

Supplementary Materials for
**Computational prediction of MHC anchor locations guides neoantigen
identification and prioritization**

Huiming Xia *et al.*

Corresponding author: Malachi Griffith, mgriffit@wustl.edu; Obi L. Griffith, obigriffith@wustl.edu

Sci. Immunol. **8**, eabg2200 (2023)
DOI: 10.1126/sciimmunol.abg2200

The PDF file includes:

Figs. S1 to S13
Legends for data files S1 to S9
Legend for movie S1
References (46, 47)

Other Supplementary Material for this manuscript includes the following:

Data files S1 to S9
Movie S1
MDAR Reproducibility Checklist

Supplementary Methods

Input data for orthogonal evaluation of predicted anchor sites

To evaluate our anchor predictions, we collected 166 protein structures (pdb format) of peptide-MHC complexes and 61 peptide-MHC-TCR complexes from the Protein Data Bank(46) by querying for structures containing macromolecules matching class I HLAs. Structures were additionally reviewed to ensure valid peptide length (8-11) and those with TCRs attached were separated into a different list for downstream analysis to allow accurate solvent-accessible surface area (SASA) calculations. The HLA-peptide structures corresponded to 33 HLA alleles with peptides of varying lengths (8 to 11mer), while the HLA-peptide-TCR structures corresponded to 12 HLA alleles. A complete list of PDB ids selected for this analysis can be found in **Data file S4**.

Orthogonal validation of predicted anchor sites by analysis of pMHC structures

The structures of peptide-MHC molecules were analyzed to infer potential anchor locations/residues. All PDB structures were analyzed in python using the MDTraj package(47). For each position of a peptide bound to an HLA, we utilized two different metrics: 1) minimum distance of non-backbone atoms to all HLA associated atoms and 2) estimated solvent-accessible surface area (SASA) of the residue. In method 1, we calculated the distances between each atom of each residue and all HLA associated atoms. Non-backbone atoms were ordered by their distance to the closest HLA-associated atom and the top 50% were used to calculate an average distance representing an entire residue (with the exception of glycine where all atoms were considered). In method 2, we directly calculated the SASA of each residue (shrake_rupley function in MDTraj), which was used to infer the likelihood of being able to be recognized by the T-cell receptor. After calculating these values for each position of the peptide, they were directly compared to the anchor prediction scores by calculating a Spearman correlation. In the case of the distance metric, we expect positions of the peptide closer to the MHC to be more likely an anchor and those further (“bulging out”) to more likely interact with the TCR. Similarly, for the SASA metric, if a peptide position is more solvently accessible (higher SASA value) we expect it to be more accessible to the TCR as well and those that are less accessible would be more likely interacting with the MHC as an anchor.

For an overall evaluation of how well our anchor predictions correlated with these metrics (distance and SASA), Spearman correlations were determined across all structural data collected. For example, for a 9-mer peptide, a Spearman correlation was calculated for the 9 anchor prediction scores from the in silico mutation exercise compared to the 9 distance or SASA estimates obtained from the structure analysis. Out of 166 peptide-MHC structures collected, correlation values for 87 were plotted by randomly selecting at most 5 structures per HLA-length combination (**Data file S5**). For comparison, we also randomly shuffled distance and SASA scores across all positions of individual peptides and calculated correlation scores against this randomized dataset. The different sets of correlation values were then fit to Gaussian distributions (**Fig. 4d**). Non-paired two sample t-tests assuming unequal variance were performed to evaluate the differences among distributions (using python SciPy `scipy.stats.ttest_ind`).

Additional analysis was performed on the 61 peptide-MHC-TCR structures collected. After randomly selecting at most 5 structures per HLA-length combination, Spearman correlations derived from 31 structures were plotted. Correlations were calculated for 1) distance from peptide to HLA versus anchor prediction scores and 2) distance from peptide to TCR versus anchor prediction scores. Once again,

the HLA-peptide distances were randomly shuffled and used as comparison and two sample t-tests were performed to evaluate the differences among distributions.

Input data for evaluating the impact of anchor site considerations

To evaluate how anchor site considerations might influence neoantigen prioritization decisions, we considered a balanced HLA allele distribution when selecting input data. We randomly sampled up to 10 corresponding TCGA samples for each HLA allele with sufficient data (at least 3 out of 4 lengths have 10 or more matching peptides). 923 TCGA-HLA combinations were chosen from a total of 9,216 TCGA samples excluding the 1,356 used for the seed anchor site prediction data set described above. The 923 TCGA-HLA combinations corresponded to TCGA patients (**Data file S8**). To further evaluate impact of anchor considerations on a patient-specific level, an additional 100 TCGA patients were selected from the original 1,356 TCGA patient samples where we had neoantigen predictions for the patient's full set of HLA alleles (**Data file S8**).

Evaluating the impact of anchor site consideration on neoantigen prioritization

To analyze the importance of positional information on prioritization of neoantigens, TCGA patient samples were used as input and run through pVACtools (version 1.5.2) using the following options: `-e 8,9,10,11, --iedb-retries 50, --downstream-sequence-length 500, --minimum-fold-change 0, --trna-cov 0, --tdna-vaf 0, --trna-vaf 0, --pass-only`. The neoantigen candidates were then filtered and prioritized according to different criteria: **A**) Basic Filter: mutant peptide IC50 < 500 nM and agretopicity > 1, **B**) Decision based on a conventional anchor assumption that anchors are located at position 2 and the C-terminal position, **C**) Decision based on computationally predicted anchor locations. Specifically, under filter A) the position of a mutation with respect to MHC anchor positions is not considered. An accepted peptide means: MT peptide IC50 < 500 nM and WT IC50 / MT IC50 > 1, otherwise the peptide is rejected. Under filter B), anchor positions are defined to be 2 and n for all n-mer peptides. Under filter C), anchor positions are defined by our computational predictions, which are allele-specific. For both filters B and C, peptides are accepted if 1) the MT IC50 < 500 and WT IC50 > 500, or 2) the MT IC50 < 500, WT IC50 < 500 and the mutation is at a non-anchor location, as defined by the anchor definition of filters B and C respectively.

For filter C, anchor positions were defined individually for each peptide using the following strategy: Per anchor calculation results from our computational workflow, each position of the n-mer peptide was assigned a score based on how binding to a certain HLA allele was influenced by mutations. These scores were then used to calculate the relative contribution of each position to the overall binding affinity of the peptide. We ranked the normalized score across the peptide in descending order (e.g. [2,9,1,3,2,8,7,6,5] for a 9-mer peptide) and started summing the scores from top to bottom. Positions that together account for 80% of the overall binding affinity change (e.g. 2,9,1) were assigned as anchor locations for impact analysis. Assuming P_n represents the normalized score of position n within the peptide, for each HLA allele, the anchor(s) is determined as following:

$$\sum_a^n P_n \geq 0.8 \text{ where } P_a > P_b \dots > P_n$$

Filtered lists were then compared for overlap and differences. We also followed the same evaluation process for different starting candidate lists of varying stringency, including: candidates filtered with a strong-binding cutoff of 100 nM, candidates filtered with a strong-binding cutoff of 50 nM and candidates filtered by their binding HLA allele's anchor patterns. For our HLA allele filtered dataset, all

HLA alleles with an exact [2, n] anchor pattern (for n-mer peptides) were considered canonical and excluded from further evaluation.

For our cohort analysis, all neoantigen candidates were considered with no additional filtering. For our patient-level analysis, neoantigen candidates were processed additionally using the `top_score_filter` (“pVACseq `top_score_filter`” command of pVACtools) to generate top neoantigen candidates for individual variants. These top candidates were compiled and the same filters A, B, and C were used to determine prioritization decisions. The percentage differences between filters were calculated based on decisions for all top candidates for each individual patient.

Supplementary Figures

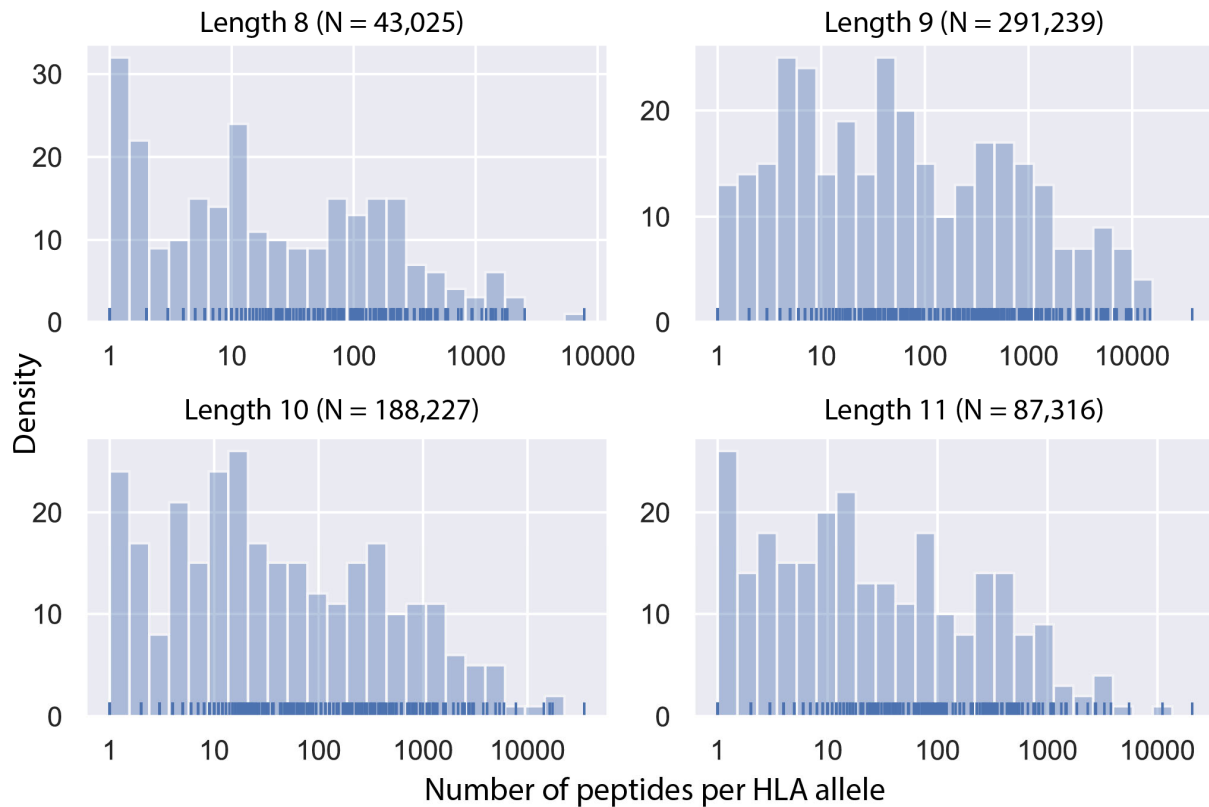


Fig. S1. Distribution of peptides collected per allele across 328 HLA alleles, split by peptide length

