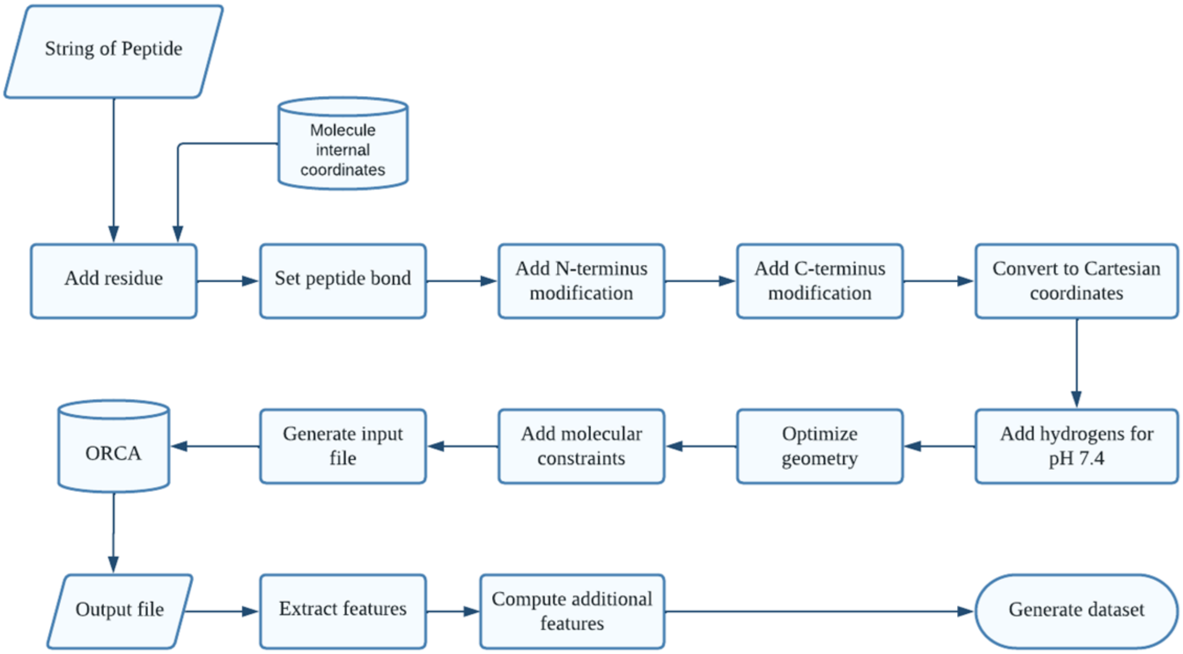
Supplementary Information

The ExoGAN generative AI framework enables extracellular vesicle-based immunotherapy

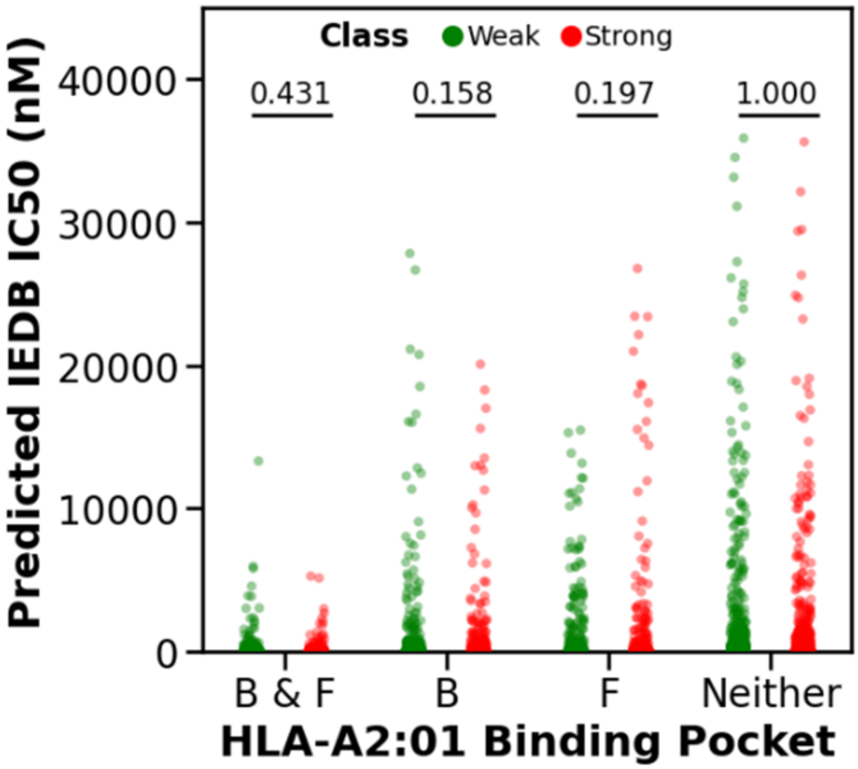
Zachary F. Greenberg[1], Tina Salehi Torabi[2] , Jiayu Huang[2], Franco Krepel[2], James A. Cahill[4,6], David A. Ostrov[3,5], Mei He[1,5*], Kiley S. Graim[2,5.6*]

1. Department of Pharmaceutics, University of Florida, Gainesville, FL 32603, USA
2. Computer and Information Sciences and Software Engineering, University of Florida, Gainesville, FL 32611 USA.
3. Department of Pathology, Immunology and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL 32610, USA.
4. Engineering School of Sustainable Infrastructure and Environmental Engineering, University of Florida, Gainesville, FL 32611, USA
5. University of Florida Health Cancer Center, Gainesville, FL 32610, USA
6. University of Florida Genetics and Genomics Institute, Gainesville, FL 32610, USA

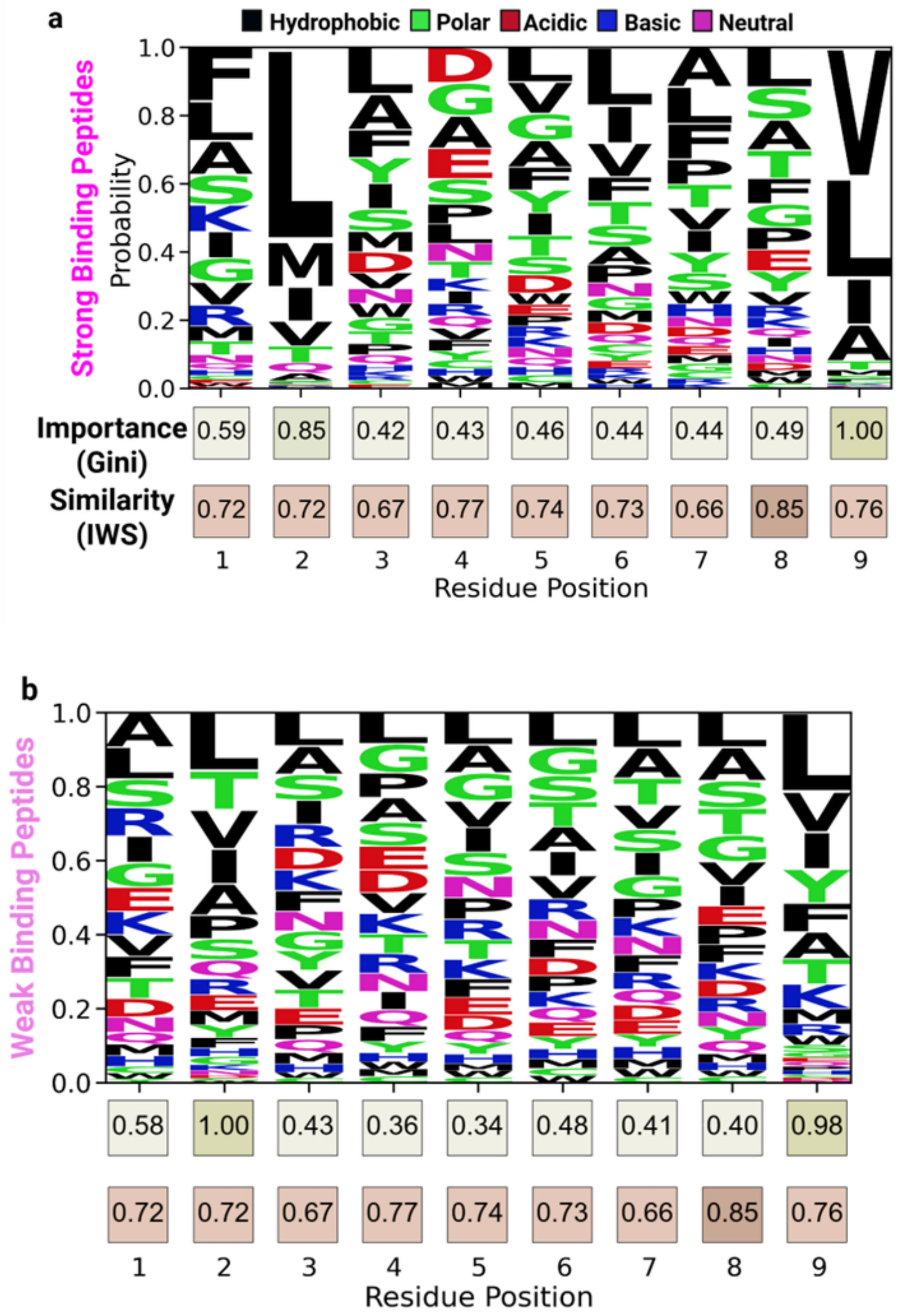
* corresponding contacts: Dr. Mei He, mhe@cop.ufl.edu; Dr. Kiley Graim, kgraim@ufl.edu


String of Peptide
Molecule internal coordinates
Add residue
Set peptide bond
Add N-terminus modification
Add C-terminus modification
Convert to Cartesian coordinates
Add hydrogens for pH 7.4
Optimize geometry
Add molecular constraints
Generate input file
ORCA
Output file
Extract features
Compute additional features
Generate dataset

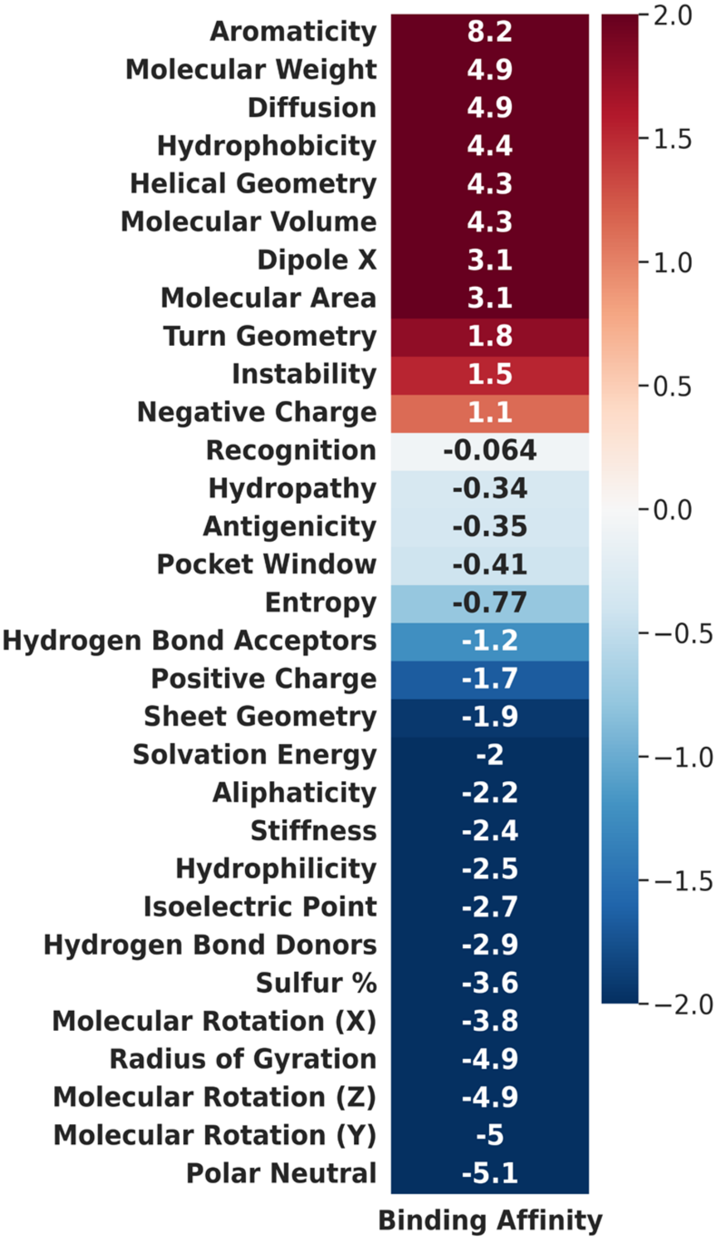
Supplementary Fig s1. Sequence preparation and physiochemical engineering workflow.



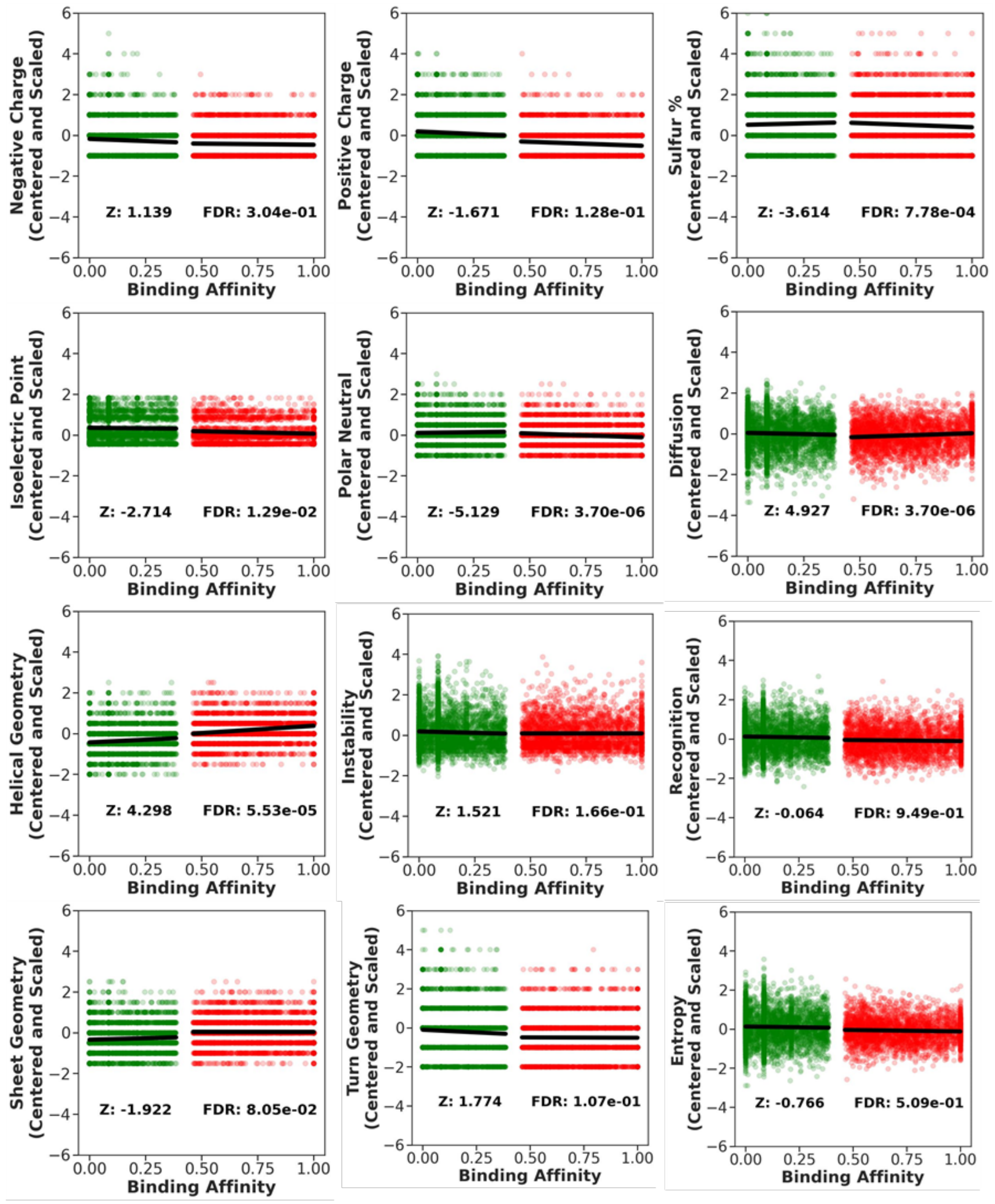
Supplementary Fig s2. Using a Dirichlet prior from the strong and weak binding peptide position weight matrices for conserved sequence prediction, we generated three sequences groups by fixing their residues for B and F pockets. These sequences were then predicted by NetMHCPan 4.1 and statistically evaluated with a Benjamini-Hochberg corrected two-sided Mann Whitney test.



