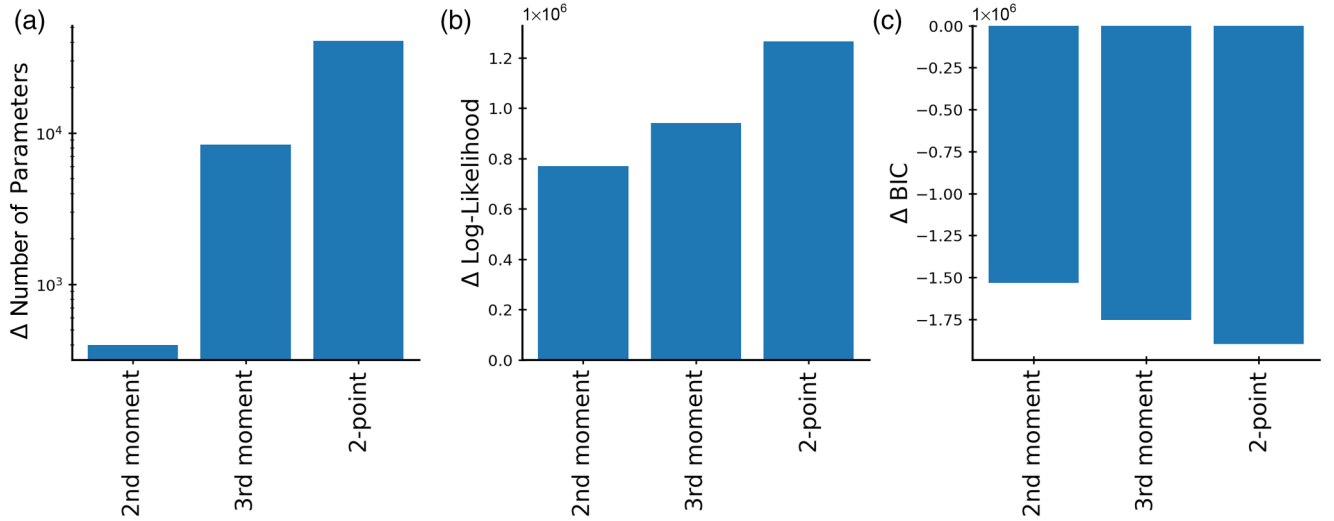


FIG. 8. More complex models are favored by the Bayesian information criteria over the independent site model. (a) Number of additional parameters relative to the independent site (first-moment) model for the energy functions in the nested set of models. (b) Increase in (log) likelihood on the training data relative to the independent site model. (c) Decrease in the Bayesian information criterion (BIC) relative to the independent site model. For all panels, values for the models on the y axis are shown as relative to the independent site (first-moment) model.

true number of independent parameters. We note that these numbers overestimate the true number of independent parameters, as the parameter matrices have certain symmetries that are ignored for this calculation. Even with this therefore conservative overestimate of the parameter count, each additional constraint term lowers the BIC as expected, favoring the use of these terms in the final model (Fig. 8).

3. Calculating entropies and statistical divergences using thermodynamic integration

How can we determine the entropy of a fitted model? From the definition of the model in terms of a Boltzmann distribution,

$$P(\sigma) = \frac{1}{Z} \exp[-E(\sigma)], \quad (\text{B15})$$

we obtain the entropy

$$S[P(\sigma)] \equiv - \sum_{\sigma} P(\sigma) \ln P(\sigma) \quad (\text{B16})$$

$$= \langle E(\sigma) \rangle_{P(\sigma)} - F, \quad (\text{B17})$$

where $F = -\ln Z$. Similarly, we can express the Kullback-Leibler divergence as

$$D_{\text{KL}}[P(\sigma)||Q(\sigma)] = \sum_{\sigma} P(\sigma) \ln \frac{P(\sigma)}{Q(\sigma)}, \quad (\text{B18})$$

$$= \langle \Delta E(\sigma) \rangle_{P(\sigma)} - \Delta F, \quad (\text{B19})$$

where $\Delta E = E_Q(\sigma) - E_P(\sigma)$ and $\Delta F = F_Q - F_P$. In both instances, we can approximate the expectation value over $P(\sigma)$ as the mean over Monte Carlo samples drawn from that distribution. But we also need to determine F , or the partition function of the fitted models.

As an exact evaluation of the partition function is computationally intractable we use thermodynamic integration [32]

to numerically approximate $F = -\ln Z$: We define the perturbed energy function

$$E_g(\sigma) = E_{\text{ref}}(\sigma) + g\Delta E(\sigma), \quad (\text{B20})$$

where $E_{\text{ref}}(\sigma)$ is a reference energy for which F_{ref} can be calculated analytically, and $0 \leq g \leq 1$ is a parameter scaling the additional energy term. For the maximum entropy model defined in Eq. (B11) we use

$$E_{\text{ref}} = \sum_{\alpha=1}^{20} \lambda_1^{\alpha} n^{\alpha}, \quad (\text{B21})$$

$$\begin{aligned} \Delta E(\sigma) = & \sum_{\alpha, \beta=1}^{20} \lambda_2^{\alpha\beta} n^{\alpha} n^{\beta} + \sum_{\alpha, \beta, \gamma=1}^{20} \lambda_3^{\alpha\beta\gamma} n^{\alpha} n^{\beta} n^{\gamma} \\ & + \frac{1}{2} \sum_{i,j=1}^9 \sum_{\alpha, \beta=1}^{20} J_{\Delta}^{\alpha\beta} s_i^{\alpha} s_j^{\beta}. \end{aligned} \quad (\text{B22})$$

Notice that $g = 1$, $E_g(\sigma) = E(\sigma)$ in Eq. (B11).

We have chosen E_{ref} to describe a model in which amino acids are chosen independently at each site, and we can see that

$$F_{\text{ref}} = -k \ln \left[\sum_{\alpha=1}^{20} \exp(\lambda_1^{\alpha}) \right]. \quad (\text{B23})$$

In contrast, ΔE includes all couplings between different residues in the k -mer. Thus, as we move along the family of models from $g = 0$ to $g = 1$, we interpolate between the independent model and the true model.

We can define the free energy F_g for the model at a specified value of the parameter g . Importantly, we have

$$\frac{dF}{dg} = -\langle \Delta E(\sigma) \rangle_{P_g(\sigma)}, \quad (\text{B24})$$

