Flexible machine learning prediction of antigen presentation for rare and common HLA-I alleles

Barbara Bravi[a], Jérôme Tubiana[b], Simona Cocco[a], Rémi Monasson[a], Thierry Mora[a], and Aleksandra M. Walczak[a]

[a]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France
[b]Blavatnik School of Computer Science, Tel Aviv University, Israel

The recent increase of immunopeptidomic data, obtained by mass spectrometry (MS) or binding assays, opens unprecedented possibilities for investigating endogenous antigen presentation by the highly polymorphic human leukocyte antigen class I (HLA-I) protein. We introduce a flexible and easily interpretable peptide presentation prediction method. We validate its performance as a predictor of cancer neoantigens and viral epitopes and use it to reconstruct peptide motifs presented on specific HLA-I molecules.

Recognition of malignant and infected cells by the adaptive immune system requires binding of cytotoxic T-cell receptors to antigens, 8-11-mer peptides presented by the Major Histocompatibilty Complex (MHC) class I coded by HLA-I alleles (Fig. 1a). Tumour-specific neoantigens are currently sought-after targets for improving cancer immunotherapy [1, 2]. Computational predictions can help select potential neoantigens and accelerate immunogenicity testing. To be useful, these predictions must be specific to each HLA type.

State-of-the-art methods [3–5], such as NetMHC [6, 7], are based on artificial neural networks trained in a *supervised* way to predict peptide presentation from known peptide-HLA association. They must be trained on large datasets, and perform best on common alleles. Their accuracy is degraded for rare or little studied HLA-I alleles which are poorly represented in databases. In that case, another approach is to train *unsupervised* models of presentation from custom elution experiments with little or no information about peptide-HLA association. For instance, MixMHCp [8, 9] can reconstruct, from unannotated peptide sequences, a mixture of generative models - one for each expressed HLA type. However, it makes simplifying assumptions about binding specificity, and is not designed to leverage available (albeit limited) annotation information from the Immune Epitope Database [10] (IEDB) to improve accuracy.

We present an alternative method for predicting peptides presented by specific HLA types, which can be trained on custom datasets. It can be applied to patient- or experiment-specific samples and treats common and rare alleles on equal footing, avoiding potential database biases. We use a Restricted Boltzmann Machine (RBM), an unsupervised machine learning scheme that learns probability distributions of sequences given as input [11–13]. The RBM estimates presentation scores for each peptide, and can generate candidate presentable peptides. It also provides a lower dimensional representation of peptides with a clear interpretation in terms of associated HLA type, which can be exploited to train classifiers of HLA association from a small number of HLA annotations. The RBM has a simple structure with one hidden layer and weights connecting the input - the peptide sequence - to the hidden layer. The weights, along with the biases acting on both input and hidden units, are learned from a list of presented peptides (see Methods, Fig. 1b, Supplementary Figs. 1, 2, 3). The RBM probability is interpreted as a probabilistic score of antigen presentation by any of the 6 HLA-I in the experiment. We applied standard alignment routines to reduce sequences to the most common length of 9 residues (see Methods).

As a first validity check, we built an RBM model to predict presentation by a single common allele, HLA-A*02:01, from MS data annotated with this HLA restriction in IEDB. We tested the model's ability to discriminate presented peptides from "generic" peptides (randomly drawn from the human proteome). We calculated a Receiver Operating Characteristic (ROC) curve from RBM presentation scores assigned to a test set of presented versus generic peptides (Fig. 1c). The Area Under the Curve of the receiving operating characteristic, AUC = 0.973, proves the RBM predictive power at recovering presented antigens. To further probe the applications of RBM to neoantigen discovery, we considered the missense mutations in 5 ovarian and melanoma cancer cell lines, filtered by expression levels and by the NetMHC-predicted affinity to HLA*A:02:01 [14]. We attributed RBM presentation scores to all 8-11-mer peptides harbouring these mutations. RBM performed comparably to NetMHCpan-4.0 [7], the state-of-the-art method for antigen prediction: 15 out of 16 neoantigens that were MS-validated as HLA-I bound are ranked by the RBM in the

top 4% among generic peptides (mean percentile 1.4% versus 1.7% for NetMHC), and the top 10% among the mutated peptides (mean 4.2% versus 2.5% for NetMHC) considered for the corresponding cell line, Fig. 1c-d and Methods.

Next, we applied the RBM to complex data on which it can show its efficiency: multi-allelic eluted samples where peptides may be bound to up to 6 different HLA-I proteins. The goal is to reconstruct binding specificity motifs, *i.e.* classify epitopes by HLA preference. To test performance in a case where the ground truth is known, we trained RBMs on 10 synthetic "single-individual" samples pooling IEDB antigens presented by 6 HLA-I proteins per individual, covering 43 different alleles in total (12 HLA-A, 17 HLA-B, 14 HLA-C, see Methods). Each HLA-I recognizes specific sequence "motifs" (preferred residues at each position) in the presented peptides. RBM hidden units allow for the mapping of peptide sequences onto a lower dimensional "representation space" (see Methods) which organizes them into clusters by motif, reflecting different HLA-binding specificities present in the sample (Fig. 1b). The data-derived RBM parameters underlying that representation play a key role in capturing what amino acids contribute to these binding motifs (Supplementary Fig. 4).

The low-dimensional representation of peptides (Fig. 1b) suggests an efficient way to classify them, using the annotation of a small number of sequences with their HLA preference. In practice these annotated sequences may come from other experiments or public databases. In our synthetic example, we randomly selected 10% of peptides and labeled them with their associated HLA-I. Using these annotations, we trained a linear classifier to predict the HLA-I restriction of individual peptides from their RBM representation (see Methods, Fig. 1b, Supplementary Figs. 1, 5). We refer to this architecture, which combines the RBM (trained on unannotated peptides) *and* the HLA-I type classifier (trained on a few annotated ones), as RBM-MHC. The performance of RBM-MHC at predicting HLA association, as measured by AUC = 0.993, is excellent (see Fig. 1e and Methods). The RBM representation is crucial for achieving high prediction performance. Training a linear HLA-I classifier directly on the annotated sequences (rather than their RBM representation) yields a much poorer performance (AUC=0.724). On the other hand, a completely unsupervised clustering algorithm (K-means [15]) applied to unlabeled data in RBM representation space ('RBM-km') performed well (AUC=0.935), while K-means applied directly to sequences did not (AUC=0.645, Fig. 1e and Methods).

RBM-MHC outperforms MixMHCp [8,9] in terms of both accuracy and stability across the 10 synthetic datasets, in part thanks to the 10% labeled data it exploits. The RBM also captures global sequence correlations that MixMHCp's independent-site models miss, which are important for correctly classifying antigens across alleles with similar binding motifs. The major drop in MixMHCp performance occurs precisely in datasets that mix same-supertype alleles [16] (HLA-B*27:05 and HLA-B*39:01 belonging to supertype B27, HLA-B*40:01 and HLA-B*44:03 belonging to B44, see Supplementary Table 1). RBM-MHC also performs better than NetMHCpan-4.0 [7], which is trained on a separate, IEDB-derived dataset to predict binding to each of the 6 alleles in our datasets. The gain in performance is slightly more pronounced for 9-mers-only samples than for 8-11-mers (Supplementary Figs. 6, 7), suggesting room for improvement in accounting for variable lengths. Furthermore, the structure and interpretation of the RBM representation space allows for the generation of new, artificial peptides from the model with controled HLA-binding specificity (see Supplementary Note, Supplementary Fig. 14).

To assess the relevance of our approach in a clinical setting, we considered single-patient, melanoma-associated immunopeptidomic datasets [17–19], complemented with patient HLA typing and Whole Exome Sequencing (WES) of tumorous cells, where in total 11 neoantigens were identified. We tested the the RBM-MHC approach for motif reconstruction. Since in this case true peptide-HLA associations are not known, model performance is evaluated through the correlation between the predicted motifs and motifs reconstructed from IEDB monoallelic data (Fig. 2a-b, Supplementary Figs. 8, 9, 10). RBM-MHC, MixMHCp and NetMHCpan4.0 perform comparably, with the only exception of Mel15, partially explained by the limited number of available labels (for HLA-A*68:01 and HLA-B*35:03). RBM-MHC and its unsupervised version RBM-km, when trained on the patient dataset (see Methods) systematically predicted the correct HLA association for all neoantigens. They also assigned a top 1% score among all WES mutants to 8 out of 11 identified neoantigens (Supplementary Table 2).

Beyond their usefulness for predicting HLA specificity, RBM hidden units serve as a tool to discover additional features *within* the same HLA-binding specificity. A compelling case is the fast identification of subclasses in HLA-B*51:01 ligands corresponding to the two structurally alternative binding modes validated in [9] (Fig. 2c, Supplementary Fig. 4).

In conclusion, RBM-MHC efficiently exploits the statistical information of unannotated, moderate-size datasets of antigen presentation available from clinical studies, and combines it with small amounts of antigen annotations (as few as < 100, typical of poorly represented HLA-I alleles) to deliver accurate and stable

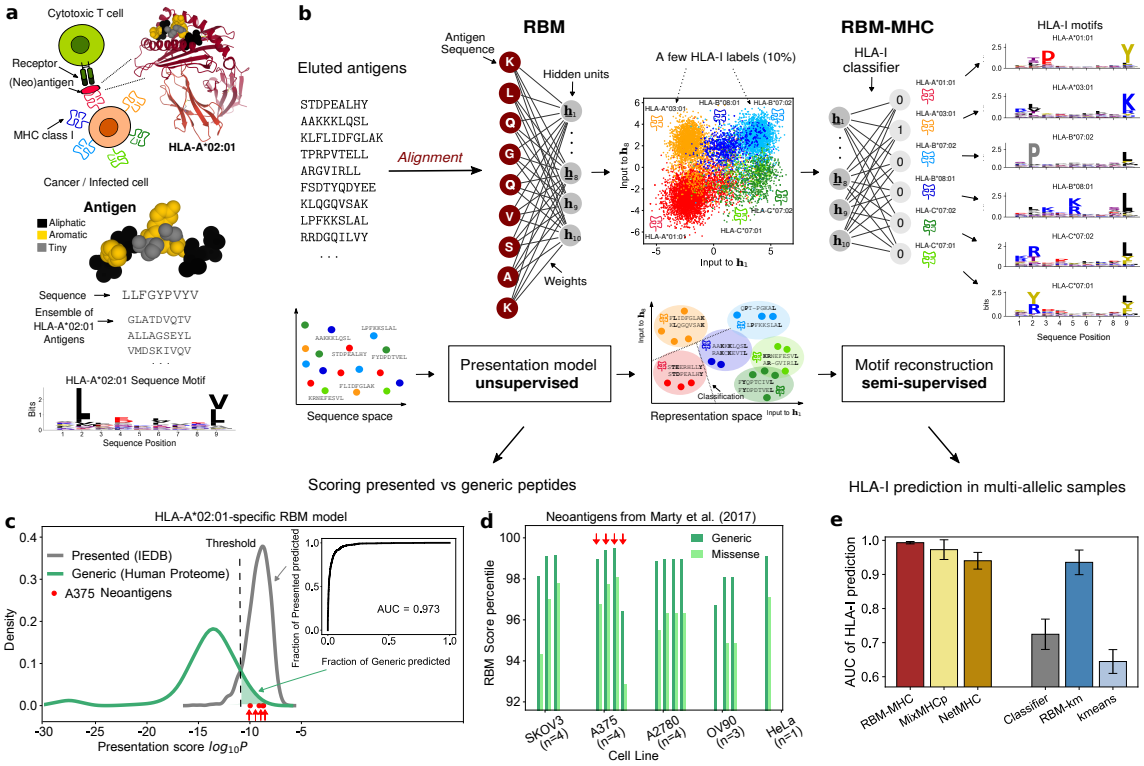


Figure 1: **RBM-MHC approach to HLA-I antigen presentation.** **a**, Antigens binding to HLA match specific "sequence motifs" represented by logos (example of Tax-HLA-A*02:01 complex structure, PDB-ID:1BD2, Mol* image [22]). **b**, RBM-MHC workflow. RBM projects sequences onto a representation in terms of "hidden units" (we selected 2 for illustration) by the set of weights. Data: IEDB-derived dataset for Haplotype 1, Supplementary Table 1. **c**, RBM scores for presented vs generic peptides on HLA-A*02:01 and classification performances assessed by ROC. **d**, Score percentiles on the same allele of validated neoantigens from 5 cancer cell lines [14]. **e**, Performance (AUC) of 6 methods for HLA-I prediction on on 9-mers in 10 synthetic-individual samples, each carrying 6 HLA-I covering A,B,C allels (see Supplementary Table 1). Bars are standard deviations over 10 datasets.

predictions of HLA-I binding. Our results show that the approach is useful for systematic applications with newly produced large-scale datasets [5]. We plan to develop an extension to HLA class II presentation. As the MS amino acid frequency distribution is potentially affected by biases of the technique, we also designed a sequence re-weighting scheme to avoid overfitting and better score ligands tested *in vitro* (Supplementary Note, Supplementary Fig. 13). We stress that the method is not limited to MS datasets. As a timely application of the approach to SARS-CoV-2, we showed that affinity-trained RBM models recover with high probability SARS-CoV-2 HLA-I epitopes validated *in vitro* [20] (Supplementary Fig. 11). MS-trained RBM-MHC outperforms NetMHC versions [6,7] used in Ref. [20] for rare HLA-C alleles (Fig. 2d), although the validated peptides were selected by NetMHC itself. RBM-MHC also assigns high presentation scores to SARS-CoV-2 epitopes that are homologous to experimentally validated SARS-CoV cytotoxic T-cell epitopes [21], on par with NetMHC4.0 [6] (Fig. 2e) and NetMHCpan-4.0 [7] (Supplementary Fig. 12). In general, our method is designed to be trained on custom samples, which could be of relevance to produce sample-specific insights about the complexity of endogenous HLA-I presentation.