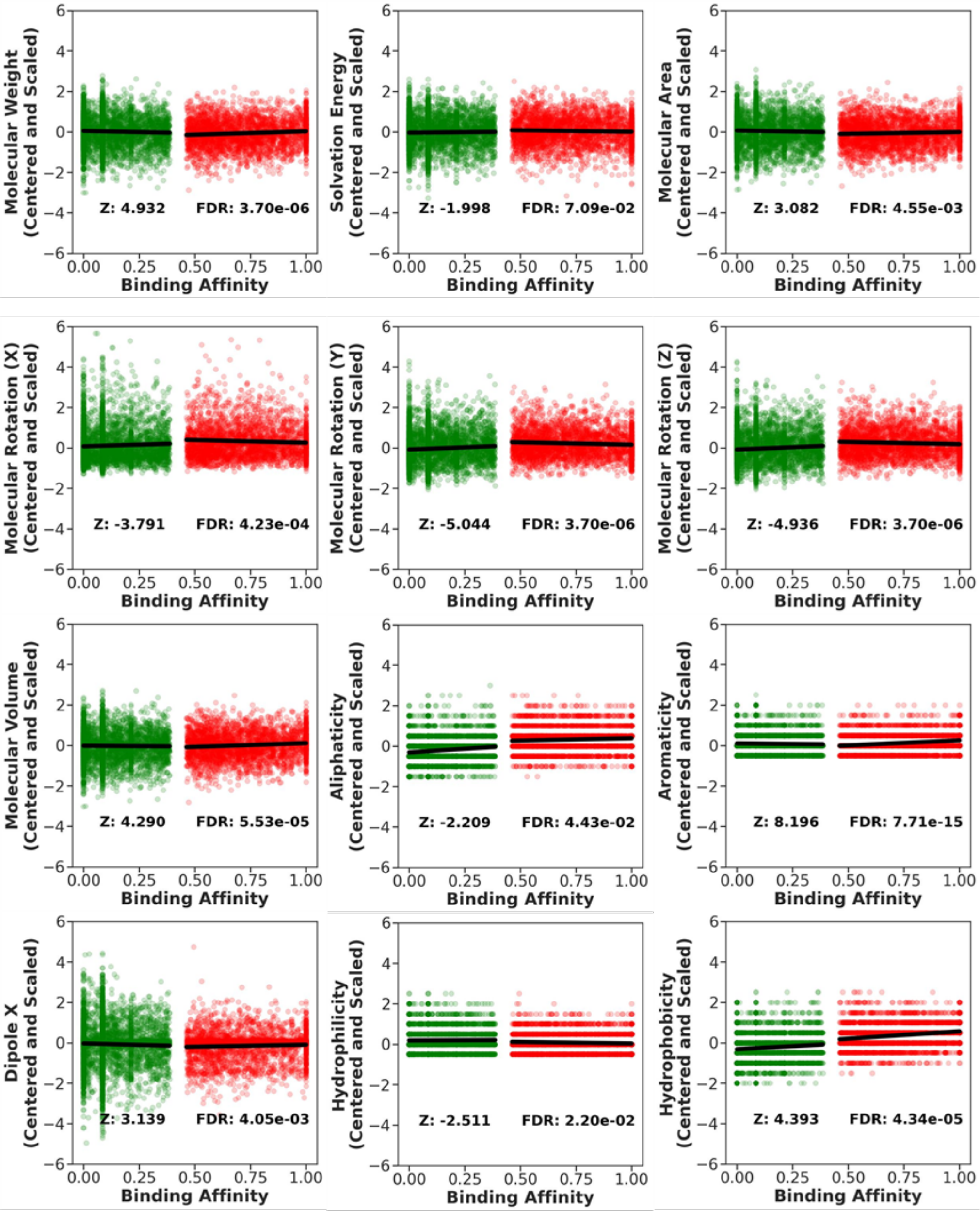
Supplementary Fig s3. **(a) Residue importance for antigens strongly binding HLA*A02:01.** In IEDB's curated data for HLA*A02:01, a total of 7860 antigens with harmonized binding affinities, as computed by Nielsen et al. [38], are listed. Based on Nielsen et al. we binned the dataset into strong and weak binding antigens to assess residue importance for strong binding antigens. **(b) Residue importance for antigens weakly binding HLA*A02:01.** Similar to **(a),** we separated out the weak binding antigens to assess their residue importance to bind HLA*A02:01.



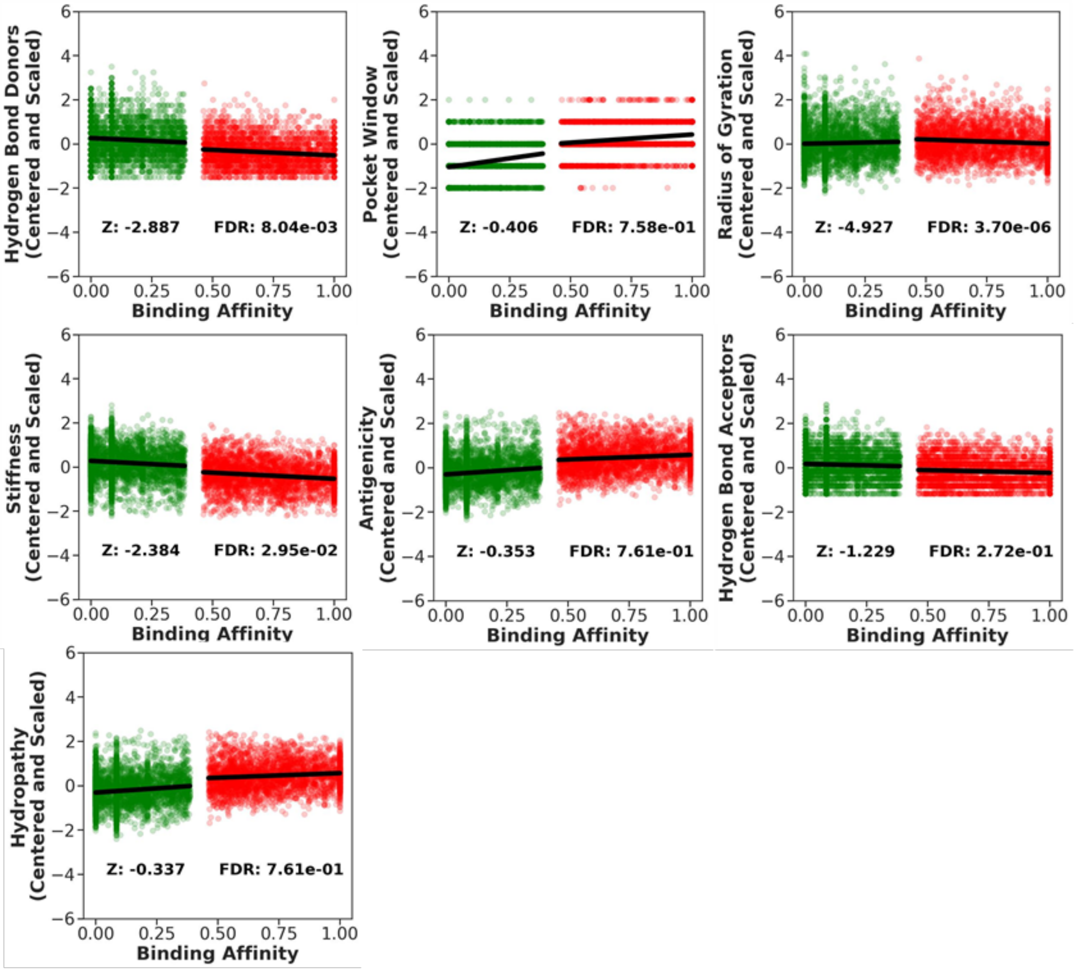
Supplementary Fig s4. Fisher’s Z transformation was applied onto our feature engineering set to assess physiochemical feature importance of a peptide’s binding affinity to HLA-A*02:01.



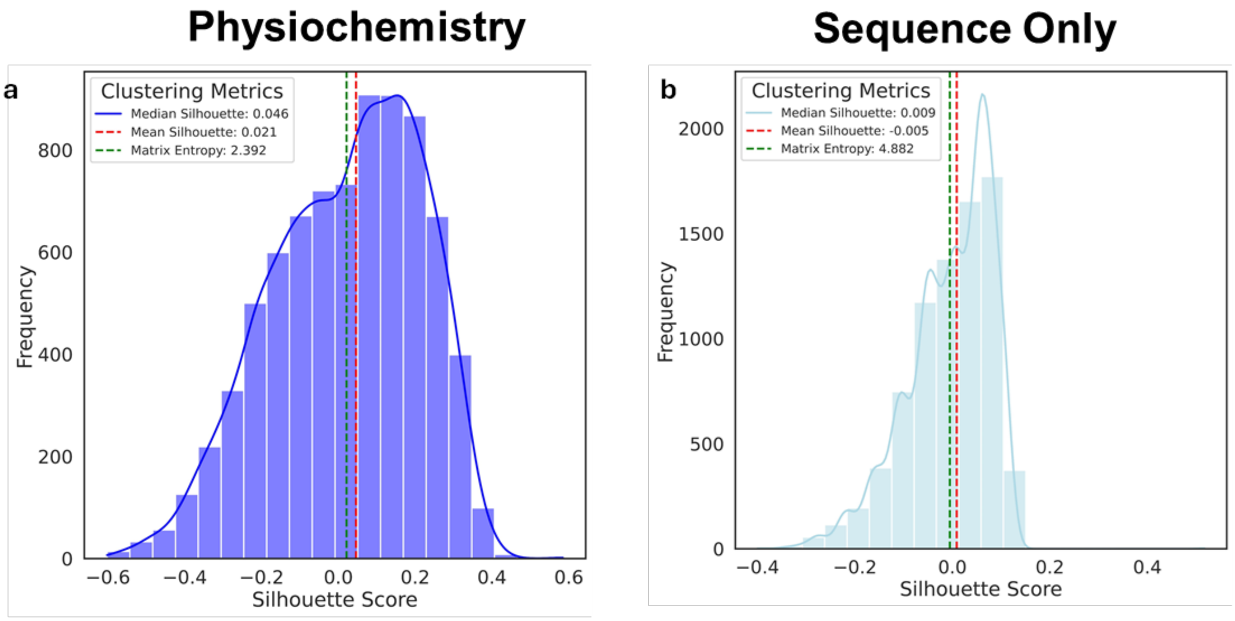
Supplementary Fig s5. We plotted the correlation between our scaled physiochemical features and each peptide's binding affinity to evaluate correlative statistical significance using a Benjamini-Hochberg corrected Fisher's Z-test. More physiochemical correlative plots are shown in Fig s6 and Fig s7.



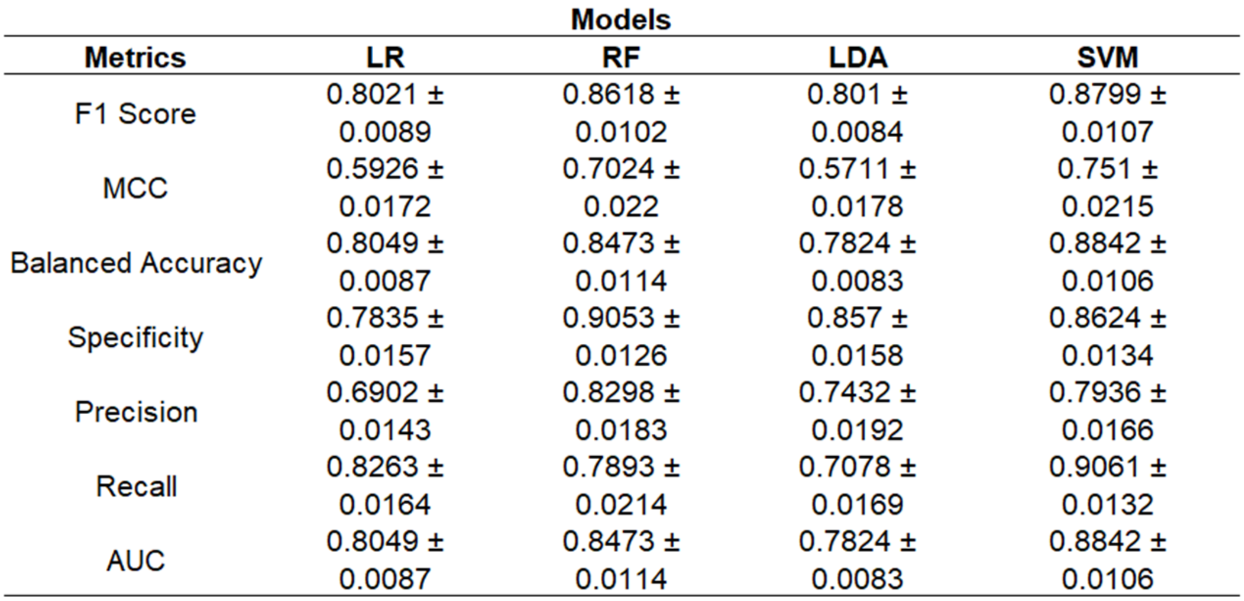
Supplementary Fig s6. We plotted the correlation between our scaled physiochemical features and each peptide’s binding affinity to evaluate correlative statistical significance using a Benjamini-Hochberg corrected Fisher’s Z-test. More physiochemical correlative plots are shown in Fig s7.



Supplementary Fig s7. Physiochemical correlative plots. We plotted the correlation between our scaled physiochemical features and each peptide's binding affinity to evaluate correlative statistical significance using a Benjamini-Hochberg corrected Fisher's Z-test.

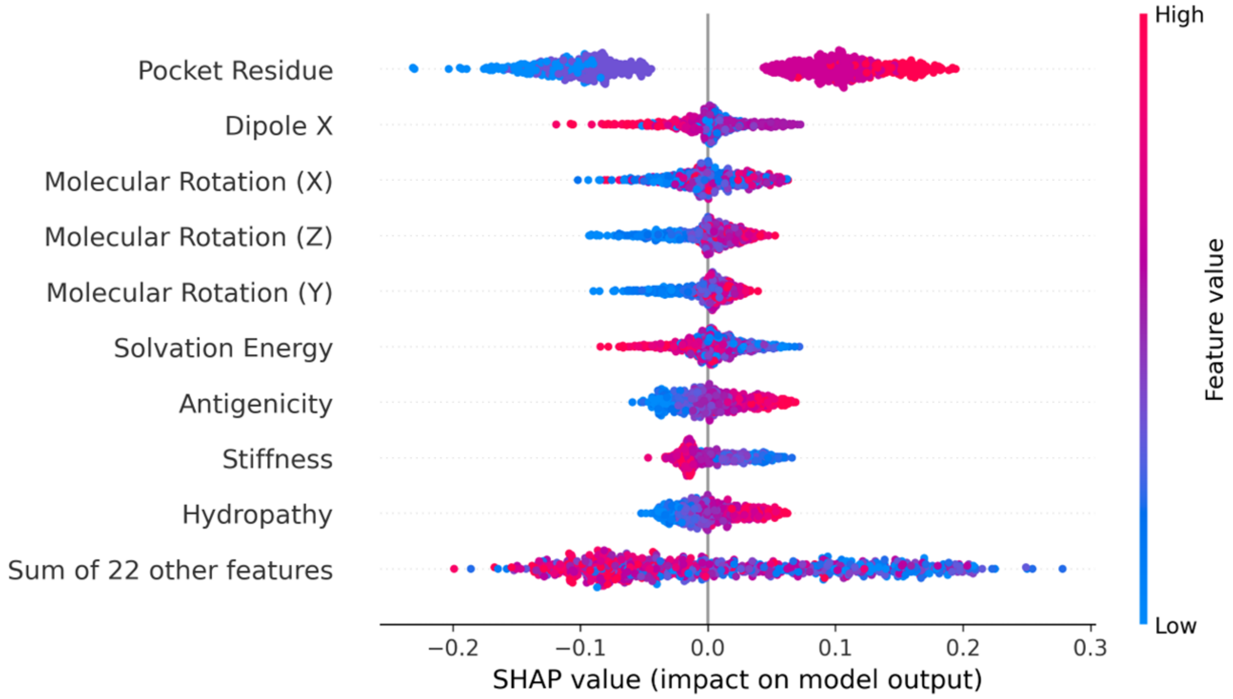


Supplementary Fig s8. **a. and b**. Quantifying feature set clustering performance by computing the per-sample silhouette distribution, median and mean silhouette score, and correlation matrix entropy.