Histograms summarizing the distribution of predicted strong binding peptides collected for 328 HLA alleles. Peptides are plotted according to their respective lengths. The total number of peptides for each k-mer is shown as N. The x-axis represents the total number (X) of k-mer peptides for individual HLA alleles while the y-axis represents the number of HLA alleles with X number of peptides matched.

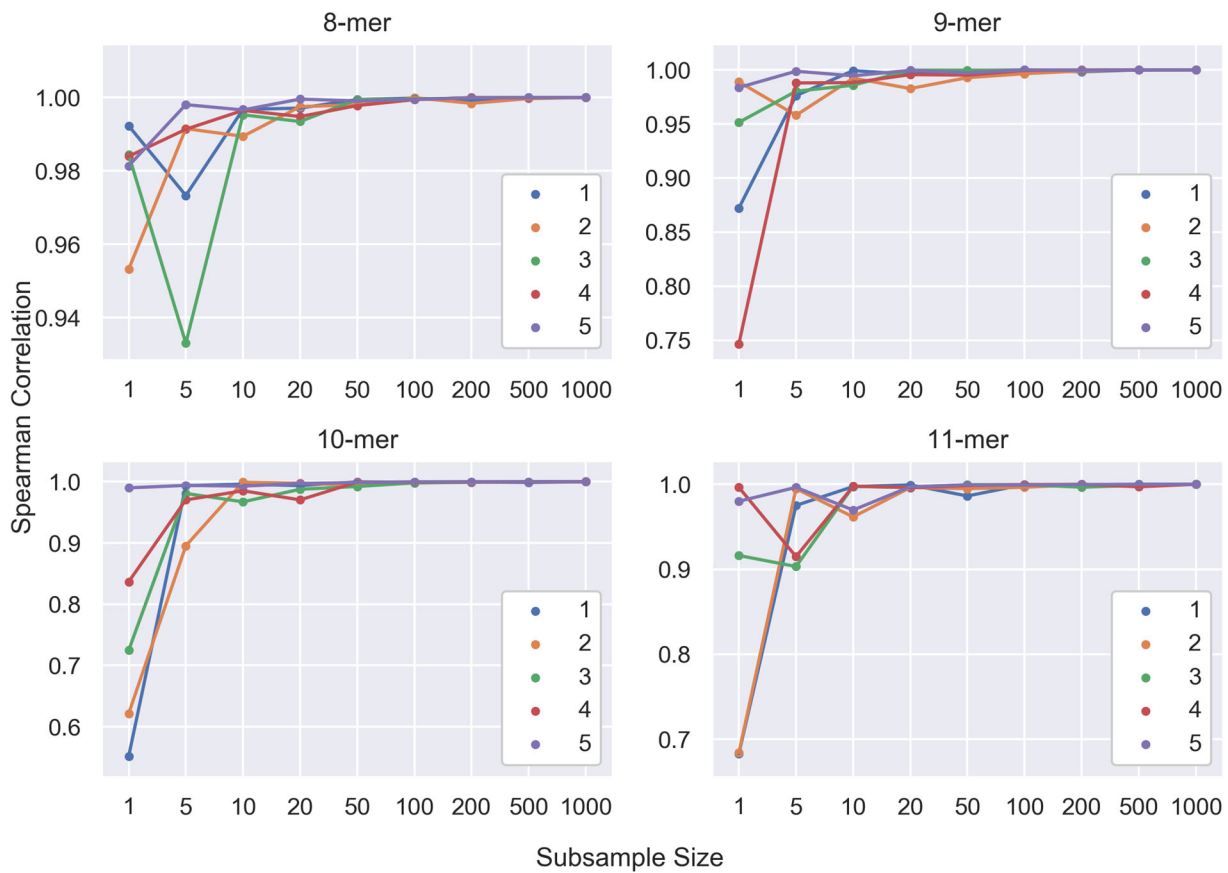


Fig. S2. Saturation analysis for evaluating subsample size of peptides needed for simulation analysis

Results of saturation analyses performed for 8-11-mer peptides using spearman correlation are depicted as line plots with colors indicating random iterations. For each k-mer, 5 random rounds were performed with subset sizes of 1, 5, 10, 20, 50, 100, 200, 500 and 1,000. Correlations were calculated between each subset size and the largest subset (N=1,000) with the assumption that 1,000 samples were adequate to represent a ground truth for this purpose.

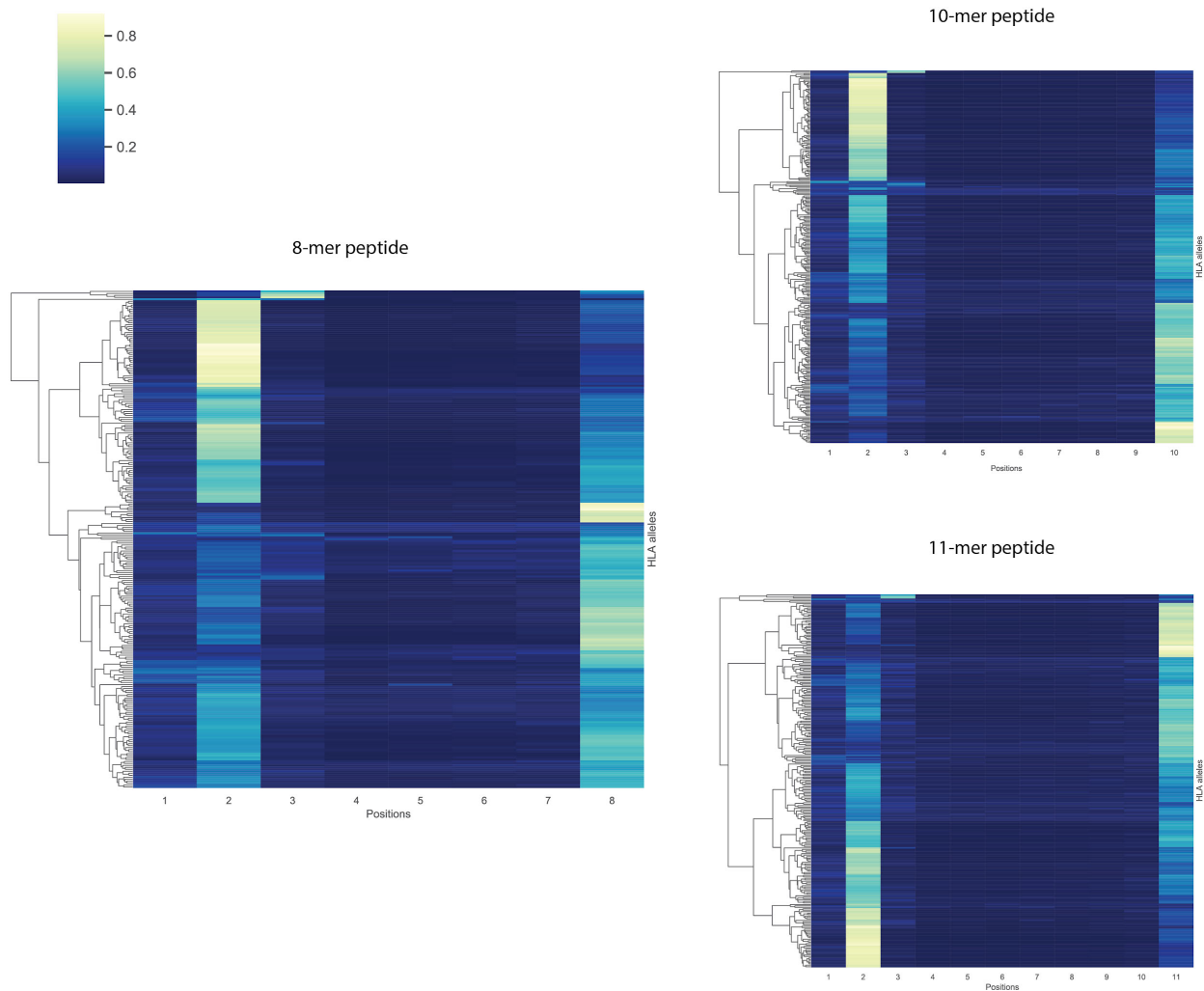


Fig. S3. Hierarchical clustering of anchor prediction scores across all 8, 10, and 11-mer peptides assembled

Heatmaps depict anchor prediction scores clustered using hierarchical clustering with average linkage across all 328 HLA alleles for which 8-mer, 10-mer and 11-mer peptide data were collected. For the individual heatmaps, the x-axis represents the k peptide positions and the y-axis represents the HLA alleles for which k-mer peptides were available.