Acknowledgment

We thank Benjamin Greenbaum for suggesting to look at SARS-CoV-2, and helpful exchanges with Clément Roussel, Andrea Di Gioacchino, Carlos Olivarez, Hannah Carter, David G Feller, Shelly Kalaora, Yardena Samuels, David Hoyos, Jayon Lihm. This work was partially supported by Stand Up to Cancer, the European Research Council Consolidator Grant n. 724208, the ANR-17 RBMPro and ANR-19 Decrypted CE30-0021-01 grants. J.T. was supported by a postdoctoral fellowship from the Human Frontier Science Program Organization (reference number: LT001058/2019).

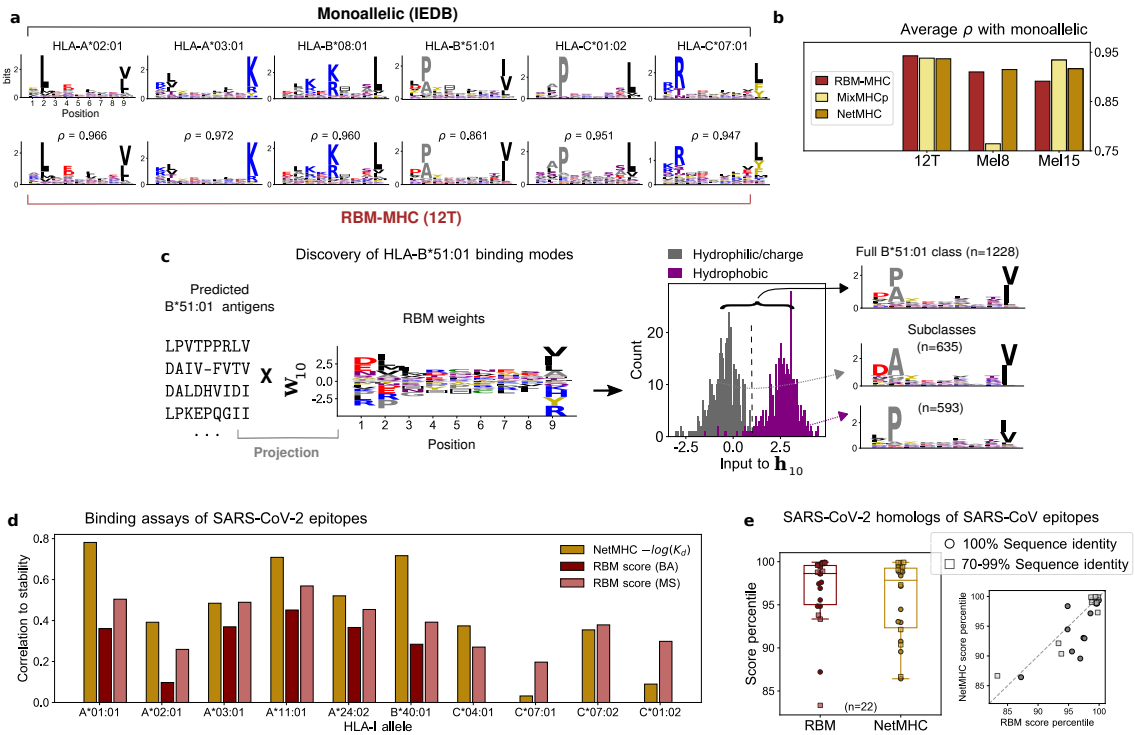


Figure 2: **Predictions for cancer samples and SARS-CoV-2 epitopes**. **a**, Sequence logos of clusters found with RBM-MHC trained on melanoma-associated sample 12T [17, 18] and IEDB monoallelic data; ρ shows Pearson correlation between respective amino acid frequencies. **b**, Performances (average ρ over clusters) for 3 samples from [17–19]. MixMHCp merges Mel8 antigens specific to the 2 HLA-C into the same cluster, causing a drop in the average ρ (see Supplementary Fig. 9). **c**, RBM distinguishes 2 subclasses in HLA-B*51:01-binding antigens. The HLA-B*51:01-peptide bond can be established [9]: *(i)* via the interaction of the HLA-B*51:01 residue 62 with peptide position 2, requiring a hydrophilic residue there, typically Alanine (A), and a polar or negatively charged residue, typically Aspartic Acid (D), at position 1; *(ii)* with R62 sidechain facing the solvent, requiring co-occurrence at positions 1-2 of hydrophobic residues, typically Proline (P), at position 2). Inspection of the inputs to the 10th hidden unit ($\boldsymbol{h}_{10}$), found by projecting peptides predicted by RBM-MHC as HLA-B*51:01-specific onto the corresponding weights (Methods, eq. 4), reveals a bimodal distribution, enabling the discrimination of the "hydrophilic/charged" pattern *(i)* from the "hydrophobic" pattern *(ii)*, as recapitulated by the sequence logos. **d**, Correlation of experimental stability of SARS-CoV-2 epitopes [20] with scores from NetMHC ($logK_d$ - dissociation constant [nM]) and RBM trained on Binding Affinity (BA) and Mass Spectroscopy (MS) data. **e**, Score percentiles (relative to all SARS-CoV-2 9-mers peptides) of homologs of dominant SARS-CoV epitopes. Boxplots indicate median, upper and lower quartiles.

Methods

RBM-MHC algorithm. We use RBM-MHC to predict antigen sequences presented on specific HLA types. The RBM-MHC architecture (Fig. 1b and Supplementary Fig. 1) is composed of two parts, which are trained in successive steps. The first, a Restricted Boltzmann Machine, is learned in an unsupervised way, while the second, a HLA-I classifier, is fully supervised but only by a minimal amount of known antigen-HLA-I associations (semi-supervised).

Restricted Boltzmann Machine (RBM). The RBM learns an underlying probability model for the antigen sequences in the training datasets, in our case the HLA-I presented antigens. A RBM [11, 12] consists of one layer of N^v observed units $\boldsymbol{v} = \{v_i\}_{i=1}^{N^v}$ and one layer of N^h hidden units $\boldsymbol{h} = \{h_\mu\}_{\mu=1}^{N^h}$ connected by weights $\boldsymbol{W} = \{w_{i\mu}\}$ (see Supplementary Fig. 1). Mathematically, the model is defined by a joint probability over presented antigen sequences and hidden units:

$$P(\boldsymbol{v},\boldsymbol{h}) \sim \exp\left(\sum_{i=1}^{N^v} g_i(v_i) - \sum_{\mu=1}^{N^h} \mathscr{U}_\mu(h_\mu) + \sum_{i,\mu} h_\mu w_{i\mu}(v_i)\right) \tag{1}$$

where $g_i(v_i)$ are local potentials acting on observed units, $\mathscr{U}_\mu(h_\mu)$ are local potential on hidden units and the weights $w_{i\mu}$ couple hidden and observed units. The parametric form of $\mathscr{U}_\mu(h_\mu)$ is chosen as a double Rectified Linear Units (dReLu):

$$\mathscr{U}_\mu(h) = \frac{1}{2}\gamma_{\mu,+}h_+^2 + \frac{1}{2}\gamma_{\mu,-}h_-^2 + \theta_{\mu,+}h_+ + \theta_{\mu,-}h_- \qquad h_+ = \max(h,0) \quad h_- = \min(h,0) \tag{2}$$

containing parameters $(\gamma_{\mu,+}, \gamma_{\mu,-}, h_+, h_-)$ to infer from data during training (see below). The dReLu was shown to outperform other choices of potential such as Gaussian [13, 23], guaranteeing that correlations beyond pairwise in data are captured. The probability distribution over the presented antigen sequences one is interested in modelling is recovered as the marginal probability over hidden units:

$$P(\boldsymbol{v}) = \int \prod_{\mu=1}^{N^h} dh_\mu P(\boldsymbol{v},\boldsymbol{h}) \sim \exp\left(\sum_{i=1}^{N^v} g_i(v_i) + \sum_{\mu=1}^{N^h} \Gamma_\mu\big(I_\mu(\boldsymbol{v})\big)\right) \tag{3}$$

where $\Gamma_\mu\big(I(\boldsymbol{v})\big) = \log \int dh\, e^{-\mathscr{U}_\mu(h)+hI}$. We define $I_\mu(\boldsymbol{v})$, the input to hidden unit μ coming from the observed sequence $\boldsymbol{v}$, as the sum of the weights entering that particular hidden unit:

$$I_\mu(\boldsymbol{v}) = \sum_i w_{i\mu}(v_i) \tag{4}$$

During the modelling step called "training", the weights $w_{i\mu}$, the local potentials $g_i(v_i)$ and the parameters specifying $\mathscr{U}_\mu(h_\mu)$ are inferred from data by maximizing, through stochastic gradient ascent, the average log-likelihood $\mathscr{L}^{RBM} = \langle \log P(\boldsymbol{v})\rangle_{data}$ of sequence data $\boldsymbol{v}$, as previously described [13,23]. This leads to inferring the RBM probability distribution that optimally reproduces the statistics of the training dataset. During training, the contribution of sequences to $\mathscr{L}^{RBM}$ can be re-weighted following a re-weigthing scheme we have conceived (see Supplementary Note) to correct for amino-acid frequency biases in the training dataset, as the ones introduced by mass spectrometry (MS) detection. A regularization, *i.e.* a penalty term over the weight parameters is introduced to prevent overfitting. The function maximized during training then becomes:

$$\mathscr{L}^{RBM} - \frac{\lambda_1^2}{2qN^v}\sum_\mu \Big(\sum_{i,v} |w_{i\mu}(v)|\Big)^2 \tag{5}$$

where $q = 21$ is the number of symbols appearing in observed sequences (20 amino acids and the alignment gap). This regularization, which was first introduced in [13], plays effectively the role of a L_1 regularization, imposing sparsity of weights, with a strength that is adapted to increasing magnitude of weights, hence favouring homogeneity among hidden units (see [13] for more detailed explanations). Examples of inferred sets of weights at different regularization strength λ_1^2 are provided in Supplementary Figs. 3-4. The package used for training, evaluating and visualising RBMs is an updated version of the one described in Refs. [13, 23, 24]. The package was ported to Python3 and the optimization algorithm was changed from SGD to RMSprop (*i.e.* to ADAM without momentum with parameters $l_r = 5x10^{-3}$, $\beta_1 = 0$, $\beta_2 = 0.99$, $\varepsilon = 10^{-3}$). The adaptive learning rate of RMSprop/ADAM result in larger updates for the fields and weights attached to rare amino acids, and hence speeds up convergence.

The HLA-A*02:01-specific RBM presentation model is trained with 5 hidden units and $\lambda_1^2 = 0.001$, the multi-allelic RBM models are trained with 10 hidden units and $\lambda_1^2 = 0.001$. The (hyper-)parametric search to select optimal λ_1^2 regularization and the number of hidden units was made from the trend of the log-likelihood on a held-out validation dataset, the aim being to achieve a good fit but to avoid overfitting (Supplementary Figs. 2a-b).

HLA-I Classifier. The HLA-I classifier part of the RBM-MHC takes as input $\boldsymbol{I}(\boldsymbol{v}) = \{I_\mu(\boldsymbol{v})\}_{\mu=1}^{N^h}$ of the annotated peptide sequences and gives a categorical output, *i.e.* the HLA-I specificity of the peptide $\boldsymbol{v}$ (Fig. 1b and Supplementary Fig. 1). $c = 1, ..., N^c$ denotes the HLA-I type. Typically for single-individual samples $N^c = 6$, from the 2 HLA-A, HLA-B, HLA-C. The classifier is trained by minimizing a loss function chosen to be a categorical cross-entropy $\mathscr{S}$:

$$\mathscr{S} = -\sum_{\boldsymbol{v}} \sum_{c=1}^{N^c} y_c(\boldsymbol{v}) \log(\hat{y}_c(\boldsymbol{v})) \tag{6}$$

where the $\boldsymbol{v}$ sum runs only over the sequences labeled with their HLA-I association. $y_c(\boldsymbol{v})$ is the label assigned to $\boldsymbol{v}$ for supervised training in one-hot encoding, *i.e.* $y_c(\boldsymbol{v}) = 1$ only for the c standing for its associated HLA type and zero otherwise so $\sum_c y_c(\boldsymbol{v}) = 1$ (it is normalized over categories). The choice of one-hot encoding is justified by the fact that, for the sake of discriminating motifs, we select peptides associated to only 1 HLA type in the database (mainly from monoallelic sources, see below). $\hat{\boldsymbol{y}}(\boldsymbol{v})$ is the categorical output predicted by the HLA-I classifier for $\boldsymbol{v}$, calculated from the softmax activation function:

$$\hat{\boldsymbol{y}}(\boldsymbol{v}) = \text{softmax}(\boldsymbol{X} \cdot \boldsymbol{I}(\boldsymbol{v}) + \boldsymbol{b}) \tag{7}$$

where $\boldsymbol{X}$ are the classifier weights, connecting input to output layer (see Supplementary Fig. 1), and $\boldsymbol{b}$ are local biases adjusted during training. Element-wise softmax is defined as:

$$\text{softmax}(\boldsymbol{z})_c = \frac{e^{\boldsymbol{z}_c}}{\sum_{c'} e^{\boldsymbol{z}_{c'}}} \tag{8}$$

The activation function softmax has the advantage of giving predictions normalized over categories, thus each element $\hat{y}_c(\boldsymbol{v})$ can be interpreted as the probability that sequence $\boldsymbol{v}$ belongs to class c. Our numerical implementation relies on Theano and Keras Python libraries. Training is performed in minibatches of 64 sequences, by the Adam optimizer for 1000 epochs, retaining the model that gives the best accuracy on a held-out partition of the ∼30% of the training dataset. The choice of RBM (hyper-)parameters (see above) also ensures a high accuracy of classification (Supplementary Fig. 2c). Accuracy of classification is measured there as an Area Under the Curve (AUC), which is different from the cross-entropy optimized during training (Eq. 6) and is defined below, in the section "RBM-MHC performance at motif reconstruction in multi-allelic samples". This AUC value is only minimally affected when reducing the RBM training dataset and is close to 1 (indicating the maximal accuracy) already with the small number of labeled sequences used, *i.e.* 10% of data. (The AUC clearly increases further when increasing this amount, see Supplementary Fig. 5).