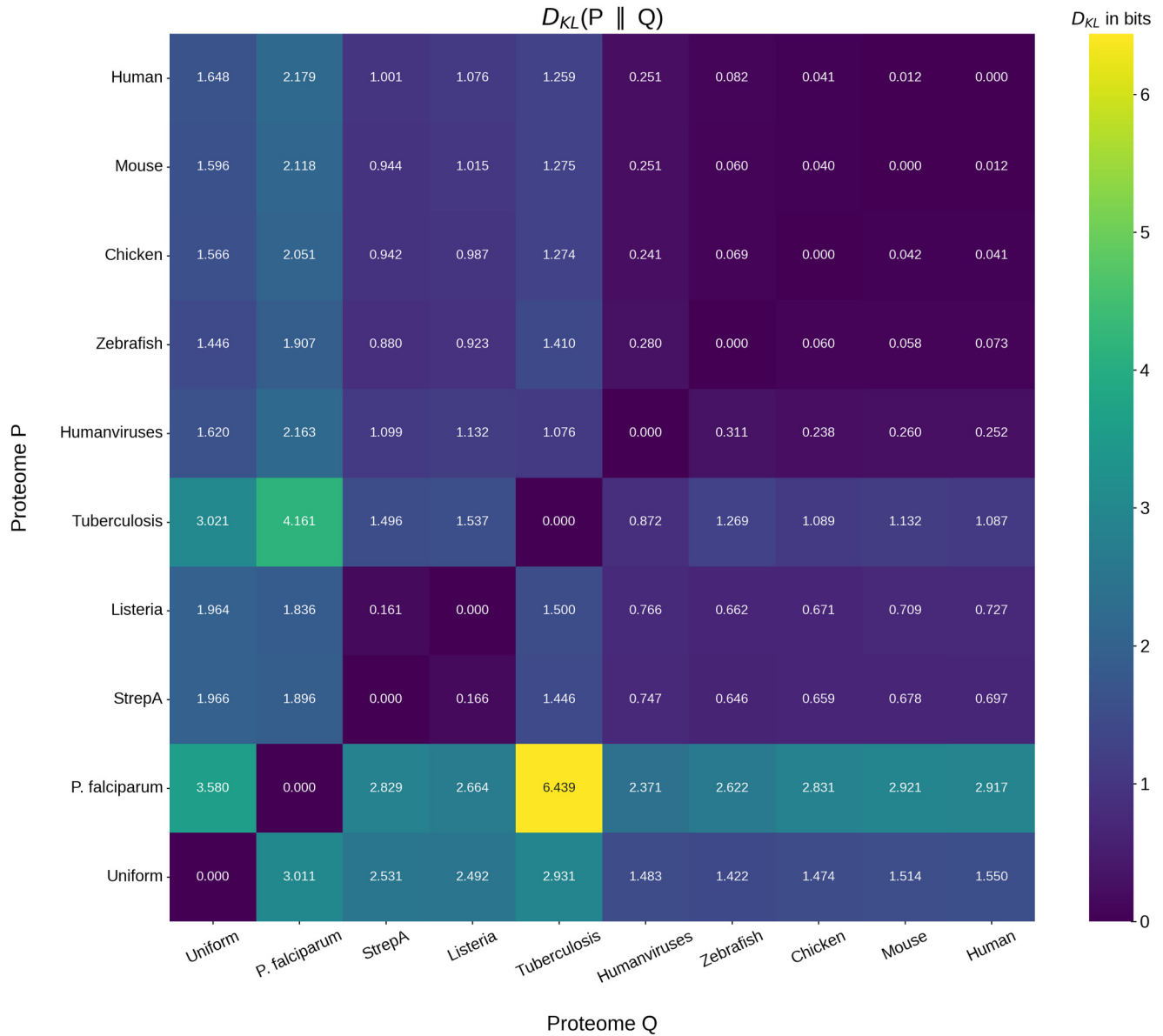


FIG. 9. Pairwise divergence between peptides from different proteomes. Pairwise Kullback-Leibler (KL) divergences between peptide distributions of different proteomes. For each proteome we show the statistical distance calculated according to the full two-point model with all of the constraints. The first column and the last column here correspond to the y and x coordinates, respectively, for the diamond shaped points in Fig. 2(a). The MDS in Fig. 2(b) is calculated using the symmetrized version of this matrix.

where we emphasize again that the right-hand side can be approximated as the mean over Monte Carlo samples. Thus, we can calculate the free energy of the real model at $g = 1$ using “thermodynamic integration”

$$F(g = 1) = \int_0^1 dg \frac{dF}{dg} + F_{\text{ref}}. \quad (\text{B25})$$

In practice, we draw Monte Carlo samples from the perturbed models with evenly spaced $g \in [0, 1]$, use these samples to evaluate dF/dg , and then evaluate the integral by Simpson’s rule.

4. Symmetrized divergence and MDS

Using the formula in Eq. (B19) for each pair of proteomes, we calculate the asymmetric divergence matrix M

shown in Fig. 9. To perform MDS, we first calculated the symmetrized version of this matrix, also known as the Jeffrey’s divergence, by calculating $J = (M + M^T)/2$. MDS seeks a low-dimensional representation of the data in which the distances respect well the distances in the original high-dimensional space. In general, MDS is a technique used for analyzing dissimilarity data. It attempts to model dissimilarities as distances in a Euclidean space. Metric MDS was calculated using scikitlearn’s implementation found at Ref. [66].

5. Model parameters

Our inferred model parameters for every model used in this paper can be found by downloading the data per the instructions in Ref. [67]. The zip directory has a subdirectory

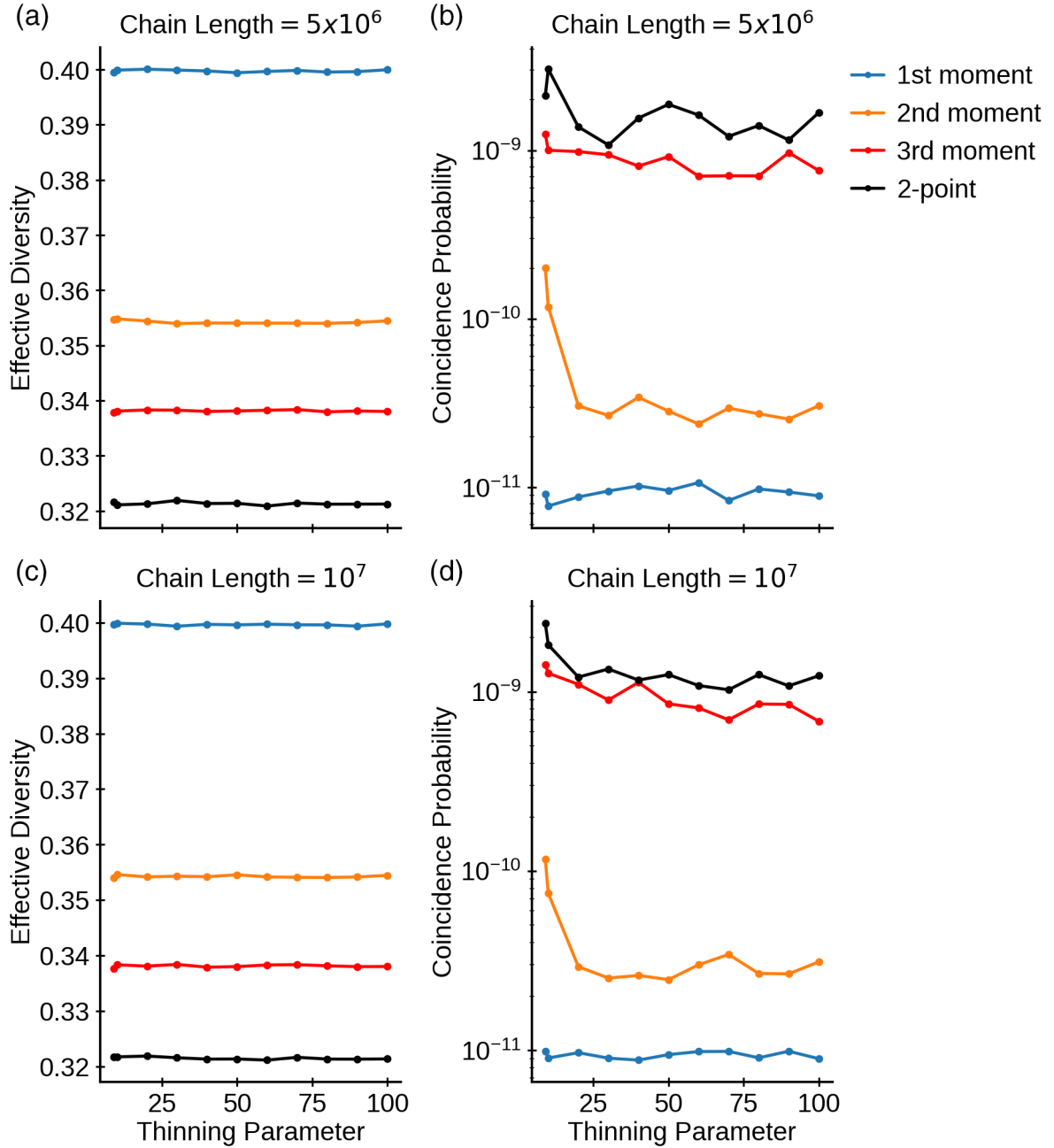


FIG. 10. Statistics on samples are robust to hyperparameters. Sampling from the models is used to estimate entropy and coincidence. Estimates of effective diversity (a), (c) and coincidence probability (b), (d) for different hyperparameters used to draw samples from the models using the metropolis-Hastings algorithm [68,69] (see Appendix B 6). Data is shown here for the family of models trained on the human proteome.

“generated_data” and each model’s parameters are stored in a separate *_params.npz file.

6. Model sampling

Samples were drawn from the model using the Markov chain Monte Carlo (MCMC)-based metropolis-Hastings algorithm [68,69]. Samples for the original models were drawn using a chain length of 5×10^6 and thinning parameter of 60. Samples for the filtered models with MHC were drawn using a longer chain length of 10^7 (so that enough samples remained after the filter) and thinning parameter of 60. The estimates

calculated via sampling did not depend significantly on the choice of hyperparameters (Fig. 10).