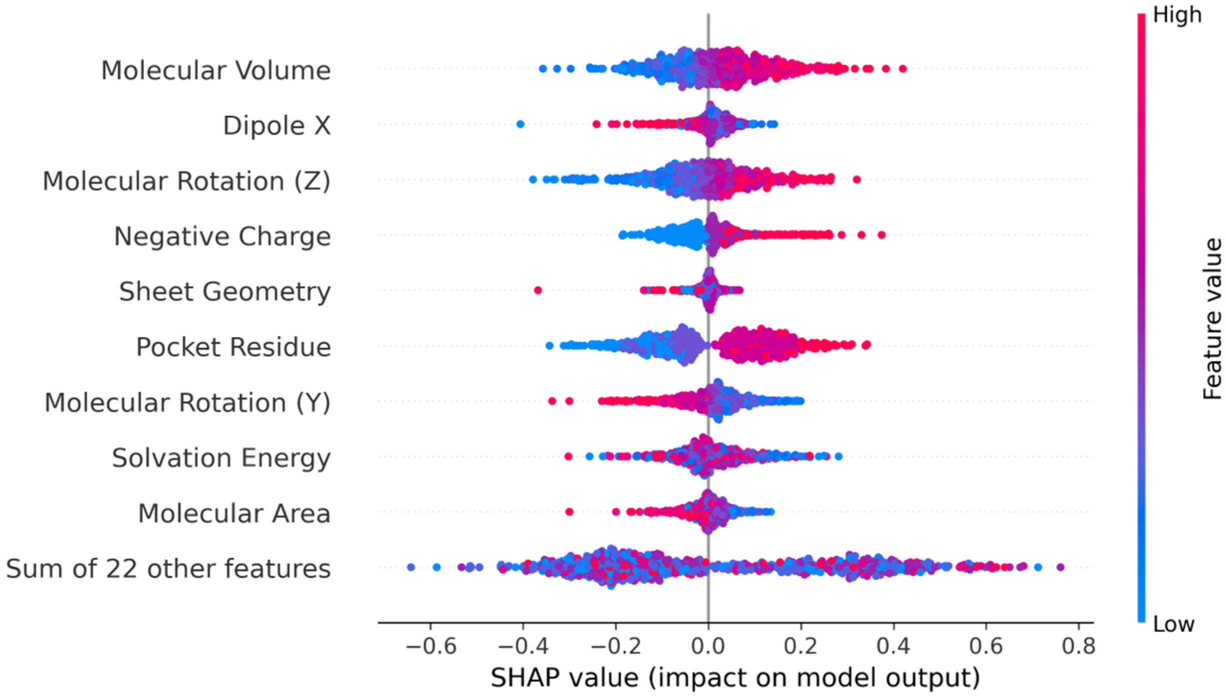


	Models			
Metrics	LR	RF	LDA	SVM
F1 Score	0.8021 ± 0.0089	0.8618 ± 0.0102	0.801 ± 0.0084	0.8799 ± 0.0107
MCC	0.5926 ± 0.0172	0.7024 ± 0.022	0.5711 ± 0.0178	0.751 ± 0.0215
Balanced Accuracy	0.8049 ± 0.0087	0.8473 ± 0.0114	0.7824 ± 0.0083	0.8842 ± 0.0106
Specificity	0.7835 ± 0.0157	0.9053 ± 0.0126	0.857 ± 0.0158	0.8624 ± 0.0134
Precision	0.6902 ± 0.0143	0.8298 ± 0.0183	0.7432 ± 0.0192	0.7936 ± 0.0166
Recall	0.8263 ± 0.0164	0.7893 ± 0.0214	0.7078 ± 0.0169	0.9061 ± 0.0132
AUC	0.8049 ± 0.0087	0.8473 ± 0.0114	0.7824 ± 0.0083	0.8842 ± 0.0106

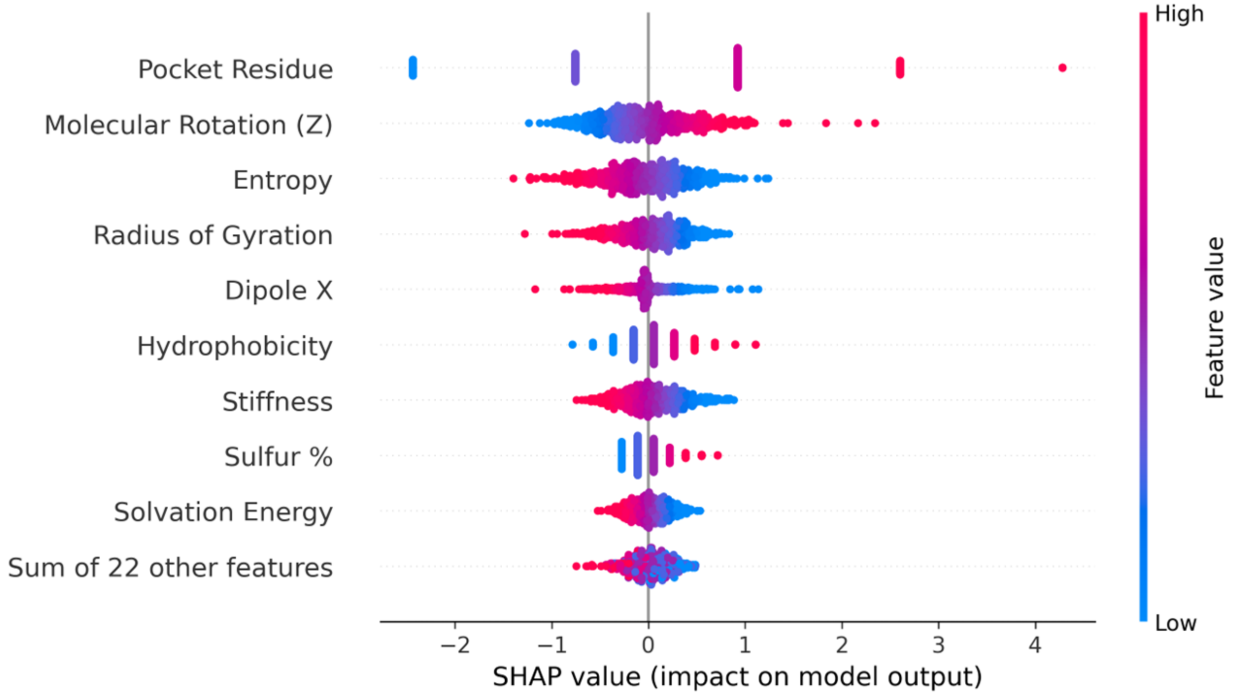
Supplementary Table s1. Classical model performance summary using physiochemical features to predict strong and weak binding peptides (StratifiedKFold = 10). LR is logistic regression, RF is random forest, LDA is linear discriminant analysis, and SVM is support vector machine.



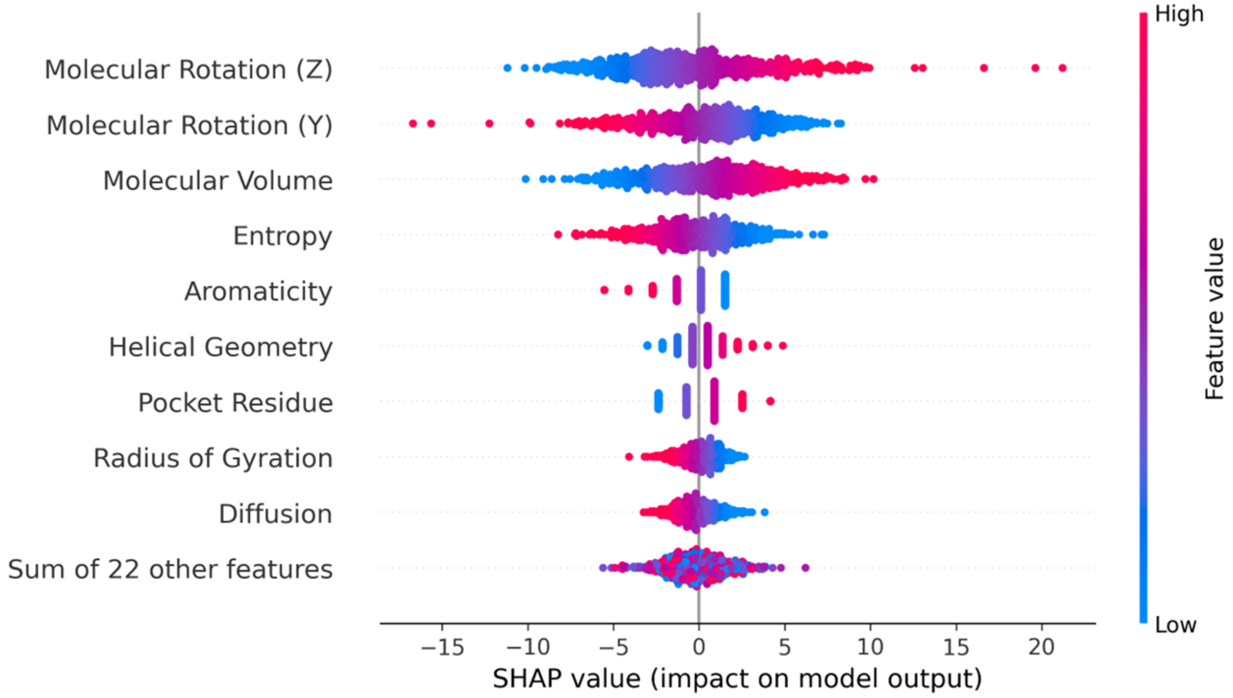
Supplementary Figure s9. Random Forest SHAP analysis. For Random Forest trained on the peptide physiochemistry feature set, a SHAP waterfall plot was visualized to assess predictor importance of predictor classifying strong and weak peptide binders.



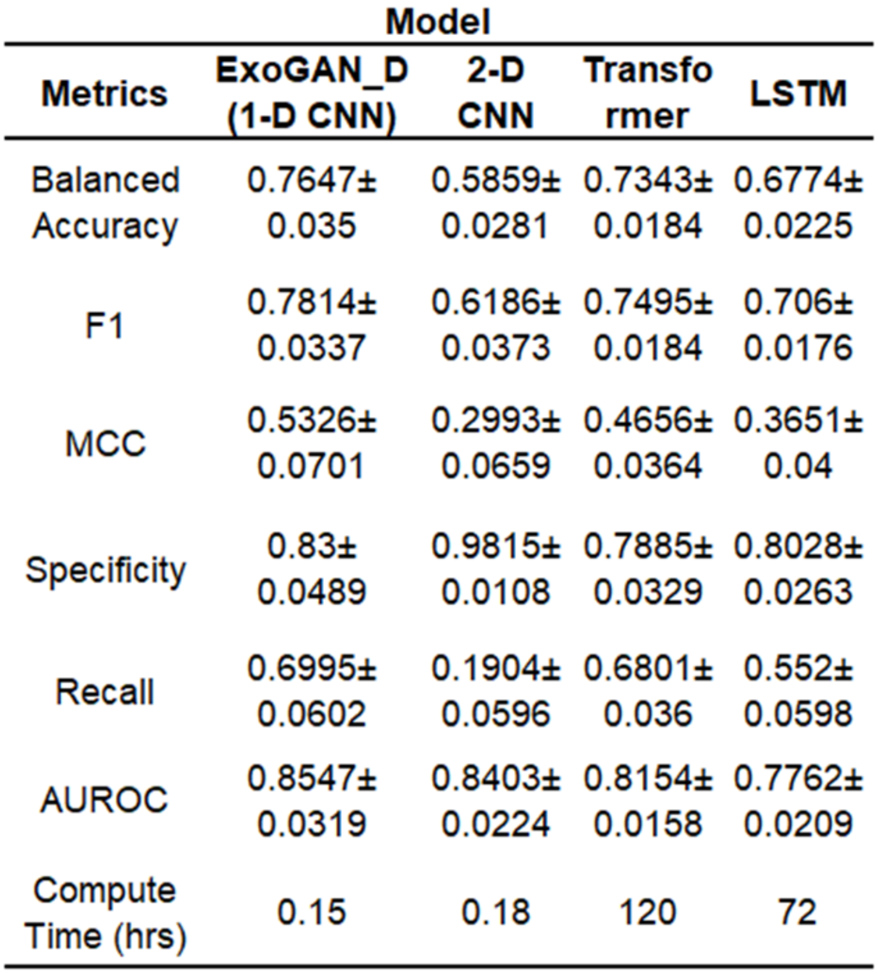
Supplementary Figure s10. Support Vector Machine SHAP analysis. For Support Vector Machine trained on the peptide physiochemistry feature set, a SHAP waterfall plot was visualized to assess predictor importance of predictor classifying strong and weak peptide binders.



Supplementary Figure s11. Logistic Regression SHAP analysis. For Logistic Regression trained on the peptide physiochemistry feature set, a SHAP waterfall plot was visualized to assess predictor importance of predictor classifying strong and weak peptide binders.

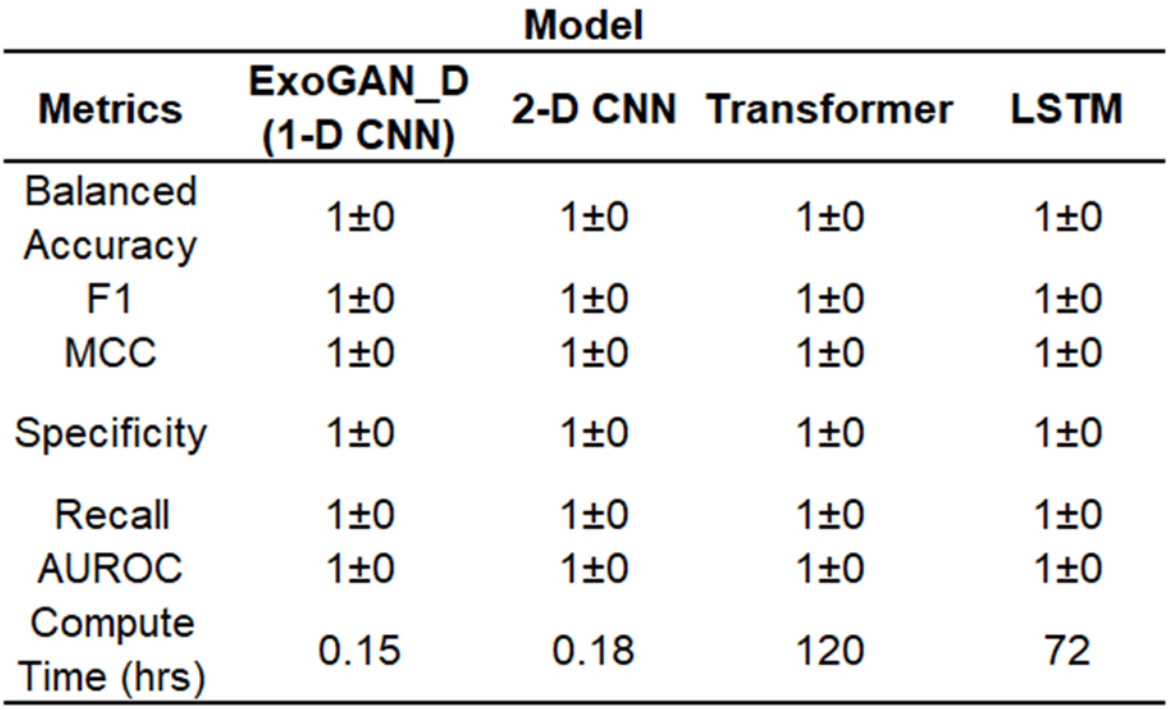


Supplementary Figure s12. Linear Discriminant Analysis SHAP analysis. For Linear Discriminant Analysis trained on the peptide physiochemistry feature set, a SHAP waterfall plot was visualized to assess predictor importance of predictor classifying strong and weak peptide binders