Combining the probability functions estimated by RBM and HLA-I classifier, we define for each sequence $\boldsymbol{v}$ a global score $\mathscr{L}(\boldsymbol{v})$:

$$\mathscr{L}(\boldsymbol{v}) = \mathscr{L}^{RBM}(\boldsymbol{v}) + \mathscr{L}^{Cl}(\boldsymbol{v}) \tag{9}$$

where $\mathscr{L}^{RBM}(\boldsymbol{v}) = \log P(\boldsymbol{v})$ is the RBM log-likelihood assigned to sequence $\boldsymbol{v}$ and $\mathscr{L}^{Cl}(\boldsymbol{v})$ is the classifier score, defined from the vector of predicted class probabilities $\hat{y}_c(\boldsymbol{v})$ as $\mathscr{L}^{Cl}(\boldsymbol{v}) = \sum_{c=1}^{N^c} \hat{y}_c(\boldsymbol{v}) \log(\hat{y}_c(\boldsymbol{v}))$. $\mathscr{L}^{Cl}(\boldsymbol{v})$ is the negative entropy of classification, so it contributes to $\mathscr{L}(\boldsymbol{v})$ with higher values when the confidence with which a HLA-I class is predicted is higher.

Unsupervised clustering. *K-means algorithm.* Given a set of points $\boldsymbol{x}$, K-means [15, 25] finds the centroids of $c = 1, ..., N^c$ clusters and assigns each $\boldsymbol{x}$ to the cluster whose centroid has the minimal distance to $\boldsymbol{x}$. If we indicate by $d_c(\boldsymbol{x})$ the distance between point $\boldsymbol{x}$ and a cluster c, we can express the probability that $\boldsymbol{x}$ belongs to cluster c as:

$$\hat{y}_c^{km}(\boldsymbol{x}) = \frac{e^{-d_c(\boldsymbol{x})}}{\sum_{c'=1}^{N^c} e^{-d_{c'}(\boldsymbol{x})}}, \tag{10}$$

as is a common choice in the "soft" version of K-means. With the approach we refer to as RBM-km we apply K-means to sequence representations in the space of inputs to hidden units (*i.e.* $\boldsymbol{x} = \boldsymbol{I}(\boldsymbol{v})$) instead of

sequences themselves (*i.e.* $\boldsymbol{x} = \boldsymbol{v}$). From such a probabilistic clustering prediction, we define the classification score $\sum_{c=1}^{N^c} \hat{y}_c^{km}(\boldsymbol{x}) \log(\hat{y}_c^{km}(\boldsymbol{x}))$. Our implementation of K-means relies on the routine available in the Python package Scikit-learn [26].

MixMHCp algorithm. Consistently with the other approaches discussed, we assume that the expected number of clusters N_c is known, and we implement the MixMHCp clustering [8,9]. First we build Position Weight Matrices (PWM) for each of the N^c clusters found by MixMHCp; each PWM describes the single-site amino acid frequencies $f_i^c(v_i)$ in cluster c, with $c = 1,...,N^c$. For a sequence $\boldsymbol{v}$ a set of presentation scores (one per cluster) can be defined from the log-likelihood under the corresponding PWM:

$$\mathscr{L}_c^{MM}(\boldsymbol{v}) = \sum_{i=1}^{L} \log f_i^c(v_i) \qquad c = 1,...,N^c, \tag{11}$$

where the superscript *MM* stands for "MixMHCp" and L is the length to which all sequences are reduced by the alignment procedure in MixMHCp ($L = 9$). The final score of a sequence $\boldsymbol{v}$ is taken as the maximal log-likelihood among the N^c clusters, $\mathscr{L}^{MM}(\boldsymbol{v}) = \max_c \mathscr{L}_c^{MM}(\boldsymbol{v})$, which is used jointly to predict the HLA association.

Annotation of HLA-I binding motifs. The first step in clustering approaches, either by MixMHCp or K-means, is fully unsupervised and consists of optimally assigning peptides to N^c clusters. Clusters then need to be annotated with the corresponding HLA-binding specificity, among the N^c ones known to be expressed in the sample cells from *e.g.* HLA typing. For this second step, we consider the fraction of data for which we have labels (the same used for training the HLA-I classifier) and we estimate from them a PWM for each HLA type, in such a way as to obtain a set of reference motifs. We next estimate the PWM of each cluster and we label the cluster with the HLA association of the reference motif that minimizes the difference to the cluster PWM. Note that this mapping could give the same HLA type associated to several clusters and other HLA types not associated to any cluster, indicating a poor classification performance.

Antigen sequence alignment. To work with fixed-length sequences when learning a presentation model for datasets containing 8-11-mers, we first reduce these peptides of variable length to a homogeneous length that we set to the typical length of antigens presented by HLA class I proteins, *i.e.* 9 residues (the length of the antigen binding core). Our alignment routine consists of first building a reference profile of length 9 by estimating the PWM describing the subset of 9-mers in the dataset of interest. Next we align the other peptides to this profile by the insertion of a gap in 8-mers and deletions in 10-11 mers, using the functions *seqprofile* and *profalign* (with default options) of the Matlab Bioinformatics Toolbox (release R2018b). Typical samples of interest pool together peptides of different binding specificity to HLA-I, each of these being characterized by well defined sequence motifs. On one hand, the best alignment for each peptide needs to be based on the PWM of 9-mers sharing the same HLA preference, which best describes the respective sequence motif. On the other hand, identifying these subsamples sharing the same specificity requires first identifying the motif, therefore we propose an iterative procedure where the sequence alignment is re-iterated after training the classifier in such a way that the steps of classification and alignment inform each other. Peptides of core length 9 do not need to be re-aligned. The overall workflow with peptides of length 8-11 is as follows:

1. Align all sequences to a PWM profile of length 9.

2. Use this first alignment to train RBM-MHC (*i.e.* RBM and HLA-I classifier).

3. Re-align peptides based on the putative class predicted by the HLA-I classifier. First, for each HLA-I class, we put together the 10% of labeled data and the peptides classified in that class (weighted by the probability of classification) and we build on this sample a Hidden Markov Model (HMM) with the routines *hmmprofstruct* and *hmmprofestimate* (with default parameters) of the Matlab Bioinformatics Toolbox. We use these HMM profiles (essentially capturing the pattern of single-site residue conservation) as seeds of each class's alignment.

4. Use the HMM alignment score to correct classification errors coming from the initial, suboptimal alignment. We take every unlabeled peptide and we compute the HMM alignment score to each of the classes' seeds (by the *hmmprofalign* function). If the alignment score of a peptide is higher for a class that is different from the original, classifier-predicted one, we change the peptide alignment to the class with the higher alignment score, in such a way as to re-classify the peptide more accurately in the subsequent iteration.

5. Repeat the RBM and HLA-I classifier training after the re-alignment.

In the 10 "synthetic-individual" datasets considered for testing motif reconstruction (Supplementary Table 1), we have observed a systematic improvement of the classification performance (measured by the indicators provided below) after 1 re-alignment step. For 2 or more iterations, there are cases in which the classification performance is degraded, therefore in our routine we re-align and re-train once.

Dataset collection from IEDB and preparation. *Mass spectrometry datasets.* A set of 8-11-mer peptides documented with HLA-A*02:01 restriction in the Immune Epitope Database (IEDB, last release [10]), for a total of 5607 monoallelic-source sequences filtered as outlined below, was used to train the HLA-A*02:01-specific RBM presentation model. We randomly selected a subset with the 80% of these sequences as training dataset and kept the remaining 20% as test set to evaluate the model's predictions in terms of probabilistic scores of presentation (Fig. 1c). We randomly selected 5000 peptides from the human proteome (as in UniProt database [27]) with length distribution matching the one of presented peptides to serve as a set of peptides predominantly not presented on the cell surface ("generic").

For motif reconstruction, we considered 10 "synthetic-individual" datasets of antigens with known HLA-I specificity to assess the RBM-MHC classification performance by comparing RBM-MHC predictions against the known HLA-I specificity. These datasets are built by collecting all IEDB antigens associated to 2 haplotypes, *i.e.* combinations of an A, a B, a C allele observed to co-occur in the human population (see Supplementary Table 1) in such a way as to preserve linkage. Information on haplotypes co-occurrence was found at allelefrequencies.net [28]. We prioritized HLA-specific peptides from monoallelic sources [4] in such a way that the assignment of HLA-binding specificity is unambiguous and does not rely on additional *in silico* predictions. The search was performed as follows. The full set of curated HLA-I ligands was downloaded from IEDB (file *mhc_ligand_full.csv* from http://www.iedb.org/database_export_v3.php, accessed in September 2019). We searched in IEDB for linear, human peptides of length 8-11 residues (as most of the HLA-I presented peptides have a length in this range). We looked for peptides eluted from cells (field in IEDB file *Assay Group* = "ligand presentation") and detected by mass spectrometry techniques (field *Method/Technique* = "cellular MHC/mass spectrometry", "mass spectrometry", "secreted MHC/mass spectrometry"). Among these data, we set the field *Allele Evidence Code* (describing the basis upon which the HLA-I assignment is made) to the most stringent option, "Single allele present". This definition indicates that antigen presenting cells are known to only express a single HLA-I molecule, as is the case for monoallelic cell lines [4, 5]. If for a given allele we found less than a minimal number (set to 300) of sequences among monoallelic-source data, we extended the search to peptides with evidence code "Allele Specific Purification", since this procedure attaches greater confidence to the HLA assignment than its inference by *in silico* methods. Only for one allele (HLA-B*39:01) we did extend the search to all other MS data, thus including the evidence code "Inferred by motif or alleles present". To apply the RBM-MHC method to patient-derived immunopeptidomic datasets, when we cannot retrieve a HLA annotation in IEDB for the 10% of their peptides, we add to each of them labeled peptides for the 6 HLA-I (given by the patient's HLA typing) collected from IEDB as above. The RBM-MHC training set is this minimally extended dataset, to guide the learning of the HLA-I classifier by the labeled peptides, whose predictions are used to reconstruct HLA-I motifs in the original dataset. We have not attempted to identify motifs on dataset Mel5 from [19] as the patient's HLA typing was uncomplete (lacking the HLA-C alleles).

Binding assay datasets. Records of binding assays from IEDB (as of March 2020) were filtered following [5], *i.e.* selecting peptides annotated with a quantitative measure of binding dissociation constants < 500 nM and excluding assay types that led to documented discrepancies between predicted and effective affinity ("purified MHC/direct/radioactivity/dissociation constant KD", "purified MHC/direct/fluorescence/half maximal effective concentration (EC50)", "cellular MHC/direct/fluorescence/ half maximal effective concentration (EC50)").

MS-based validation of model predictions for neoantigen discovery. To evaluate the predictive power of the HLA-A*02:01-specific RBM model for neoantigen discovery, we acquired the list of missense mutations for ovarian and melanoma cancer cell lines SKOV3, A2780, OV90, HeLa and A375 from the Cosmic Cell Lines Project [29] as described in Ref. [14], *i.e.* filtered by expression levels and by their affinity, estimated with NetMHCpan3.0 [30], to the allele HLA*A:02:01 (known to expressed in these cell lines). The number of missense mutations is as follows: 93 (A375), 60 (SKOV3), 39 (A2780), 22 (OV90), 28 (HeLa), resulting in 38 possible peptides of length 8-11 containing each of them. We used the RBM log-likelihood $\mathscr{L}^{RBM}$ as probabilistic score of presentation by HLA*A:02:01 to assign to all 8-11 residue long mutated peptides. Since the MS database for the 5 cancer lines produced by Ref. [14] was annotated by comparison to the consensus human proteome, it contained only wildtype peptides whose mutated version was in the list of putative neoantigens. As long as a mutation does not affect an anchor site, the mutated version should

preserve a high probability of presentation, thus we have excluded from the analysis of Fig. 1d neoantigens arising from anchor-site mutations of MS-derived peptides.