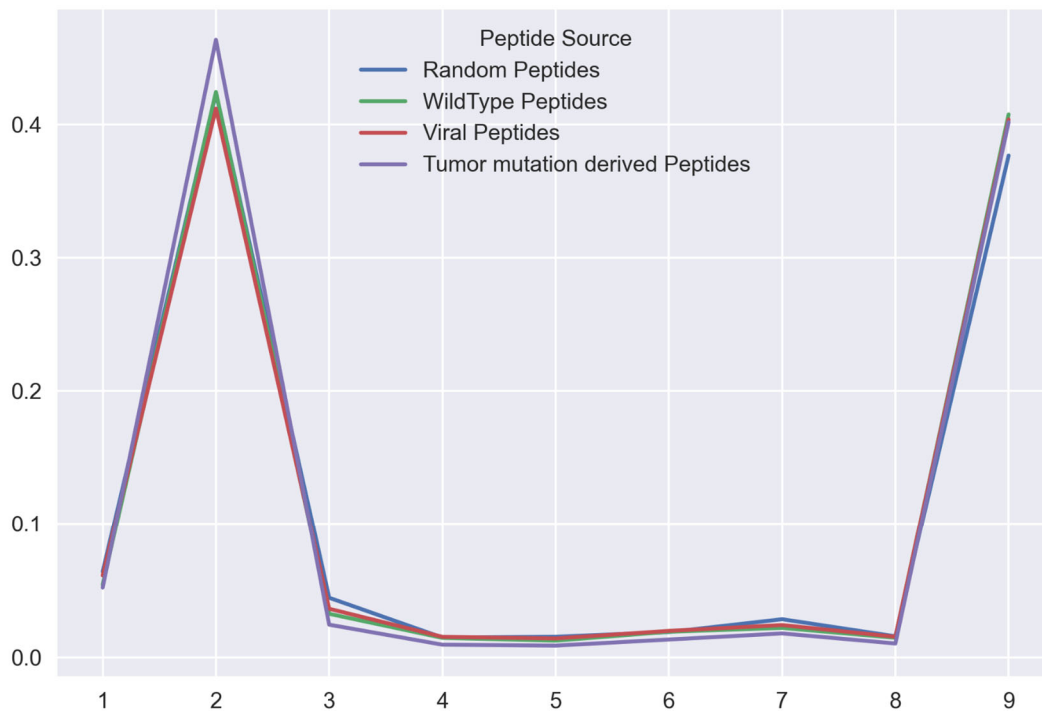
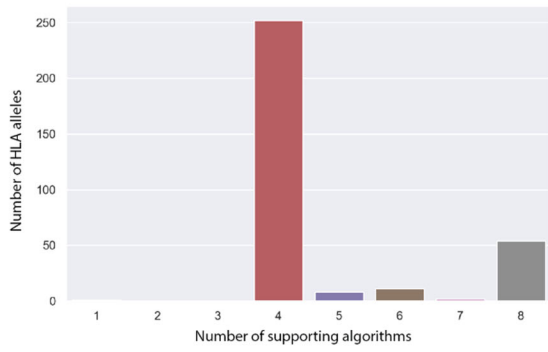


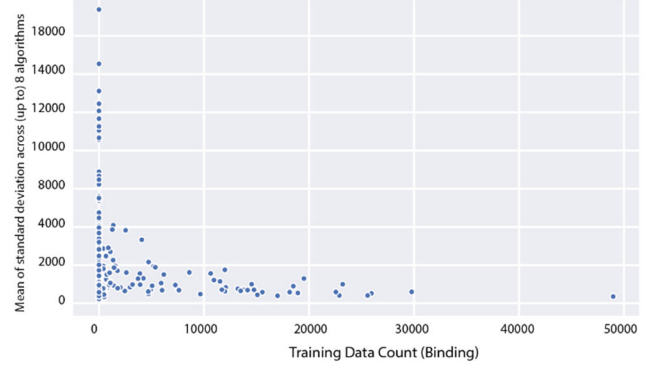
Fig. S4. Comparison of anchor pattern across different seed peptide sources using HLA-A*02:01

Anchor pattern for HLA-A*02:01 calculated from 1,000 peptides for each individual peptide source. We compared results from our seed dataset of peptides (tumor mutation derived) to three qualitatively distinct additional sources of peptides: 1) 9-mer peptides generated from a random 100,000,000 length sequence of amino acids, 2) random 9-mer peptides sampled from the human reference proteome (based on known Ensembl protein sequences) and 3) random 9-mer peptides from a viral proteome (variola). From each source, we identified 1,000 peptides that were predicted to be strong binders (Methods) and mutated each position for each amino acid following our computational workflow. We performed in silico predictions for each of the mutated peptides and calculated anchor probabilities based on all peptides collected. Normalized anchor probabilities are plotted on the y-axis and the peptide positions are listed on the x-axis.

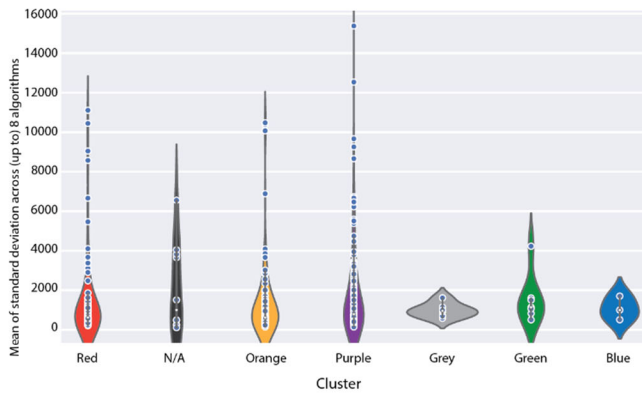
a Number of supporting algorithms for 328 HLA alleles analyzed



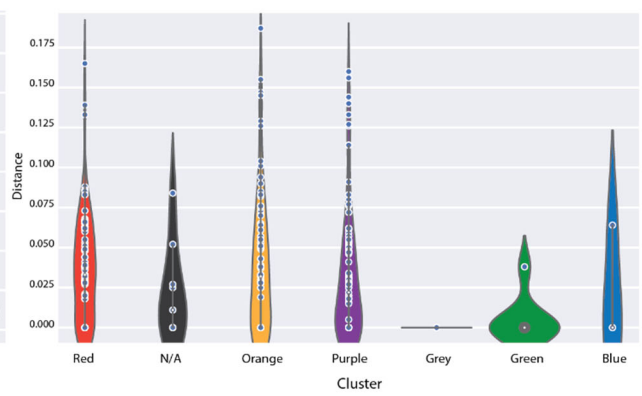
b



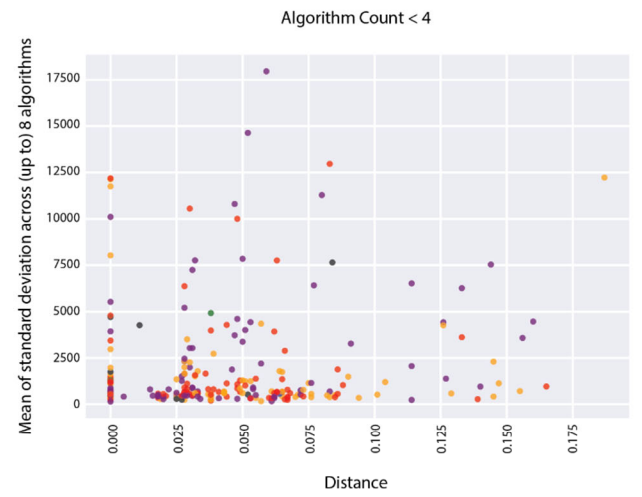
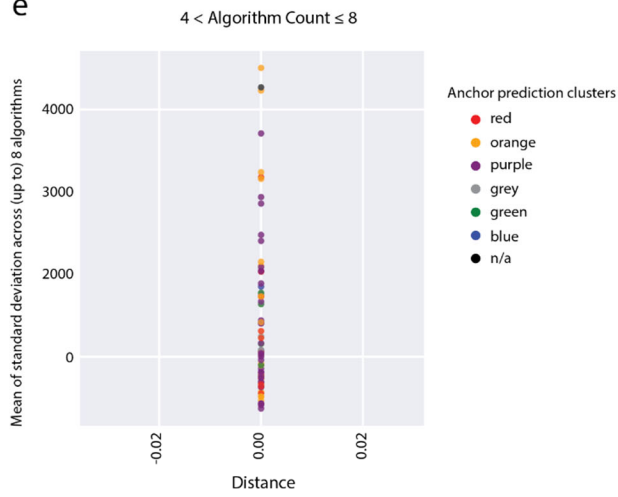
c