7. Coincidence probability

Coincidence probability was estimated using the Simpson’s index [70,71] on samples drawn from the models:

$$D = \sum_{i=1}^S \frac{n_i(n_i - 1)}{N(N - 1)}, \quad (\text{B26})$$

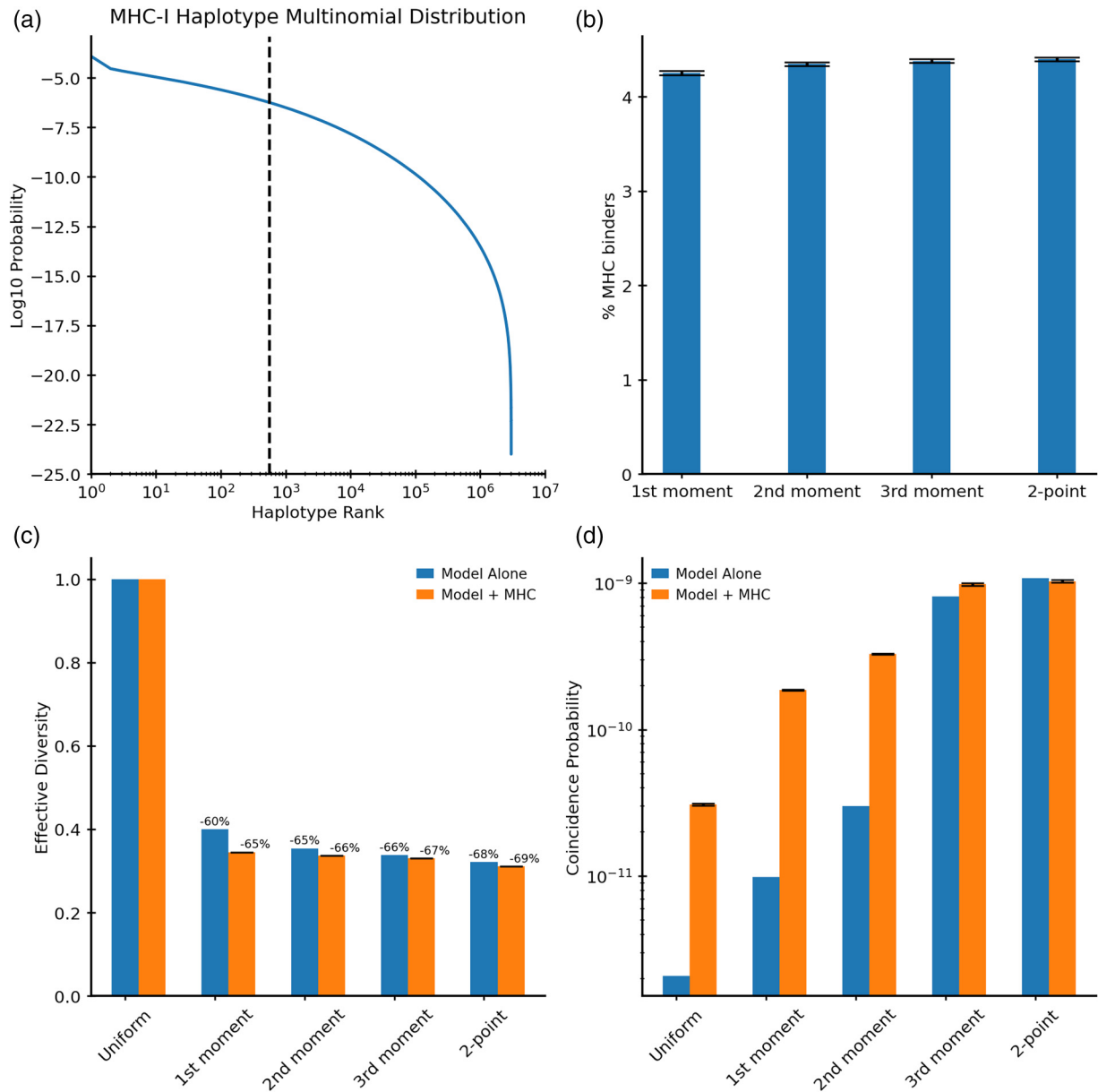


FIG. 11. Joint maximum entropy and MHC models. (a) Ranked probability distribution of MHC-I haplotypes, as described in Appendix C. Haplotypes that ranked within the top 500 most probable were used in this analysis (cutoff shown as dotted black line). (b) Percentage of peptides in samples from each model that were predicted by netMHC-4.0 [74] to be strong binders to at least one allele in the haplotype. Error bars show SEM across the top 500 haplotypes. (c) Effective diversity for each model alone (blue) and each model joint with MHC-I presentation (orange). Relative decrease is shown relative to the uniform for the appropriate category. Error bars for joint models (orange) shown SEM across the top 500 haplotypes. (d) Probability of coincidence for each model alone (blue) and each model joint with MHC-I presentation (orange). Error bars for joint models (orange) shown SEM across the top 500 haplotypes.

where N is the total number of peptides in the sample population, S is the number of unique peptides in the sample, and n_i is the number of occurrences of the i th peptide in the sample.

8. Effective diversity

Effective diversity is calculated as $D = e^S/20^9$, where S is the estimated Shannon entropy [for models, calculated using Eq. (B17)] and D is the Shannon diversity (also known as perplexity) relative to that of a uniform distribution over 9-mers.

APPENDIX C: COMBINED MHC AND MAXIMUM ENTROPY DISTRIBUTION

1. MHC haplotypes

To characterize canonical human haplotypes, we used previous work [33] that characterized a haplotype multinomial distribution using data from the National Marrow Donor Program database, which provides information on 1 242 890 donors with full MHC-I linkage information [72,73]. Following this previous work [33], we assume the MHC-I haplotype will consist of two each of HLA-A, HLA-B, and HLA-C

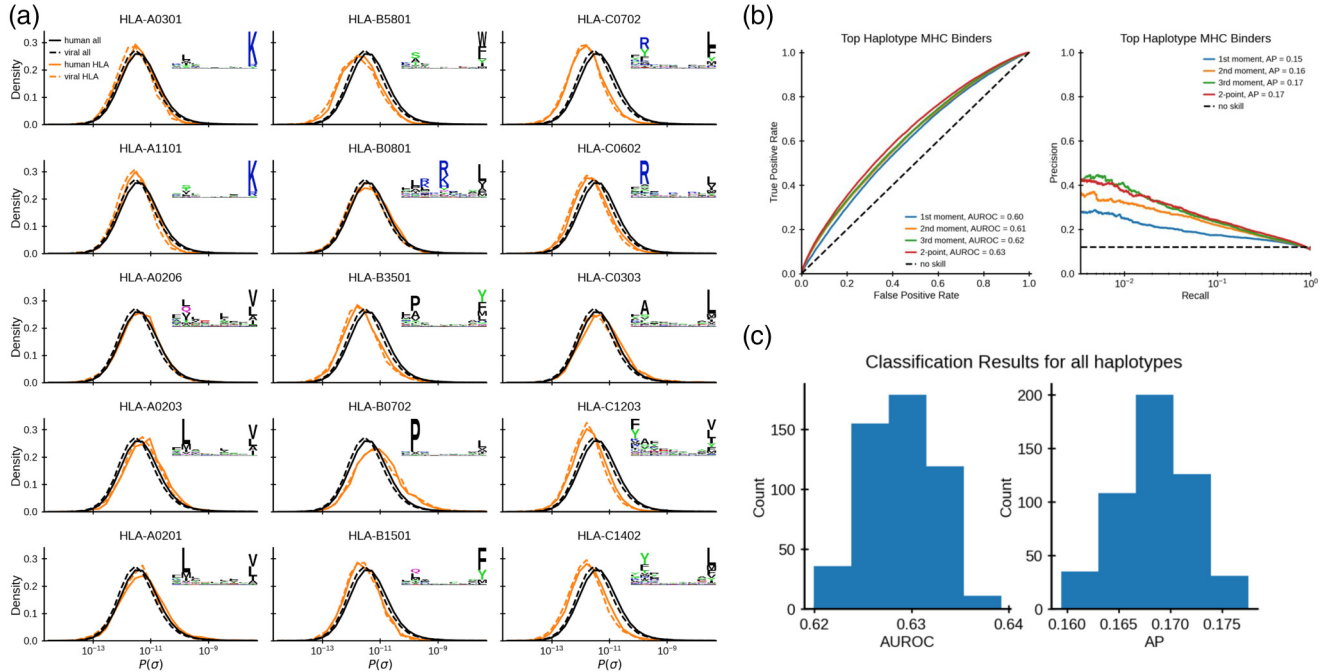


FIG. 12. Classification of HLA restricted peptides. (a) Each subplot displays the distribution of model probabilities for HLA restricted peptides predicted by netMHC-4.0 [74] for the top 5 HLA-A/B/C alleles (in terms of number of training sequences). The model probabilities are predicted by Eq. (2) with parameters inferred from the human proteome. For each HLA allele, the plot displays the distributions both of predicted self-peptide binders (solid orange lines) and predicted binders from human viruses (dashed orange lines). Each plot also shows the distributions of all peptides regardless of HLA restriction for reference (black solid/dashed lines), alongside a sequence logo of all predicted binders that highlights which residues are enriched in anchor positions. (b) ROC and PR Curves for classifier using model energies (as in Fig. 2) for restricted test set of peptides that bind to an allele in the most common HLA haplotype [rank = 1 in Fig. 11(a)]. (c) Distribution of AUROC and AP values for same classification scheme, across each of the top 500 ranking haplotypes.

for a total of six MHC-I genes. We used the previously derived multinomial distribution with an independent frequency model without MHC-I linkage for each HLA-A, HLA-B, and HLA-C gene. As described in the previous work [33], the number of heterozygous HLA-I genes is given by N_H , where $N_H \in [0, 1, 2, 3]$. The probability of a haplotype $H = [A_1, A_2, B_1, B_2, C_1, C_2]$ is given by

$$p(H) = 2^{N_H} \prod_{h \in H} p_h, \quad (C1)$$

where p_h corresponds to the marginal probability of HLA-I h within haplotype H . A visualization of this distribution can be seen in Fig. 11(a). For many of the analyses in this paper, we used the top 500 most probable haplotypes as determined by Eq. (C1). Classification of HLA presented peptides using specific allele or haplotype filtered data is shown in Fig. 12.