The studies from which single-patient samples Mel15, Mel8, Mel5, 12T were retrieved [17–19], listed a total of 11 tumour-presented neoantigens, using techniques that allow for the detection of variants of native proteins by MS. For some of these neoantigens, immunogenicity was also validated *in vitro* by identification of neoantigen-specific T-cell responses. Their HLA association was predicted by NetMHC. The prediction was to a large extent confirmed by *in vitro* validation of immunogenicity, which relied on antigen presenting cells that were positive to the predicted HLA-I. As the 11 neoantigens are 9-10 residues long, following Ref. [8] we considered 9-mer and 10-mer peptides overlapping the missense mutations observed in the patient's WES, which resulted in a list of thousands of putative neoantigens (see Supplementary Table 2). The RBM-MHC training set *per se* can thus take into account only 9-10 mers from patients' samples; patients' neoantigens to validate were not included in the training dataset. In this multi-allelic case where antigens may be presented by several HLA types (as opposed to the allele-specific case of the previous paragraph) we used the global score accounting for the probability of presentation (by RBM) as well as the confidence of the HLA assignment (either by the HLA-I classifier in RBM-MHC or by K-means clustering in RBM-km) of equation 9.

RBM-MHC performance at motif reconstruction in multi-allelic samples. We chose the Area Under the Curve (AUC) of the Receiving Operating Characteristic (ROC) as a metric for classification (*i.e.*, HLA assignment) performance, estimated for each approach as follows. RBM-MHC, through the HLA-I classifier, outputs for each peptide a probability to belong to each HLA-I class. MixMHCp [8, 9] models probabilistically data by a mixture of independent models, thus it describes each sequence by a vector of "responsibilities", describing the probabilities to belong to each cluster (*i.e.* HLA-I class). NetMHCpan4.0 [7] predicts peptide binding values from either the training on eluted data (EL option) or the binding affinity data (BA option) and it estimates from these values presentation scores and percentile rank scores. Low values of percentile rank define binders (following the authors' recommendations, peptides with percentile rank <2% and <0.5% are considered HLA-I weak and strong binders respectively). Having built the samples from MS data, we considered NetMHC predictions from the EL option (NetMHCpan4.0-EL). We compare the $N^c = 6$ HLA-I classes by pairs. We consider the probability of belonging to a certain class that each method would assign to "positives" of that class (peptides binding to the respective HLA-I allele) and "negatives" (peptides binding to other alleles): when the classification performance is good, the former has values close to one, the latter has values close to zero. Varying the threshold between false positive and false negative distributions, we build the ROC curve. We take the area under the ROC curve (AUC) as measure of the ability to discriminate the two classes (AUC = 1 means perfect discrimination, AUC = 0.5 means chance). To obtain one cumulative indicator, we average over the AUCs of these pairwise comparisons. For the approaches partially or fully supervised (RBM-MHC and classifier-only), the AUC is measured only from data *not* in the 10% used for the supervised learning step.

To further assess classification performance, we looked also at the HLA type of each peptide predicted based on: the highest responsibility value among the N_c values for MixMHCp; the lowest percentile rank for NetMHCpan-4.0 (EL option); the highest class probability estimated by the HLA-I classifier for RBM-MHC. In this way, when comparing peptides of different classes two by two, we can count true positives of classification (tp), false positives (fp), true negatives (tn) and false negatives (fn). Once these quantities are defined, additional classification performance indicators are: Accuracy=$(tp+tn)/(tp+fp+tn+fn)$, Precision=$tp/(tp+fp)$, Specificity=$tn/(tn+fp)$, Sensitivity=$tp/(tp+fn)$. MixMHCp [8, 9] was run using default options, which include an additional "trash cluster". When measuring these classification performance indicators, the assignment to the trash cluster is considered among the false negatives. The comparison of classification performance as measured by these indicators is shown in Supplementary Fig. 7.

Evaluation of model predictions for SARS-CoV-2 epitopes. We downloaded the protein-coding regions of SARS-CoV-2 genome from GenBank, NCBI Reference Sequence: NC_045512.2. We extracted all the 9-mers contained in the SARS-CoV-2 proteome, giving a list of 9656 candidate cytotoxic T-cell epitopes (HLA-I antigens). Ref. [20] (available at https://www.immunitrack.com/free-coronavirus-report-for-download/) tested *in vitro* the 94 top-scoring epitopes according to NetMHC4.0 [6] for each of the HLA-I alleles A*01:01, A*02:01, A*03:01, A*11:01, A*24:02, B*40:01, C*04:01, C*07:01, C*07:02 and according to NetMHCpan4.0 [7] for C*01:02. 159 peptides were identified as high-stability binders (*i.e.* with stability above the threshold of 60% of the reference peptide value for the corresponding HLA-I allele). To probe our method as a predictor for such binders, we learned a series of allele-specific RBM presentation models (5 hidden units, $\lambda_1^2 = 0.001$) for the 10 HLA-I alleles considered in this study. We decided to train

a series of allele-specific models to minimize the uneven accuracy across alleles, emerging especially in multi-allelic models, due to the different abundances of peptides with different HLA-I preferences available as training data. (The use of multi-allelic models is intended as tool to characterize unannotated samples through motif reconstruction). Here, we prioritized training datasets from binding assays in IEDB (see above) to customize the presentation score toward the identification of high-affinity ligands. These types of data are almost absent for the 4 HLA-C mentioned, and in these cases models were learned from MS data only but a re-weighting of frequencies, aimed at correcting for MS biases in detection, was applied (see Supplementary Note). All peptides tested for binding to a given HLA-I in Ref. [20] were scored by the corresponding RBM model (see Fig. 2d, where, as a comparison, results achieved by training all RBMs on MS data with re-weigthing are also reported). We next estimated their score percentile relative to scores assigned to all candidate epitopes and we assessed that tested binders were predominantly assigned high scores, able to a good extent to discriminate them from the tested non-binders (see Supplementary Fig. 11).

As an additional test, we considered the SARS-CoV-2 cytotoxic T-cell epitopes identified as potentially associated to high immune responses by Ref. [21], who mapped the dominant, experimentally validated SARS-CoV-derived epitopes to the corresponding regions in the SARS-CoV-2 proteome. We scored the 22 epitopes in this list with complete (100%) or moderate-high (> 70%) sequence similarity to the homologous SARS-CoV epitope, which should preserve high likelihood of presentation. Since homologous SARS-CoV epitopes were experimentally tested in binding assays, scores were assigned by the same (affinity-trained) models as above covering the HLA restrictions reported in Ref. [21] (HLA-A*02:01 to the largest extent, HLA-A*24:02, HLA-B*40:01). In addition, we checked that scores estimated by MS-trained RBM models with re-weigthing lie in the same range, see Supplementary Fig. 12b.

References

1. Yarchoan M, Johnson BA, Lutz ER, Laheru DA, Jaffee EM. Targeting Neoantigens to Augment Antitumour Immunity. Nature Reviews Cancer. 2017;17(4):209–222.

2. Garcia-Garijo A, Fajardo CA, Gros A. Determinants for Neoantigen Identification. Frontiers in Immunology. 2019;10.

3. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A Comprehensive Review and Performance Evaluation of Bioinformatics Tools for HLA Class I Peptide-Binding Prediction. Briefings in Bioinformatics. 2019;bbz051.

4. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-Allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017;46(2):315–326.

5. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A Large Peptidome Dataset Improves HLA Class I Epitope Prediction across Most of the Human Population. Nature Biotechnology. 2020;38(2):199–209.

6. Andreatta M, Nielsen M. Gapped Sequence Alignment Using Artificial Neural Networks: Application to the MHC Class I System. Bioinformatics (Oxford, England). 2016;32(4):511–517.

7. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology. 2017;199(9):3360–3368.

8. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I Motifs across HLA Peptidomes Improves Neo-Antigen Predictions and Identifies Allostery Regulating HLA Specificity. PLOS Computational Biology. 2017;13(8):e1005725.

9. Gfeller D, Guillaume P, Michaux J, Pak HS, Daniel RT, Racle J, et al. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. The Journal of Immunology. 2018;201(12):3705–3716.

10. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 Update. Nucleic Acids Research. 2019;47(D1):D339–D343.

11. Smolensky P. Information Processing in Dynamical Systems: Foundations of Harmony Theory ; CU-CS-321-86. Computer Science Technical Reports. 1986;315.

12. Hinton GE. Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation. 2002;14(8):1771–1800.

13. Tubiana J, Cocco S, Monasson R. Learning Protein Constitutive Motifs from Sequence Data. eLife. 2019;8:e39397.

14. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, et al. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell. 2017;171(6):1272–1283.e15.

15. Lloyd SP. Least Square Quantization in PCM. IEEE Transactions on Information Theory. 1982;28(2):129–137.

16. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA Class I Supertypes: A Revised and Updated Classification. BMC Immunology. 2008;9(1):1.

17. Kalaora S, Barnea E, Merhavi-Shoham E, Qutob N, Teer JK, Shimony N, et al. Use of HLA Peptidomics and Whole Exome Sequencing to Identify Human Immunogenic Neo-Antigens. Oncotarget. 2016;7(5):5110–5117.

18. Kalaora S, Wolf Y, Feferman T, Barnea E, Greenstein E, Reshef D, et al. Combined Analysis of Antigen Presentation and T-Cell Recognition Reveals Restricted Immune Responses in Melanoma. Cancer Discovery. 2018;8(11):1366–1375.

19. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct Identification of Clinically Relevant Neoepitopes Presented on Native Human Melanoma Tissue by Mass Spectrometry. Nature Communications. 2016;7(1):1–16.

20. Immunitrack, Intavis. Covid19 Intavis_Immunitrack Stability Dataset 1; 2020.

21. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host & Microbe. 2020;p. S1931312820301669.

22. Sehnal D, Rose AS, Koča J, Burley SK, Velankar S. Mol*: Towards a Common Library and Tools for Web Molecular Graphics. In: Proceedings of the Workshop on Molecular Graphics and Visual Analysis of Molecular Data. MolVA '18. Brno, Czech Republic: Eurographics Association; 2018. p. 29–33.

23. Tubiana J, Monasson R. Emergence of Compositional Representations in Restricted Boltzmann Machines. Physical Review Letters. 2017;118(13):138301.

24. Tubiana J, Cocco S, Monasson R. Learning Compositional Representations of Interacting Systems with Restricted Boltzmann Machines: Comparative Study of Lattice Proteins. Neural Computation. 2019;31(8):1671–1717.

25. Forgy EW. Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications. Biometrics. 1965;21(3):768–769.

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.

27. Consortium TU. UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Research. 2019;47(D1):D506–D515.

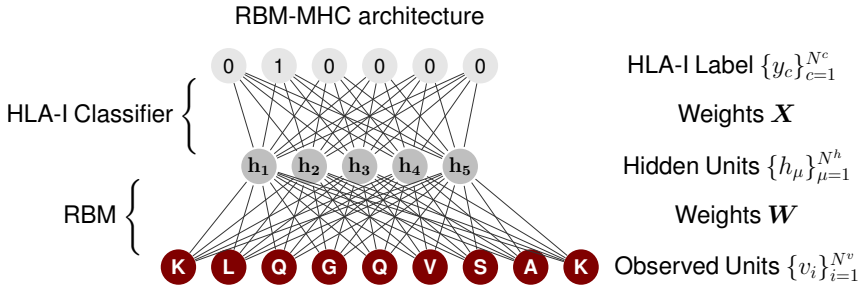
28. González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, da Silva ALS, et al. Allele Frequency Net 2015 Update: New Features for HLA Epitopes, KIR and Disease and HLA Adverse Drug Reaction Associations. Nucleic Acids Research. 2015;43(D1):D784–D788.

29. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, et al. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. Current Protocols in Human Genetics. 2016;91(1):10.11.1–10.11.37.