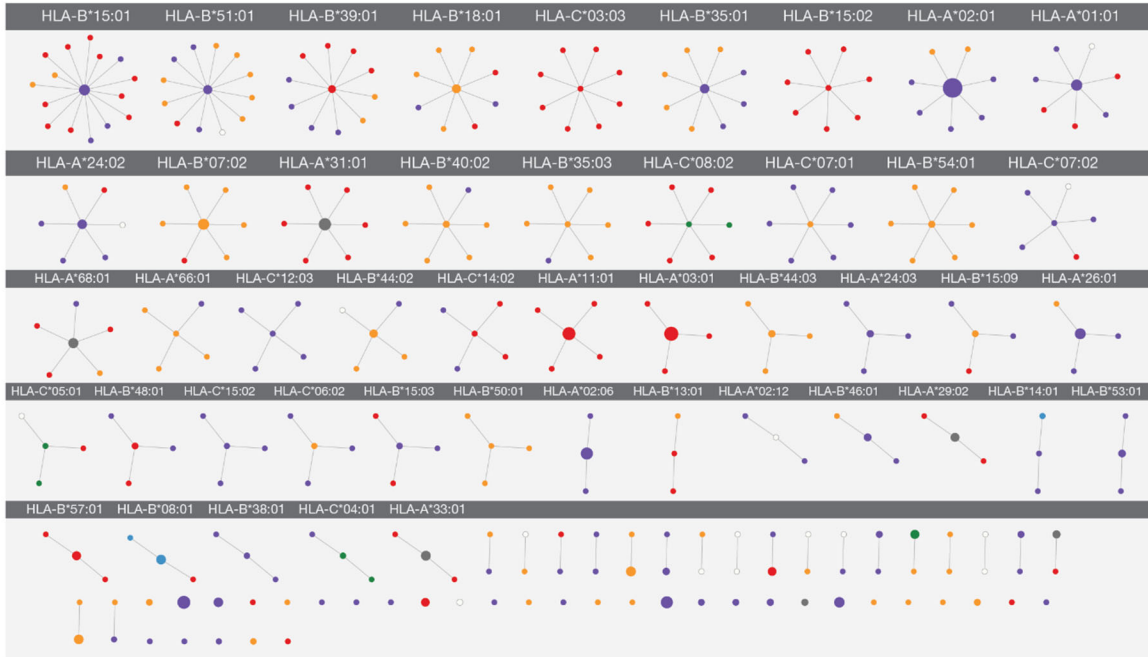
d



e



f



g

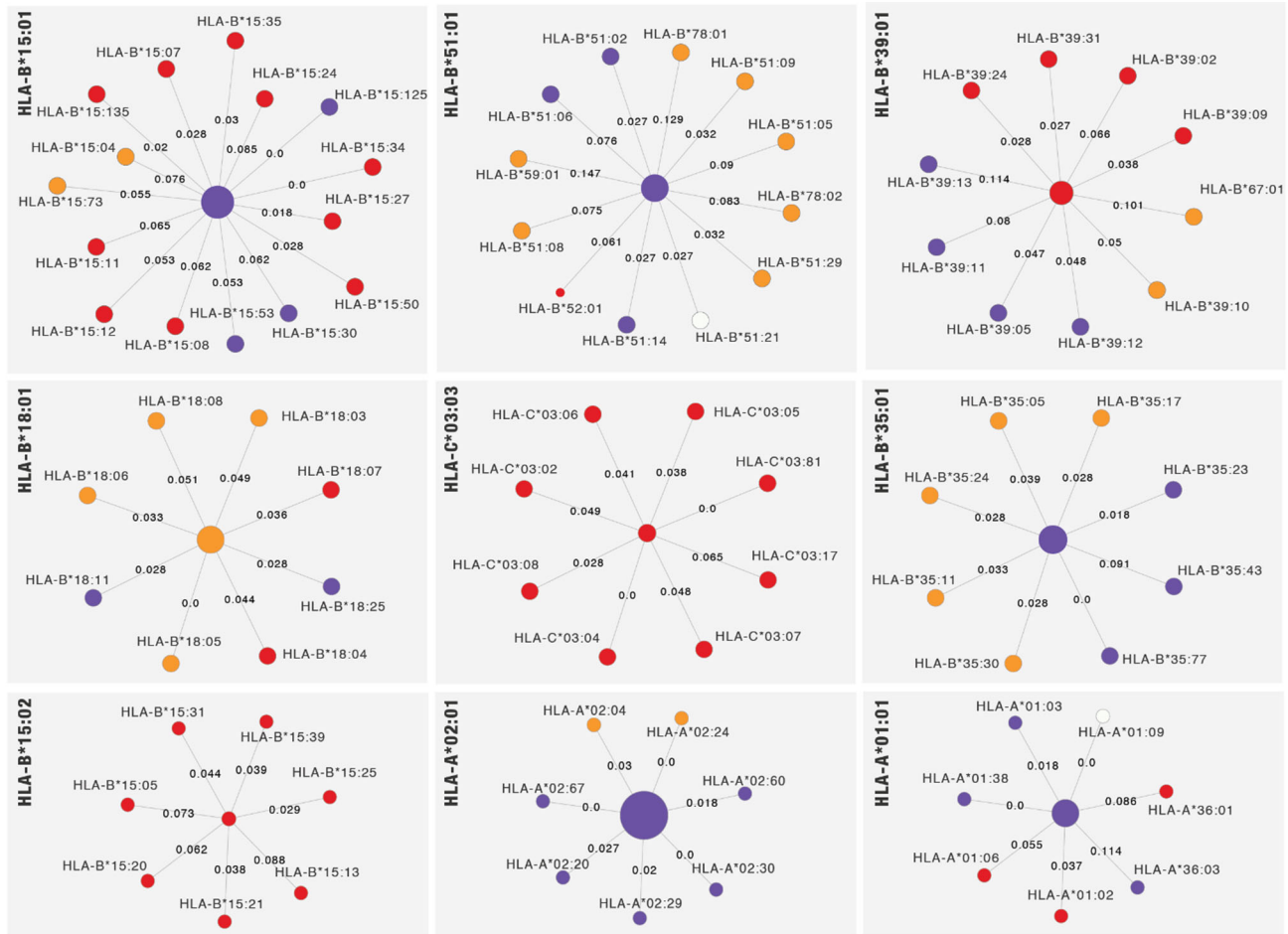


Fig. S5. Analysis of potential for supporting algorithm bias across 328 HLA alleles

a, Bar plot showcasing the number of HLA alleles being supported by the algorithms used. A maximum of 8 algorithms are available with the majority of the HLA alleles being supported by at least 4 algorithms. The specific breakdown is as following: out of the 328 HLA alleles, 54 HLA alleles are supported by all 8 algorithms, 2 are supported by 7 algorithms, 11 are supported by 6 algorithms, 8 are supported by 5 algorithms, 252 are supported by 4 algorithms and 1 is supported by 1 algorithm. **b**, Scatter plot showing distribution of HLA allele anchor pattern clusters with respect to two metrics: Mean value of variances calculated based on up to 8 different algorithmic predictions (x-axis) and distance to nearest HLA allele neighbor according to NetMHCpan (y-axis). If the HLA allele has training data available, then the distance is 0 since the nearest neighbor is only used in cases where the HLA allele in query needs to be estimated based on other similar HLA alleles. Colors are used to denote the clusters as annotated in Fig. 3. **c**, Violin plot showing the distribution of mean value of variances calculated across (up to 8) algorithms across the different anchor clusters. Lower variance indicates a better prediction consistency across the different algorithms used. **d**, Violin plot showing the distribution of distances across the different anchor clusters. Distance of 0 indicates that the HLA allele either did not need a closest neighbor to estimate binding or that the HLA pseudo-sequences between the two alleles were identical. **e**, Scatter plot showing the correlation of distance to the mean variance across different HLA alleles. **f**, Overview of a network graph showing how HLA alleles in our dataset are connected to each other. Each center node has training data available with the size proportional to the size of the center node. Edges connect nodes representing neighboring HLA alleles (as defined by NetMHCpan4.0). Color of each node represents the anchor cluster assigned as in Fig. 3. **g**, Zoomed in view of the network graph in **f** for the top 10 largest networks. Weights on edges reflect the distance between the HLA alleles (output from NetMHCpan4.0). Center HLA allele is listed on the top left of each box in bold.