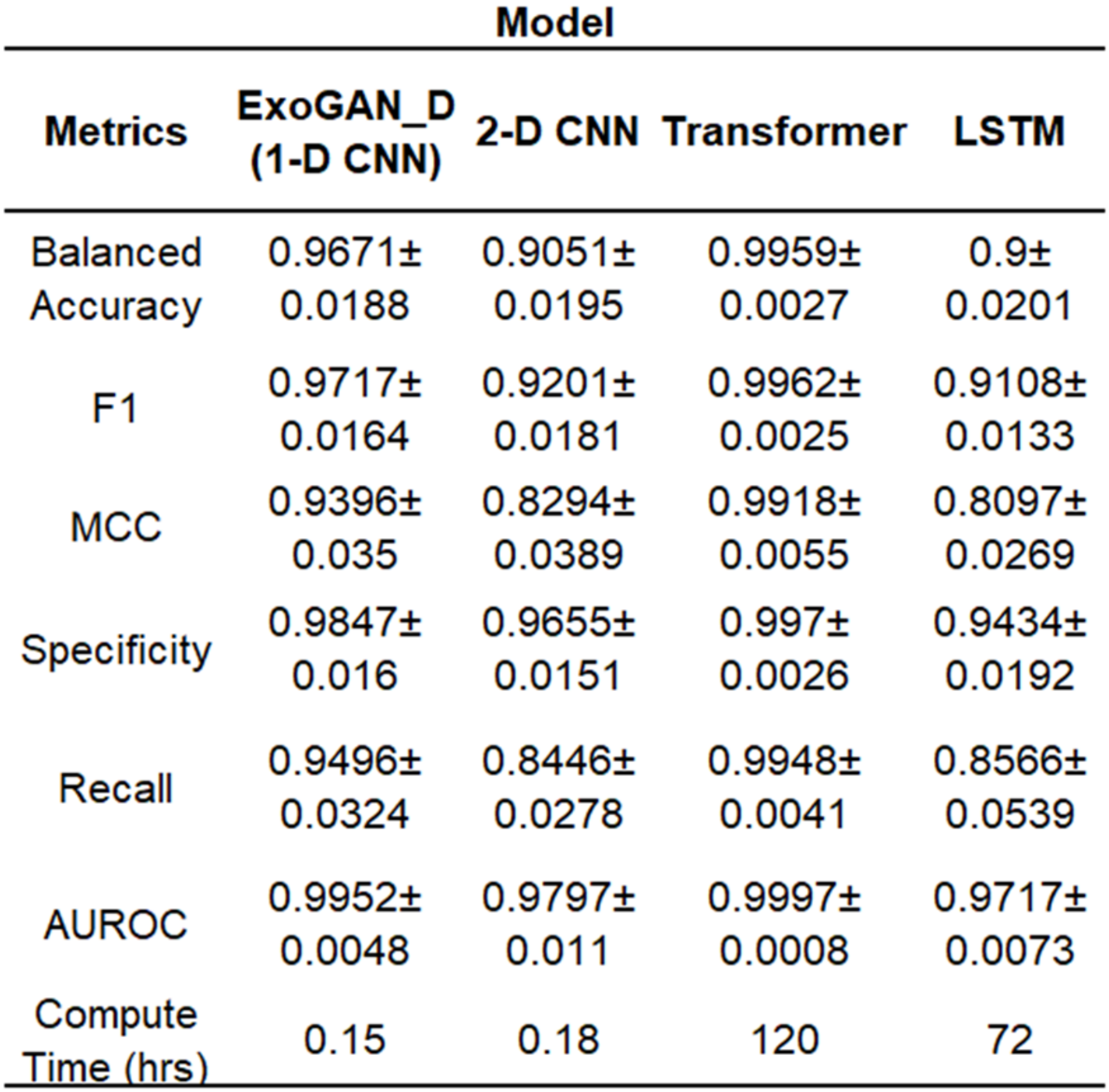
	Model			
Metrics	ExoGAN_D (1-D CNN)	2-D CNN	Transformer	LSTM
Balanced Accuracy	0.7647± 0.035	0.5859± 0.0281	0.7343± 0.0184	0.6774± 0.0225
F1	0.7814± 0.0337	0.6186± 0.0373	0.7495± 0.0184	0.706± 0.0176
MCC	0.5326± 0.0701	0.2993± 0.0659	0.4656± 0.0364	0.3651± 0.04
Specificity	0.83± 0.0489	0.9815± 0.0108	0.7885± 0.0329	0.8028± 0.0263
Recall	0.6995± 0.0602	0.1904± 0.0596	0.6801± 0.036	0.552± 0.0598
AUROC	0.8547± 0.0319	0.8403± 0.0224	0.8154± 0.0158	0.7762± 0.0209
Compute Time (hrs)	0.15	0.18	120	72

Supplementary Table s2. Model architecture summary table evaluating different architecture performances to discriminate strong and weak binding peptides using sequence information



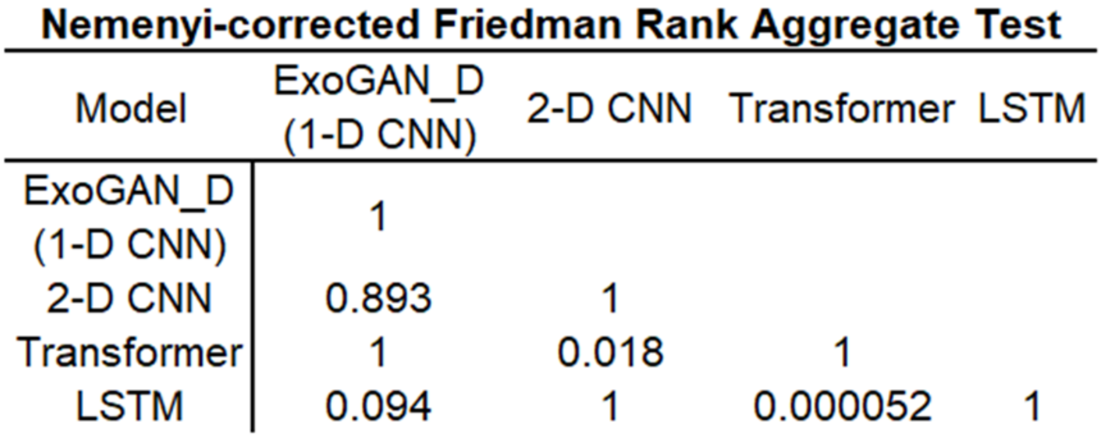
	Model			
Metrics	ExoGAN_D (1-D CNN)	2-D CNN	Transformer	LSTM
Balanced Accuracy	1±0	1±0	1±0	1±0
F1	1±0	1±0	1±0	1±0
MCC	1±0	1±0	1±0	1±0
Specificity	1±0	1±0	1±0	1±0
Recall	1±0	1±0	1±0	1±0
AUROC	1±0	1±0	1±0	1±0
Compute Time (hrs)	0.15	0.18	120	72

Supplementary Table s3. Model architecture summary table evaluating different architecture performances to discriminate strong and weak binding peptides using physiochemistry



	Model			
Metrics	ExoGAN_D (1-D CNN)	2-D CNN	Transformer	LSTM
Balanced Accuracy	0.9671± 0.0188	0.9051± 0.0195	0.9959± 0.0027	0.9± 0.0201
F1	0.9717± 0.0164	0.9201± 0.0181	0.9962± 0.0025	0.9108± 0.0133
MCC	0.9396± 0.035	0.8294± 0.0389	0.9918± 0.0055	0.8097± 0.0269
Specificity	0.9847± 0.016	0.9655± 0.0151	0.997± 0.0026	0.9434± 0.0192
Recall	0.9496± 0.0324	0.8446± 0.0278	0.9948± 0.0041	0.8566± 0.0539
AUROC	0.9952± 0.0048	0.9797± 0.011	0.9997± 0.0008	0.9717± 0.0073
Compute Time (hrs)	0.15	0.18	120	72

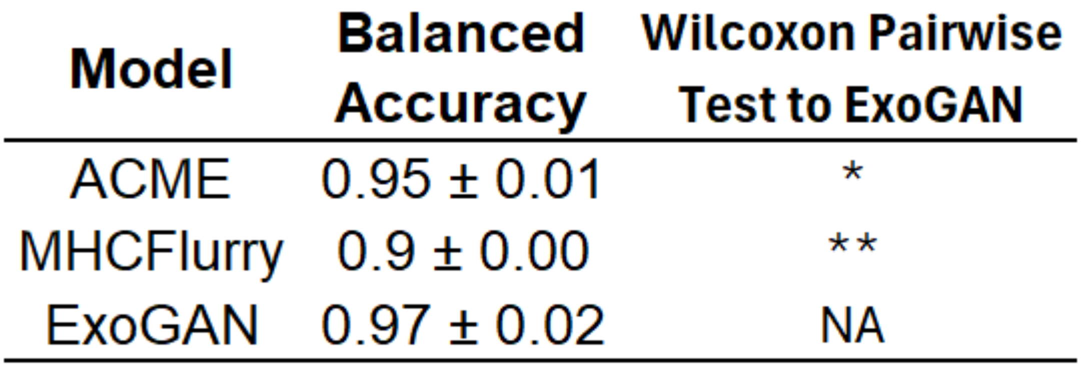
Supplementary Table s4. Classical model performance summary using physiochemical features to predict strong and weak binding peptides



Nemenyi-corrected Friedman Rank Aggregate Test

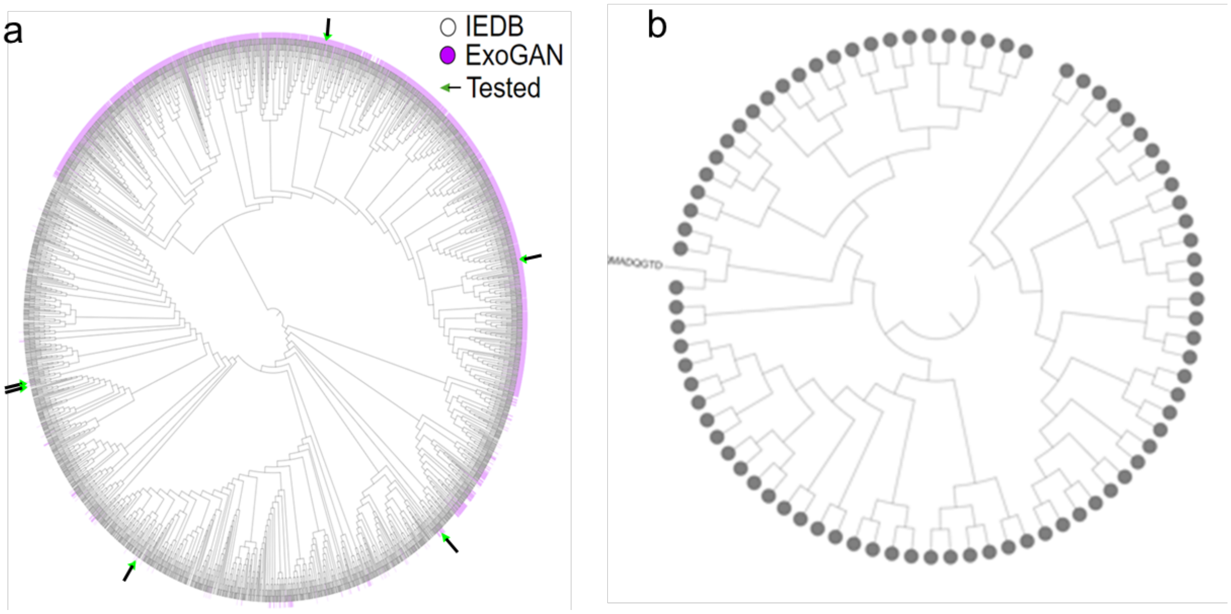
Model	ExoGAN_D (1-D CNN)	2-D CNN	Transformer	LSTM
ExoGAN_D (1-D CNN)	1			
2-D CNN	0.893	1		
Transformer	1	0.018	1	
LSTM	0.094	1	0.000052	1

Supplementary Table s5. Statistical summary table evaluating each model architecture discriminating strong and weak binding peptides using both sequence and physiochemical information

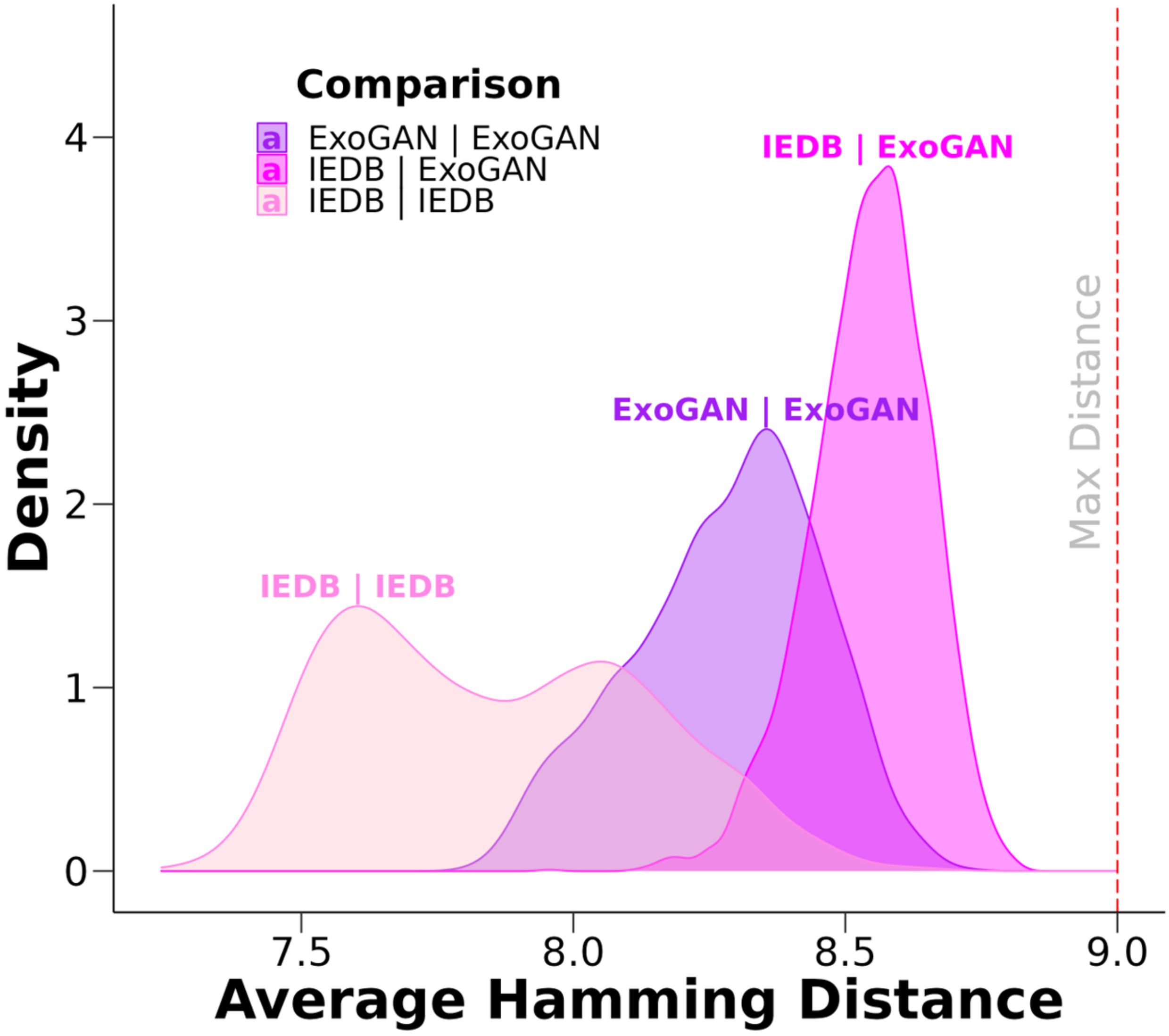


Model	Balanced Accuracy	Wilcoxon Pairwise Test to ExoGAN
ACME	0.95 ± 0.01	*
MHCFlurry	0.9 ± 0.00	**
ExoGAN	0.97 ± 0.02	NA