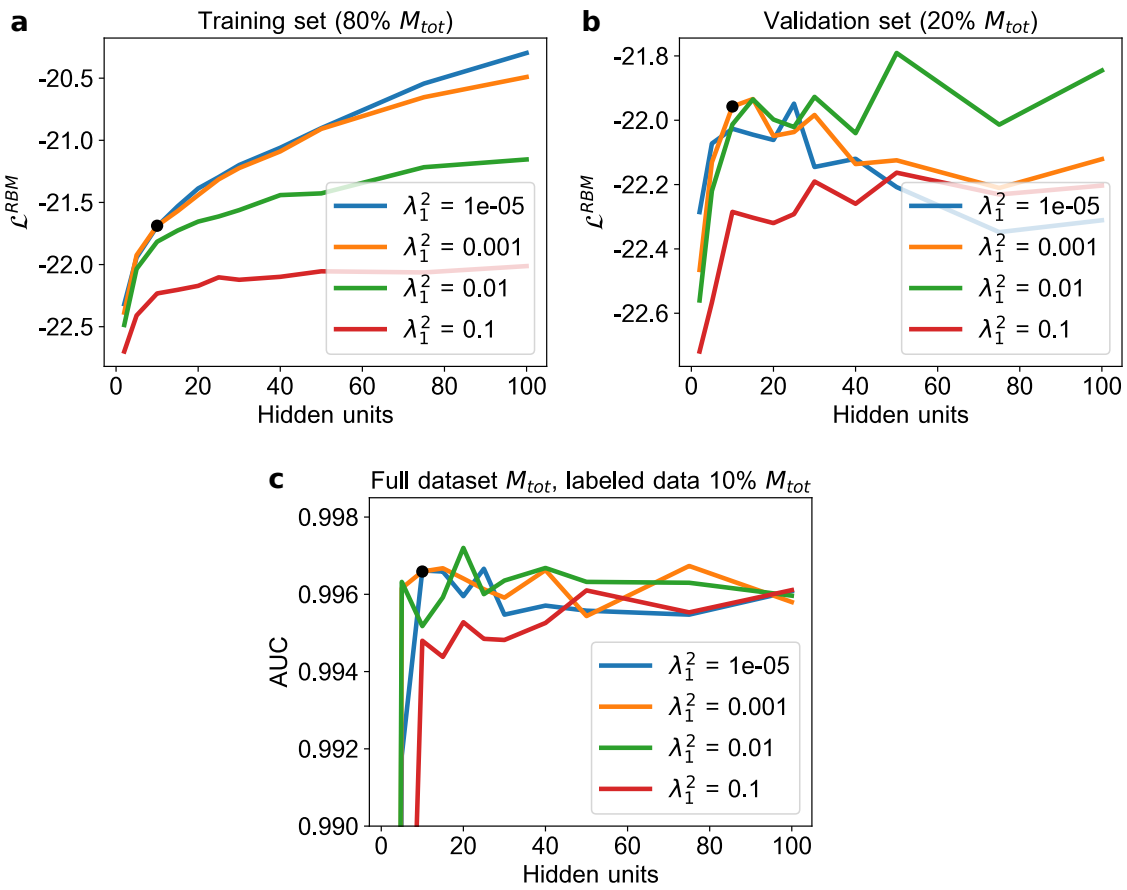
30. Nielsen M, Andreatta M. NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets. Genome Medicine. 2016;8(1):33.

Supplementary Information

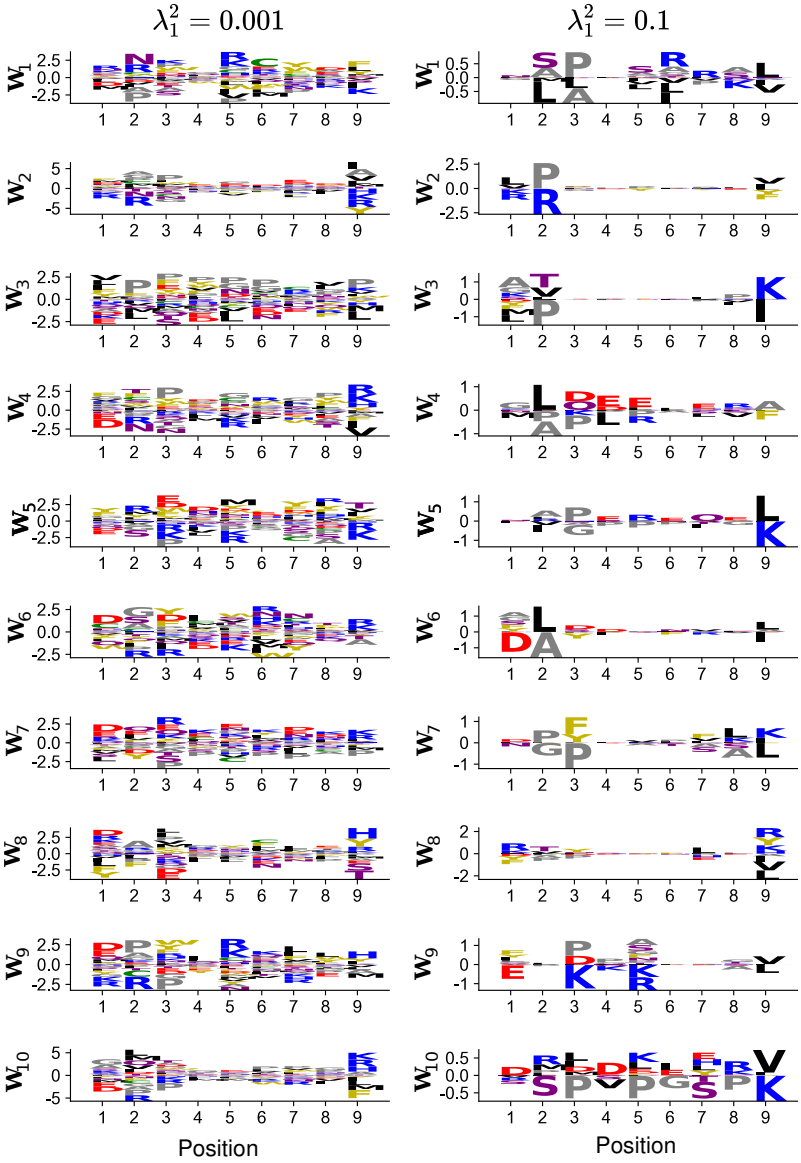
Supplementary Figures



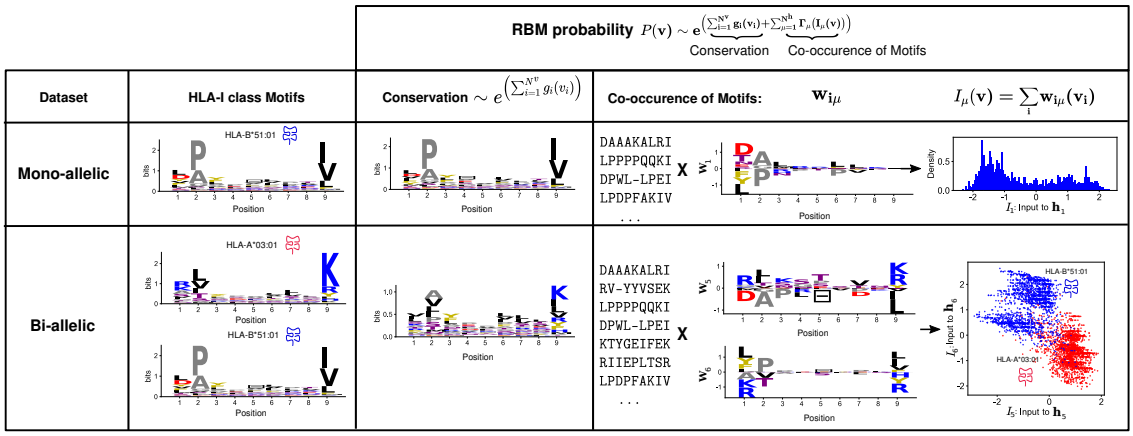
Supplementary Figure 1: **RBM-MHC graphical architecture**. The RBM architecture consists of the layer of observed units $\boldsymbol{v}$ - corresponding to the antigen sequences to model - coupled to the layer of hidden units $\boldsymbol{h}$ by weights $\boldsymbol{W}$. The HLA-I classifier has a fully connected network of weights $\boldsymbol{X}$ connecting the layer of inputs to hidden units $\boldsymbol{I}(\boldsymbol{v})$ (RBM data representations) to the output layer of HLA-I classes $\boldsymbol{y}_c$ (light gray nodes) formulated in one-hot representation (*e.g.* HLA-I class 2 is encoded by a vector of zeros with the symbol 1 only at position 2).



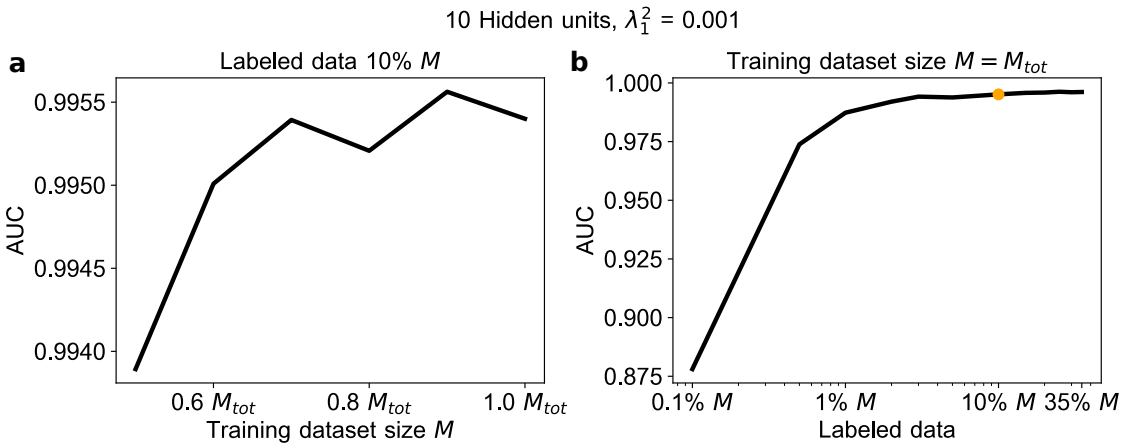
Supplementary Figure 2: **RBM model selection.** RBM average log-likelihood on the training set (**a**) and on the validation set (**b**) as a function of the number of hidden units, for different values of the regularization λ_1^2. The black dot indicates an optimal combination of parameters chosen for the results of Figs. 1e, 2a-c (10 hidden units, $\lambda_1^2 = 0.001$). **c**, AUC of HLA-I classification (see Methods) as a function of the number of hidden units for different values of the regularization strength λ_1^2. All tests are run on a "synthetic-individual" dataset (related to Haplotype 10 in Supplementary Table 1).



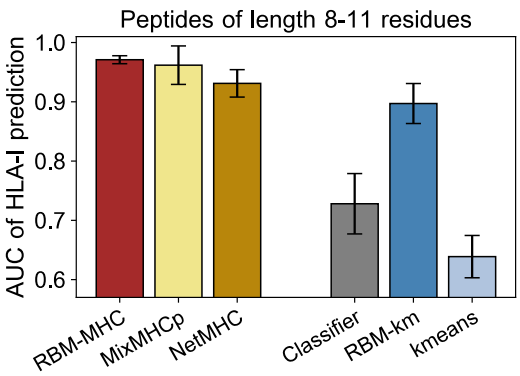
Supplementary Figure 3: **Example of RBM weights and change of representation with regularization.** The RBM learns a set of weights for each hidden unit. In each set, weights have the length of sequence (9) and for each position there is a weight value for each amino acid appearing at that position that can be either positive or negative (the height of the amino acid letter here gives its magnitude, as in the representation by [1]). Color stands for amino acid chemical/physical properties: red = negative charge (E,D), blue = positive charge (H, K, R), purple = non-charged polar (hydrophilic) (N, T, S, Q), yellow = aromatic (F, W, Y), black = aliphatic hydrophobic (I, L, M, V), green = cysteine (C), grey = other, tiny (A, G, P). Here, we show a comparison of RBM weights for 2 choices of the regularization strength λ_1^2 analyzed in Supplementary Fig. 2 (training with 10 hidden units, on the "synthetic-individual" dataset corresponding to Haplotype 10 in Supplementary Table 1). As found in [1], at low regularization ($\lambda_1^2 = 0.001$) the data representation is more intricate while a higher regularization ($\lambda_1^2 = 0.1$), by enforcing sparsity, gives non-zero weights to the amino acids and positions learnt as the most discriminative among "features" (here, HLA-I classes). The first λ_1^2 choice however retains more information about data, hence reaches a slightly higher classification performance when combined to the HLA-I classifier (see Supplementary Fig. 2c); the second λ_1^2 choice results in a more interpretable representation.