2. Joint models

To model the effects of MHC presentation, we introduce a filtered probability distribution for 9-mers in a proteome that are presented on an HLA haplotype $H = \{A_1, A_2, B_1, B_2, C_1, C_2\}$. We assume that presentation and maximum entropy probability are independent, and thus

$$\tilde{P}(\sigma) = P(\sigma)Q(\sigma, H), \quad (C2)$$

where $P(\sigma)$ is the fit maximum entropy model, and $Q(\sigma, H)$ describes the probability that peptide σ is presented on at least

one allele in a haplotype H . For each peptide σ and HLA allele h , we use netMHC-4.0 [74] to get a rank of predicted binding affinity for σ on allele h .

We define the peptide's presentability score as

$$s(\sigma, H) = \min\{\text{Rank}(\sigma, h); h \in H\}. \quad (C3)$$

We design $Q(\sigma, H)$ as a binary filter, only keeping peptides that are a strong binder to at least one of the alleles in the haplotype. We use the % Rank cutoff of 0.5% as advised by the authors of NetMHC-4.0:

$$Q(\sigma, H) \propto \begin{cases} 1 & s < 0.5\%, \\ 0 & s \geq 0.5\%. \end{cases} \quad (C4)$$

To estimate the entropy of the filtered model, we sample from the maximum entropy distribution and filter the sample using netMHC to only keep strong binders. The entropy is then calculated from the filtered, sampled set as

$$\begin{aligned} S[\tilde{P}(\sigma)] &= - \sum_{\sigma} \tilde{P}(\sigma) \ln \tilde{P}(\sigma) \\ &= - \sum_{\sigma} \tilde{P}(\sigma) (\ln Q(\sigma) + \ln P(\sigma)) \\ &= - \sum_{\sigma} \tilde{P}(\sigma) \ln P(\sigma) - \sum_{\sigma} \tilde{P}(\sigma) \ln Q(\sigma) \\ &= -F + \langle E(\sigma) \rangle_{\tilde{P}} - \ln \frac{1}{b}, \end{aligned}$$

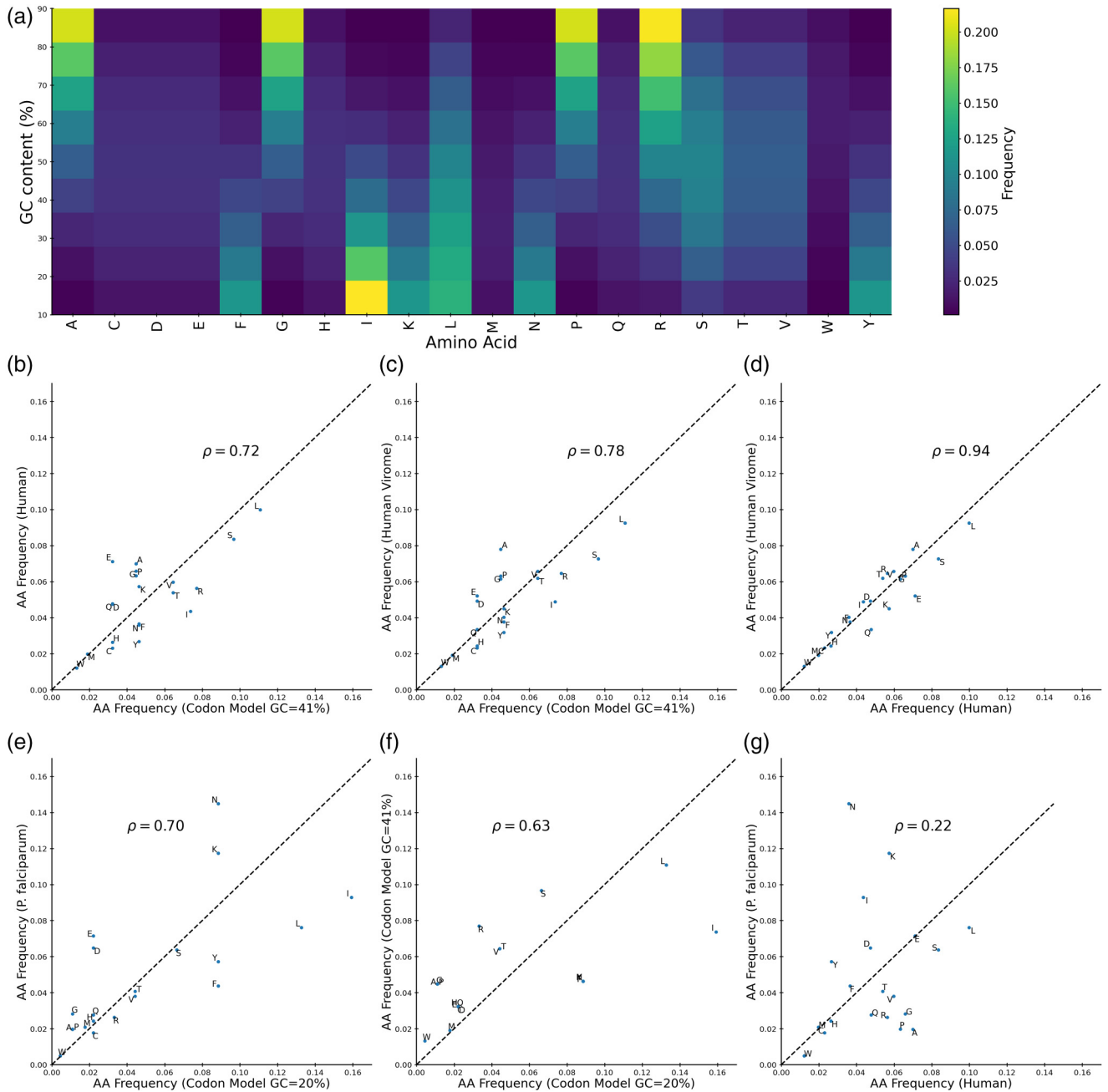


FIG. 13. Codon bias models and amino acid frequencies. (a) Predicted amino acid frequencies as a function of simulated GC content based on the codon table. (b)–(g) Scatter plots comparing amino acid frequencies in the reference proteomes and codon bias null models. The codon model GC = 20% matches the content of the AT-rich *P. falciparum*, while the codon model GC = 41% matches the content of the human proteome. The Pearson correlation is shown for each plot.

where b is the fraction of the sample that survived the binding filter ($b = \frac{n_{\text{binders}}}{N}$), as in Fig. 11(b), and $\langle E(\sigma) \rangle_{\tilde{p}}$ is a Monte Carlo estimate of the cross-entropy term relating the new distribution to the original model distribution. Intuitively, this means that a more selective filter (smaller b) leads to less entropy in the filtered distribution, and a filter that keeps 100% of the entire original sample does not change the entropy at all ($\ln 1 = 0$). Estimated entropies and coincidence probabilities for the joint models with MHC versus the models without MHC are shown in Figs. 11(c) and 11(d).

APPENDIX D: CODON BIAS MODEL

Neutral evolution model

To simulate the effect of the codon table on peptide samples, we randomly generated peptides with different GC content fractions. For a given GC fraction f , we set the frequency of “C” and “G” to $\frac{f}{2}$ and the frequency of “A” and “T” to $\frac{1-f}{2}$. Given these letter frequencies, codon frequencies are then computed by multiplying independent nucleotide frequencies for each of the three positions. The frequency