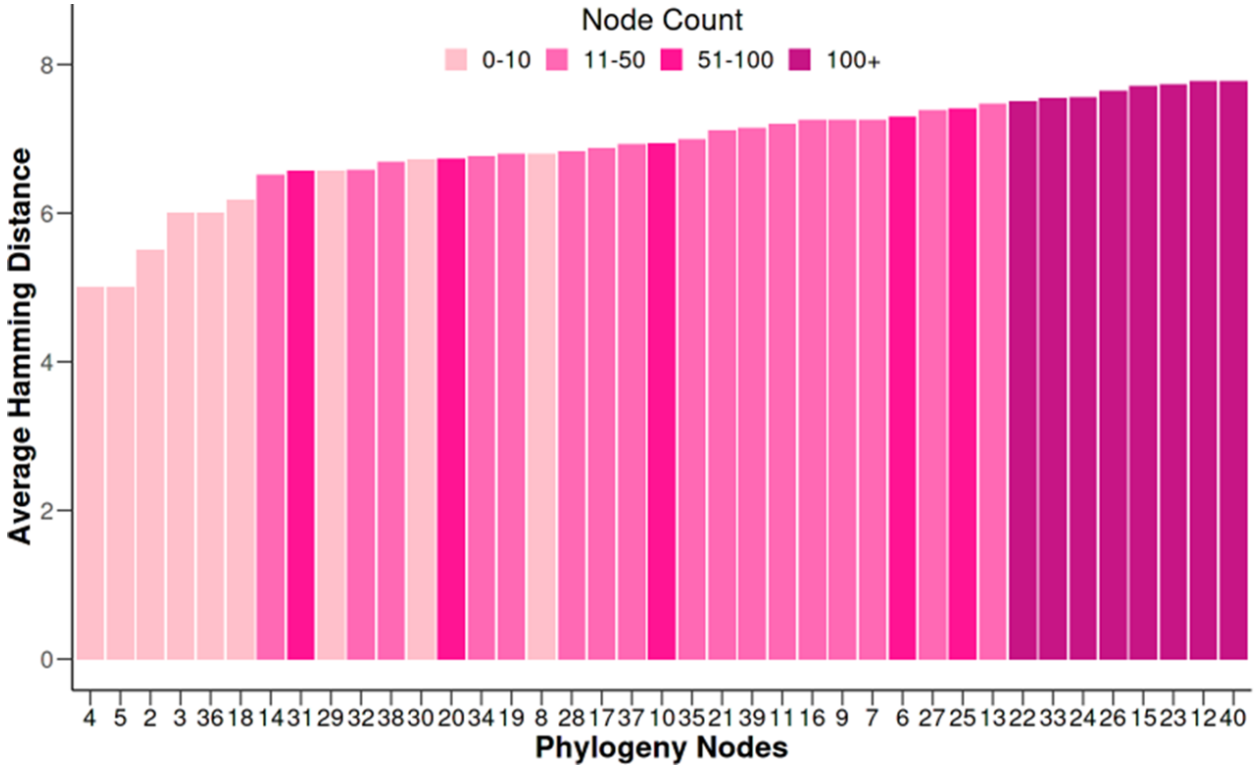
Supplementary Table s6. Statistical summary table evaluating each model to ExoGAN to discriminate HLA-A*02:01 strong and weak binding peptides.



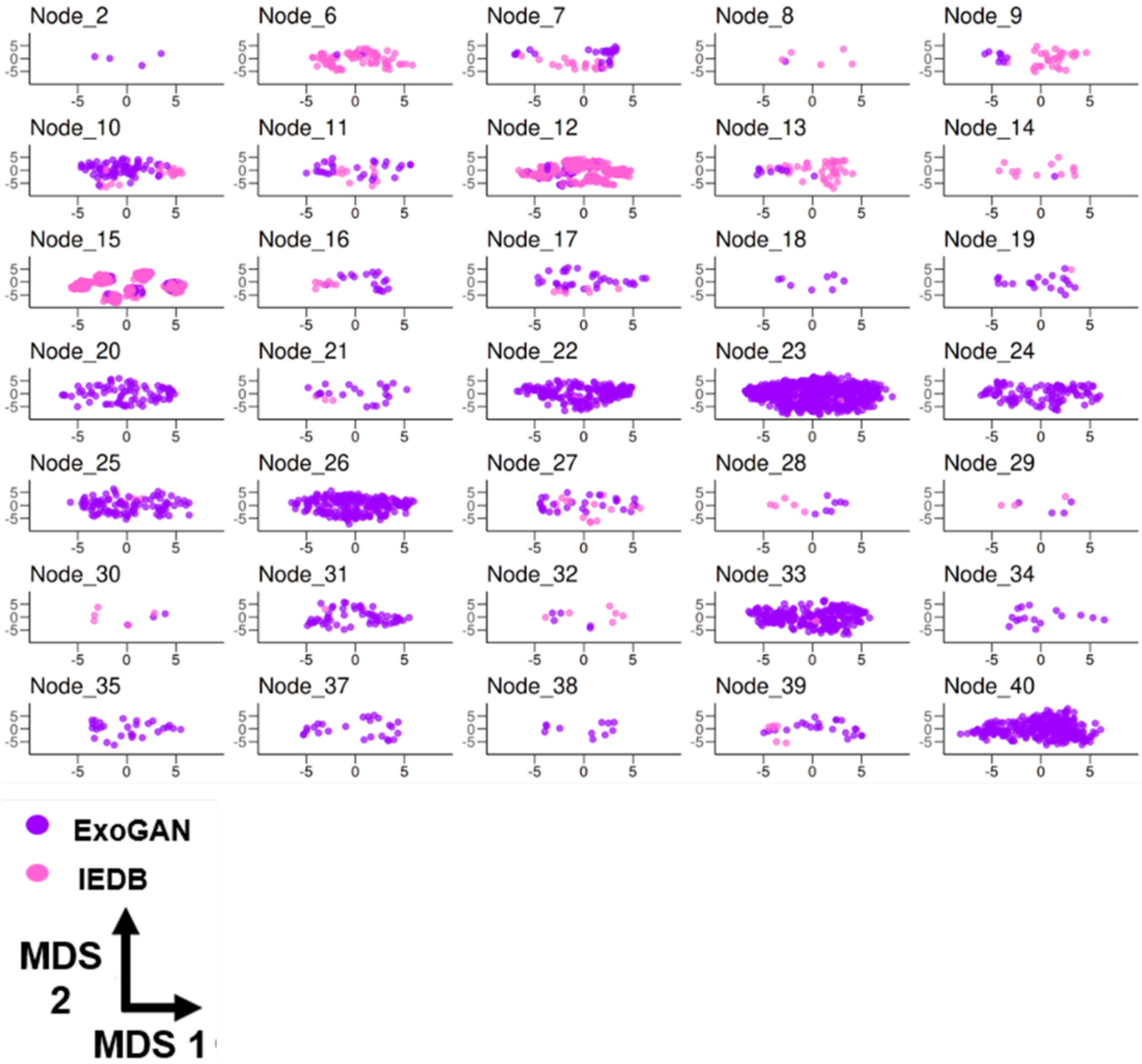
Supplementary Figure s13. (a). Phylogeny between ExoGAN generated and IEDB strong binding peptides was constructed using unweighted pair group means arithmetic (UPGMA) within a raw differences model. Colors highlight sequence location per class and arrows indicate experimental validation of selected peptides. (b). A condensed phylogeny was formed after grouping branches longer than 4.10 into their respective nodes.



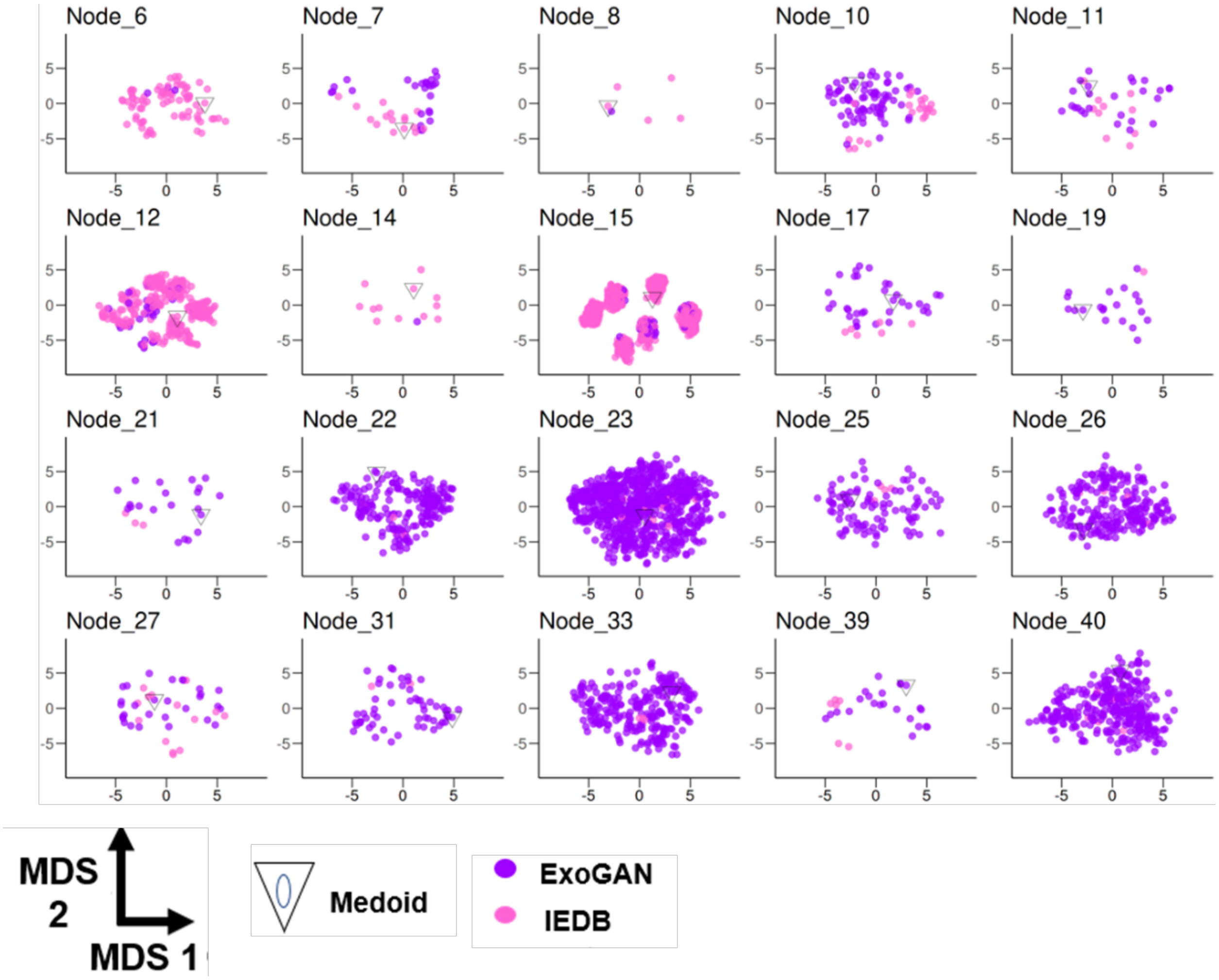
Supplementary Figure s14. A Hamming distance density plot showing the average sequence similarity of sequences compared to IEDB's strong binding peptides and ExoGAN-generated peptides. Higher values indicate more residues must change in one sequence to match the other sequence



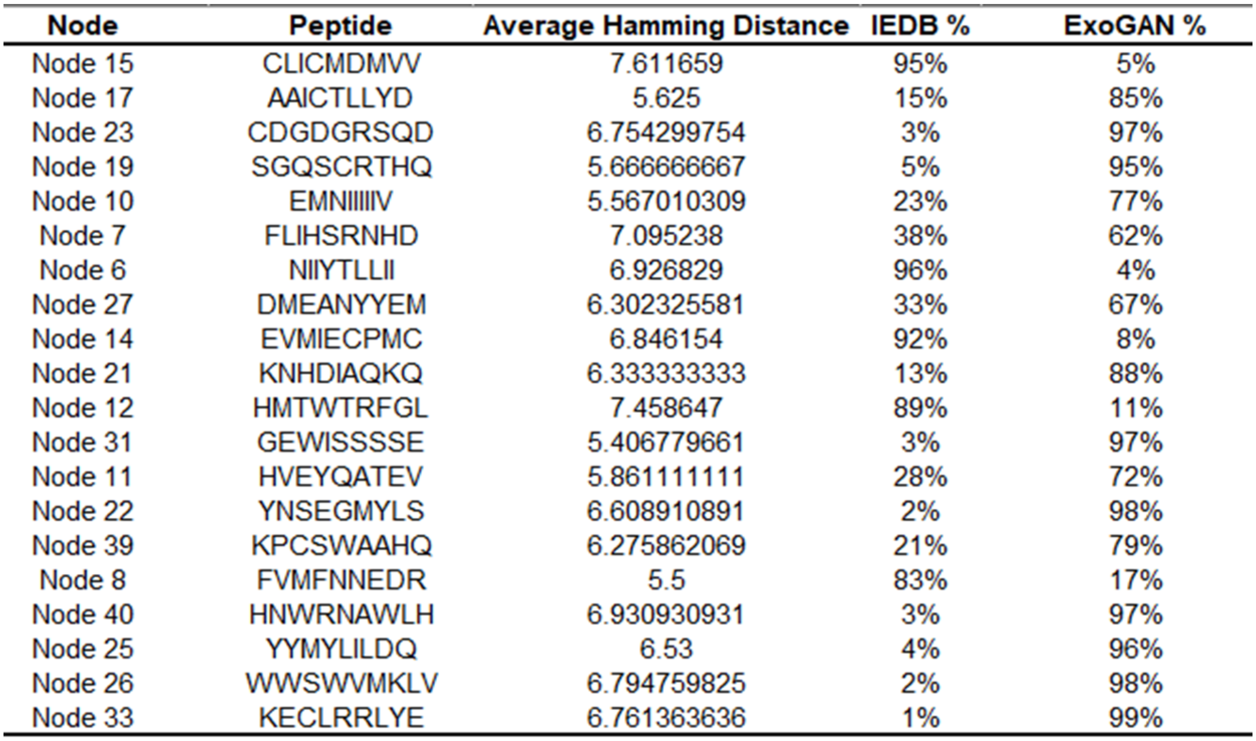
Supplementary Figure s15. A barplot sorted by the number of sequences in each node after condensing the phylogeny to assess the average sequence hamming distance in the node.



Supplementary Figure s16. Multidimensional scaling was performed on the condensed phylogeny nodes specific to each peptide's physiochemical features. Summarized counts are shown in Fig s15.

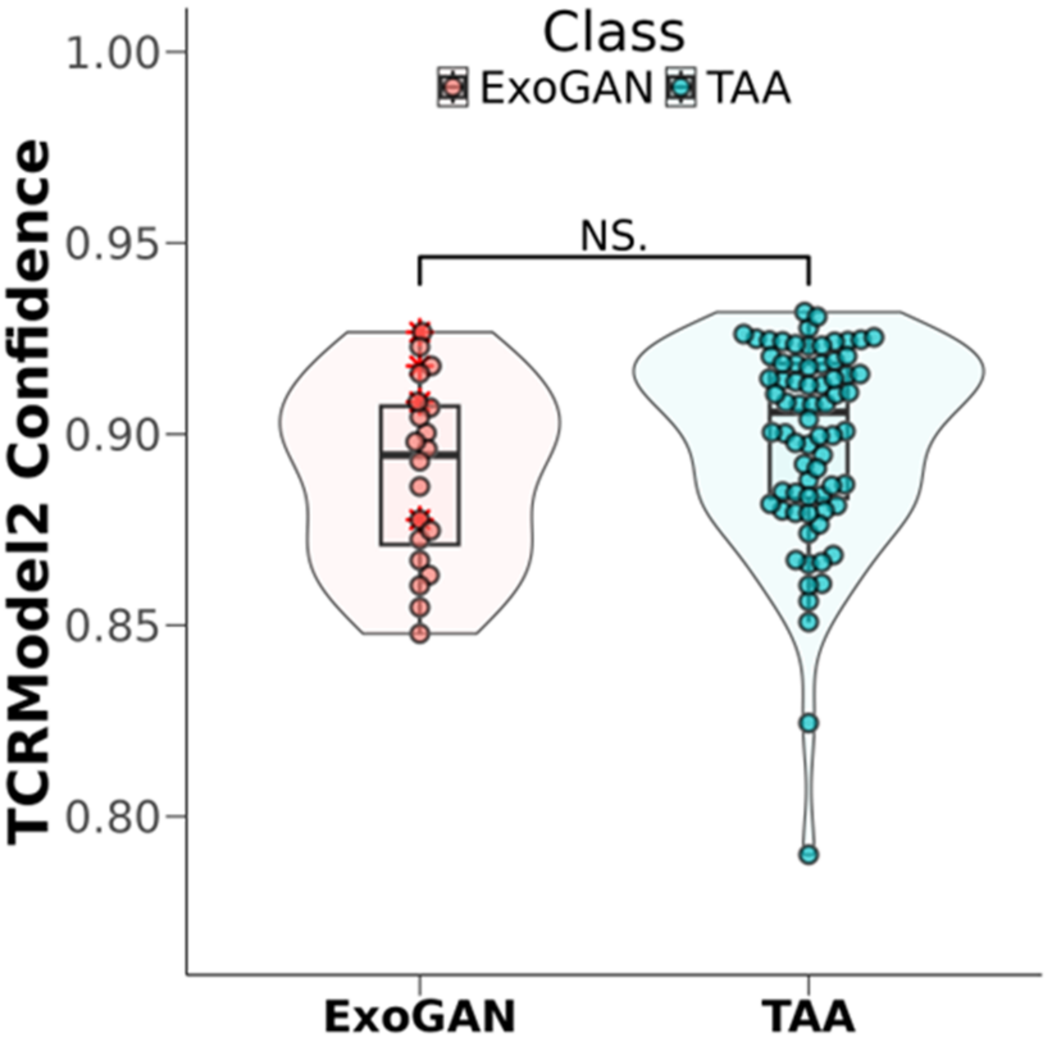


Supplementary Figure s17. Multidimensional scaling was performed on the condensed phylogeny nodes specific to each peptide's physiochemical features and then filtered for geometric overlap between ExoGAN and IEDB.