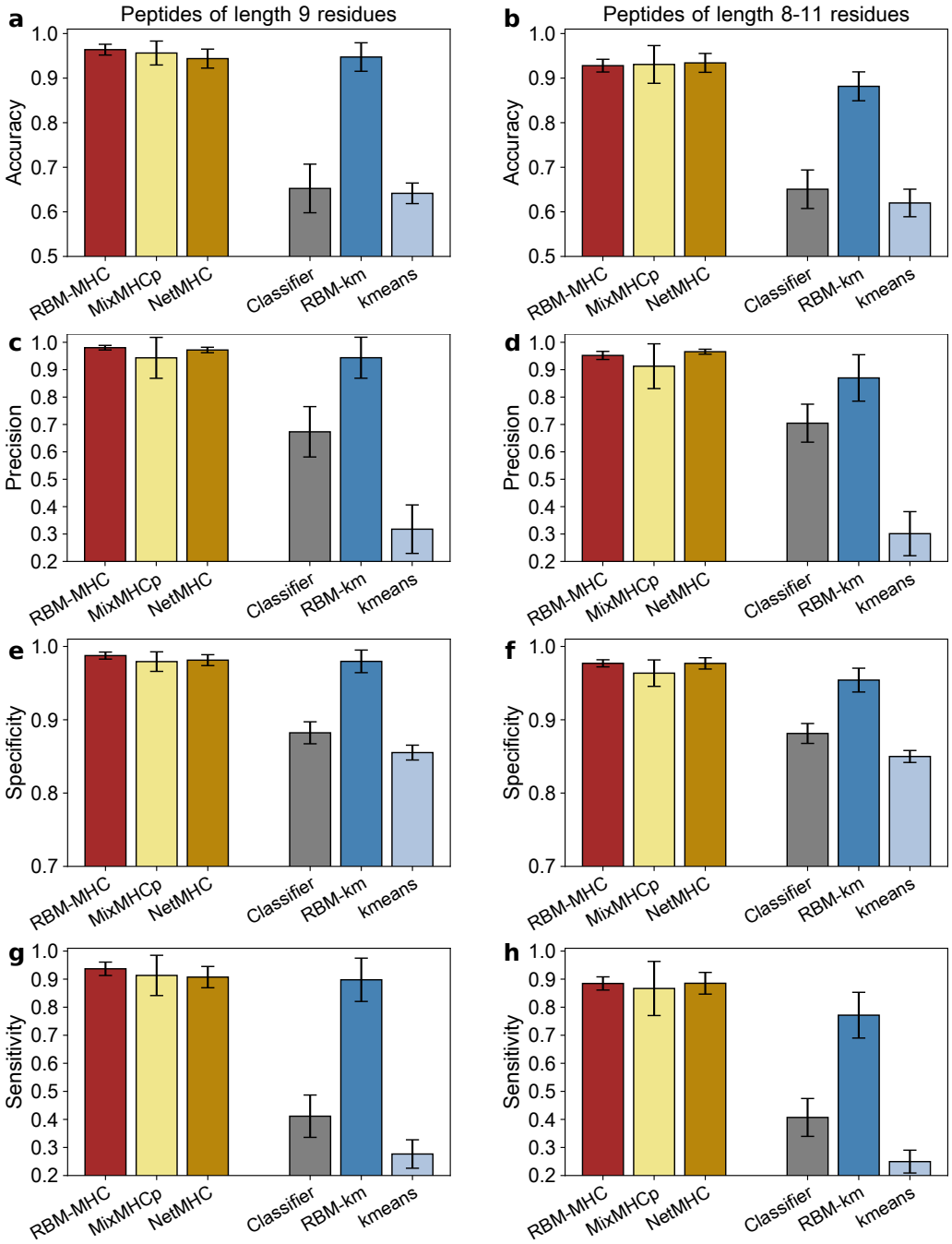
Supplementary Figure 4: **HLA-I sequence motifs are captured through RBM local potentials and weights.** The probability of presentation $P(\boldsymbol{v})$ inferred by the RBM consists of potentials $g_i(v_i)$, describing single-site amino acid conservation, and terms describing the co-occurrence of motifs through the weights $w_{i\mu}(\boldsymbol{v})$. In the first row, we consider a IEDB, MS-based dataset of antigens specific to one allele, HLA-B*51:01 (monoallelic dataset). A RBM model is trained on these data (chosen with 5 hidden units and $\lambda_1^2 = 0.1$, to enhance weights' interpretability, see Supplementary Fig. 3). With data of the same HLA-I type, the binding motif is essentially captured by conservation alone (logo representation of $e^{\sum_i(g_i(v_i))}$, central column). Weights and hidden units embed additional data features (right column). Hidden unit 1 $\boldsymbol{h}_1$ encodes the multiple binding modes of HLA-B*51:01 antigens discussed in [2] (see also Fig. 2c): the sequence property "hydrophilic/charge" has a positive projection on $\boldsymbol{W}_1$ (and hence positive inputs to $\boldsymbol{h}_1$) while the pattern "hydrophobic" has a negative one. In the second row, we train a RBM (10 hidden units and $\lambda_1^2 = 0.1$) on a bi-allelic dataset, *i.e.* where we added to the monoallelic dataset IEDB, MS-derived antigen sequences specific to the allele HLA-A*03:01 (a minimal example of a sample mixing antigens of different specificities). In this setting: *(i)* overall conservation does not allow *discrimination* of HLA-I motifs (central column), for which the RBM data representation through weights and inputs to hidden units is necessary; *(ii)* this representation is *robust* with respect to the monoallelic case. To illustrate this, we have selected sets $\boldsymbol{W}_5$ and $\boldsymbol{W}_6$ (right column). They distinguish HLA-B*51:01 antigens with sequence patterns underlying alternative binding modes (pattern "hydrophilic/charge" has significant, negative projection on $\boldsymbol{W}_5$ and is kept separate from pattern "hydrophobic", with a significant, positive projection on the other set of weights, $\boldsymbol{W}_6$). HLA-B*51:01 sequence patterns are represented as mutually exclusive with respect to the sequence motif characteristic of HLA-A*03:01 (weights of opposite sign). As a result, antigens of different HLA-I types are clearly partitioned into 2 clusters in the space of inputs to $\boldsymbol{h}_5$ and $\boldsymbol{h}_6$. The 2 subclusters in the HLA-B*51:01 one (blue points) correspond to its alternative binding modes. As conserved sites at position 9 play a role in representing HLA-B*51:01 *vs* HLA-A*03:01 specificity but not in discriminating HLA-B*51:01 binding modes, only in this bi-allelic case weights at that position have a non-negligible value.



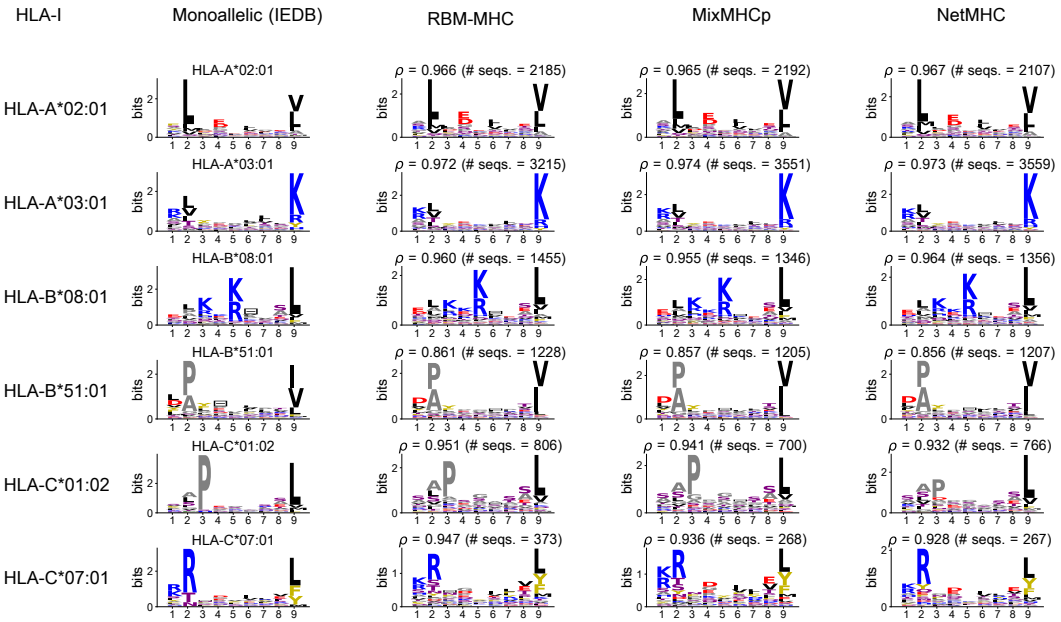
Supplementary Figure 5: **Trend of classification performance**. **a**, Average AUC of classification as a function of the fraction of data considered as training dataset of the RBM (here $M_{tot} = 10368$). **b**, Average AUC of classification as a function of the RBM training dataset percentage used for the supervised learning of the HLA-I classifier. The orange dot corresponds to the value chosen for all RBM-MHC results in Figs. 1 and 2 (i.e. labeled data = 10%M). All tests shown here are run on a "synthetic-individual" dataset (related to Haplotype 6 in Supplementary Table 1) and represent the average over 10 realizations of the training.



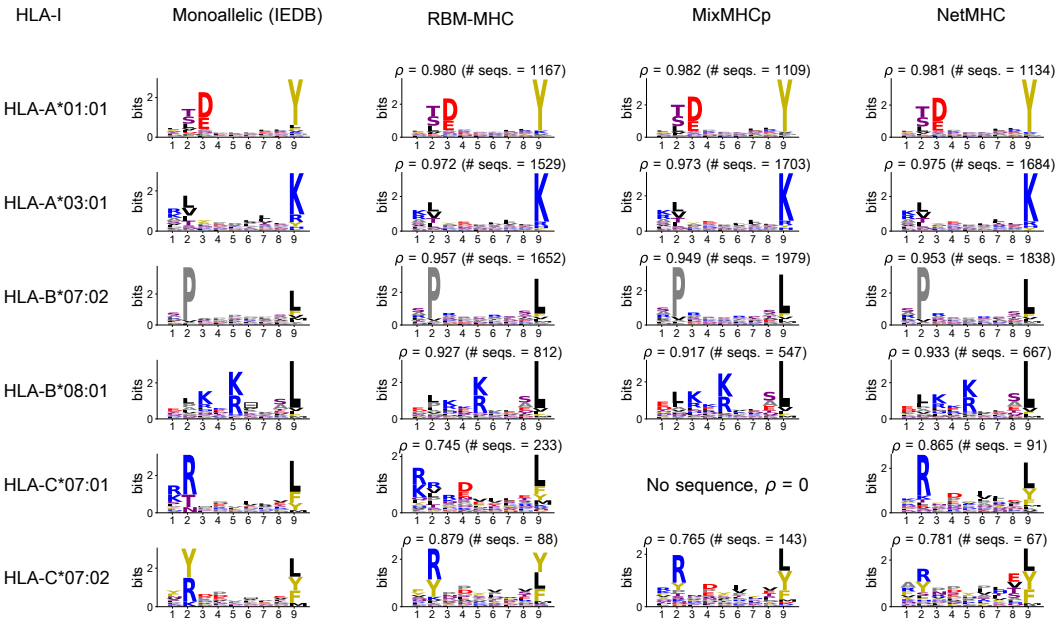
Supplementary Figure 6: **AUC of performance of predicting antigen HLA association in "synthetic-individual" samples**: same plot as in Fig. 1e for datasets containing peptides of length 8-11 residues, where a sequence alignment step is required (see Methods). "NetMHC" refers to the prediction by the version NetMHCpan4.0-EL (see Methods).



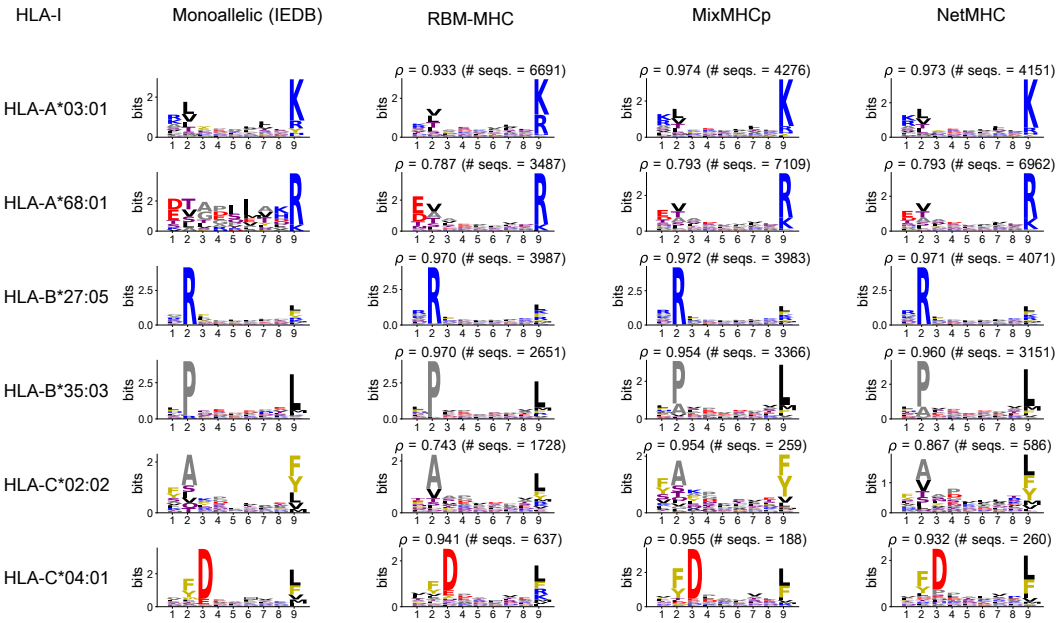
Supplementary Figure 7: **Performance of HLA association in "synthetic-individual" samples**: same plots as in Fig. 1e and Supplementary Fig. 6, with average accuracy (**a**,**b**), precision (**c**,**d**), specificity (**e**,**f**), sensitivity (**g**,**h**) (see Methods). Classification performance indicators are measured separately in datasets containing only peptides of length 9 (**a**,**c**,**e**,**g**) and in datasets containing peptides of length 8-11 (**b**,**d**,**f**,**h**).