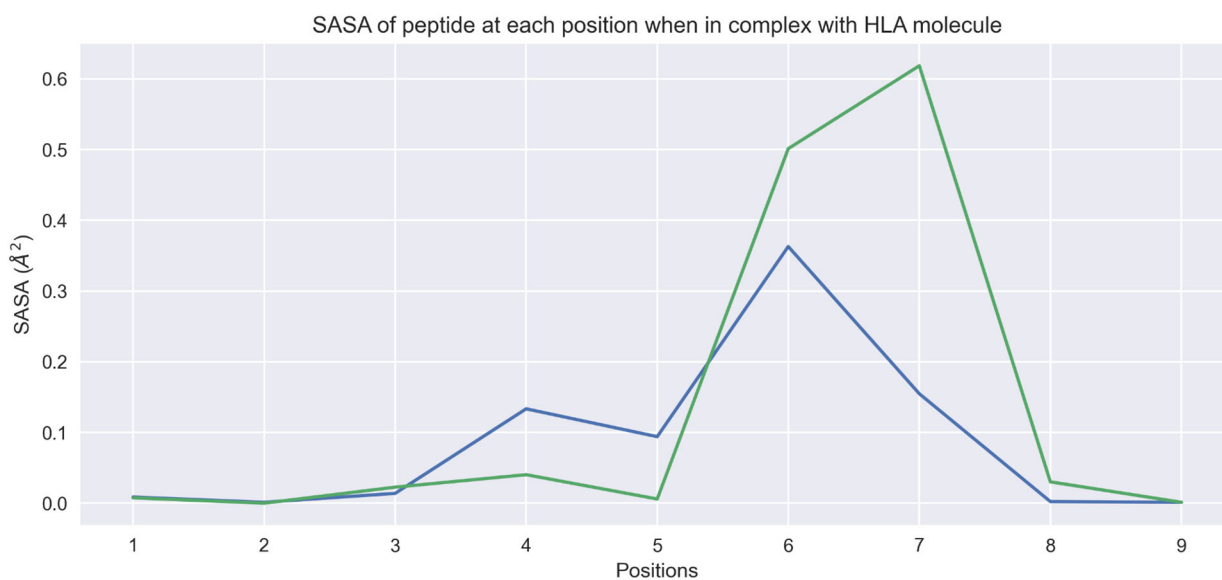
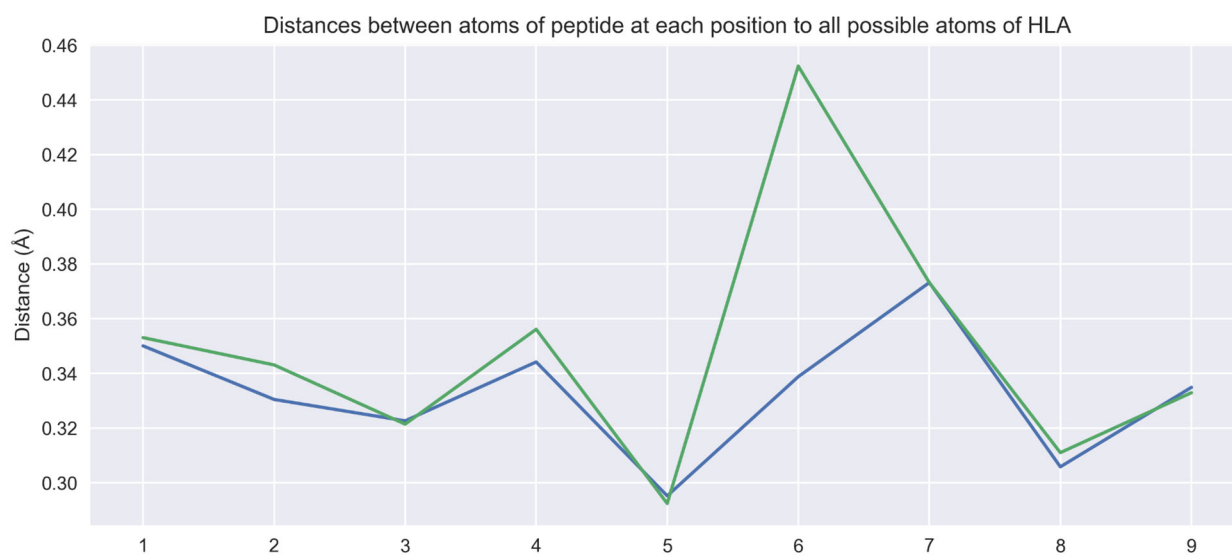


Fig. S6. Analysis of crystallography data for HLA-B*08:01 and 9-mer peptides

Results of two structures produced for HLA-B*08:01 with 9-mer peptides (blue and green lines). Top panel corresponds to distance measurements for each position while the bottom panel corresponds to SASA measurements. X-axis represents positions 1 to 9 of the peptides included.

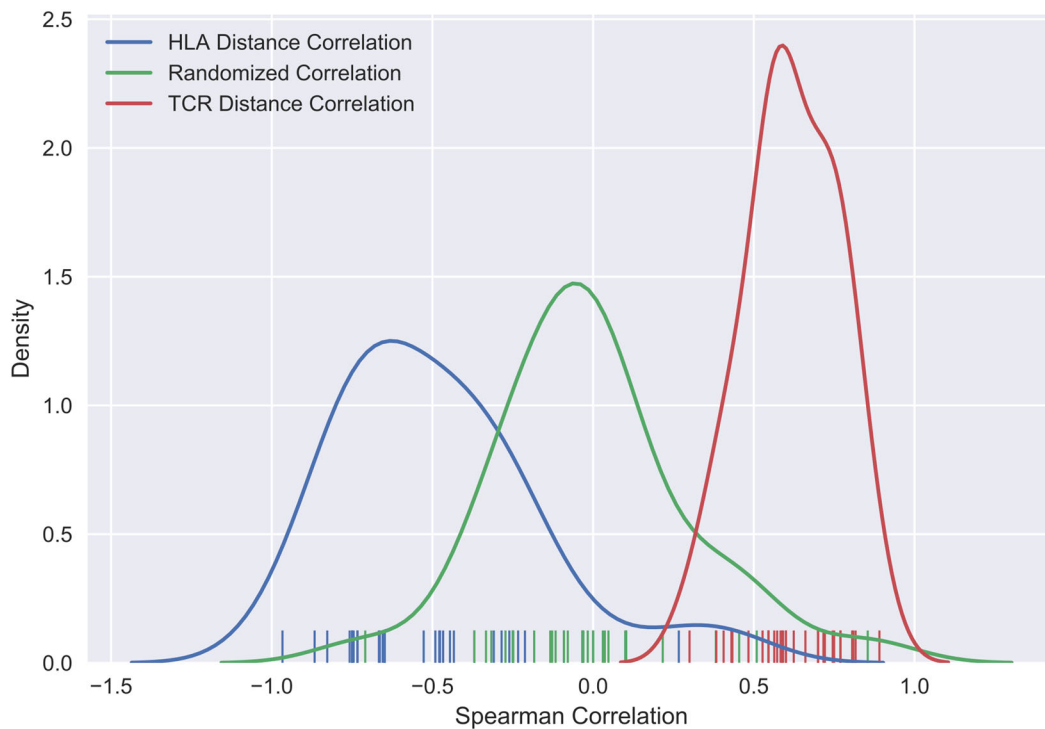


Fig. S7. Distribution of spearman correlation values comparing anchor predictions and peptide-HLA/TCR distance measurements

This analysis is based on a collection of 61 structures obtained from PDB where a specific peptide, MHC and TCR were crystallized as a three part complex. The distribution of Spearman correlation values calculated by comparing our anchor scores to distance metrics obtained from these structures is shown. The blue line represents correlations between the predicted anchor score and the HLA-peptide distance measurements (expected to be negatively correlated because a position with a strong/high anchor score will tend to be closer to the MHC groove). The red line represents correlations between the predicted anchor score and the TCR-peptide distance measurements (expected to be positively correlated because positions acting as an MHC anchor will tend to be further away from the TCR interface). The green line represents the correlations between the predicted anchor score and randomized distances.

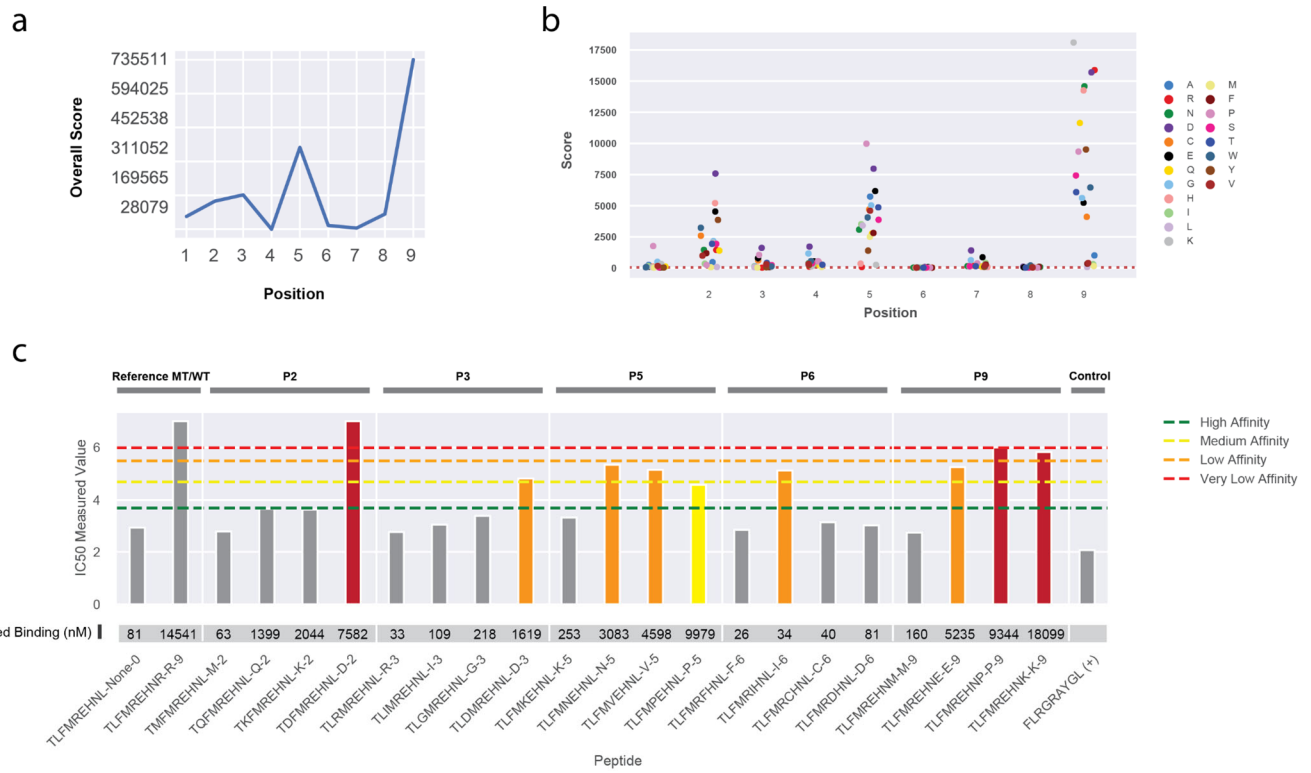


Fig. S8. Additional experimental validation data for the predicted HLA-B*08:01 anchor pattern
 Validation results for HLA-B*08:01 with a predicted 2W-3W-5W-9M anchor pattern using peptide TLFMREHNL. **a**, Overall anchor prediction scores for each position of 9-mer peptides when binding to HLA-B*08:01. **b**, Binding affinity changes (y-axis) plotted for each position of the specific 9-mer peptide TLFMREHNL. Each position of peptide was mutated to 19 other amino acids to evaluate influence on binding affinity. **c**, IC50 values measured from binding affinity assays are plotted for both unmutated and mutated peptides. Peptides are marked by their mutation positions (P2, P3, P5, P6 and P9), predicted binding affinity values. Lower IC50 values correspond to stronger binding between the peptide and HLA allele.

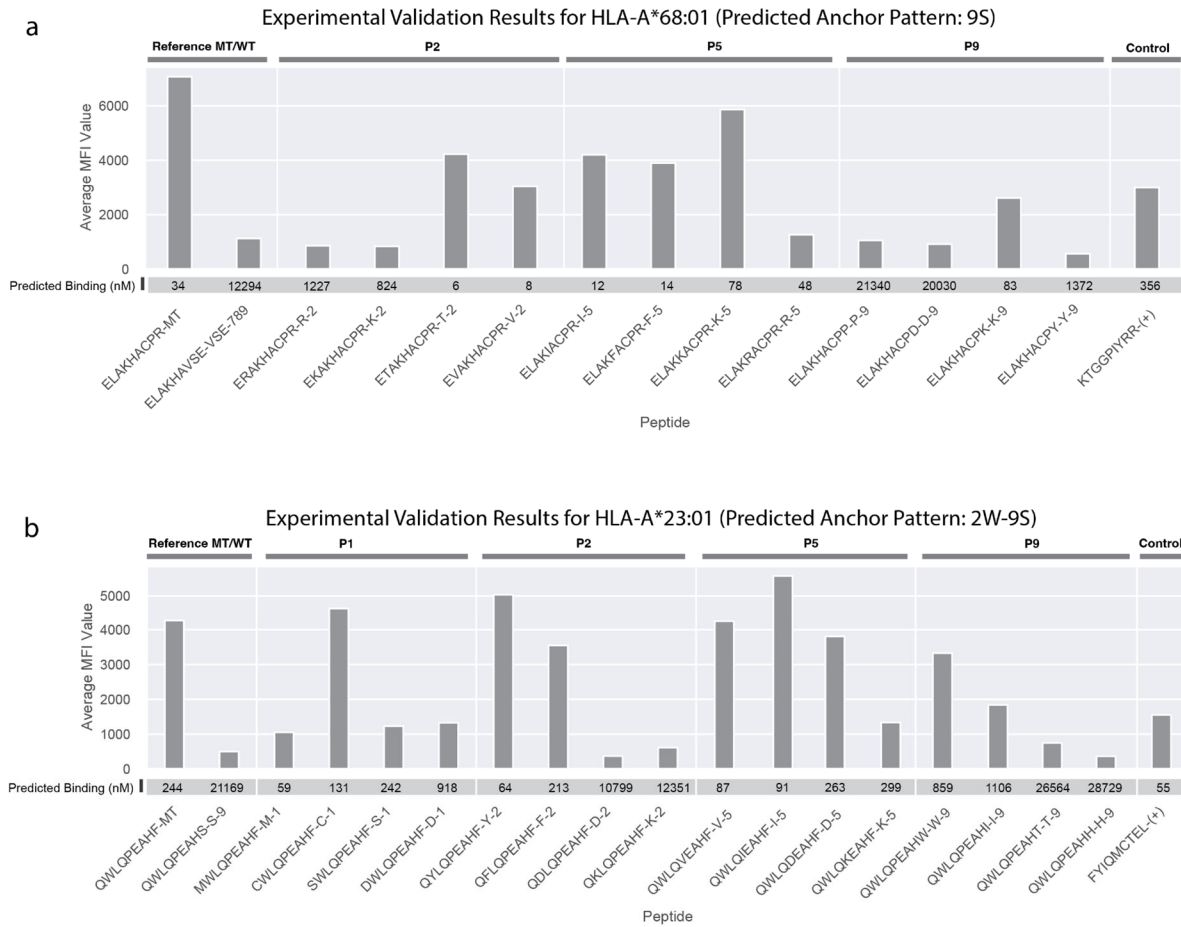
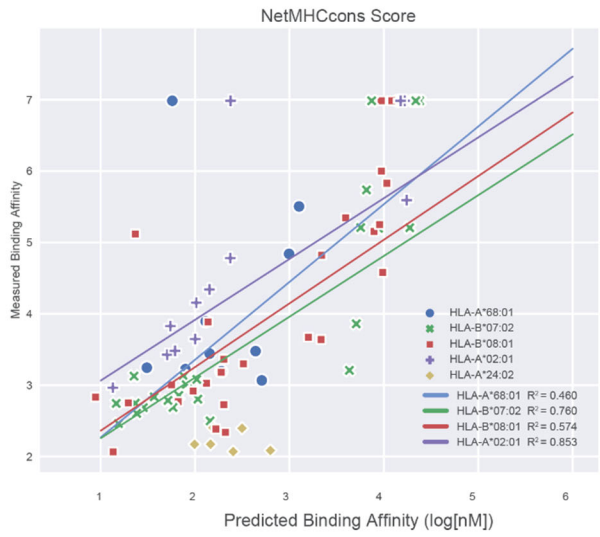
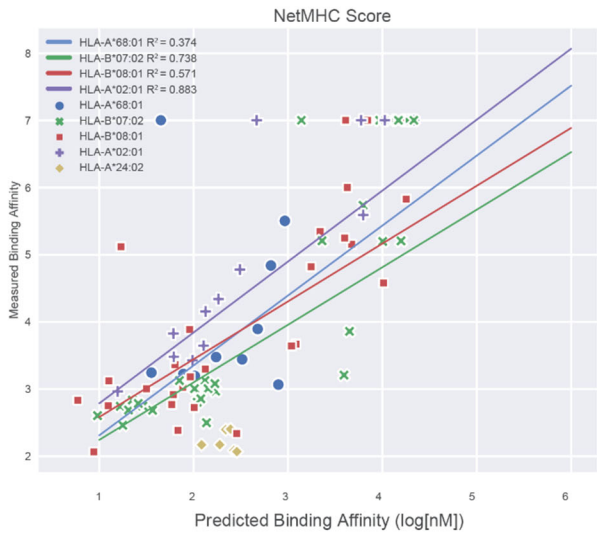
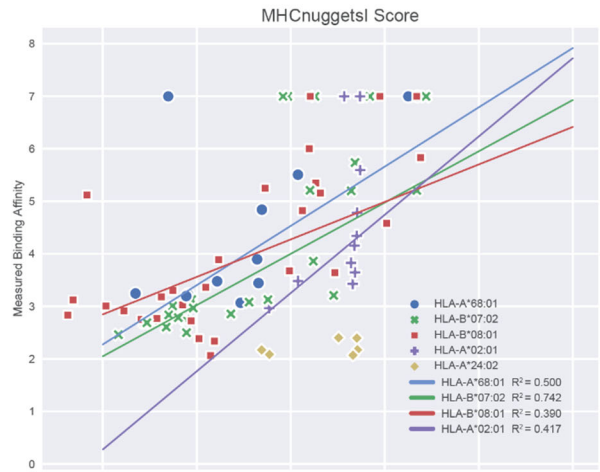
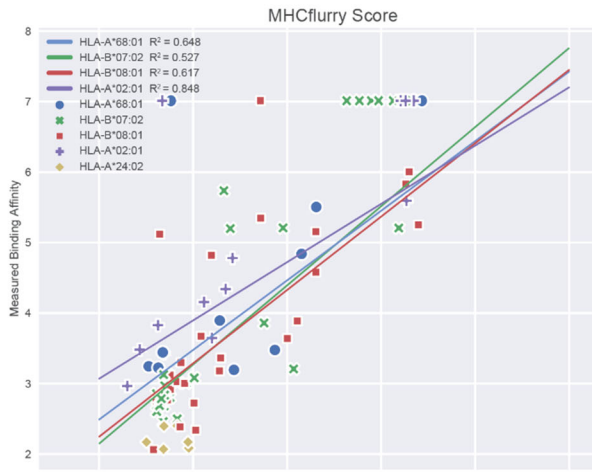


Fig. S9. Additional experimental validation data for HLA-A*23:01 and HLA-A*68:01

a, Validation results for HLA-A*68:01 with a predicted 9S anchor pattern using peptide ELAKHACPR. Median MFI values measured from cell stabilization assays are plotted for both unmutated and mutated peptides at 100 nM concentration. **b**, Validation results for HLA-A*23:01 with a predicted 2W-9S anchor pattern using peptide QWLQPEAHF. Median MFI values measured from cell stabilization assays are plotted for both unmutated and mutated peptides at 20 nM concentration. Peptides are marked by their mutation positions (P1, P2, P5, P9), and predicted binding affinity values. Higher MFI values correspond to stronger binding between the peptide and HLA allele.



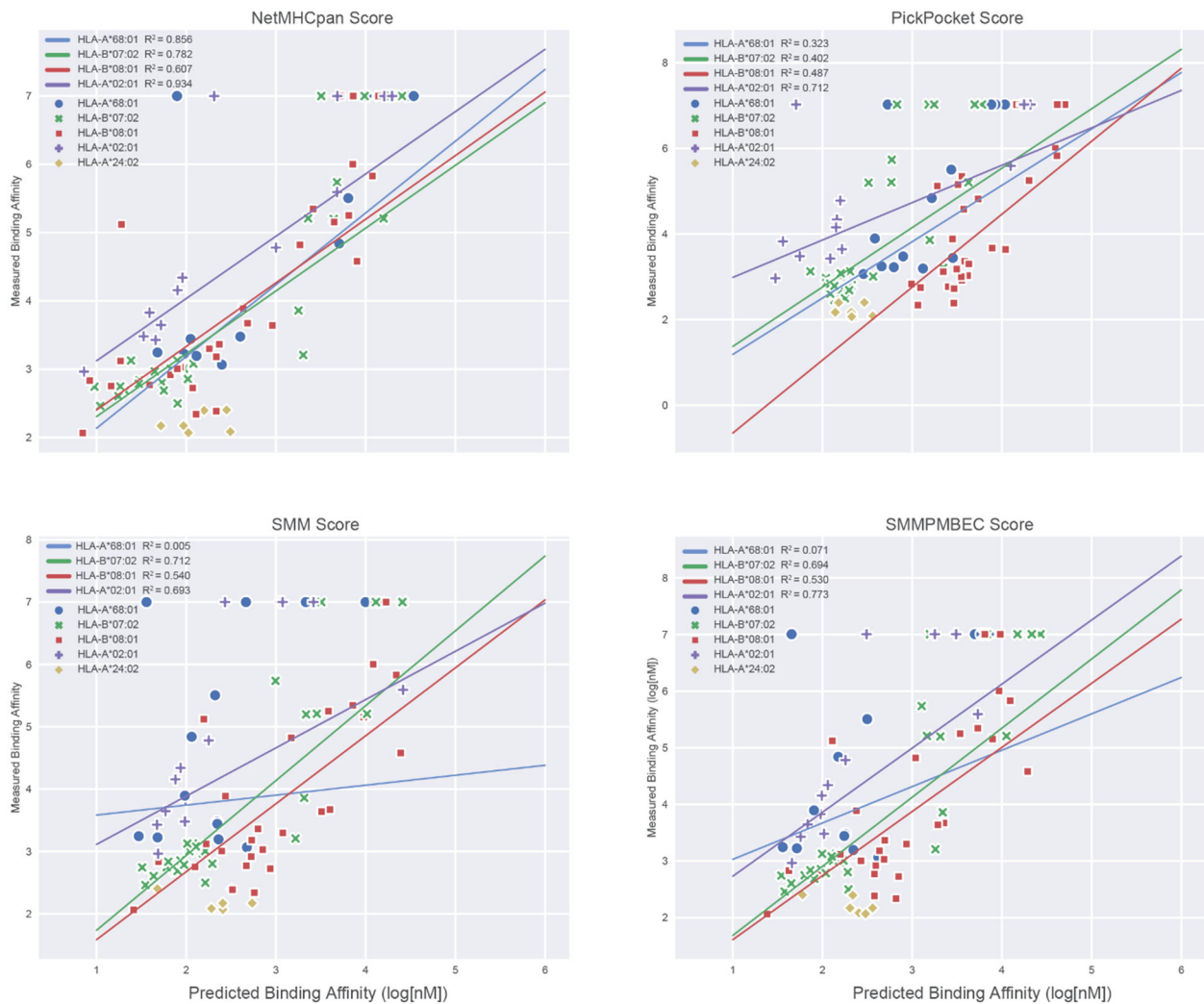


Fig. S10. Breakdown of predicted binding affinity values versus measured binding affinity by individual algorithms

Binding affinity values predicted by each individual algorithm plotted against the measured binding affinity (log[nM]). HLA-A*68:01, HLA-B*07:02, HLA-B*08:01, HLA-A*02:01 were subject to linear fitting with R² values shown (HLA-A*24:02 was excluded due to the limited range of data available). Non-binding peptide-HLA combinations were also excluded from this plot since they have no measured binding affinity value available. Prediction scores are shown for the following 8 different algorithms: MHCflurry, MHCnuggetsI, NetMHC, NetMHCcons, NetMHCpan, Pickpocket, SMM, and SMMPMBEC.

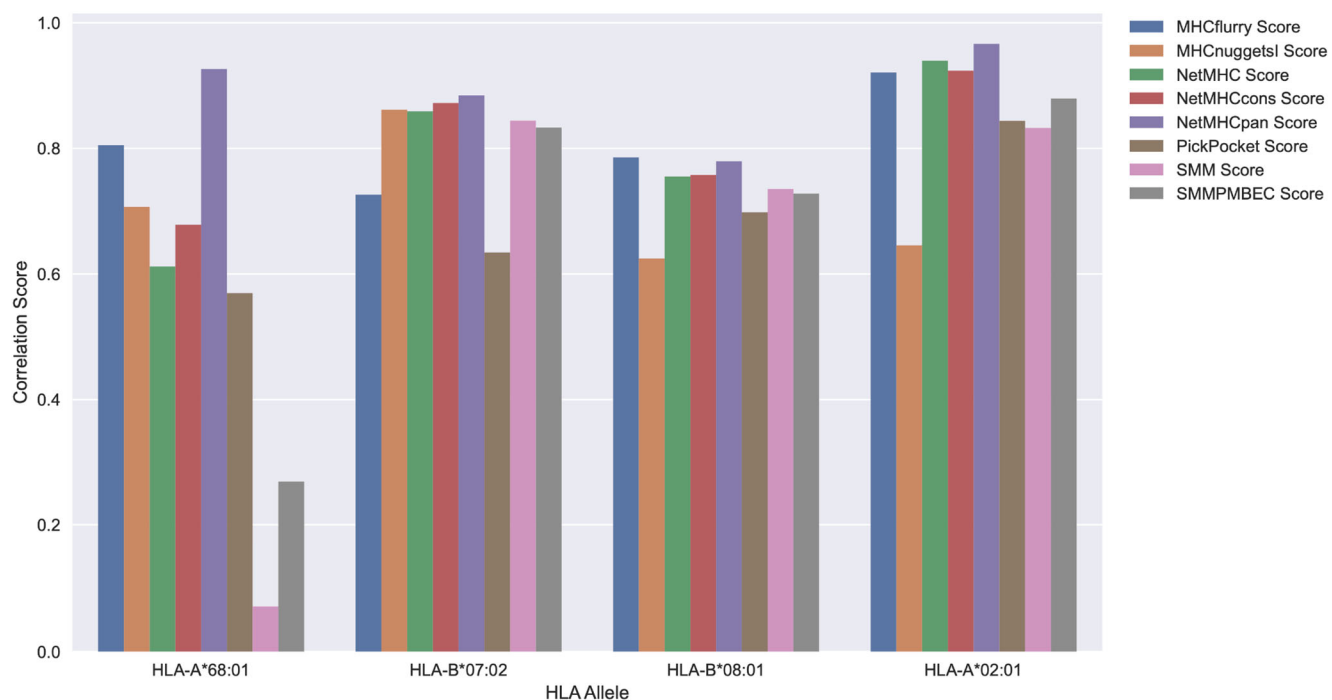


Fig. S11. Correlation scores between individual MHC binding algorithm predictions and measured binding affinities across HLA alleles

Bar plot showing Pearson correlation scores calculated between predicted binding affinity values from 8 different algorithms and the measured IC50 binding affinity values for individual HLA alleles. Plots are grouped by the specific HLA allele and individual algorithms are marked with a color legend.

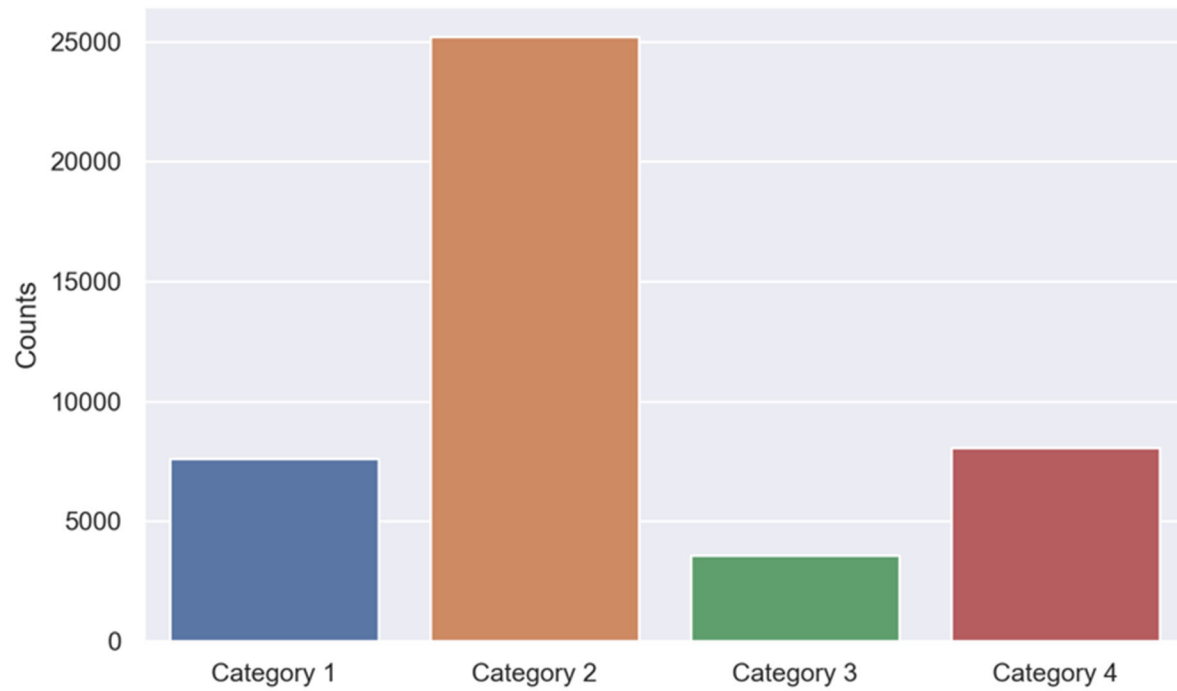


Fig. S12. Distribution of anchor scenario categories for neoantigen candidates from 923 tumor-HLA paired samples

Bar plot showing counts of neoantigen falling into each of the scenarios listed in Fig 7a by analyzing neoantigen candidates from 923 tumor-HLA paired TCGA samples.

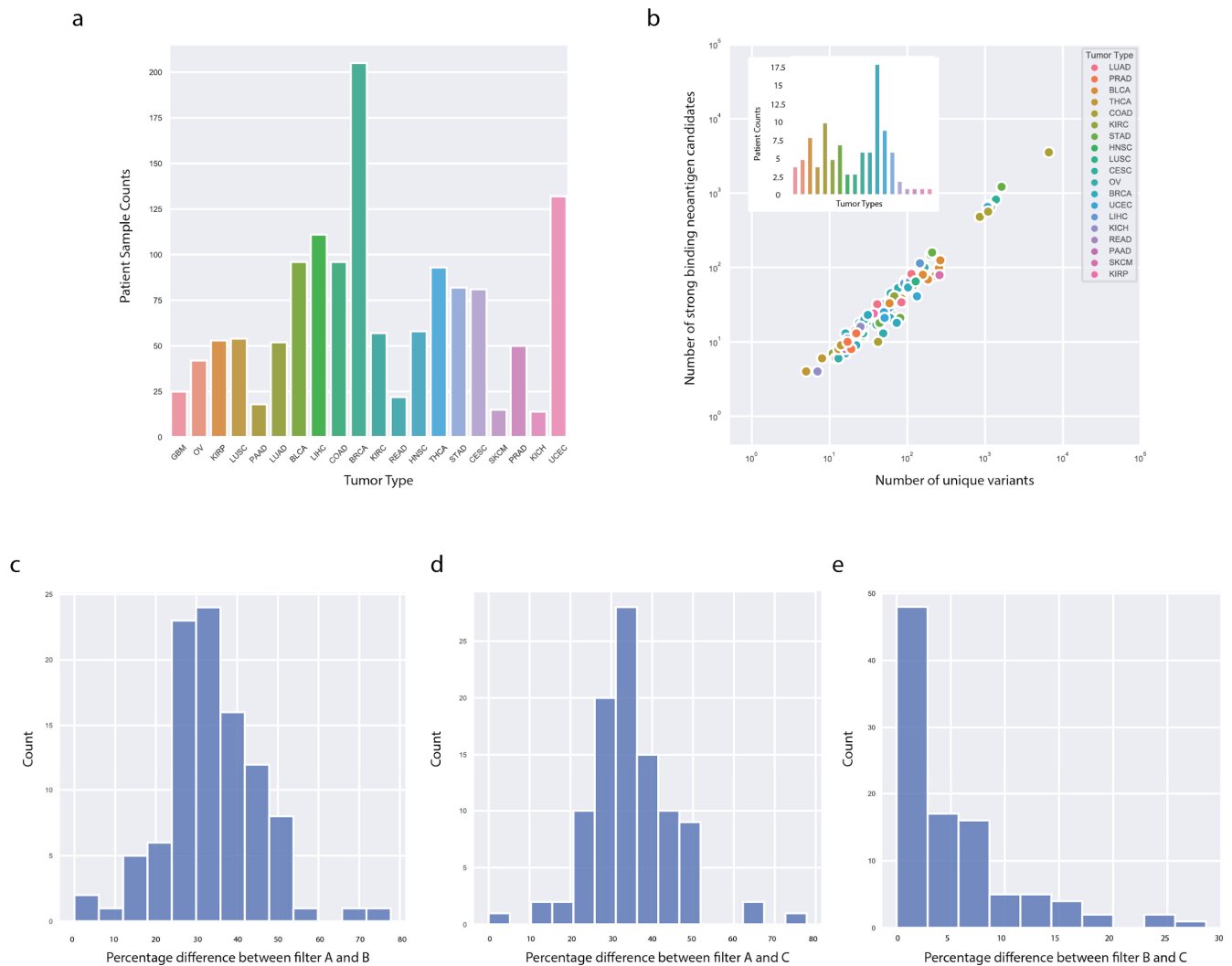


Fig. S13. Patient-level analysis for impact of anchor considerations on neoantigen prioritization
 Patient-level impact of anchor position information on neoantigen prioritization decisions using 100 randomly selected samples from a pool of 1,356 TCGA samples. **a**, The distribution of tumor sample types for a TCGA sample pool ($n = 1,356$) is shown as a bar plot. The x-axis represents the different cancer types included and the y-axis shows the number of patient samples. **b**, A scatterplot shows the number of unique variants plotted against the number of predicted strong binding neoantigen candidates. Neoantigen candidates were compiled such that each variant had its top neoantigen candidate selected. The candidates were subsequently filtered based on a 500 nM binding affinity cutoff for each patient. The inner bar plot shows the distribution of tumor types for the randomly selected 100 patients. The legend shows the color labeling for each tumor type and is consistent between outer scatterplot and inner bar plot. **c**, **d**, **e**: Histogram plots showing distribution of neoantigen prioritization decision differences between filters A, B and C (Methods). Differences were normalized with respect to each individual patient's neoantigen counts and presented as the percentage of peptides that would be classified differently under each anchor interpretation scheme in terms of its neoantigen candidacy. The x-axis shows the range of percentage differences between filters while the y-axis shows patient sample counts for respective bins.

Description of Supplementary Data Files

Data file S1. Seed dataset of strong binding neoantigen candidates

We identified 609,807 strong binding peptides for 328 HLA alleles from clinical and TCGA datasets. These served as a seed dataset for our anchor prediction workflow where a maximum of 10 peptides were selected at random for each HLA-peptide length combination. Peptide sequences, corresponding strong binding HLA alleles and the predicted median binding affinities across up to 8 different algorithms are included. The full dataset with all individual algorithmic prediction scores is also made available at http://genomedata.org/anchor_predictions/.

Data file S2. Anchor predictions for 328 HLA alleles

Prediction results from our computational workflow sorted by HLA alleles and peptide lengths. Overall scores are listed by peptide positions and represent the level of binding affinity change when mutated at the particular location. The 9-mer peptide data section contains additional information matching individual HLA alleles to their color-coded cluster in Figure 3.

Data file S3. HLA Summary Information

Information regarding all HLA alleles analyzed, including: 1) Nearest neighbor as defined by NetMHCpan, 2) Distance to nearest neighbor, 3) mean standard deviation of scores across all algorithms predicting for that allele using all available peptides in the seed dataset, 4) anchor cluster predicted as shown in Fig. 3., 5) training data available for all HLA alleles based on NetMHCpan4.0 (both elution and binding data) 6) number of algorithms able to generate predictions for each HLA allele, and 7) the detailed list of algorithms included in the previous column count. Note that if an HLA allele had training data, then it did not need to use a nearest neighbor for estimating binding and hence distance is 0 with nearest neighbor listed as itself.

Data file S4. HLA PDB data table

All protein data bank structures, collected for orthogonal validation of anchor prediction results, are listed. Table includes: the specific HLA allele and peptide pair, the PDB identifiers, predicted binding affinities and anchor cluster codes. The first tab of the spreadsheet contains information for structures collected that only contain the HLA-peptide complex and the second tab contains those collected that contain HLA-peptide-TCR complexes.

Data file S5. Orthogonal validation correlation data

A subset of X-ray crystallography structures was used in demonstrating the distribution of correlation scores for the distance and SASA metrics against prediction scores. Information on the subset data used such as peptide sequence, HLA allele, distance/SASA correlations and respective p values are included. The first tab of the spreadsheet contains information on the subset plotted from HLA-peptide only structures and the second tab contain information on the subset plotted from HLA-peptide-TCR complexes.

Data file S6. Summary of all in vitro and cell based experimental validation data

Validation experiments were performed on a total of 136 peptide-HLA combinations. These experiments include both IC50 binding assays (measured binding affinity) as well as cell-based stabilization assays (average MFI value). A summary of the entire dataset as well as information on predicted binding affinities and measured binding categories are included.

Data file S7. Breakdown of individual algorithm predictions and their correlation with validation data

Individual algorithm predictions (across 8 different algorithms) and their correlation with data from IC50 binding assays (measured binding affinity) were calculated. The spreadsheet includes two sheets, one with the individual algorithm scores for each peptide-HLA combination and the other with Pearson correlation coefficients for each algorithm across HLA alleles.

Data file S8. List of TCGA samples for impact analysis

A subset of TCGA samples were chosen using HLA-balance-based selection for our overall cohort-level impact analysis (Tab 1). An additional 100 TCGA samples were selected at random to further evaluate patient-level impact of anchor considerations (Tab 2). Specific TCGA sample names are included.

Data file S9. Raw data file

Supplementary Video

Movie S1. Demonstration of orthogonal validation using distance and SASA metrics from x-ray crystallography structures

Video of the X-ray crystallography structure of HLA-B*08:01 and 9-mer peptide FLRGRAYGL. The video highlights the three components of the complex, including HLA (green), peptide (pink) and B2M (blue). It also provides a zoomed in view of the MHC binding groove, showcasing atoms surrounding the peptide in spheres and sticks for spatial perspective.