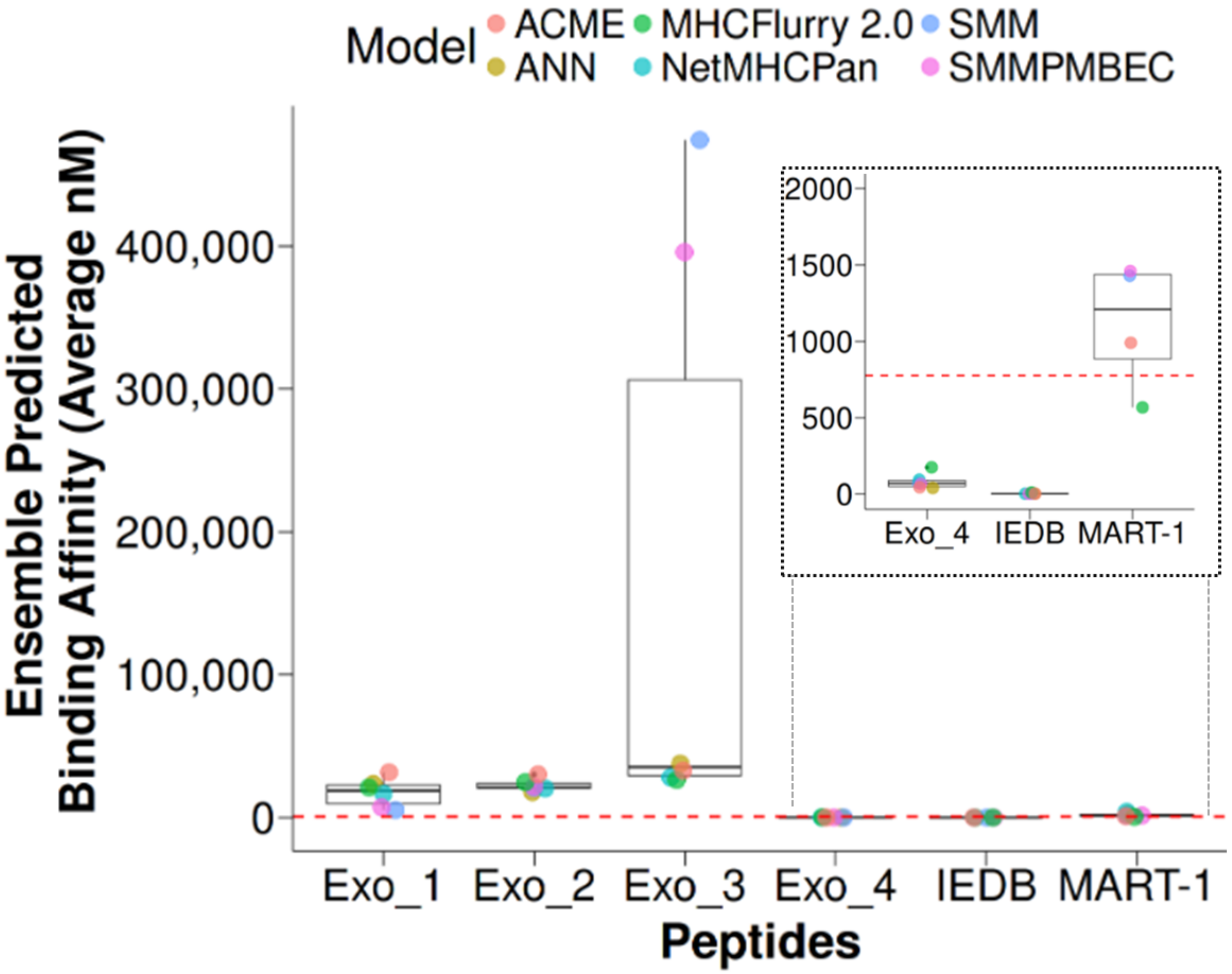


Node	Peptide	Average Hamming Distance	IEDB %	ExoGAN %
Node 15	CLICMDMVV	7.611659	95%	5%
Node 17	AAICTLLYD	5.625	15%	85%
Node 23	CDGDGRSQD	6.754299754	3%	97%
Node 19	SGQSCRTHQ	5.666666667	5%	95%
Node 10	EMNIIIIIV	5.567010309	23%	77%
Node 7	FLIHSRNHD	7.095238	38%	62%
Node 6	NIIYTLLII	6.926829	96%	4%
Node 27	DMEANYYEM	6.302325581	33%	67%
Node 14	EVMIECPMC	6.846154	92%	8%
Node 21	KNHDIAQKQ	6.333333333	13%	88%
Node 12	HMTWTRFGL	7.458647	89%	11%
Node 31	GEWISSSSE	5.406779661	3%	97%
Node 11	HVEYQATEV	5.861111111	28%	72%
Node 22	YNSEGMYLS	6.608910891	2%	98%
Node 39	KPCSWAAHQ	6.275862069	21%	79%
Node 8	FVMFNNEDR	5.5	83%	17%
Node 40	HNWRNAWLH	6.930930931	3%	97%
Node 25	YYMYLILDQ	6.53	4%	96%
Node 26	WWSWVMKLV	6.794759825	2%	98%
Node 33	KECLRRLYE	6.761363636	1%	99%

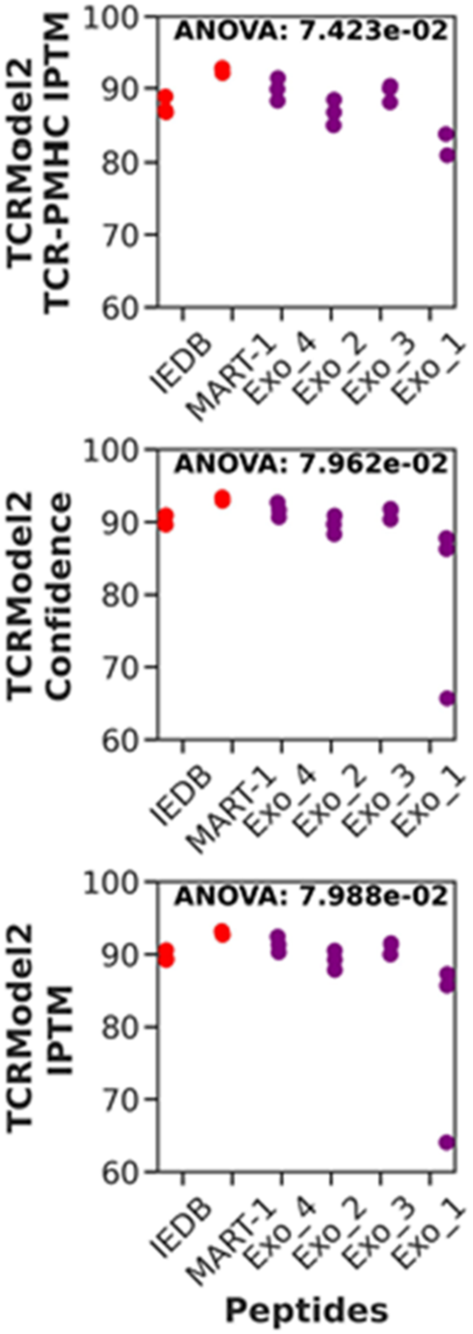
Supplementary Table s7. The representative peptide for each cluster from the PAM clustering, their average hamming distance, and percentage of peptide types within the cluster (IEDB, ExoGAN).



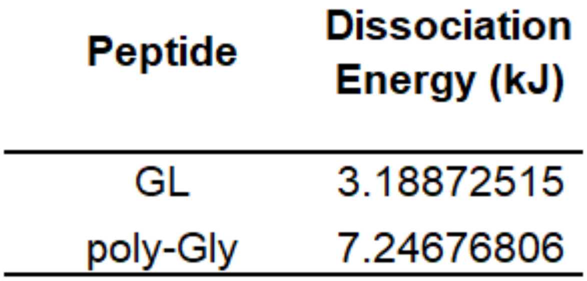
Supplementary Figure s18. TCRModel2 confidence to predict HLA*A2:01 presentation of ExoGAN's representative peptides compared to known tumor associated antigens (TAA), using the DMF5 allele. Red stars indicate peptide sequences selected for further validation.



Supplementary Figure s19. Gold standard MHC-I binding prediction models were used to predict generated peptide affinity to HLA*A02:01.

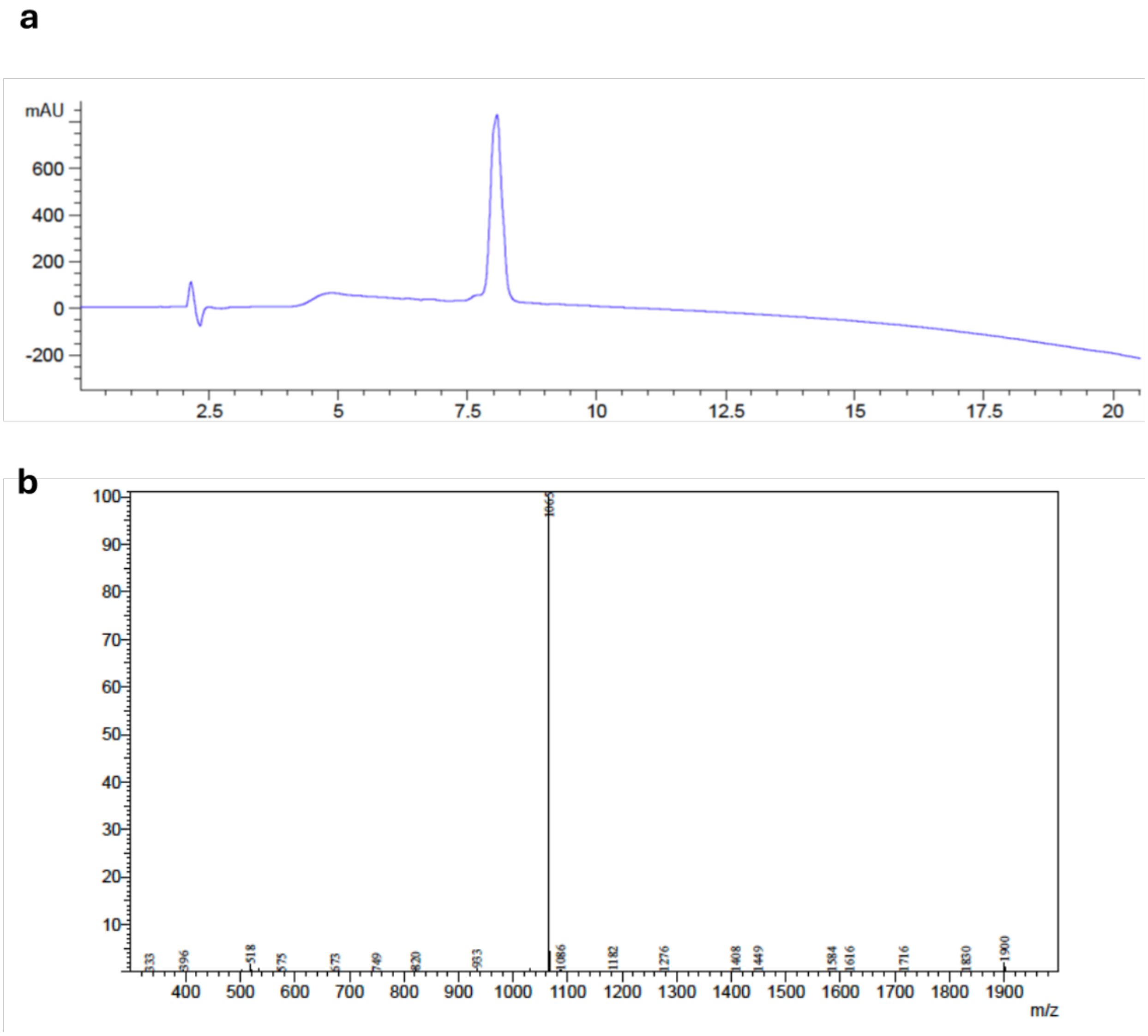


Supplementary Figure s20. TCRModel2 performance metrics assessing peptide presentation by HLA*A2:01 (n=3 top models).