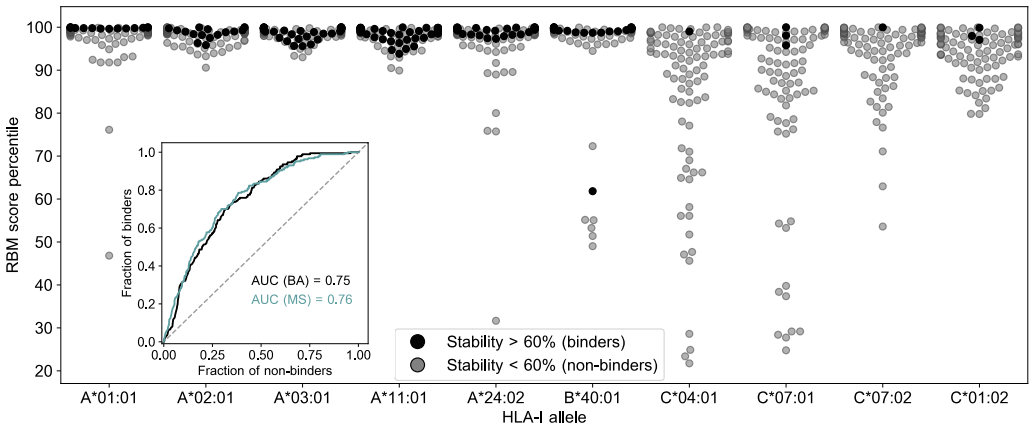
Supplementary Figure 8: **Reconstructed sequence motifs in multi-allelic sample 12T** from [3, 4], sample size = 9262 sequences (restricted to length 8-11). Sequence motifs recovered by RBM-MHC are compared to the ones from monoallelic data in IEDB and the ones recovered by MixMHCp and NetMHCpan4.0-EL in the same sample. The quality of motif reconstruction is measured by ρ, the coefficient of correlation between single-site amino acid frequencies of the monoallelic IEDB motif and of the predicted one. The average ρ over the 6 HLA-I motifs recovered by the 3 methods is compared in Fig. 2b. Motifs of length 9 in monoallelic data are obtained by aligning sequences in each HLA-I class separately, using standard alignment routines as described in Methods. When comparing these motifs to the ones predicted by the different methods, we keep the alignment step consistent with the choice made for monoallelic data, *i.e.* for each method we align separately, applying the same standard alignment routines, the sequences assigned to the same HLA-I class.



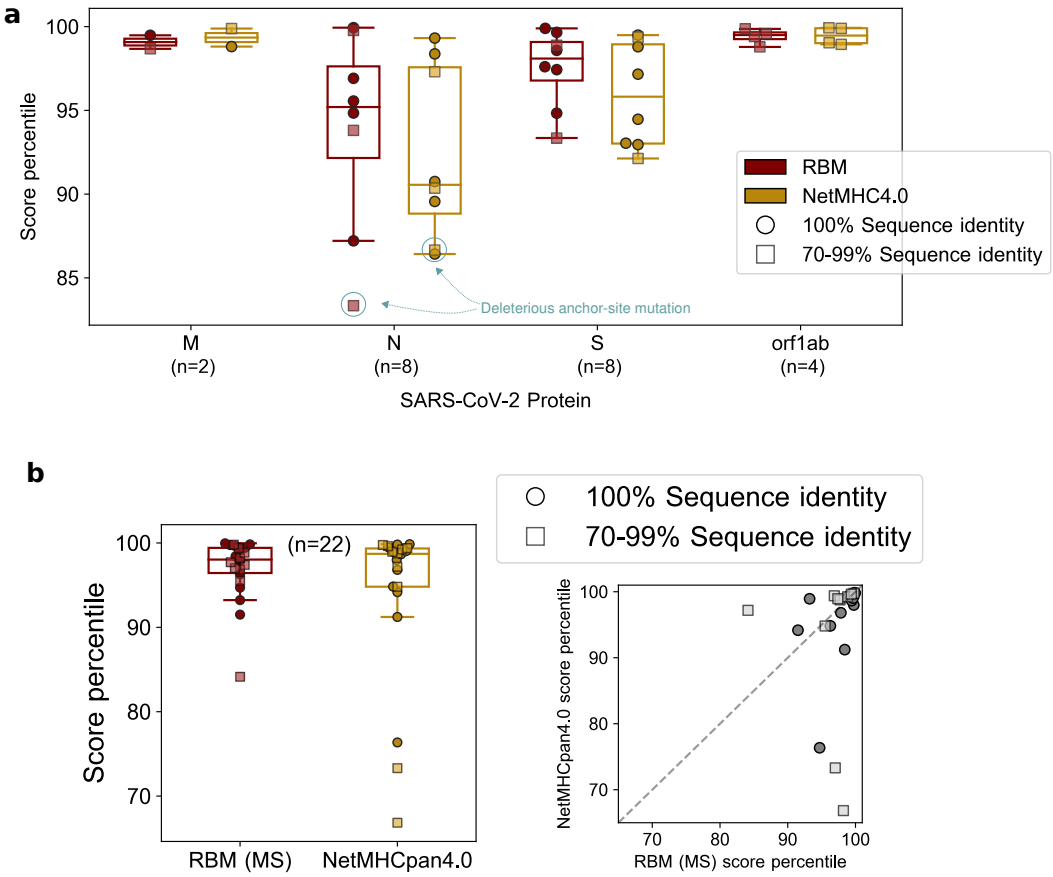
Supplementary Figure 9: **Motif reconstruction from multi-allelic sample Mel8** from [5], sample size = 5481 sequences (see caption of Supplementary Fig. 8).



Supplementary Figure 10: **Motif reconstruction in sample Mel15** from [5], sample size = 19181 sequences (see caption of Supplementary Fig. 8).



Supplementary Figure 11: **RBM prediction of SARS-CoV-2 epitopes tested *in vitro*.** RBM score percentile of putative SARS-CoV-2 ligands predicted by NetMHC (NetMHC4.0 and NetMHCpan4.0) for 10 HLA-I alleles, divided into "binders" and "non-binders" on the basis of the binding stability assigned *in vitro* by [10] (respectively > 60% and < 60% of reference values). RBM models were trained on binding assay (BA) data for HLA-A and HLA-B, on mass spectrometry (MS) data for HLA-C (see Methods). The average RBM score percentile for binders is consistently above 98 for each HLA-I allele. We quantify the RBM ability to distinguish binders from non-binders by the Receiver Operating Characteristic curve (inset), which describes, as a function of a score percentile threshold, how the fraction of binders and non-binders is recovered by the RBM score. Its Area Under the Curve (AUC) is significantly above the random expectation of 0.5. As a comparison, we added the AUC obtained using RBM models all trained on MS data, giving a comparable performance. An unbiased comparison to NetMHC is here difficult because peptides were pre-selected ahead of the experiments from the top 1% ranks estimated by NetMHC.



Supplementary Figure 12: **RBM scores to SARS-CoV-2 homologs of tested, dominant SARS-CoV epitopes.** **a**: Same data as in Fig. 2e, where epitopes are divided by SARS-CoV-2 protein, with S = Spike glycoprotein, E = Envelope protein, M = Membrane glycoprotein, N = Nucleocapsid phosphoprotein. The lowest RBM score percentile (consistent with NetMHC4.0 one, see also in Fig. 2e) corresponds to an anchor-site mutation of the SARS-CoV homolog expected to be deleterious, as it leads from a highly abundant amino acid at position 9 to a low-frequency one. **b**: Same figure as in Fig. 2e with scores obtained by MS-trained RBM models. Here the comparison is made with the NetMHC algorithm trained on MS datasets, *i.e* NetMHCpan4.0 (run with the option EL = Eluted).

Supplementary Tables

Supplementary Table 1: **Detailed benchmark of RBM-MHC prediction performance** (provided as a separate Excel file). Allelic composition of "synthetic-individual" samples retrieved from IEDB (see Methods) and measures of AUC, accuracy, precision, sensitivity, specificity for each sample.

Neoantigen		**Rank**			**HLA-I prediction**		
Sequence	**HLA-I**	**RBM-MHC**	**RBM-km**	**MixMHCp**	**RBM-MHC**	**RBM-km**	**MixMHCp**
Sample 12*T* - 15750 candidates							
DAN**S**FLQSV	B*51:01	27	33	49	✓	✓	✓
GV**Y**PMPGTQK	A*03:01	7	5	5188	✓	✓	✓
Sample *Mel*8 - 1340 candidates							
SPGPVKLE**L**	B*07:02	1	1	1	✓	✓	✓
Sample *Mel*15 - 24880 candidates							
GRIAF**F**LKY	B*27:05	55	73	97	✓	✓	✓
LPIQYEPVL	B*35:03	8	13	3	✓	✓	✓
KLKLP**I**IMK	A*03:01	60	54	14347	✓	✓	✗
GRTGAGKSFL	B*27:05	306	313	222	✓	✓	✓
K**L**ILWRGLK	A*03:01	1805	1868	1337	✓	✓	✓
ASWVVPIDI**K**	A*03:01	8848	9019	12418	✓	✓	✗
Sample *Mel*5 - 25807 candidates							
ETS**K**QVTRW	A*25:01	137	113	102	✓	✓	✓
YIDE**R**FERY	A*01:01	99	73	29	✓	✓	✓

Supplementary Table 2: **Ranking and HLA-I restriction prediction for neoantigens from cancer-related samples** (12T from [3,4] and Mel8, Mel5 and Mel15 from [5]). The mutated residue in each neoantigen sequence is marked as a bold letter. All neoantigen candidates (whose number is reported in the header) were ranked according to the presentation scores assigned by RBM-MHC (see equation 9 in Methods) in decreasing order. RBM-MHC predictions are compared to the same ranking by the two unsupervised approaches, MixMHCp and RBM-km (both trained on the same single-patient dataset). For each method, we have highlighted in blue ranks in the top 1%, *i.e.* higher than the 99% of the full set of neoantigen candidates. This comparison shows that: the RBM-km attains a performance comparable to RBM-MHC (8/11 ranked among top 1%, expected HLA restriction systematically retrieved), emphasizing the key role of RBM modeling for predictive power; MixMHCp provides accurate predictions (4/11 times gives higher rank than RBM-MHC) but in a less robust way, missing 2 neoantigens among the top 1% and 2 HLA-I restrictions that RBM-MHC and RBM-km assess well. The same ranks estimated by NetMHCpan4.0 and MixMHCpred, available respectively in [6, 7] (apart from the neoantigen GVYPMPGTQK, detected in a more recent analysis [4]), are overall higher than RBM-MHC ones, largely due to training on large-scale, optimized datasets. However, NetMHCpan4.0 fails at placing in the top 1% the same 3 neoantigens for which the RBM-MHC fails, and so does MixMHCpred for 2 of them. Hence, the RBM-MHC bears the comparison to these methods, given that it is trained on a moderate-size, patient-derived dataset instead of a large-scale, optimized dataset as NetMHCpan4.0 and MixMHCpred.

Supplementary Note

Sequence re-weighting scheme. An important question is whether a RBM presentation model, trained on datasets obtained by mass spectrometry (MS), is able to properly score the probability of presentation of ligands that are less frequently detected by MS. The most evident case are ligands containing cysteine, an amino acid that can undergo chemical modifications by oxidation typically not included in standard MS searches. We introduce a procedure for re-weighting sequences as a general strategy to correct for biases in the training dataset, such as MS detection biases. We define a weight for each sequence $\boldsymbol{v}$ in the training set:

$$\alpha(\boldsymbol{v}) = \prod_{i=1}^{N^v} \frac{f^{back}(v_i)}{f^{MS}(v_i)}, \tag{12}$$

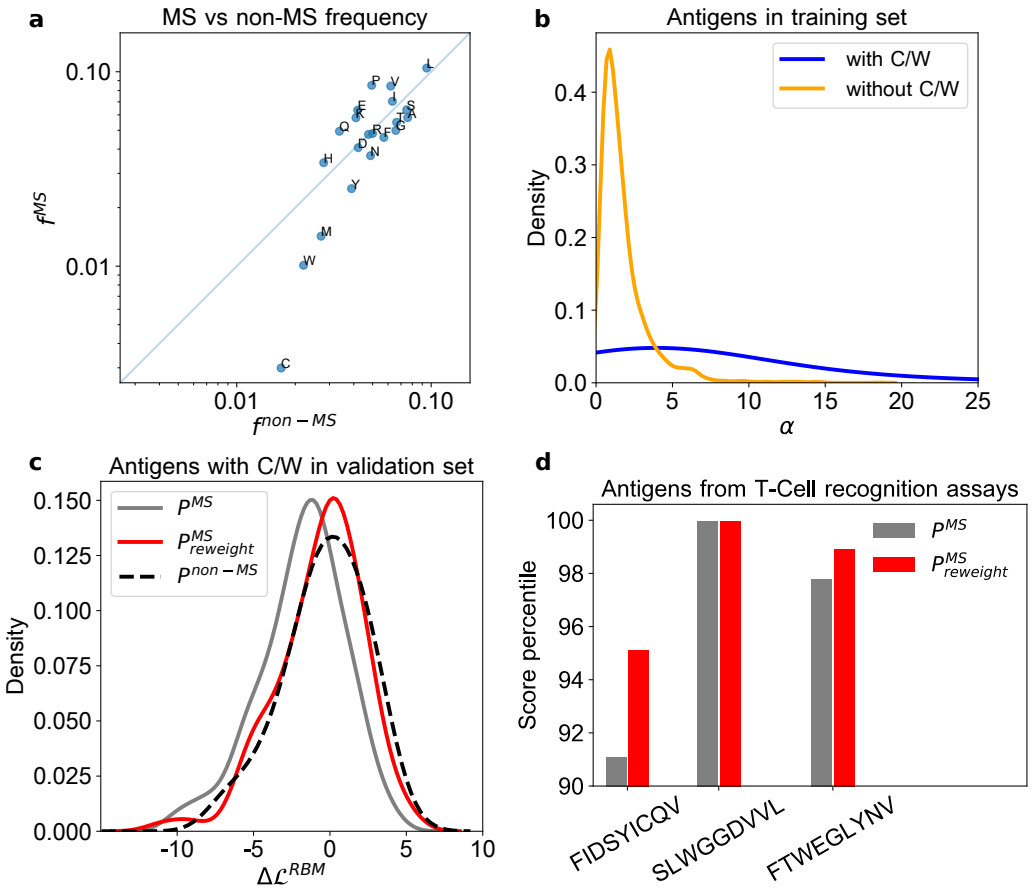
where f^{MS} denotes typical amino acid frequencies in MS (as can be estimated from IEDB data) while f^{back} is a background amino acid frequency distribution that does not suffer from the same biases as the MS one. These weights $\alpha(\boldsymbol{v})$ are incorporated into the average log-likelihood to maximize during RBM training:

$$\mathscr{L}^{RBM} = \frac{1}{\sum_{\boldsymbol{v}} \alpha(\boldsymbol{v})} \sum_{\boldsymbol{v}} \alpha(\boldsymbol{v}) \log(P(\boldsymbol{v})). \tag{13}$$

When modelling multiple specificities through mixtures of PWMs (as by MixMHCp and MixMHCpred, see [7]), an analogous correction can be implemented by rescaling the amino acid frequency composing each PWM by the background frequency. Following Ref. [7], f^{back} can be chosen as the frequency of amino acids in the human proteome or as the frequency of amino acids in IEDB antigens detected by other techniques than MS (we shall refer to these as f^{non-MS}). More precisely, f^{non-MS} is estimated from frequencies at non-anchor sites (position 4 to position 7 of 9-mers), excluding from the search all alleles that: *(i)* have fewer than 100 peptides associated to them; *(ii)* are present in neither MS nor non-MS datasets; *(iii)* show specificity at positions between 4 and 7.