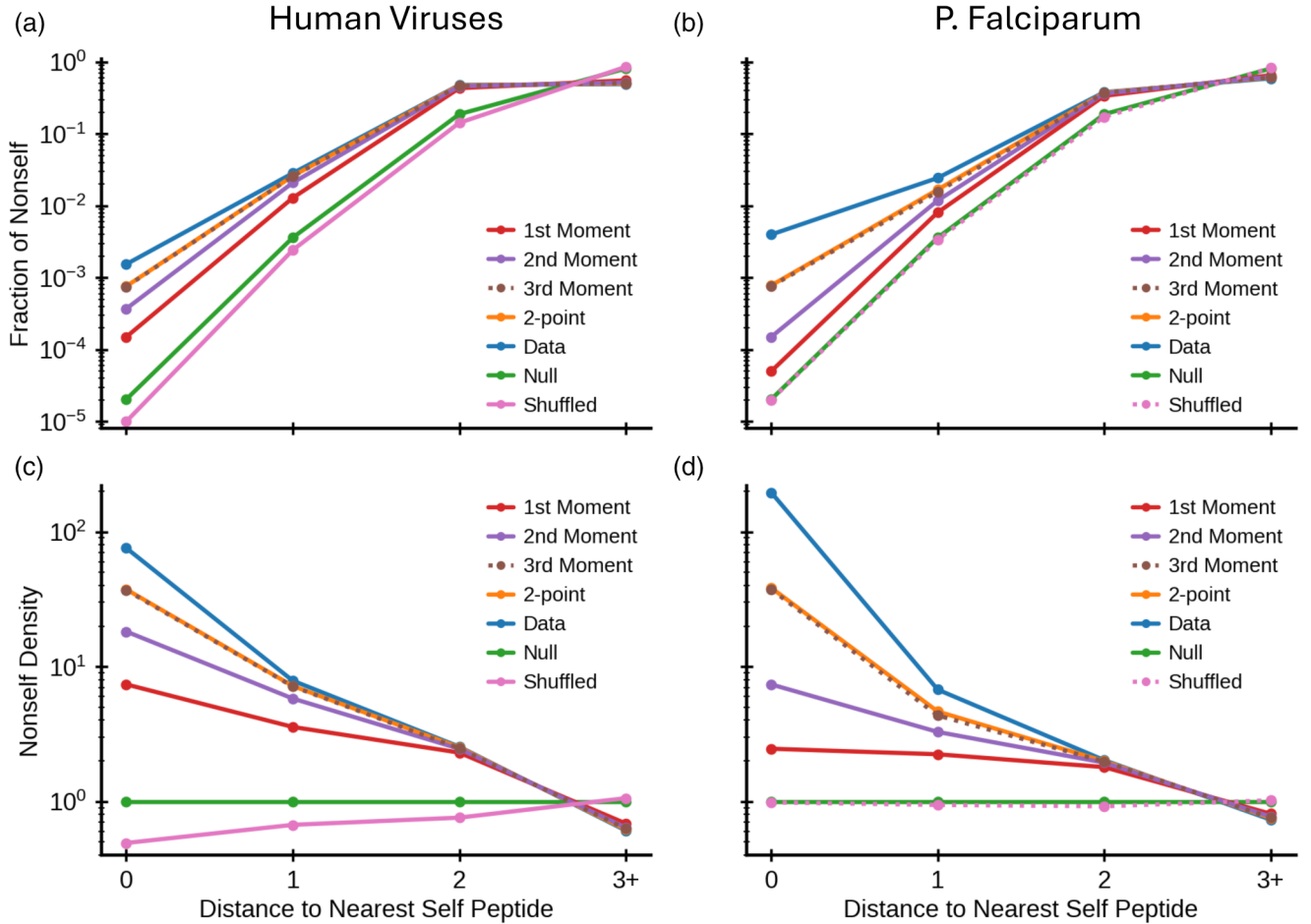


FIG. 14. Distance to nearest self-peptide as a function of model constraints. (a), (b) Distribution of distances to the nearest self-peptide for peptides from human viruses (a) and *P. falciparum* (b) as found in the data (blue), in predictions from the family of models (red, purple, brown, orange), in a null model with a uniform distribution over all 20^9 9-mers (green), and model with nonuniform but shuffled probabilities for amino acid usage (pink). (c), (d) Relative density of distances to the nearest self-peptide for the same sets of peptides. Colors as in panels (a), (b).

associated with an amino acid is then the sum of each of the degenerate codons that can create that amino acid using the standard codon table. For example, the amino acid V can be made from the codons [GTT, GTC, GTA, GTG]. With GC content at 70%, this would lead to the following calculation of frequency for V:

$$f_V = (0.35)(0.15)(0.15) + (0.35)(0.15)(0.35) + (0.35)(0.15)(0.15) + (0.35)(0.15)(0.35) = 0.0525,$$

which corresponds to value in the third row (70%) and 18th column (V) in Fig. 13(a).

APPENDIX E: DISTANCE DISTRIBUTIONS AND IMMUNOGENIC EPITOPES

1. Hamming distance distributions

We aligned peptides to their nearest self-peptide by Hamming distance using an exact algorithmic approach based on hashing-based lookups [75], leveraging publicly available code in pyrepseq [76]. We show these distributions for various peptide sets, sampled from different models and organisms

(Figs. 14–16). We note that owing to the size of the proteomes (Table I) error bars on the density of such peptides at different distances are small and are thus omitted from the figures. For example, the length of the *P. falciparum* proteome is $\sim 10^6$ amino acids, and a simple calculation of $\sqrt{(p * (1 - p))/N}$ gives standard errors of $\sim 10^{-4}$ at $p \sim 10^{-2}$.

2. Epitope immunogenicity

To understand how the immunogenicity of foreign epitopes depends on their distance to the nearest self-peptides we analyzed data from the IEDB [42]. This database contains a collection of experimental data about T-cell activation in response to different peptides. The experiments recorded in the database involve assays, such as the ELISPOT assay, in which T cells from a donor sample are cultured in the presence of an antigen. Antigenicity refers to the ability of a molecule to bind specifically to immune receptors such as antibodies or T-cell receptors, whereas immunogenicity is the ability to elicit an actual immune response; assays like ELISPOT measure T-cell immunogenicity, as they detect functional cytokine release following antigen recognition. This assay allows assessing the reactivity of antigen-specific T cells by counting cells that

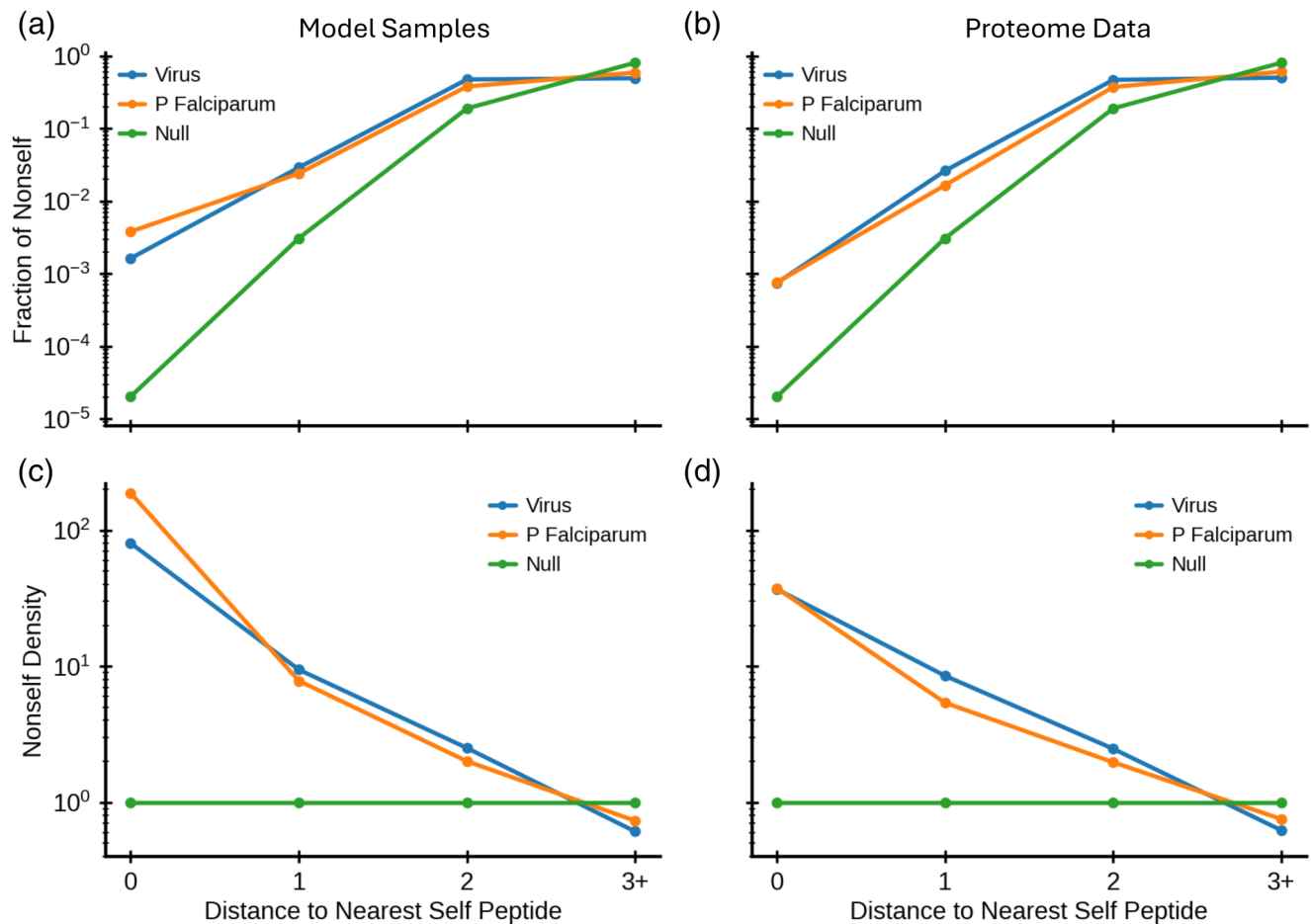


FIG. 15. Distance to nearest self-peptide for model samples versus actual data. (a), (b) Distribution of distances to the nearest self-peptide for peptides from samples drawn from models (a) and actual data (b). (c), (d) Relative density of distances to the nearest self-peptide for the same sets of peptides. Despite the *P. falciparum* proteome having a relatively large KL divergence to the human proteome, its nearest neighbor distribution looks very similar to that of the human virome.

activate in response to the antigen and secrete cytokines. The ability of an epitope to elicit an immune response as measured by such an assay is commonly called the immunogenicity of the epitope.

To perform our analyses, we downloaded the full T-cell reactivity dataset from IEDB [77]. We extracted all epitopes presented on human MHC class I, with a length of nine amino acids and not annotated as coming from the host. We excluded epitopes of foreign origin tested in the context of autoimmune disease, allergy, and cancer, as these might introduce a sampling bias toward “special” peptides close to self but immunogenic. We define epitopes as immunogenic, if their recorded qualitative score was “Positive,” “Positive-Low,” or “Positive-High.” In cases, for which multiple independent measurements were available we only retained epitopes with concordant immunogenicity assessment across studies.

We aligned foreign epitopes to their nearest self-peptide by Hamming distance as described above. The number of immunogenic and nonimmunogenic peptides for each distance bin are shown in Table II.

We performed a χ^2 -test between peptides that are distance ≤ 2 and peptides that are distance > 2 as implemented in the Python package statsmodels v0.12.2. For Fig. 4(a), $651/2550 = 25.5\%$ of the “close” peptides were immuno-

genic, compared to only $700/3515 = 19.9\%$ of the “far” peptides. This revealed a statistically significant drop in immunogenicity for far peptides (at Hamming distance 3+) compared to close ones at distance 2 or less. This suggests the immune system naturally recognizes foreign epitopes close to self, even when farther peptides, more dissimilar from self, are available to be recognized. The same qualitative finding is also supported by Ref. [51], in which the analysis of epitope sequence similarity was restricted to T-cell exposed motifs. IEDB reports results from multiple different experimental techniques, and we found evidence that data might not be comparable across experimental approaches. Specifically, we found that assays differed greatly in the fraction of reported immunogenic epitopes (Fig. 17). The analysis reported in

TABLE II. IEDB T-cell epitopes.

Distance	Count	Immunogenic count
0	22	1
1	107	28
2	2421	622
3	3496	694
4+	14	6

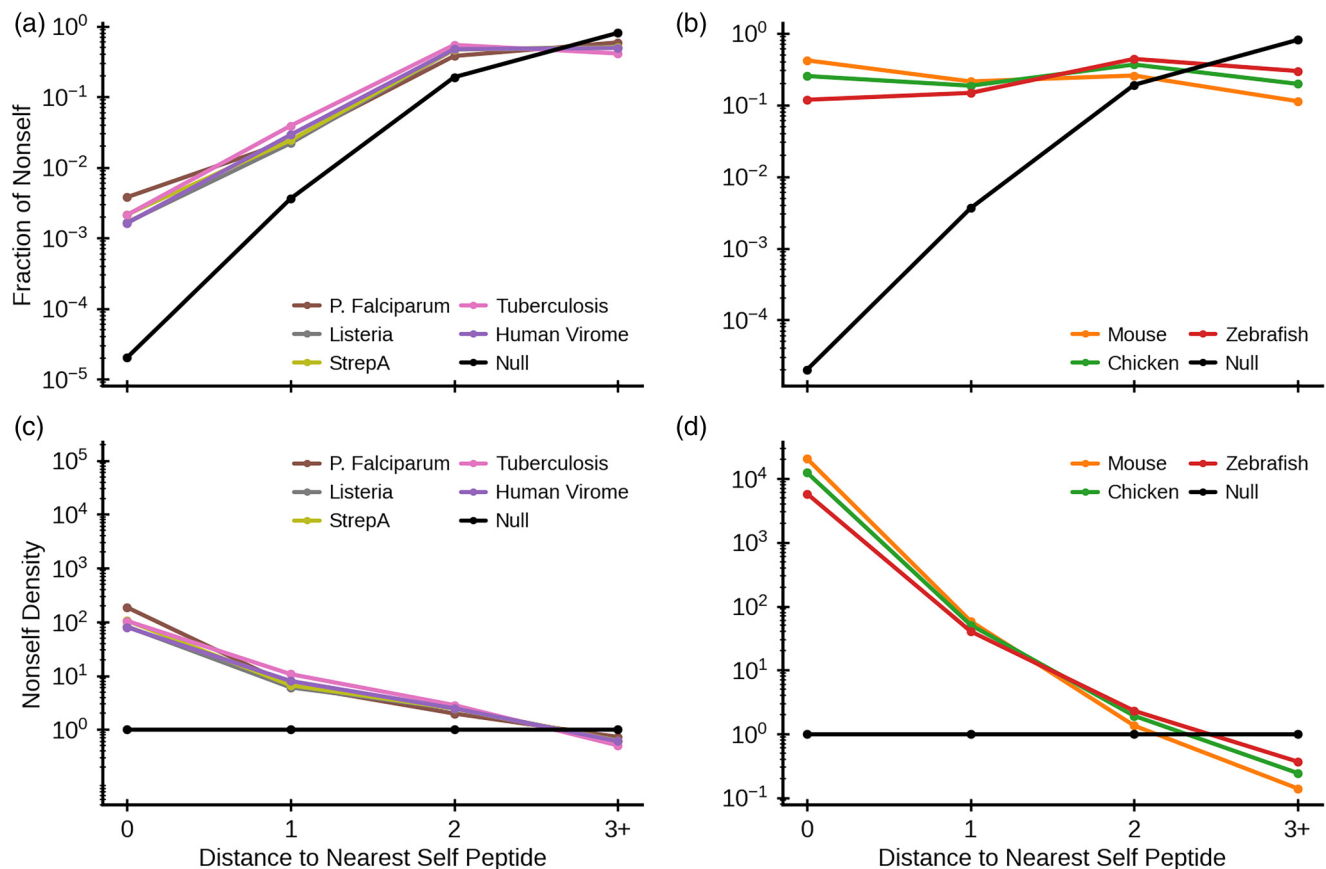


FIG. 16. Distance to nearest self-peptide for different proteomes. (a), (b) Distribution of distances to the nearest self-peptide for peptides from different pathogen (a) and vertebrate (b) proteomes, and a null model with a uniform distribution over all 20^9 9-mers (black). (c), (d) Relative density of distances to the nearest self-peptide for the same sets of peptides. Colors as in panels (a), (b). The maximum entropy models are sufficient to closely predict the overlap between human and pathogen proteomes. However, the overlap between human and other vertebrate proteomes (c), (d) is much higher due to the existence of many phylogenetically related proteins, which is not captured in our model.

Fig. 4 report results from a single assay, the ELISPOT assay, which is the most commonly reported within the database. An analysis of data from other assays shows less clear signal [Figs. 17(b)–17(f)]. Interestingly, we find that epitopes with exact matches in the self-proteome are largely nonimmunogenic when assessed by ELISPOT but find multiple instances of such epitopes annotated as immunogenic for other assays. This is in line with experimental evidence that ELISPOT assays are uniquely sensitive to peripheral tolerance mechanisms, as supported by increased responses following depletion of regulatory T cells [78]. We also find that low T-cell cross-reactivity distance of epitopes with single amino acid substitutions makes epitopes more likely (although not statistically significant) to be assessed as immunogenic by non-ELISPOT assays [Fig. 17(c)], in contrast to the expected, reverse trend observed in ELISPOT data [Fig. 4(b)].

3. *Plasmodium falciparum* epitopes

While the distributions of vertebrate and viral proteomes are largely similar to the human proteome (Fig. 2), the proteome of the malaria parasite *P. falciparum*, an organism known to have an unusually AT-rich genome [35], has a larger divergence. To understand how the immunogenicity of foreign epitopes from this parasite depends on their distance to the

nearest self-peptides we analyzed data from IEDB annotated as being from human host with “Parent Species” = “*Plasmodium falciparum*.” We filtered to MHC Class-I epitopes as above, and additionally for any MHC annotation that was missing in the database, we manually looked at the original reference and kept the epitope if the original contributing literature specified that the epitope was for a class-I HLA. This resulted in 151 positive epitopes, 60 of which were based on ELISPOT measurements.

A limitation of our analysis is that it does not directly address the high diversity of the parasite. Our analysis relies on the reference proteome on Uniprot for *P. falciparum*, and on the annotated epitopes from IEDB. Unfortunately, we did not have independent information on different strains of the parasite, which could provide important and distinct evolutionary information in different regions/settings. However, we do not expect the main results to differ between different strains, as the models are learning more general sequence features that are largely shared across strains of the parasite.

4. *Plasmodium falciparum* protein abundance

Protein abundance data was downloaded from Pax-DB [48] at Ref. [79]. This data was mapped to IEDB data using the Uniprot identifier map [80].

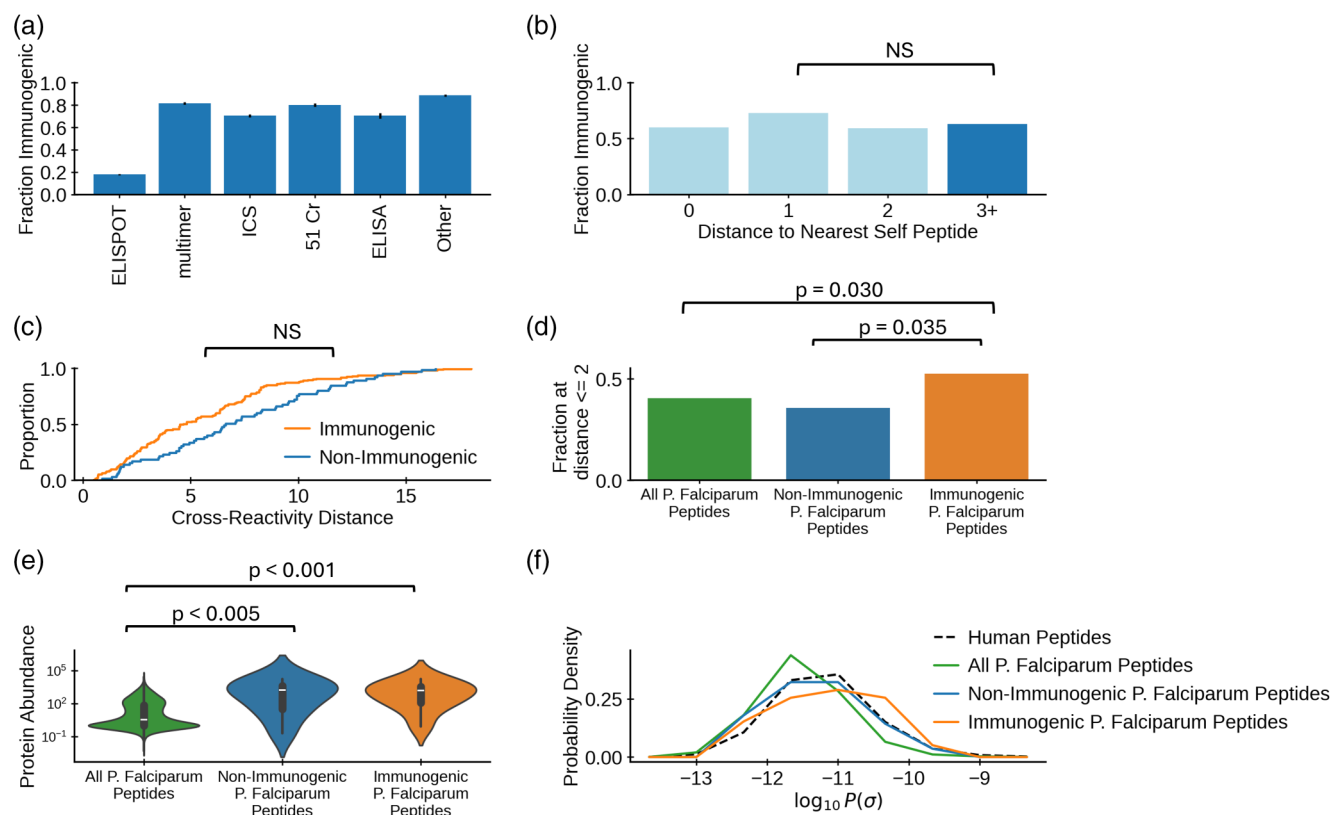


FIG. 17. Immunogenicity for ELISPOT vs other assays. (a) Fraction of peptides annotated to be immunogenic differs strongly between assays, potentially confounding analyses of immunogenicity. Assays: ELISPOT—enzyme-linked immunosorbent spot assay, multimer—pMHC multimer binding assay, ICS—intracellular cytokine staining assay, 51 Cr—Chromium-51 release assay, ELISA—enzyme-linked immunosorbent assay. We thus split the analysis into peptides measured via ELISPOT alone (Fig. 4) and all available assays (b)–(f). (b) Immunogenicity (as ascertained using all available assays) as a function of Hamming distance to the nearest self-peptide. Immunogenicity did not differ significantly between distance bins (χ^2 test). (c) Empirical cumulative distribution functions of T-cell cross-reactivity distances [7] for immunogenic (orange) and nonimmunogenic (blue) epitopes differing from the nearest self-peptide by a single amino acid. For hamming-distance-one peptides, cross-reactivity distances for immunogenic epitopes (as ascertained using all available assays) do not differ significantly from nonimmunogenic epitopes (Kolmogorov–Smirnov Test). (d) Fraction of different sets of peptides from the *P. falciparum* proteome within distance 2 of the nearest self-peptide. Fraction with distance ≤ 2 differs between immunogenic peptides (orange), nonimmunogenic peptides (blue), and all peptides (green) (χ^2 test). (e) Average protein abundance of the source protein of different sets of peptides from the *P. falciparum* proteome. Abundance significantly differed between the source proteins for IEDB epitopes (blue and orange) and the background distribution of all peptides (green). Abundance did not differ between immunogenic (orange) and nonimmunogenic peptides (blue) (All assays, Mann-Whitney *U*-test). (f) Distribution of (log) probabilities given by the human proteome maximum entropy model for different sets of peptides from the *P. falciparum* proteome. Data: Foreign epitopes from IEDB [42] presented by MHC-I tested by any assay as described in Appendix E. Protein abundance data is from Pax-DB [48] (see Appendix E).

5. Cross-reactivity distance

For hamming-distance-one epitopes, we calculated cross-reactivity distance as previously [7]. Publicly available code to compute this metric can be found at Ref. [81]. Cross-reactivity distances were compared

between immunogenic and nonimmunogenic hamming-distance-one epitopes using their empirical cumulative distributions and significance was tested using a two-sample Kolmogorov–Smirnov as implemented in *scipy*’s *ks_2samp* function.

- [1] T. N. Schumacher and R. D. Schreiber, Neoantigens in cancer immunotherapy, *Science* **348**, 69 (2015).
- [2] W. Yu *et al.*, Clonal deletion prunes but does not eliminate self-specific $\alpha\beta$ CD8+ T lymphocytes, *Immunity* **42**, 929 (2015).
- [3] M. Łuksza *et al.*, A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy, *Nature (London)* **551**, 517 (2017).
- [4] H. Matsushita *et al.*, Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting, *Nature (London)* **482**, 400 (2012).
- [5] A. Ribas and J. D. Wolchok, Cancer immunotherapy using checkpoint blockade, *Science* **359**, 1350 (2018).
- [6] D. K. Wells *et al.*, Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction, *Cell* **183**, 818 (2020).

- [7] M. Łuksza *et al.*, Neoantigen quality predicts immunoediting in survivors of pancreatic cancer, *Nature (London)* **606**, 389 (2022).
- [8] L. A. Rojas *et al.*, Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer, *Nature (London)* **618**, 144 (2023).
- [9] P. Mehta *et al.*, A high-bias, low-variance introduction to machine learning for physicists, *Phys. Rep.* **810**, 1 (2019).
- [10] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborova, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [11] T. C. Butler, M. Kardar, and A. K. Chakraborty, Quorum sensing allows T cells to discriminate between self and nonself, *Proc. Natl. Acad. Sci. USA* **110**, 11833 (2013).
- [12] F. P. Legoux *et al.*, CD4+ T cell tolerance to tissue-restricted self antigens is mediated by antigen-specific regulatory T cells rather than deletion, *Immunity* **43**, 896 (2015).
- [13] M. M. Davis, Not-so-negative selection, *Immunity* **43**, 833 (2015).
- [14] I. M. N. Wortel, C. Kesmir, R. J. De Boer, J. N. Mandl, and J. Textor, Is T cell negative selection a learning algorithm? *Cells* **9**, 690 (2020).
- [15] E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.* **106**, 620 (1957).
- [16] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, A maximum entropy approach to natural language processing, *Comput. Linguist.* **22**, 39 (1996).
- [17] A. Radford *et al.*, Language models are unsupervised multitask learners, *OpenAI blog* **1**, 9 (2019).
- [18] A. Madani *et al.*, Large language models generate functional protein sequences across diverse families, *Nat. Biotechnol.* **41**, 1099 (2023).
- [19] Y. Nagano *et al.*, Contrastive learning of T cell receptor representations, *Cell Syst.* **16**, 101165 (2025).
- [20] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, *Nature (London)* **440**, 1007 (2006).
- [21] O. Moaz, G. Tkačik, M. Esteki, R. Kiani, and E. Schneidman, Learning probabilistic neural representations with randomly connected circuits, *Proc. Natl. Acad. Sci. USA* **117**, 25066 (2020).
- [22] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, Successes and failures of simplified models for a network of real neurons, *arXiv:2112.14735*.
- [23] L. Meshulam and W. Bialek, Statistical mechanics for networks of real neurons, *Rev. Mod. Phys.* **97**, 045002 (2025).
- [24] A. Lapides, B. Giraud, L. Liu, and G. Stormo, A maximum entropy formalism for disentangling chains of correlated sequence positions, *Proc. IMS* **236** (1998).
- [25] W. Bialek and R. Ranganathan, Rediscovering the power of pairwise interactions, *arXiv:0712.4397*.
- [26] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing, *Proc. Natl. Acad. Sci. USA* **106**, 67 (2009).
- [27] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, *Rep. Prog. Phys.* **81**, 032601 (2018).
- [28] D. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation, *PLoS ONE* **6**, e28766 (2011).
- [29] W. Bialek *et al.*, Statistical mechanics for natural flocks of birds, *Proc. Natl. Acad. Sci. USA* **109**, 4786 (2012).
- [30] J. Lehmann, A. Libchaber, and B. D. Greenbaum, Fundamental amino acid mass distributions and entropy costs in proteomes, *J. Theor. Biol.* **410**, 119 (2016).
- [31] D. H. Ackley, G. E. Hinton, and J. Sejnowski, A learning algorithm for Boltzmann machines, *Cognitive Science* **9**, 147 (1985).
- [32] J. Marchi *et al.*, Size and structure of the sequence space of repeat proteins, *PLoS Comput. Biol.* **15**, e1007282 (2019).
- [33] D. Hoyos *et al.*, Fundamental immune–oncogenicity trade-offs define driver mutation fitness, *Nature (London)* **606**, 172 (2022).
- [34] T. M. Cover, *Elements of Information Theory* (John Wiley & Sons, New York, NY, 1999).
- [35] W. L. Hamilton *et al.*, Extreme mutation bias and high AT content in *Plasmodium falciparum*, *Nucleic Acids Res.* **45**, 1889 (2017).
- [36] C. Maguire *et al.*, Molecular mimicry as a mechanism of viral immune evasion and autoimmunity, *Nat. Commun.* **15**, 9403 (2024).
- [37] J. J. Calis *et al.*, Properties of MHC class I presented peptides that enhance immunogenicity, *PLoS Comput. Biol.* **9**, e1003266 (2013).
- [38] M. Yarmarkovich, J. M. Warrington, A. Farrel, and J. M. Maris, Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity, *Cell Rep. Med.* **1**, 100036 (2020).
- [39] S. K. Mohanty, S. K. Mohanty, and K. S. Leela, *Textbook of Immunology* (J&P Medical Ltd., Vienna, 2013).
- [40] H. B. Barlow *et al.*, Possible principles underlying the transformation of sensory messages, *Sens. Commun.* **1**, 217 (2012).
- [41] W. Bialek, Thinking about the brain, in *Physics of Biomolecules and Cells. Physique Des Biomolécules et Des Cellules*, edited by F. Flyvbjerg, F. Jülicher, P. Ormos, and F. David (Springer, Berlin, Heidelberg, 2002), pp. 485–578.
- [42] R. Vita *et al.*, The immune epitope database (IEDB): 2018 update, *Nucleic Acids Res.* **47**, D339 (2019).
- [43] F. Prugnolle *et al.*, Pathogen-driven selection and worldwide HLA class I diversity, *Curr. Biol.* **15**, 1022 (2005).
- [44] C. C. Brown and A. Y. Rudensky, Spatiotemporal regulation of peripheral T cell tolerance, *Science* **380**, 472 (2023).
- [45] Y. K. Kohanim, A. Tendler, A. Mayo, N. Friedman, and U. Alon, Endocrine autoimmune disease as a fragility of immune surveillance against hypersecreting mutants, *Immunity* **52**, 872 (2020).
- [46] J.-B. Lalanne and P. François, Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism, *Proc. Natl. Acad. Sci. USA* **112**, 1898 (2015).
- [47] P. François and G. Altan-Bonnet, The case for absolute ligand discrimination: Modeling information processing and decision by immune T cells, *J. Stat. Phys.* **162**, 1130 (2016).
- [48] Q. Huang, D. Szklarczyk, M. Wang, M. Simonovic, and C. von Mering, Paxdb 5.0: Curated protein quantification data suggests adaptive proteome changes in yeasts, *Mol. Cell. Proteomics* **22**, 100640 (2023).
- [49] J. T. George, D. A. Kessler, and H. Levine, Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides, *Proc. Natl. Acad. Sci. USA* **114**, E7875 (2017).

- [50] N. Vrisekoop, J. P. Monteiro, J. N. Mandl, and R. N. Germain, Revisiting thymic positive selection and the mature T cell repertoire for antigen, *Immunity* **41**, 181 (2014).
- [51] B. Koncz *et al.*, Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself, *Proc. Natl. Acad. Sci. USA* **118**, e2100542118 (2021).
- [52] C. P. Ratliff, B. G. Borghuis, Y.-H. Kao, P. Sterling, and V. Balasubramanian, Retina is structured to process an excess of darkness in natural scenes, *Proc. Natl. Acad. Sci. USA* **107**, 17368 (2010).
- [53] M. Polonsky *et al.*, Induction of CD4 T cell memory by local cellular collectivity, *Science* **360**, eaaj1853 (2018).
- [54] H. W. Kessels *et al.*, The impact of self-tolerance on the polyclonal CD8+ T cell repertoire, *J. Immunol.* **172**, 2324 (2004).
- [55] B. D. Greenbaum, S. Cocco, A. J. Levine, and R. Monasson, Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses, *Proc. Natl. Acad. Sci. USA* **111**, 5054 (2014).
- [56] P. Šulc *et al.*, Repeats mimic pathogen-associated patterns across a vast evolutionary landscape, *Cell Genomics* **5**, 101011 (2025).
- [57] J. D. Farmer, N. H. Packard, and A. S. Perelson, The immune system, adaptation, and machine learning, *Physica D* **22**, 187 (1986).
- [58] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, How a well-adapted immune system is organized, *Proc. Natl. Acad. Sci. USA* **112**, 5950 (2015).
- [59] A. Mayer, V. Balasubramanian, A. M. Walczak, and T. Mora, How a well-adapting immune system remembers, *Proc. Natl. Acad. Sci. USA* **116**, 8815 (2019).
- [60] H. W. Lin and M. Tegmark, Criticality in formal languages and statistical physics, *Entropy* **19**, 299 (2017).
- [61] https://github.com/jonlevi/self_not_self.
- [62] <https://github.com/andim/peptidome>.
- [63] The Uniprot-Consortium, UniProt: The universal protein knowledgebase in 2021, *Nucleic Acids Res.* **49**, D480 (2021).
- [64] G. Schwarz, Estimating the dimension of a model, *Ann. Stat.* **6**, 461 (1978).
- [65] M. Stone, Comments on model selection criteria of Akaike and Schwarz, *J. R. Stat. Soc. B* **41**, 276 (1979).
- [66] <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html#sklearn.manifold.MDS>.
- [67] https://github.com/jonlevi/self_not_self?tab=readme-ov-file#downloading-the-data.
- [68] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087 (1953).
- [69] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97 (1970).
- [70] E. H. Simpson, Measurement of diversity, *Nature (London)* **163**, 688 (1949).
- [71] P. R. Hunter and M. A. Gaston, Numerical index of the discriminatory ability of typing systems: An application of Simpson's index of diversity, *J. Clin. Microbiol.* **26**, 2465 (1988).
- [72] F. F. Gonzalez-Galarza *et al.*, Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools, *Nucleic Acids Res.* **48**, D783 (2019).
- [73] L. Gragert, A. Madbouly, J. Freeman, and M. Maiers, Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry, *Human Immunology* **74**, 1313 (2013).
- [74] M. Andreatta and M. Nielsen, Gapped sequence alignment using artificial neural networks: Application to the MHC class I system, *Bioinformatics* **32**, 511 (2016).
- [75] T. Chotisayuth and A. Tiffeau-Mayer, Lightning-fast adaptive immune receptor similarity search by symmetric deletion lookup, *arXiv:2403.09010*.
- [76] <https://github.com/andim/pyrepseq>.
- [77] http://www.iedb.org/downloader.php?file_name=doc/tcell_full_v3.zip.
- [78] A. Bonertz *et al.*, Antigen-specific Tregs control T cell responses against a limited repertoire of tumor antigens in patients with colorectal carcinoma, *J. Clin. Invest.* **119**, 3311 (2009).
- [79] <https://pax-db.org/dataset/5833/3599693021/>.
- [80] <https://pax-db.org/downloads/5.0/paxdb-uniprot-links-v5.0.zip>.
- [81] <https://github.com/LukszaLab/NeoantigenEditing>.