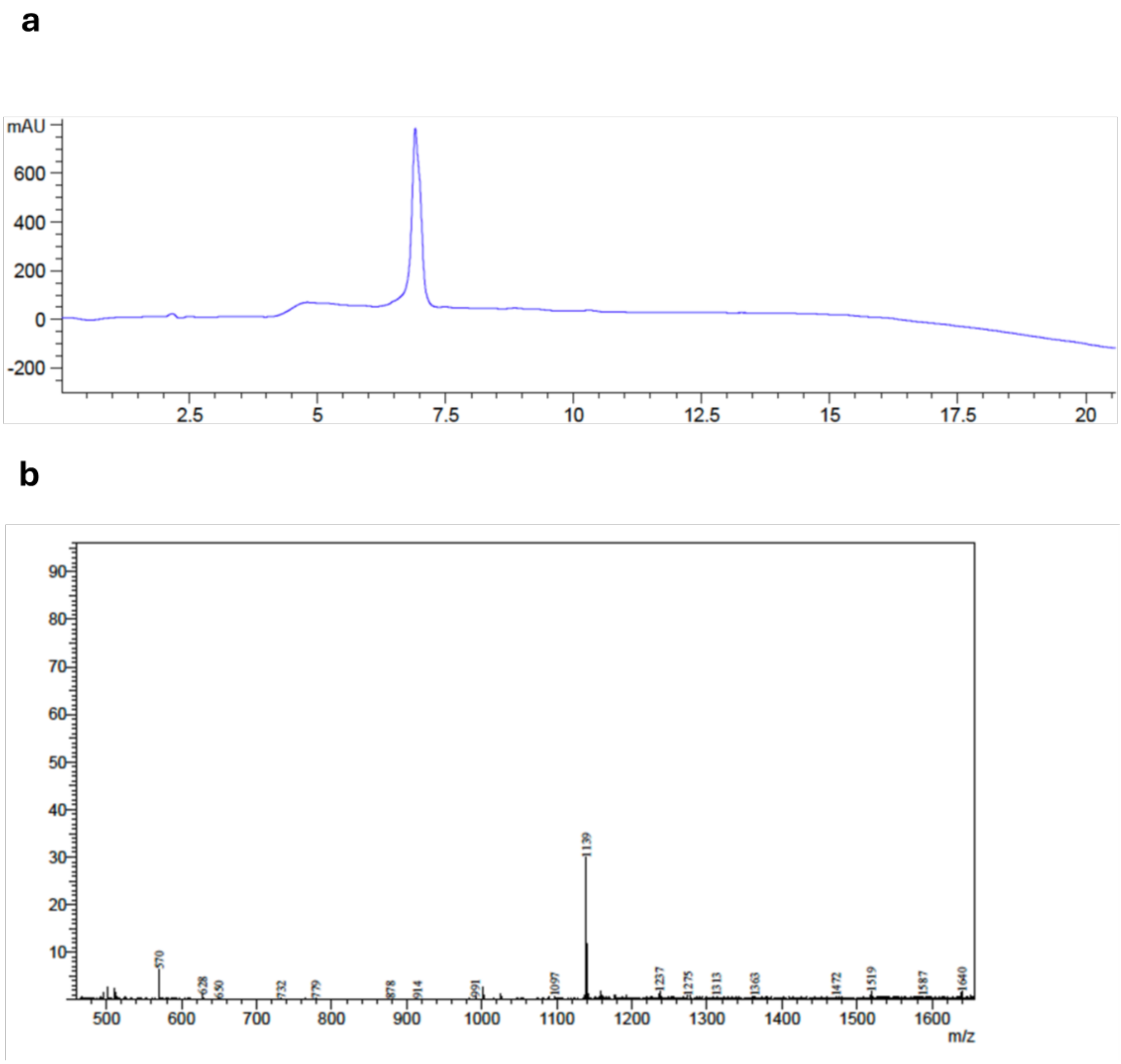


Peptide	Dissociation Energy (kJ)
GL	3.18872515
poly-Gly	7.24676806

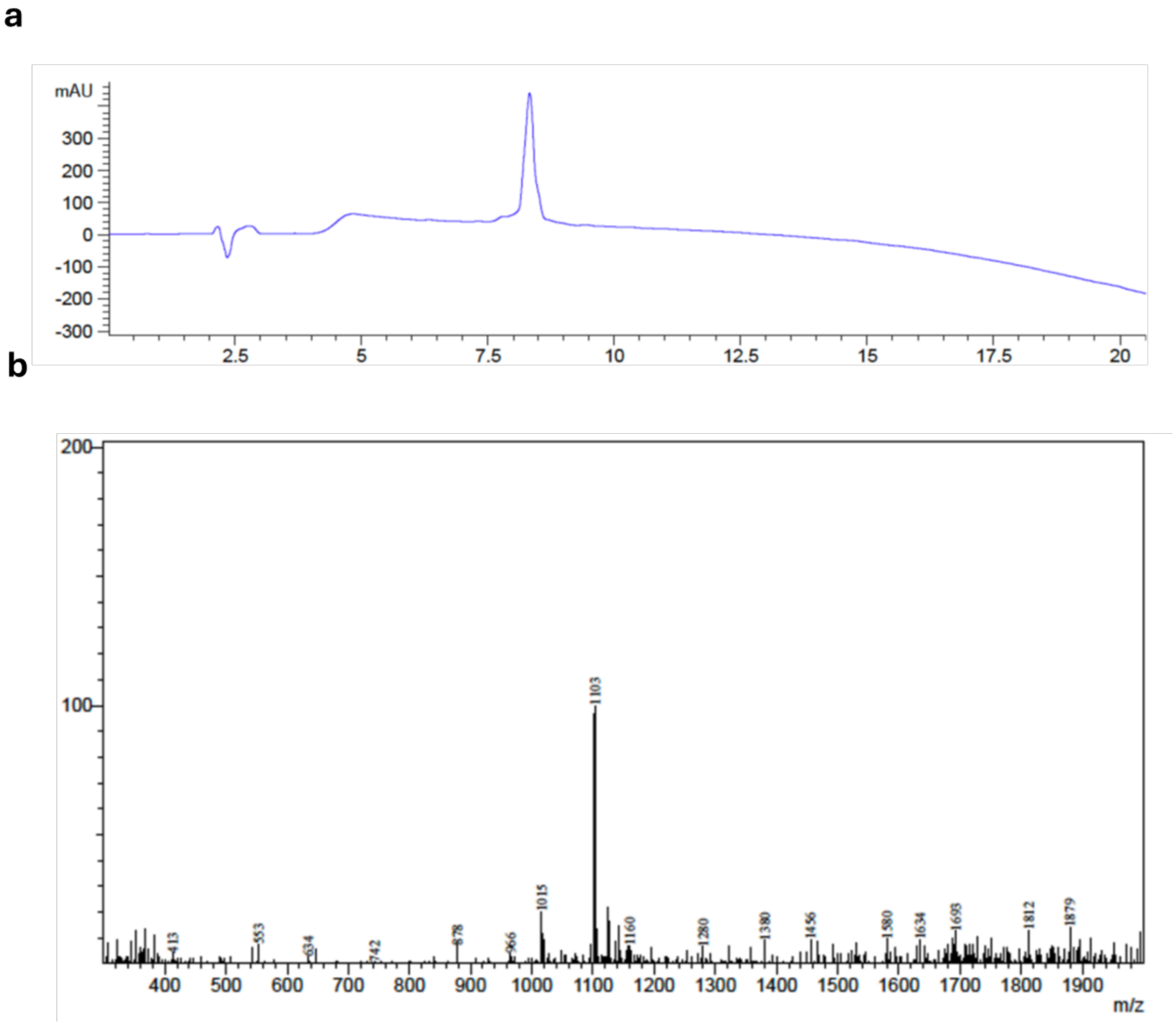
Supplementary Table s8. SMD analysis for the dipeptide (GL) and poly-Gly peptide to leave HLA-A2:01 5 nm away.



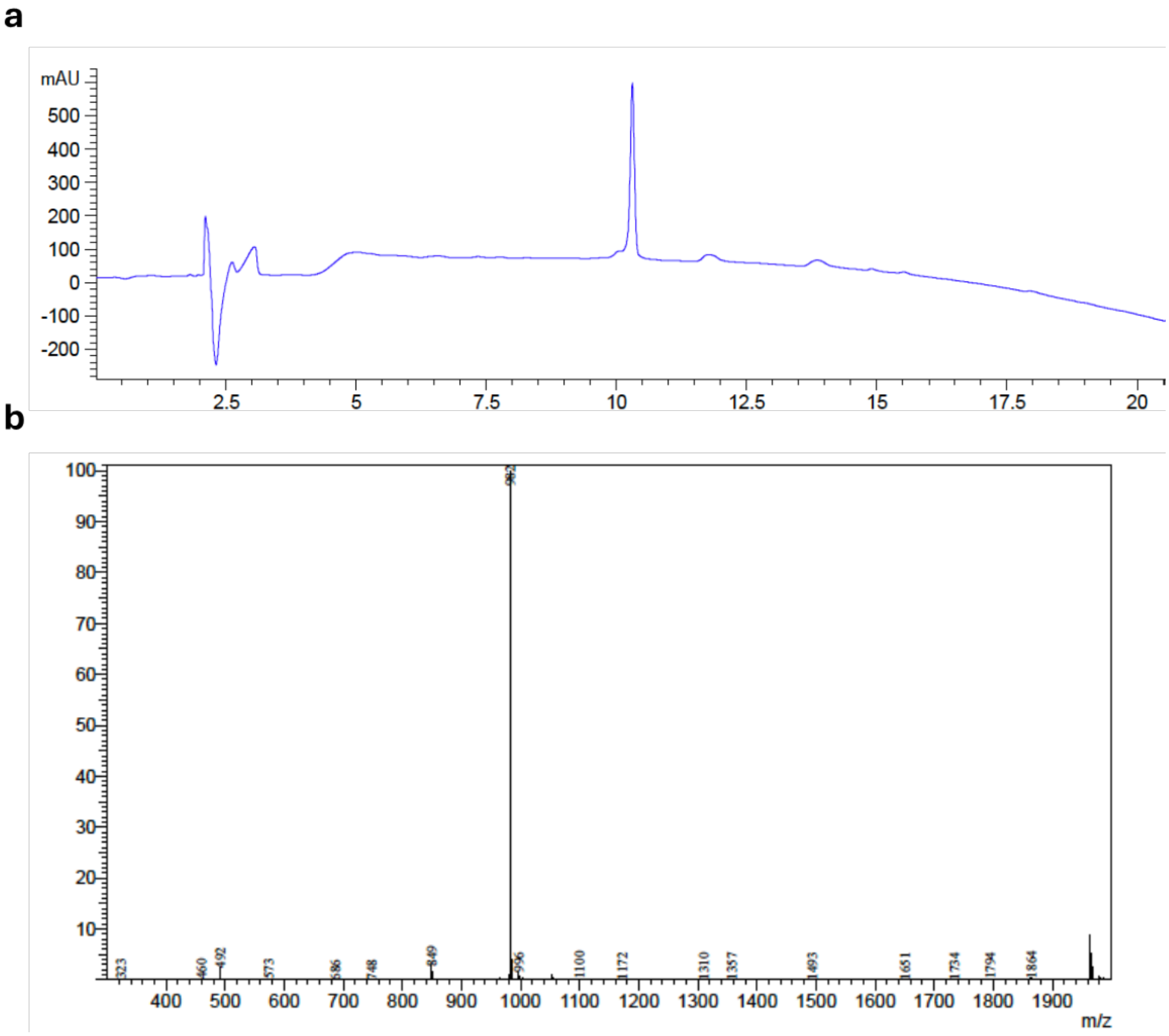
Supplementary Figure s21. FLIDLAFLI synthesis. a. HPLC chromatogram and **b**. Mass spectra for FLIDLAFLI.



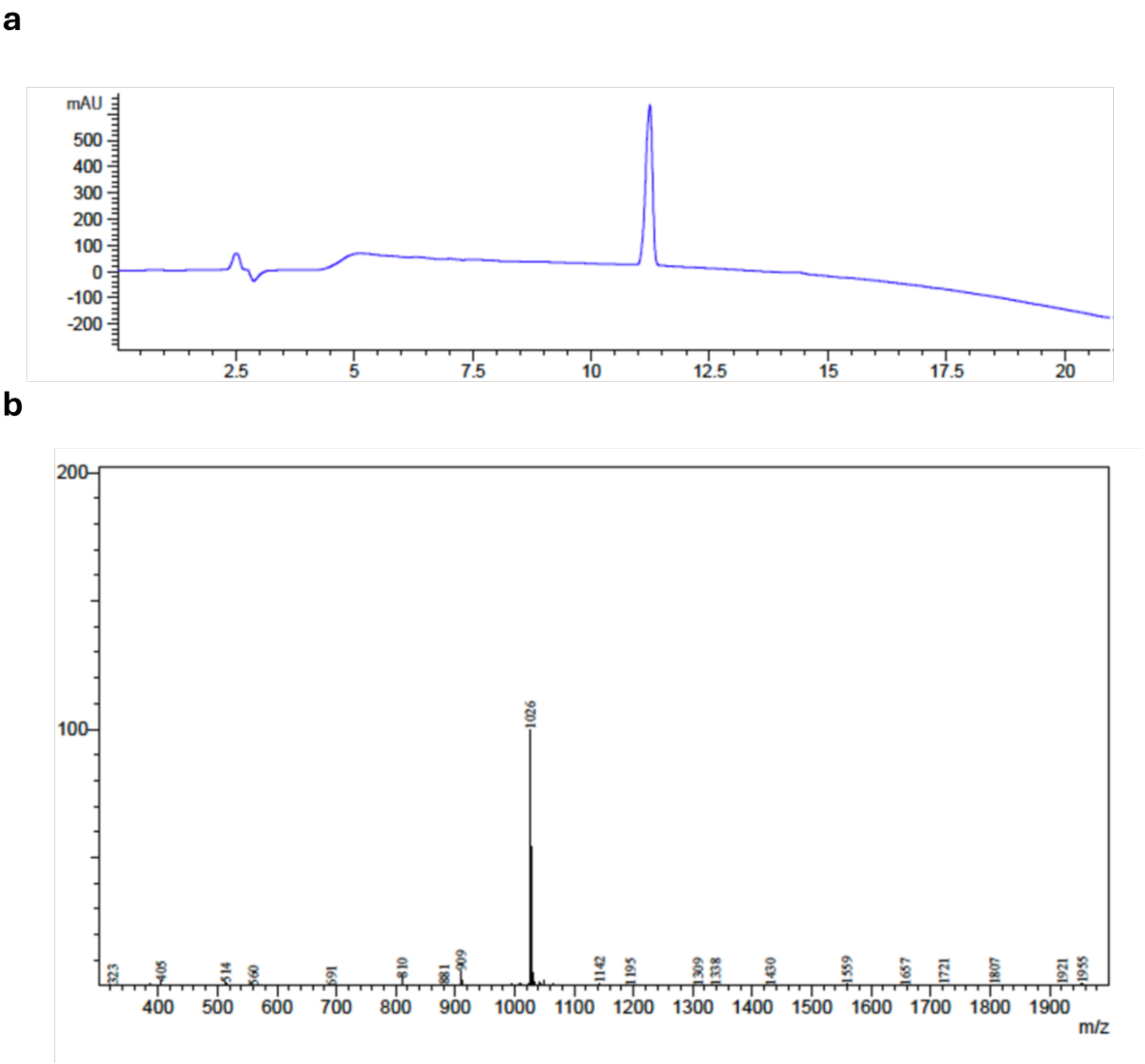
Supplementary Figure s22. FLIHSRHND synthesis. a. HPLC chromatogram and **b**. Mass spectra for FLIHSRHND.



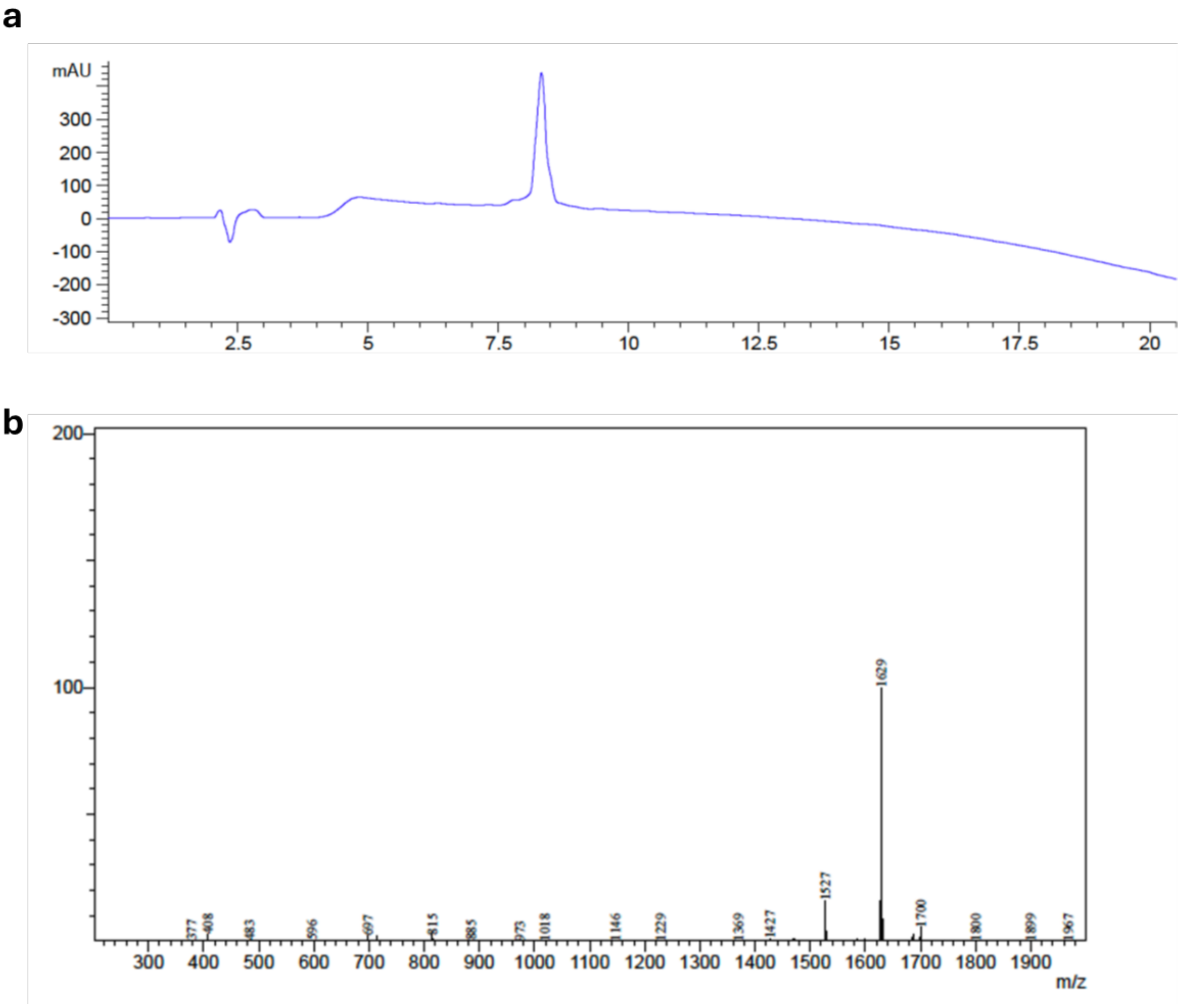
Supplementary Figure s23. HHMNMSMSK synthesis. a. HPLC chromatogram and **b**. Mass spectra for HHMNMSMSK.