In Supplementary Fig. 13a, the comparison of f^{non-MS} and f^{MS} for the 20 amino acids provides a clear indication of the MS detection bias in relation to the amino acids cysteine (C) and tryptophan (W), the frequency difference between MS detection and other techniques being more than 100% of MS frequency itself. The C/W frequency in presented antigens is therefore expected to be underestimated by MS, suggesting that ligands containing C/W, whose binding affinity to HLA-I could be successfully tested *in vitro*, would be missed by MS. For the purpose of illustration, we test how the re-weighting procedure could correct for such bias on a presentation model built from IEDB, MS data for the allele HLA-A*01:01, with $f^{back} = f^{non-MS}$. We first compute the weights α (Eq. (12)) for sequences in the training dataset. In Supplementary Fig. 13b, we show separately their distribution for sequences with and without C/W: for the former, α have particularly high values, in such a way as to weight more their statistical information and to compensate for the underestimation of C/W frequencies in MS. The effect of re-weighting can be then assessed on a validation dataset, in particular on its sequences with C/W, looking at the distribution of their presentation scores relative to the average scores over the full validation set (see Supplementary Fig. 13c). When considering a RBM model trained without re-weighting (giving the probability P^{MS}), these sequences would be assigned a score lower than the average score of the validation set, as a straight consequence of the low amount of ligands with C/W in the training set. Once the RBM model is learned with re-weighting (giving the probability $P^{MS}_{reweight}$), the score assigned to these sequences is brought closer to the average value and in particular to overlap with the range of presentation score values that would be assigned by a RBM model (P^{non-MS}) trained on IEDB data for the same HLA allele from techniques different from MS (*i.e.* data that do not underestimate the occurrence of C/W in presented antigens). The re-weighting scheme can also improve the ability of the RBM model to discriminate antigens containing C/W of validated immunogenicity from generic peptides, by assigning them presentation scores of higher rank. We show this in Supplementary Fig. 13d, where we consider a re-weighted version of the HLA-A*02:01-specific model of Figs. 1c-d in the main text. While there (as well as in the analysis of Figs. 1e and 2a-c) the re-weighting was not necessary, as we validated the model on the same type of data (MS-detected) as the training set (MS-based), in Supplementary Fig. 13d we show that the re-weigthing is useful for scoring by a MS-based model peptides from binding assays and that contain amino acids underdetected by MS. Similarly, another scenario where we applied the re-weigthing scheme is the prediction, by MS models, of Sars-cov-2 epitopes tested *in vitro* for binding, as in Fig. 2d.

The re-weighting procedure just described can be activated as an option in the RBM-MHC software package.

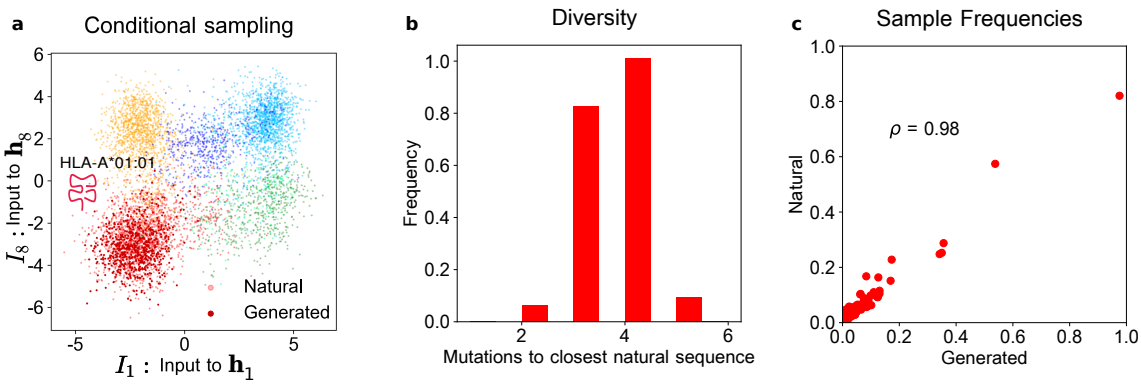


Supplementary Figure 13: **Sequence re-weighting scheme**. **a**: Amino acid frequencies in IEDB antigens detected by MS (f^{MS}) *vs* all other techniques (f^{non-MS}): cysteine (C) and tryptophan (W) frequencies are the lowest both for MS and non-MS antigens and in particular their MS estimate has the largest deviation towards very low values compared to the non-MS one, with $f^{MS} \leq 0.01$. **b-c**: Test of the re-weighting procedure in a RBM model for presentation by HLA-A*01:01 trained on a set of MS data in IEDB (with ligand length in the range of 8-11 residues). **b**: Distribution of weights αs for ligands in the training dataset with and without C/W. **c**: Distribution of differences of presentation log-likelihoods defined as $\Delta\mathcal{L}^{RBM}(\boldsymbol{v}) = \mathcal{L}^{RBM}(\boldsymbol{v}) - \langle\mathcal{L}^{RBM}\rangle_{val}$, where $\mathcal{L}^{RBM}(\boldsymbol{v}) = logP(\boldsymbol{v})$, $\langle\cdot\rangle_{val}$ stands for the average over the full validation dataset and $\boldsymbol{v}$ are sequences in the latter containing C/W. The 3 curves correspond to choosing either $P = P^{MS}$ (from the RBM model trained on MS data without re-weighting), or $P = P^{MS}_{reweight}$ (from the RBM model trained on MS data with re-weighting) or $P = P^{non-MS}$ (from the RBM model trained on data obtained by techniques different from MS). **d**: Effect of re-weighting applied to the RBM model for presentation by HLA-A*02:01 of Figs. 1c-d (trained on IEDB, monoallelic MS-detected ligands of length 8-11 residues). Experimentally validated immunogenic peptides from T-cell recognition assays containing C/W (FIDSYICQV from [8], SLWGGDVVL and FTWEGLYNV from [9]) are assigned an equal or higher RBM presentation score by the re-weighted model ($P^{MS}_{reweight}$) compared to the one without re-weighting (P^{MS}). Thanks to re-weighting, score percentiles, relative to the score distribution of the 5000 generic peptides considered in Figs. 1c-d, are all brought above 95%.

RBM generative performance guided by the HLA-I classifier. The RBM model is generative. As it is based on fitting an entire probability distribution to a given dataset, sampling from this distribution allows one to generate new candidate antigens. The binding specificity to HLA-I of such generated sequences can be controlled by conditioning (fixing) the RBM probability on the values of inputs to hidden units coding for the desired specificity. The search for these values is guided by the HLA-I classifier. This procedure directly builds on the idea of sampling while conditioning on structural, functional, phylogenetic features emerging in RBM representations of protein families [1]. Schematically, the steps of this HLA-specific sampling are as follows:

1. we select a HLA-I class c (*e.g.* c = HLA-A*01:01) and we find $\boldsymbol{I}^o$ such that $\hat{y}_c(\boldsymbol{I}^o)$ is close to 1 (that is, there is high confidence of class c prediction);
2. we estimate $\boldsymbol{h}^o = \langle \boldsymbol{h} \rangle$ from the conditional probability $P(\boldsymbol{h}|\boldsymbol{I}^o)$ and we sample new sequences $\boldsymbol{v}$ from the probability of observed sequences conditioned on $\boldsymbol{h}^o$, $P(\boldsymbol{v}|\boldsymbol{h} = \boldsymbol{h}^o)$;
3. to further explore the region encoding for the specificity to HLA-I class c, we randomly move from $\boldsymbol{I}^o$ to $\boldsymbol{I}^n = \boldsymbol{I}^o + \delta \boldsymbol{I}$ ($\delta \boldsymbol{I}$s drawn from a Gaussian $\mathcal{N}(0, \sigma)$);
4. we accept the move with a probability $\pi \sim \exp(-\beta(e^n - e^o))$, where $e^n = -log(\hat{y}_c(\boldsymbol{I}^n))$ and $e^o = -log(\hat{y}_c(\boldsymbol{I}^o))$. The parameter β (akin to an inverse temperature in physics) basically controls how stringent the selection for sequences predicted in class c with a high probability is.
5. We set $\boldsymbol{I}^o = \boldsymbol{I}^n$ to proceed with new moves as in step 3. Every arbitrary amount of moves, we generate new configurations by conditional sampling as in step 2.

Supplementary Fig. 14a shows examples of $\boldsymbol{I}^n$ values (2 of its components for simplicity) covered in this search with $\beta = 50$ and $\sigma = 0.01$ (dark red points). By this sampling procedure, we generate samples of artificial peptides that would be predicted to be presentable and specifically presented by the selected HLA-I protein complex. Such samples explore a broad diversity of peptides, that are typically 3-4 mutations away from the closest natural ones (Supplementary Fig. 14b) while preserving the profile of amino acid abundances constrained by a given binding specificity (Supplementary Fig. 14c).



Supplementary Figure 14: **Generation of HLA-specific artificial peptides**. **a**: 2-dimensional visualization of the RBM representation space of inputs to hidden units $\boldsymbol{I}$ - as in Fig. 1b. Inputs for antigen sequences with HLA-A*01:01 restriction (points in light red color) mark a region encoding for this binding specificity. We explore more systematically this region by making small, stochastic moves in this space (points in dark, red color) and by accepting them only if the classifier would predict with high confidence that the sampled antigens specifically bind to HLA-A*01:01 (see text). In this example, every 100 steps we condition on the current values of inputs to hidden units and sample 1000 configurations for a total of $M_g = 11000$ generated sequences. **b**: Distribution of the diversity of the generated sequences with the selected HLA-A*01:01 specificity, quantified as number of mutations away from the closest natural sequence of the same specificity. **c**: Scatter plot comparing positional amino acid frequency of natural and generated antigen sequences with same binding specificity, HLA-A*01:01.

References

1. Tubiana J, Cocco S, Monasson R. Learning Protein Constitutive Motifs from Sequence Data. eLife. 2019;8:e39397.

2. Gfeller D, Guillaume P, Michaux J, Pak HS, Daniel RT, Racle J, et al. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. The Journal of Immunology. 2018;201(12):3705–3716.

3. Kalaora S, Barnea E, Merhavi-Shoham E, Qutob N, Teer JK, Shimony N, et al. Use of HLA Peptidomics and Whole Exome Sequencing to Identify Human Immunogenic Neo-Antigens. Oncotarget. 2016;7(5):5110–5117.

4. Kalaora S, Wolf Y, Feferman T, Barnea E, Greenstein E, Reshef D, et al. Combined Analysis of Antigen Presentation and T-Cell Recognition Reveals Restricted Immune Responses in Melanoma. Cancer Discovery. 2018;8(11):1366–1375.

5. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct Identification of Clinically Relevant Neoepitopes Presented on Native Human Melanoma Tissue by Mass Spectrometry. Nature Communications. 2016;7(1):1–16.

6. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology. 2017;199(9):3360–3368.

7. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I Motifs across HLA Peptidomes Improves Neo-Antigen Predictions and Identifies Allostery Regulating HLA Specificity. PLOS Computational Biology. 2017;13(8):e1005725.

8. Rufer N, Wolpert E, Helg C, Tiercy JM, Gratwohl A, Chapuis B, et al. HA-1 and the SMCY-Derived Peptide FIDSYICQV (H-Y) Are Immunodominant Minor Histocompatibility Antigens after Bone Marrow Transplantation. Transplantation. 1998;66(7).

9. Amir AL, van der Steen DM, Hagedoorn RS, Kester MGD, van Bergen CAM, Drijfhout JW, et al. Allo-HLA–Reactive T Cells Inducing Graft-versus-Host Disease Are Single Peptide Specific. Blood. 2011;118(26):6733–6742.

10. Immunitrack, Intavis. Covid19 Intavis_Immunitrack Stability Dataset 1; 2020.

11. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host & Microbe. 2020;p. S1931312820301669.