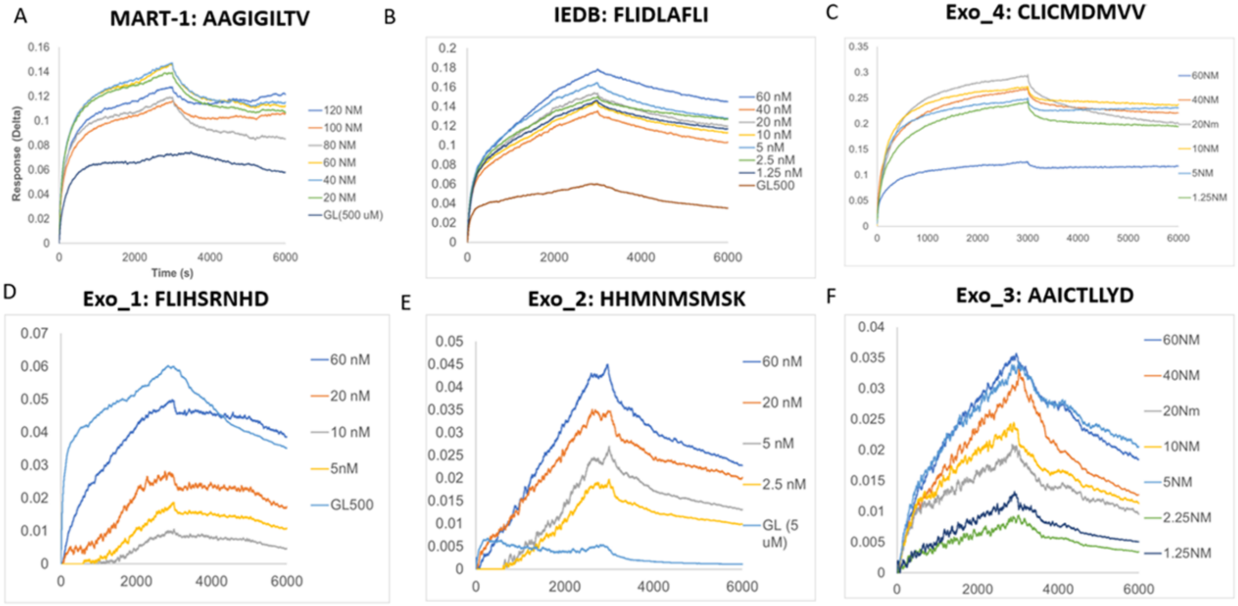
Supplementary Figure s24. AAICTLLYD synthesis. a. HPLC chromatogram and **b**. Mass spectra for AAICTLLYD.



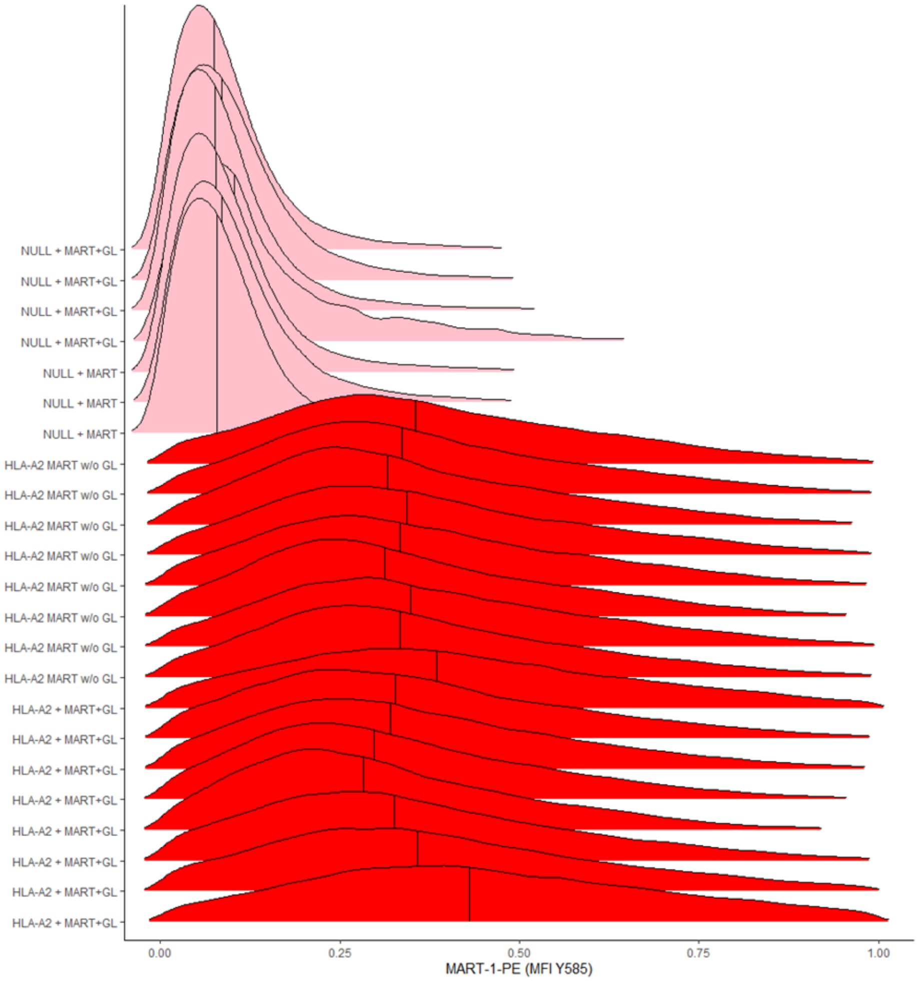
Supplementary Figure s25. CLICMDMVV synthesis. a. HPLC chromatogram and **b**. Mass spectra for CLICMDMVV.



Supplementary Figure s26. AAGIGILTV synthesis. a. HPLC chromatogram and **b**. Mass spectra for AAGIGILTV.



Supplementary Figure s27. Biolayer interferometric analysis and evaluation of binding affinity to HLA*A02:01. The dipeptide, GL, was used at 500 uM as a reference control.



Supplementary Figure s28. As reported by Saini et al, we performed a sequence exchange assay for HLA*A02:01 using with-and-without dipeptide. NULL indicates K562 cells that do not express HLA*A02:01 (K-) while HLA-A2 are K562 cells that do express HLA*A02:01 (K**+).**

Supplementary Note 1.1

Physiochemical Feature Information

Each of ExoGAN's physiochemical features were curated by either parsing through literature, using Biopython, or modifying existing equations to represent a peptide's physiochemical contribution to MHC-I's chemical potential. Below, we stratified the features into three categories: Peptide physiochemistry, peptide description, and MHC-I association. References listed detail their origination.

Peptide physiochemistry

Physiochemistry	Definition	Equation	Reference
Dipole Moment (X, Y, Z)	Magnitude of molecular charge in each cartesian direction from Coulomb repulsion	$\hat{J}_j(1)f_i(1) = f_i(1)\int \lvert\varphi_j\,(2)\rvert^2 \frac{1}{r_{12}} dr_2$	1-5
Diffusion	Represents the speed of how fast a peptide will transport in extracellular milleu	$D = \frac{K_bT}{6 \times \pi \times \eta_{blood} \times R_g}$	6
Entropy	Represents the peptide's tendency for molecular motions, normalized to length	$\frac{1}{n}\sum_{i=1}^{n}[Amino\ Acid]_i$	7
Molecular Area	Represents the molecular surface area (Solvent excluded area)	$GEPOL\ Algorithm$	3,8
Molecular Rotation (X, Y, Z)	Represents the rotational motion of a peptide	$XYZ = \frac{\pi\,(\frac{kT}{hc})^3}{(Q_{int} \times \sigma)^2}$	9
Molecular Volume	Represents the molecular volume (Solvent excluded volume)	$GEPOL\ Algorithm$	3,8
Radius of Gyration (Modified)	Represents the peptide's radius by considering its center of mass assuming all conformation states	$\sqrt{\frac{1}{2} \times MW \times B_{eff\_pep}}$	6
Stiffness	Elasticity of peptide given indexed amino acids averaged of a window	Vihinen algorithm	10-12
Solvation Energy	Represents the Gibb's Free Energy of a peptide in water	$\Delta G_S = \Delta G_{ENP} + \Delta G_{CDS}$	1-5

Peptide description

Descriptions	Definition	Equation	Reference
Aliphaticity	Residues providing interactions for hydrocarbons, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{V,I,L,M\}}$	13
Aromaticity	Residues providing interactions for non-polar, π-π interactions, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{F,W,Y,H\}}$	13
Helical Geometry	Residues providing helical secondary structure, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{V,I,Y,F,W,L\}}$	14
Hydrogen Bond Acceptors	Residues receiving hydrogen bonds, weighted by # of acceptors, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} K \times 1_{\alpha_i \in \{N,D,Q,S,T,E,H,Y\}}$	13
Hydrogen Bond Donors	Residues donating hydrogen bonds, weighted by # of donors, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} K \times 1_{\alpha_i \in \{R,K,N,Q,S,T,W,H,Y\}}$	13
Hydropathy	Kyte-Doolittle's representation of globular protein folding, normalized to length	$\frac{1}{n}\sum_{H=1}^{n} [Amino\ Acid]_H$	15
Hydrophilicity	Residues providing affinity for aqueous solutions; polar interactions, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{S,T,H,N,Q,E,D,K,R\}}$	13
Hydrophobicity	Residues providing affinity for oils/fats/lipids; non-polar interactions, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{V,I,L,F,W,Y,M\}}$	13
Instability	Residues contributing to local structural disorder	$\sum_{G=1}^{n} [Amino\ Acid]_G$	16
Isoelectric Point	pH where net electric charge of the various peptide states in the extracellular environment is zero	$\sum_{Iso=1}^{n} [Amino\ Acid]_{Iso}$	13
Negativity	Residues that provide a negative electric charge to the peptide at pH 7.4	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{D,E\}}$	13
Polar Neutral	Residues that provide polar character but no electric charge	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{S,T,C,M,N,Q\}}$	13
Positive Charge	Residues that provide a positive electric charge to the peptide at pH 7.4	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{K,R,H\}}$	13
Sheet Geometry	Residues providing sheet secondary structure, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{E,M,A,L\}}$	14
Sulfurariticity	Residues providing sulfur character, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{S,C,T,M\}}$	13
Turn Geometry	Residues providing turns/loops to the secondary structure, normalized to length	$\frac{1}{n}\sum_{i=1}^{n} 1_{\alpha_i \in \{N,P,G,S\}}$	14
Molecular Weight	Sum of amino acid molecular weights	$\sum_{i=1}^{n} [Amino\ Acid]_{MW}$	13

MHC-I association

Associations	Definition	Equation	Reference
Antigenicity (modified)	Represents antigenicity potential	$\frac{\mathrm{Hydropathy}}{Recognition}$	17
Pocket Window	Represents the hydrophobic residue bias for interaction with HLA*A02:01 in pockets a,b,c and f, normalized to the total pockets	$\frac{1}{4} \times [$ $1_{\alpha_i \in F,L} +$ $1_{b_i \in L,M,I,V} +$ $1_{c_i \in L} +$ $1_{f_i \in V,L,I,A}$ $]$	Figure 2a.
Recognition	Represents molecular visibility, normalized to length	$\frac{1}{n}\sum_{R=1}^{n}[Amino\ Acid]_R$	18

Supplementary Note 1.2 – Model descriptions for each classic model used.

To evaluate sequence, physiochemistry, and sequence and physiochemical information to classify strong and weak HLA*A02:01 antigens from IEDB's database, we list the models used and their parameters below. The parameters listed were determined through a grid search to optimize the F1 score, using a stratifiedKFold approach.

Classic Model	Information used	Parameters
Logistic Regression	Sequence	C:1, Penalty: L1
	Physiochemistry	C:0.1, Penalty: L1
	Sequence + Physiochemistry	C:1, Penalty: L1
Random Forest	Sequence	N_estimators: 5000
	Physiochemistry	N_estimators: 1000
	Sequence + Physiochemistry	N_estimators: 2500
Linear Discriminant Analysis	Sequence	Solver: svd
	Physiochemistry	Solver: svd
	Sequence + Physiochemistry	Solver: svd
Support Vector Machine	Sequence	C: 100, kernel: Linear
	Physiochemistry	C: 20, kernel: rbf
	Sequence + Physiochemistry	C: 10, kernel: Linear

Supplementary Note 1.3 – Descriptions of models used to predict HLA*A02:01 binding.

To survey generated neoantigens on existing, gold-standard, experimentally validated models to assess sequence diversity and bias impacts on predictions, we leveraged the models from IEDB’s weekly performance survey and recently published deep learning models. We describe each model used below.

Model	Prediction Architecture	Reference
ACME	Convolutional Neural Network with attention mechanism	[19]
ANN 4.0	Shallow Neural Network	[20]
MHCFlurry 2.0	Convolutional Neural Network with pruned layers	[21]
NetMHCPan 4.1	Shallow neural network with attention mechanism	[22]
SMM	Scoring Matrix	[23]
SMMPMBEC	Scoring Matrix	[24]

References

1 Caldeweyher, E. *et al.* A generally applicable atomic-charge dependent London dispersion correction. *J Chem Phys* **150**, 154122 (2019). https://doi.org/10.1063/1.5090222

2 Caldeweyher, E., Bannwarth, C. & Grimme, S.

3 Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J Phys Chem B* **113**, 6378-6396 (2009). https://doi.org/10.1021/jp810292n

4 Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys Chem Chem Phys* **8**, 1057-1065 (2006). https://doi.org/10.1039/b515623h

5 Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **7**, 3297-3305 (2005). https://doi.org/10.1039/b508541a

6 Teraoka, I.

7 Hutchens, J. O.

8 Pascual-Ahuir, J. L. & Silla, E. GEPOL: An improved description of molecular surfaces. I. Building the spherical surface set. *Journal of Computational Chemistry* **11**, 1047-1060 (1990). https://doi.org/https://doi.org/10.1002/jcc.540110907

9 Gilson, M. K. & Irikura, K. K. Symmetry numbers for rigid, flexible, and fluxional molecules: theory and applications.

10 Vihinen, M., Torkkila, E. & Riikonen, P. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics* **19**, 141-149 (1994). https://doi.org/https://doi.org/10.1002/prot.340190207

11 Smith, D. K., Radivojac P Fau - Obradovic, Z., Obradovic Z Fau - Dunker, A. K., Dunker Ak Fau - Zhu, G. & Zhu, G. Improved amino acid flexibility parameters.

12 Bowman, J. *Protein flexibility calculations with Python*, <https://www.polarmicrobes.org/protein-flexibility-calculation-with-python/> (2015).

13 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009). https://doi.org/10.1093/bioinformatics/btp163

14 Chou, P. Y. & Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* **47**, 45-148 (1978). https://doi.org/10.1002/9780470122921.ch2

15 Kyte, J. & Doolittle, R. F.

16 Guruprasad, K., Reddy, B. V. & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **4**, 155-161 (1990). https://doi.org/10.1093/protein/4.2.155

17 Fraga, S. Theoretical prediction of protein antigenic determinants from amino acid sequences. *Canadian Journal of Chemistry* **60**, 2606-2610 (1982). https://doi.org/10.1139/v82-374

18 Fraga, S. Recognition of amino acids in solution. *Journal of Molecular Structure* **94**, 251-260 (1983). https://doi.org/https://doi.org/10.1016/0022-2860(83)90283-1

19 Hu, Y. *et al.* ACME: pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics* **35**, 4946-4954 (2019). https://doi.org/10.1093/bioinformatics/btz427

20 Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* **12**, 1007-1017 (2003). https://doi.org/10.1110/ps.0239403

21 O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell systems* **11**, 42-48.e47 (2020). https://doi.org/10.1016/j.cels.2020.06.010

22 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research* **48**, W449-W454 (2020). https://doi.org/10.1093/nar/gkaa379 %J Nucleic Acids Research

23 Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method.

24 Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries.