Generation of structure-guided pMHC-I libraries using Diffusion Models

Sergio Emilio Mares [1] **Ariel Espinoza Weinberger** [2] **Nilah M. Ioannidis** [1 2 3]

Abstract

Personalized vaccines and T-cell immunotherapies depend critically on identifying peptide-MHC class I (pMHC-I) interactions capable of eliciting potent immune responses. However, current benchmarks and models inherit biases present in mass-spectrometry and binding-assay datasets, limiting discovery of novel peptide ligands. To address this issue, we introduce a structure-guided benchmark of pMHC-I peptides designed using diffusion models conditioned on crystal structure interaction distances. Spanning twenty high-priority HLA alleles, this benchmark is independent of previously characterized peptides yet reproduces canonical anchor residue preferences, indicating structural generalization without experimental dataset bias. Using this resource, we demonstrate that state-of-the-art sequence-based predictors perform poorly at recognizing the binding potential of these structurally stable designs, indicating allele-specific limitations invisible in conventional evaluations. Our geometry-aware design pipeline yields peptides with high predicted structural integrity and higher residue diversity than existing datasets, representing a key resource for unbiased model training and evaluation.

1. Introduction

Peptide–MHC class I (pMHC-I) interactions are central to adaptive immunity, enabling cytotoxic T cells to recognize and eliminate infected or cancerous cells (Chaplin, 2010). Predictors of pMHC-I binding have become essential tools for personalized T-cell immunotherapies and modern vaccine design (Saxena et al., 2025). Given the vast combinatorial diversity of $\approx 100,000$ distinct peptides, and after accounting for polymorphisms, insertions, deletions, and aberrant splicing, experimentally mapping all binding peptides is infeasible (Yewdell et al., 2003). This necessitates accurate *in silico* prediction methods to prioritize candidate peptides for immunotherapeutic design (Walz et al., 2015). Despite substantial progress, current pMHC-I prediction methods face important limitations. Most state-of-the-art models are sequence-based and trained on a large library of known binders from public databases, such as the Immune-Epitope Database (IEDB) which contains a library of $> 10^6$ pMHCs (Vita et al., 2025). These datasets predominantly originate from mass-spectrometry (MS) immunopeptidomics (Sarkizova et al., 2020) and *in vitro* binding assays and thus carry experimental biases. One well-documented bias is the under-detection of cysteine-containing peptides in standard MS workflows, which in turn causes such peptides to be under-represented in databases and often missed by trained predictors (Bruno et al., 2023). Furthermore, many benchmarks rely on similar experimental data for evaluation, potentially inflating performance by testing on peptide sequences with distributions similar to model training sets. This over-reliance on biased datasets and narrow test sets raises concerns that reported accuracy overstates real-world generalization capacity (Machaca et al., 2024).

To confront these challenges, we introduce a structure-conditioned approach using diffusion models for pMHC-I peptide generation. Our diffusion-based generative model explicitly conditions on the three-dimensional structure of the MHC-I binding groove. By leveraging this structural context, it designs peptides that are inherently compatible with a given MHC allele's binding pocket, ensuring that all generated sequences are structurally valid binders. This approach enables exploration of peptide sequence space beyond the biases present in current databases, yielding novel, previously out-of-distribution peptides guided by structural binding preferences. By generating plausible yet unconventional peptides, our structure-conditioned diffusion model expands the landscape of potential immunotherapy targets and provides new test cases for existing predictors.

[1]Center for Computational Biology, University of California, Berkeley, Berkeley, California, USA [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, California, USA [3]Chan Zuckerberg Biohub, San Francisco, California, USA. Correspondence to: Sergio Emilio Mares <sergio.mares@berkeley.edu>.

The 2^{nd} International Conference on Machine Learning in Generative AI and Biology Workshop.

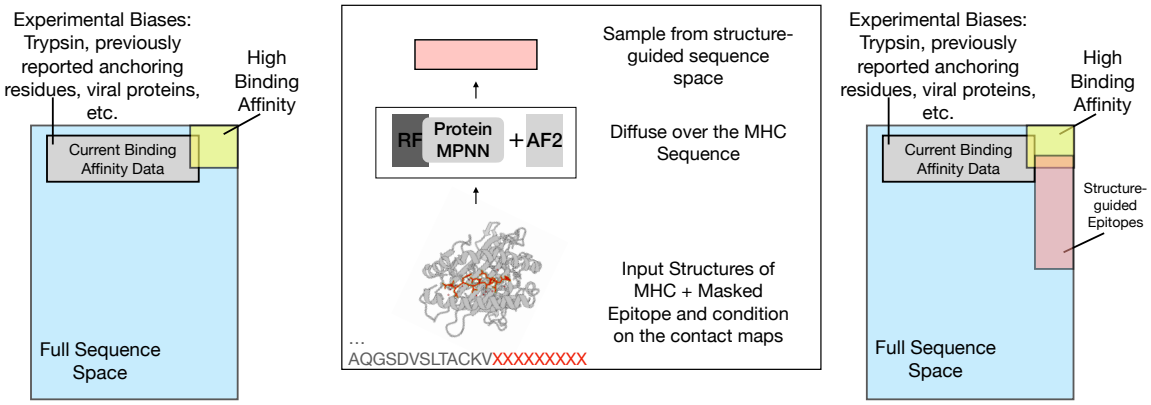

Figure 1. Overview of the structure-guided generative pipeline for designing high-affinity peptides for MHC class I molecules.

2. Data and Methods

2.1. Dataset

We selected 189 peptide-MHC Class I crystal structures from the Protein Data Bank (PDB) (Berman et al., 2003), with coverage of 20 human HLA alleles (accessed 23-04-2025) with peptides of length 9-11 amino acids. Structures were selected with resolution ≤ 3.5Å with the aim to capture hydrogen bonds and van der Waals interactions ≤ 3.5Å (Wlodawer et al., 2018). We excluded any structures with missing peptide residues, non-standard amino acids, or incomplete annotations of the HLA.

2.2. Hot-spot identification

We generated a residue–residue contact map by computing minimum Euclidean heavy-atom distances between peptide (9–11 residues) and MHC heavy chain (150–300 residues) residues, omitting hydrogens due to inconsistent resolution at ≤ 3.5Å.

Distance cut-off. Across 189 crystal structures (20 HLA-I alleles), 0.170 % of heavy-atom pairs fall within ≤ 3.5Å (95 % CI 0.166–0.175), placing this threshold at the 99.8th percentile of contact strength. Coverage rises to 6.9% at 5Å. Because the 2.2–3.2Å window matches canonical hydrogen-bond/salt-bridge geometries (≥ 4 kcal mol^{-1}) (Gilli & Gilli, 2018; Kastritis & Bonvin, 2013), we label residues with ≤ 3.5Å contacts as *core hotspots* and those with ≤ 5Å contacts as *high-contact* positions, fixing both sets during diffusion-based sequence design.

2.3. Generative Pipeline

The pipeline (Fig. 1) begins with the crystallized MHC-peptide structure. Then, we mask the peptide sequence in each PDB structure, preserving the MHC scaffold and hot-spot interactions. RFdiffusion was ran with 50 time steps and fixed structure of the MHC. RFdiffusion (Watson et al., 2023) generates the sequence space over 50 iterations, conditioned by hotspot residues to maintain high-affinity contacts. We chose a window of a 9-11-residues, and no imposed symmetry and sampling temperature of 0.5. Each RFdiffusion-generated backbone is next threaded through ProteinMPNN (Dauparas et al., 2022) to optimize side-chain identities while keeping the MHC scaffold fixed. We isolate the peptide chain coordinates, supply the full complex backbone to ProteinMPNN `--complex` mode, and sample $N = 64$ sequences per backbone at a sampling temperature of 0.5. Sequences are ranked by their negative log-likelihood (NLL); the top five NLL sequences are retained for structural evaluation. The top ProteinMPNN sequences are folded with **AlphaFold2-Multimer** (AF2) while the original MHC α- and β-chains are kept intact. AF2 is run with `num_recycles` = 6, `model_preset` = *multimer* (Jumper et al., 2021). Peptides whose AlphaFold predictions scored pLDDT > 0.8 were retained as 'high-confidence' designs.

2.4. Models Evaluated

Binding affinities are predicted by MHC-Flurry (O'Donnell et al., 2020), NetMHCSpan (McInnes et al., 2018), HLApollo (Thrift et al., 2024), HLAthena (Sarkizova et al., 2020), MixMHCpred (Bassani-Sternberg et al., 2017), MHCNuggets (Shao et al., 2020), and ESMCBA, a fine-tuned ESM-Cambrian model (ESM Team, 2024).

2.5. Additional Benchmarking Datasets

We employed three additional benchmarking datasets: 1. IEDB database with peptides after 2020, eliminating data leakage for most models training data. 2. A constructed dataset with preservation of the anchoring residues follow-

ing the same distribution of peptides observed in the public dataset, and the randomly generated rest of the sequence. 3. A list of 9-11-mers auto-regressively generated with ESM2 (ESM Team, 2024), starting from an initial random token and sampling each residue from the model's predictive distributions. We ensured independence to the dataset by removing any overlapping peptides between this generated pipeline and the public databases.

3. Results

3.1. Allele-aware motif similarity.

To investigate peptide binding motif similarities across HLA alleles, we constructed Position Weight Matrices (PWMs) from allele-dependent peptides associated with high-confidence PDB structures (pLDDT >0.8). Pairwise Jensen-Shannon (JS) divergences were computed between peptide profiles of structures within the same allele, revealing varying degrees of motif consistency. JS divergence shows the similarities among structures and allele groups (Supp. Fig. S1). The generated peptides follow similar distributions according to the MHC allele, and show differences to super alleles (eg. HLA-A vs HLA-B structures).

3.2. Experimental Validation with orthogonal unbiased data

EpiScan is a high-throughput, cell-based platform that presents bar-coded peptide libraries on the surface of HLA-I molecules and quantifies their relative display levels by deep sequencing. Because it bypasses mass-spectrometry and in vitro binding assays, EpiScan provides an unbiased measurement of peptide presentation (Bruno et al., 2023). However, EpiScan's study tested >100,000 peptides and found <400 peptides that bound to HLA-A*02:01 and <1500 for HLA-B*57:01. Although a high-throughput experimental approach, it only probed a narrow slice of the sequence space. Our structure-guided diffusion library complements it by generating tens of thousands of anchor-compatible peptides that explore under-sampled regions, covering the gap in diversity and bias.

Position-wise conservation. We report per-position Pearson R between designed libraries and public peptides (Fig. 2). Anchor positions (P2 and PΩ) reach $R > 0.4$ and $R > 0.8$ for our library and EpiScan's data respectively, whereas solvent-exposed positions drop to $R \approx 0.2$, corroborating the selective conservation of biophysically constrained sites (Supp. Fig. S2).

Enrichment in HLA-A*02:01 and HLA-B*57:01 EpiScan enriches proline at P2 and P9, whereas our library restores the expected hydrophobic peak at P2, P9 and P1, recapitulating the unbiased data enrichment in HLA-A*02:01 (Supp. Fig. S3).

Figure 2. Position-wise Pearson correlations of amino acid preferences between our generated peptides, public database motifs, and unbiased EpiScan data.

In HLA-B*57:01, we observe a similar trend where the enrichment is recapitulated in both methods and exposing the bias found in MS data (Supp. Fig. S6).

3.3. Full ROC curve performance across peptide classes

For each predictor, we constructed Receiver Operating Characteristic (ROC) curves across four distinct peptide evaluation classes and computed the Area Under the ROC (AUROC) to assess discriminative performance (Fig. 3).

Performance on Experimentally Validated Binders

To establish a baseline, we evaluated each predictor's ability to distinguish experimentally validated binders from other peptides. All methods performed moderately, achieving AUROCs from 0.66 (Apollo) to 0.81 (ESMCBA), with most predictors clustering tightly around 0.74–0.81. These high AUROCs demonstrate robust recognition of known canonical binding motifs by current sequence-based methods, aligning well with their training data.

Performance on Random-Sampled Peptides

To evaluate predictor specificity, we tested their ability to discriminate randomly sampled peptides from experimentally validated peptides with strong binding. Most predictors indeed assigned high scores to random peptides, achieving desirable AUROCs ranging from 0.86 (HLApollo) - 1.00 (NetMHCpan). These results confirm that predictors are robustly specific, effectively distinguishing random peptides from biologically plausible sequences.

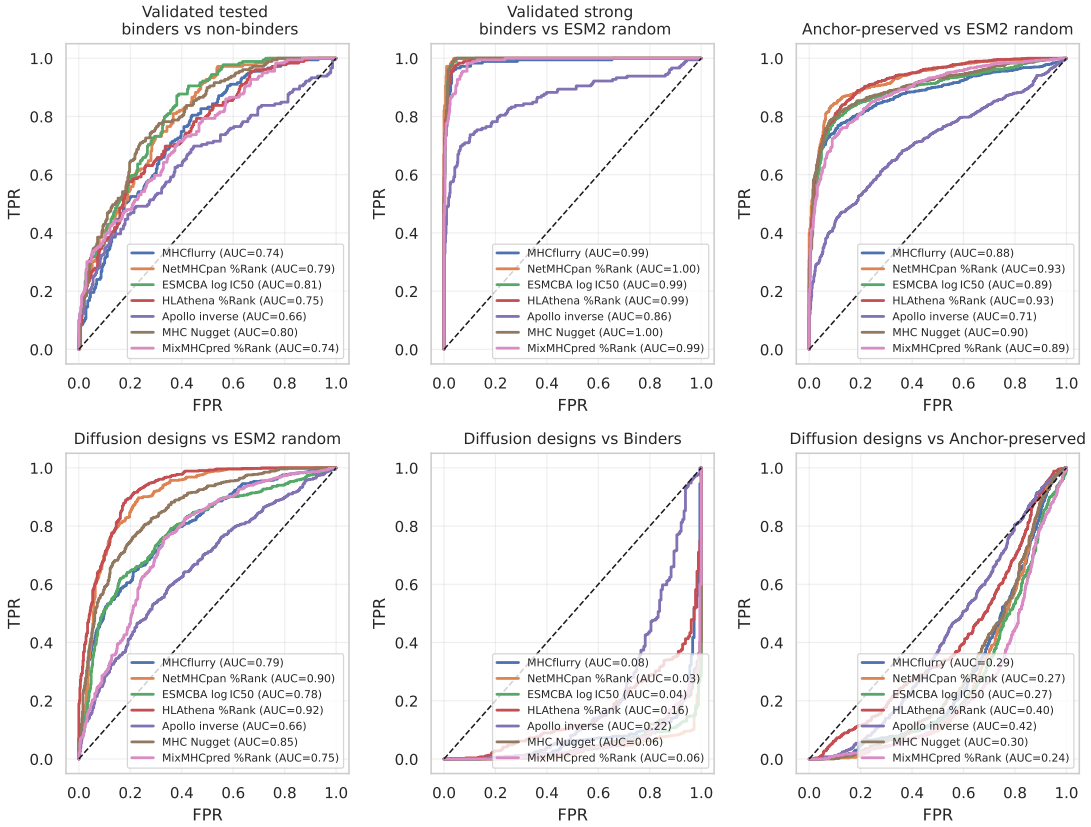

Figure 3. ROC curves and AUROC values for seven binding affinity predictors across four peptide evaluation classes: experimentally validated binders, permutation-test controls, random-sampled negatives, and structure-guided diffusion designs.

Performance on Permutation-Test Peptides

To test predictors' robustness to subtle sequence perturbations, we next assessed performance on anchor-preserved permutation-test peptides. Ideally, predictors should assign these control sequences low binding scores (AUC significantly above 0.5), reflecting their non-physiological sequence context. However, we observed AUROCs ranging from 0.71 (HLApollo) to 0.93 (HLAthena and NetMHCpan), indicating that predictors naively assigned these peptides relatively high scores. This result reveals a troubling sensitivity to subtle global sequence context perturbations beyond anchor positions, highlighting a critical vulnerability in current predictive approaches.

Performance on Structure-Guided Diffusion Designs

Finally, we evaluated model performance on peptides explicitly designed to structurally complement the MHC binding pocket (structure-guided designs with pLDDT > 0.8). We first evaluated them against the random generated peptides. All methods' AUROCs show clear ability to distinguish the diffusion peptides from random noise. However, all methods achieved poor discriminative performance against validated strong binders, with AUROCs ranging from 0.06 (MHC Nugget) to 0.22 (Apollo). This performance clearly exposes a significant blind spot: current predictors are largely unable to recognize structurally plausible peptides, highlighting critical limitations in their generalization capabilities and underscoring the need for structurally aware training data.

4. Discussion and limitations

Our study addresses critical limitations in pMHC-I binding prediction by leveraging diffusion models conditioned on atomic-level interactions, effectively avoiding biases inherent in mass-spectrometry and binding assay data. The structure-aware generative method introduced here challenges current models, highlighting their inability to recognize structurally valid, novel peptides that are out-of-distribution for their experimental training datasets.

Our workflow and generated library of 10^5 peptides serves both as a benchmark for future predictive models and as a method to expose data biases. While validation of our designed peptides primarily relied on the unbiased EpiScan dataset, which is limited to four alleles, our method is generalizable to a broader range of HLA types. Addition-

ally, incorporating TCR-binding predictions will extend our methodology's relevance to comprehensive immunotherapy design. Future research should prioritize broader experimental work and integration of TCR interactions to fully capture the complexity of immune responses.

References

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H.-Y., Gannon, P. O., Kandalaft, L. E., Coukos, G., and Gfeller, D. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Computational Biology*, 13(8):e1005725, 2017. doi: 10.1371/journal.pcbi.1005725.

Berman, H., Henrick, K., and Nakamura, H. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10:980, 2003. doi: 10.1038/nsb1203-980. URL https://doi.org/10.1038/nsb1203-980. Accessed via https://www.wwpdb.org.

Bruno, P. M., Timms, R. T., Abdelfattah, N. S., Leng, Y., Lelis, F. J. N., Wesemann, D. R., Yu, X. G., and Elledge, S. J. High-throughput, targeted mhc class i immunopeptidomics using a functional genetics screening platform. *Nature Biotechnology*, 41(7):980–992, July 2023. doi: 10.1038/s41587-022-01566-x. URL https://doi.org/10.1038/s41587-022-01566-x.

Chaplin, D. D. Overview of the immune response. *Journal of Allergy and Clinical Immunology*, 125(2, Supplement 2):S3–S23, 2010. ISSN 0091-6749. doi: 10.1016/j.jaci.2009.12.980. URL https://www.sciencedirect.com/science/article/pii/S0091674909028371.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R., Milles, L., Wicky, B., Courbet, A., de Haas, R., Bethel, N., Leung, P., Huddy, T., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A., King, N., and Baker, D. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022. doi: 10.1126/science.add2187.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Website, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian. Accessed December 4, 2024.

Gilli, P. and Gilli, G. Hydrogen bonds: Simple after all? *Biochemistry*, 57(9):1357–1369, 2018. doi: 10.1021/acs.biochem.7b01166.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, M., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Kastritis, P. L. and Bonvin, A. M. J. J. On the binding affinity of macromolecular interactions: Daring to ask why proteins interact. *Journal of the Royal Society Interface*, 10(79):20120835, 2013. doi: 10.1098/rsif.2012.0835.

Machaca, V., Goyzueta, V., Cruz, M. G., Sejje, E., Pilco, L. M., López, J., and Túpac, Y. Transformers meets neoantigen detection: a systematic literature review. *Journal of Integrative Bioinformatics*, 21(2):20230043, 2024. doi: 10.1515/jib-2023-0043. URL https://doi.org/10.1515/jib-2023-0043.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

O'Donnell, T., Rubinsteyn, A., and Laserson, U. Mhcflurry 2.0: Improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48, 2020. doi: 10.1016/j.cels.2020.06.009.

Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., Justesen, S., Stevens, J., Lane, W. J., Eisenhaure, T., Zhang, G. L., Clauser, K. R., Hacohen, N., Carr, S. A., Wu, C. J., and Keskin, D. B. A large peptidome dataset improves hla class i epitope prediction across most of the human population. *Nature Biotechnology*, 38(2):199–209, 2020. doi: 10.1038/s41587-019-0322-9.

Saxena, M., Marron, T. U., Kodysh, J., Finnigan, Jr, J. P., Onkar, S., Kaminska, A., Tuballes, K., Guo, R., Sabado, R. L., Meseck, M., O'Donnell, T. J., Sebra, R. P., Parekh, S., Galsky, M. D., Blasquez, A., Gimenez, G., Bicak, M., Bozkus, C. C., Delbeau-Zagelbaum, D., Rodriguez, D., Acuna-Villaorduna, A., Misiukiewicz, K. J., Posner, M. R., Miles, B. A., Irie, H. Y., Tiersten, A., Doroshow, D. B., Wolf, A., Mandeli, J., Brody, R., Salazar, A. M., Gnjatic, S., Hammerbacher, J., Schadt, E., Friedlander, P., Rubinsteyn, A., and Bhardwaj, N. Pgv001, a multipeptide personalized neoantigen vaccine platform: Phase

i study in patients with solid and hematologic malignancies in the adjuvant setting. *Cancer Discovery*, 15(5):930–947, May 2025. doi: 10.1158/2159-8290.CD-24-0934. URL https://doi.org/10.1158/2159-8290.CD-24-0934.

Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I. K. A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., Riemer, A. B., Velculescu, V. E., Anagnostou, V., Pagel, K. A., and Karchin, R. High-throughput prediction of MHC class i and class ii neoantigens with MHCnuggets. *Cancer Immunology Research*, 8(3):396–408, 2020. doi: 10.1158/2326-6066.CIR-19-0464.

Thrift, W. J., Lounsbury, N. W., Broadwell, Q., Heidersbach, A., Freund, E., Abdolazimi, Y., Phung, Q. T., Chen, J., Capietto, A.-H., Tong, A.-J., Rose, C. M., Blanchette, C., Lill, J. R., Haley, B., Delamarre, L., Bourgon, R., Liu, K., and Jhunjhunwala, S. Towards designing improved cancer immunotherapy targets with a peptide-mhc-i presentation model, hlapollo. *Nature Communications*, 15(1):10752, 2024. doi: 10.1038/s41467-024-54887-7.

Vita, R., Blazeska, N., Marrama, D., Members, I. C. T., Duesing, S., Bennett, J., Greenbaum, J., De Almeida Mendes, M., Mahita, J., Wheeler, D. K., Cantrell, J. R., Overton, J. A., Natale, D. A., Sette, A., and Peters, B. The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443, 2025. doi: 10.1093/nar/gkae1092.

Walz, S., Stickel, J. S., Kowalewski, D. J., Schuster, H., Weisel, K., Backert, L., Kahn, S., Nelde, A., Stroh, T., Handel, M., Kohlbacher, O., Kanz, L., Salih, H. R., Rammensee, H.-G., and Stevanović, S. The antigenic landscape of multiple myeloma: mass spectrometry (re)defines targets for t-cell–based immunotherapy. *Blood*, 126(10):1203–1213, September 2015. ISSN 0006-4971. doi: 10.1182/blood-2015-04-640532. URL https://doi.org/10.1182/blood-2015-04-640532.

Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., Borst, A., Ragotte, R., Milles, L., Wicky, B., Hanikel, N., Pellock, S., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S., Raghavan, A., Chow, C., Carter, L., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Wlodawer, A., Dauter, Z., and Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS Journal*, 285(2):191–208, 2018. doi: 10.1111/febs.14108. URL https://doi.org/10.1111/j.1742-4658.2007.06178.x.

Yewdell, J. W., Reits, E., and Neefjes, J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature Reviews Immunology*, 3(12):952–961, 2003. doi: 10.1038/nri1250.

A. Supplemental Figures

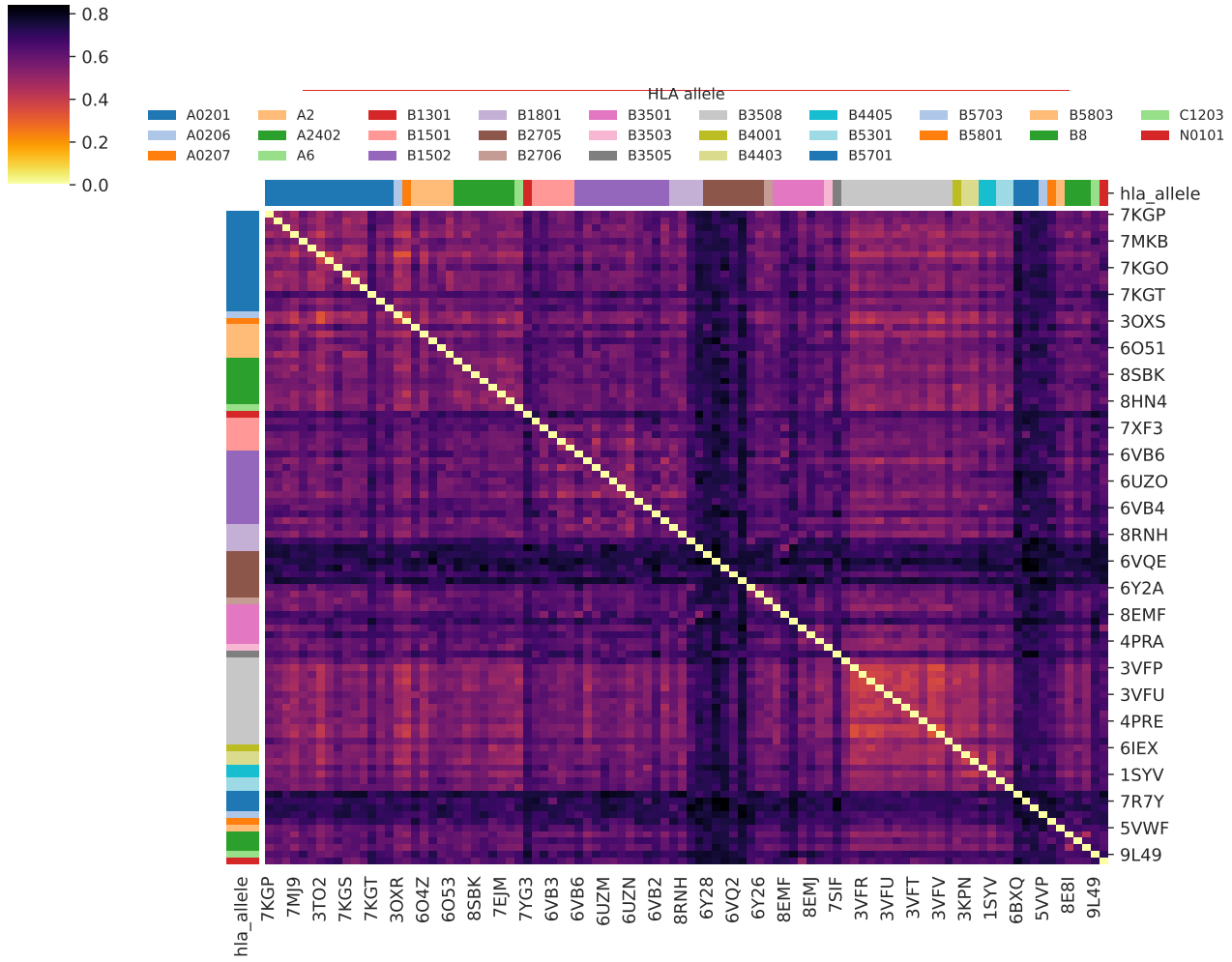

Figure S1. JS divergence matrix between 9-mer PWMs extracted from high-confidence (pLDDT ≥ 0.8) peptides bound in each crystal structure. Axes list PDB IDs; side-bars encode the restricting HLA class I allele.

vf

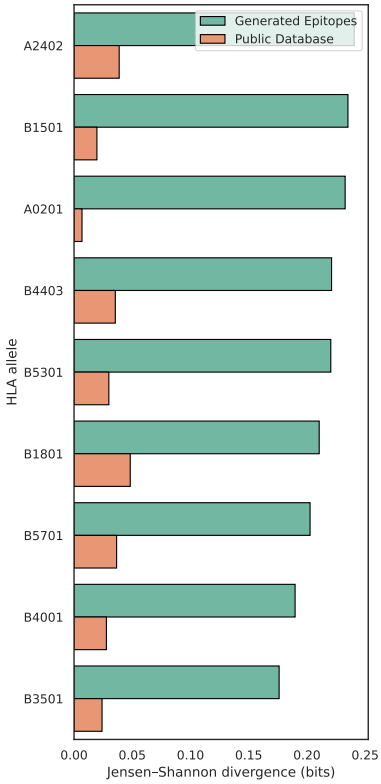

Figure S2. JS Divergence of the Structurally-aware peptides against the Publicly available peptides

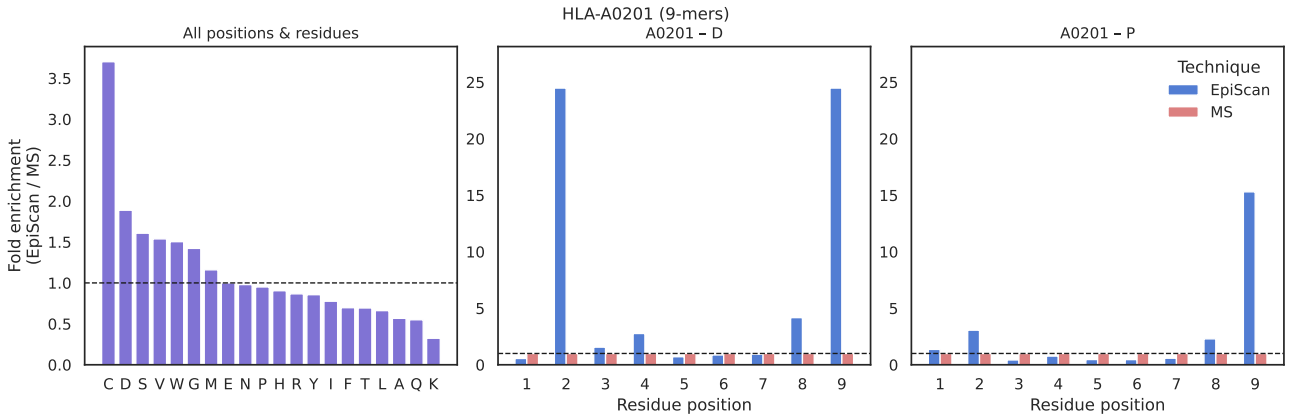

Figure S3. EpiScan results for HLA-A*02:01.

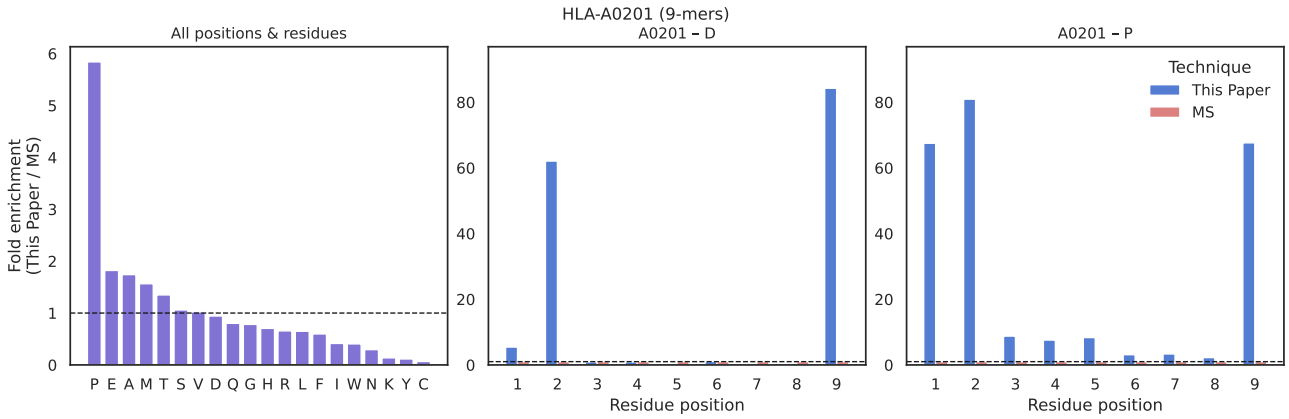

Figure S4. Results from this study for HLA-A*02:01.

Figure S5. Comparison of EpiScan and this study's results for peptide-MHC interactions for HLA-A*02:01.

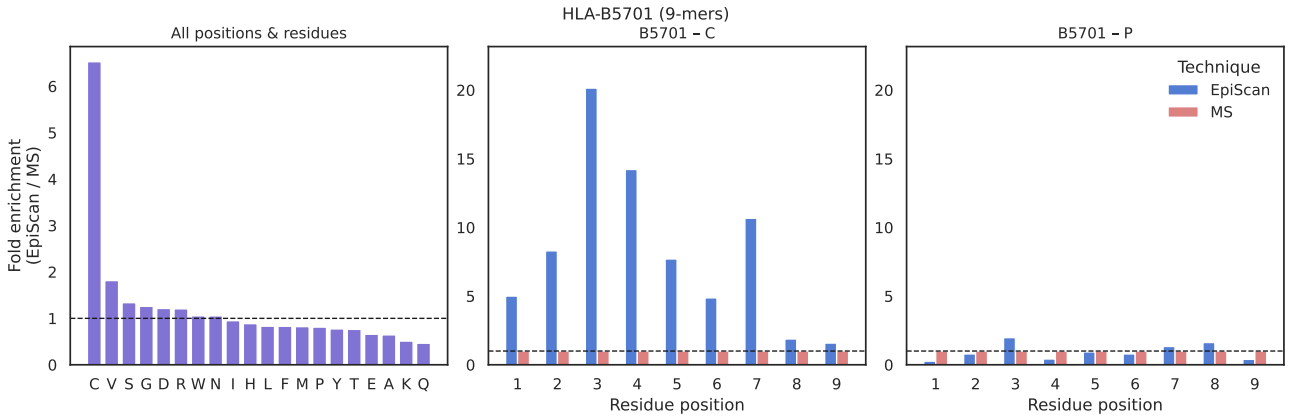

Figure S6. EpiScan results for HLA-B*57:01.

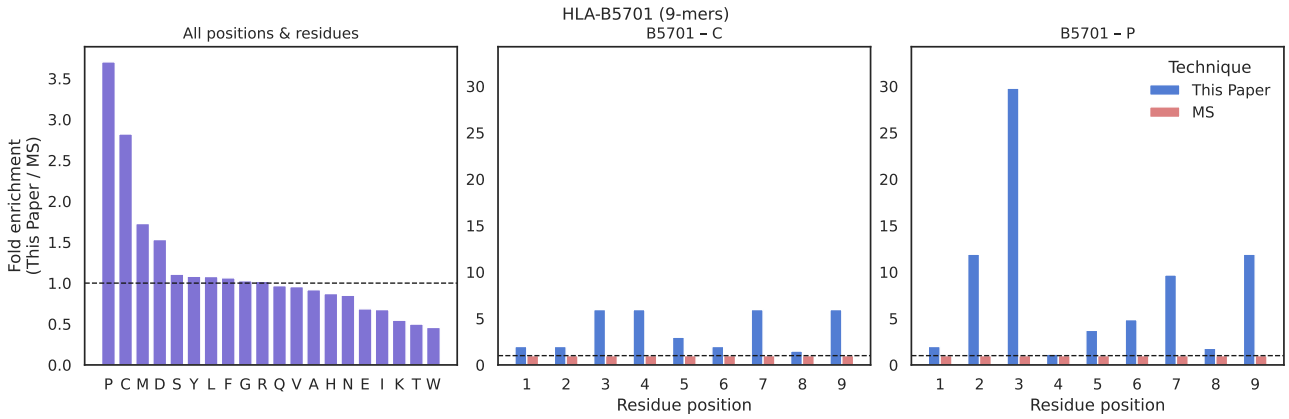

Figure S7. Results from this study for HLA-B*57:01.

Figure S8. Comparison of EpiScan and this study's results for peptide-MHC interactions for HLA-B*57:01.

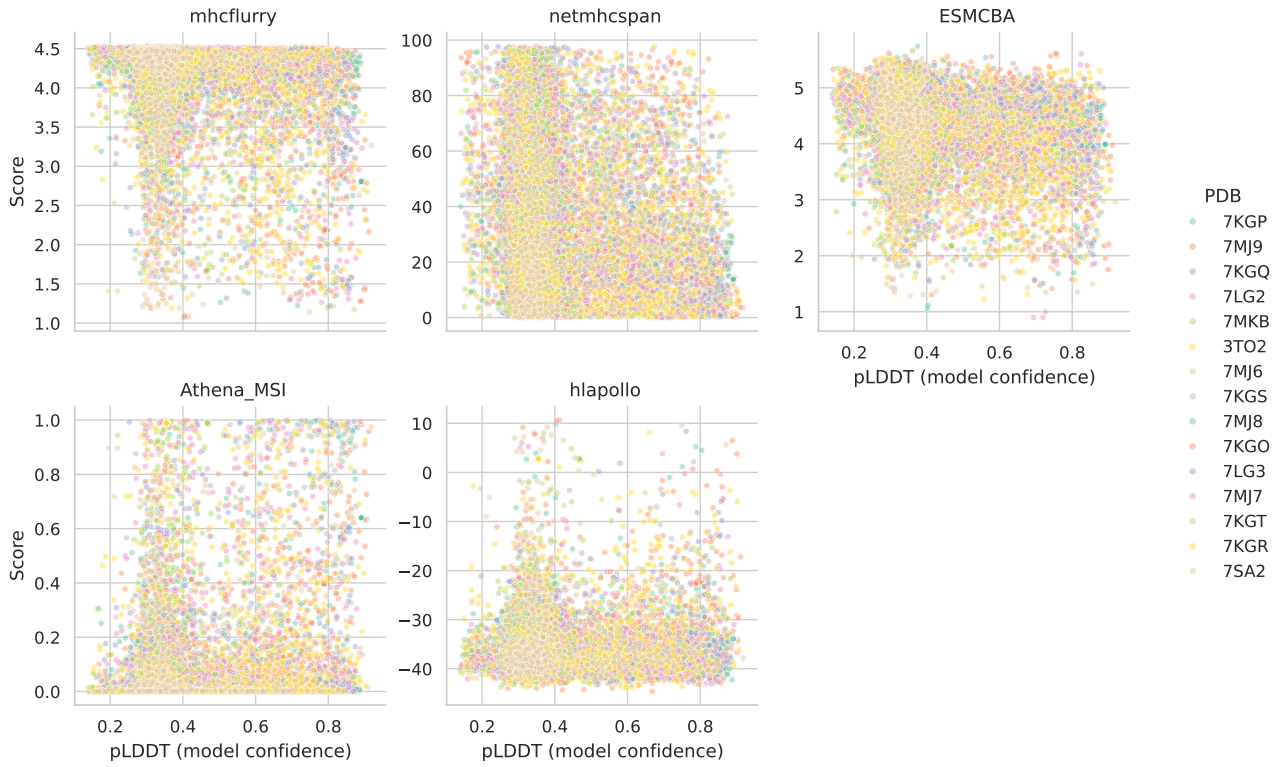

Figure S9. Scatterplots of the raw scores for each of the predictors across the confidence levels of each peptide.

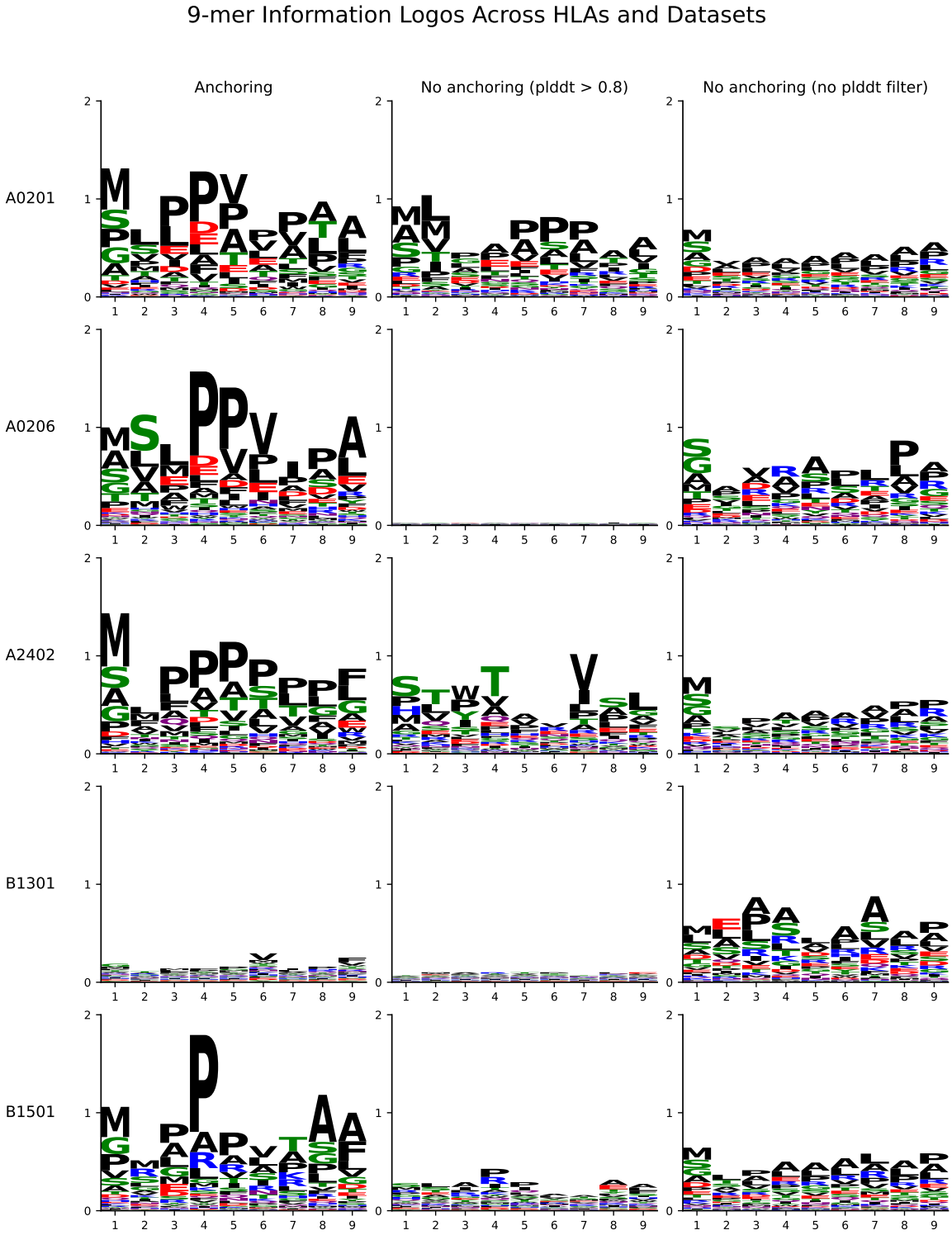

Figure S10. Sequence logo of the amino acids found for the structured-guided libraries

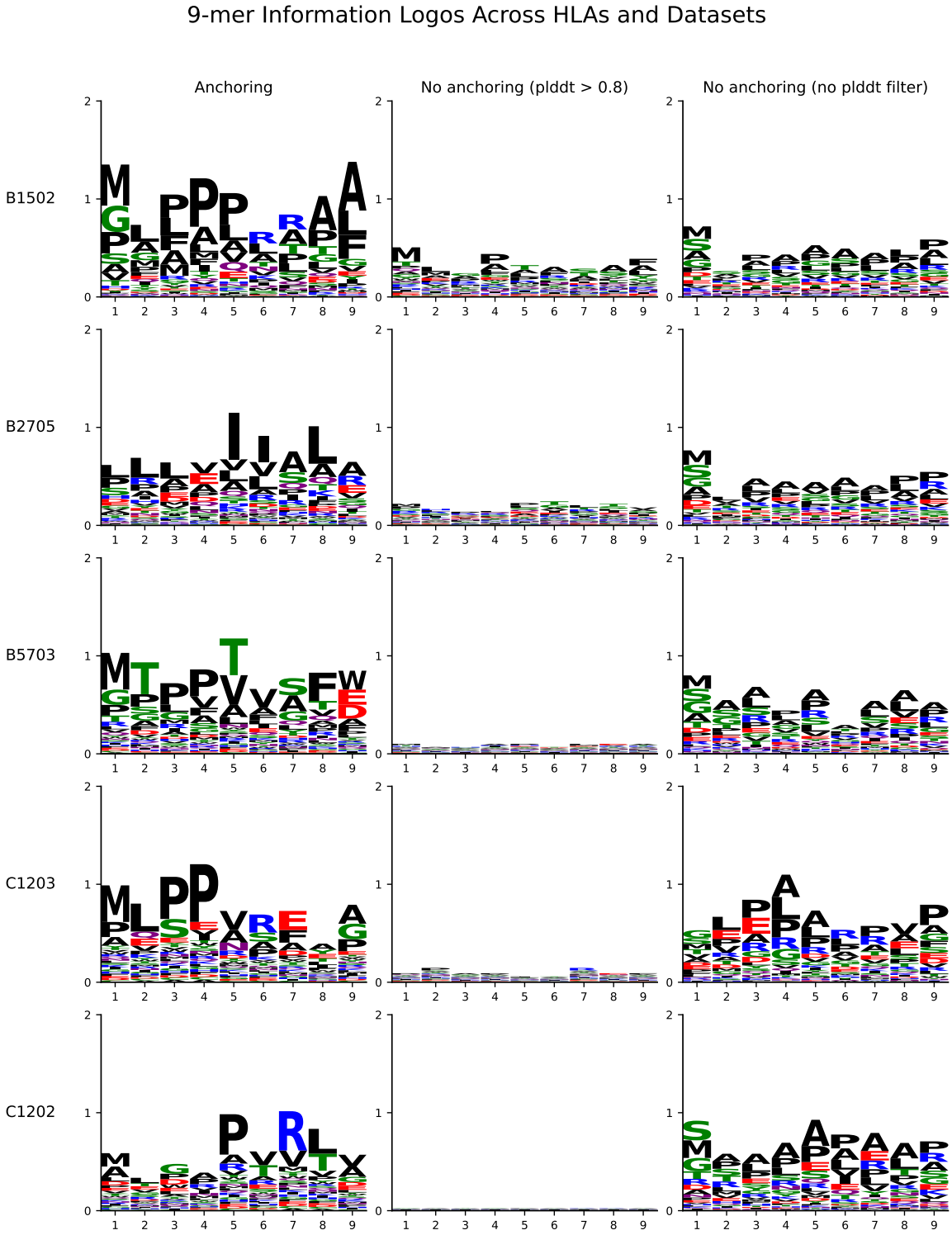

Figure S11. Sequence logo of the amino acids found for the structured-guided libraries.

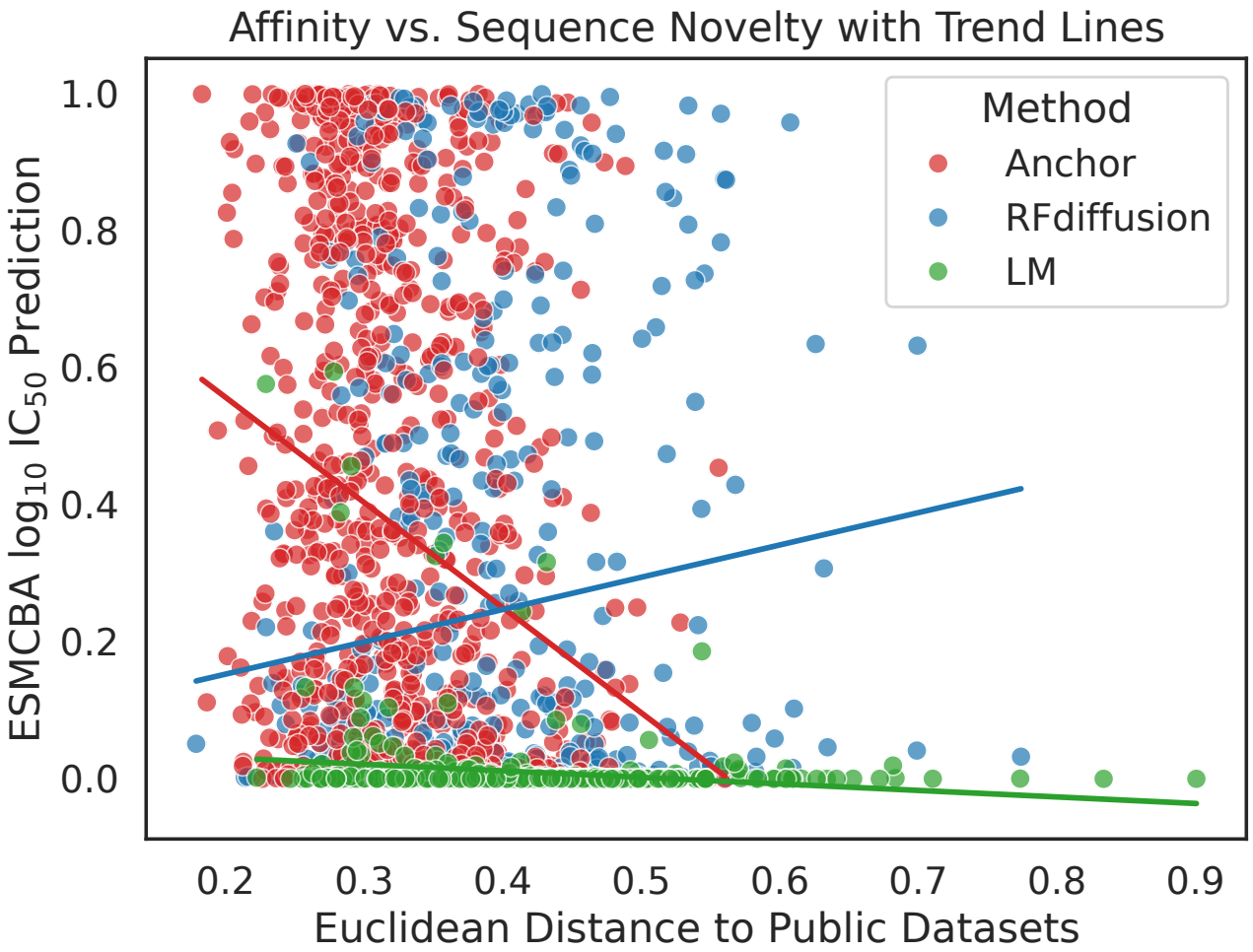

Figure S12. Eucledian Distance to contrast similarity to Public Databases.

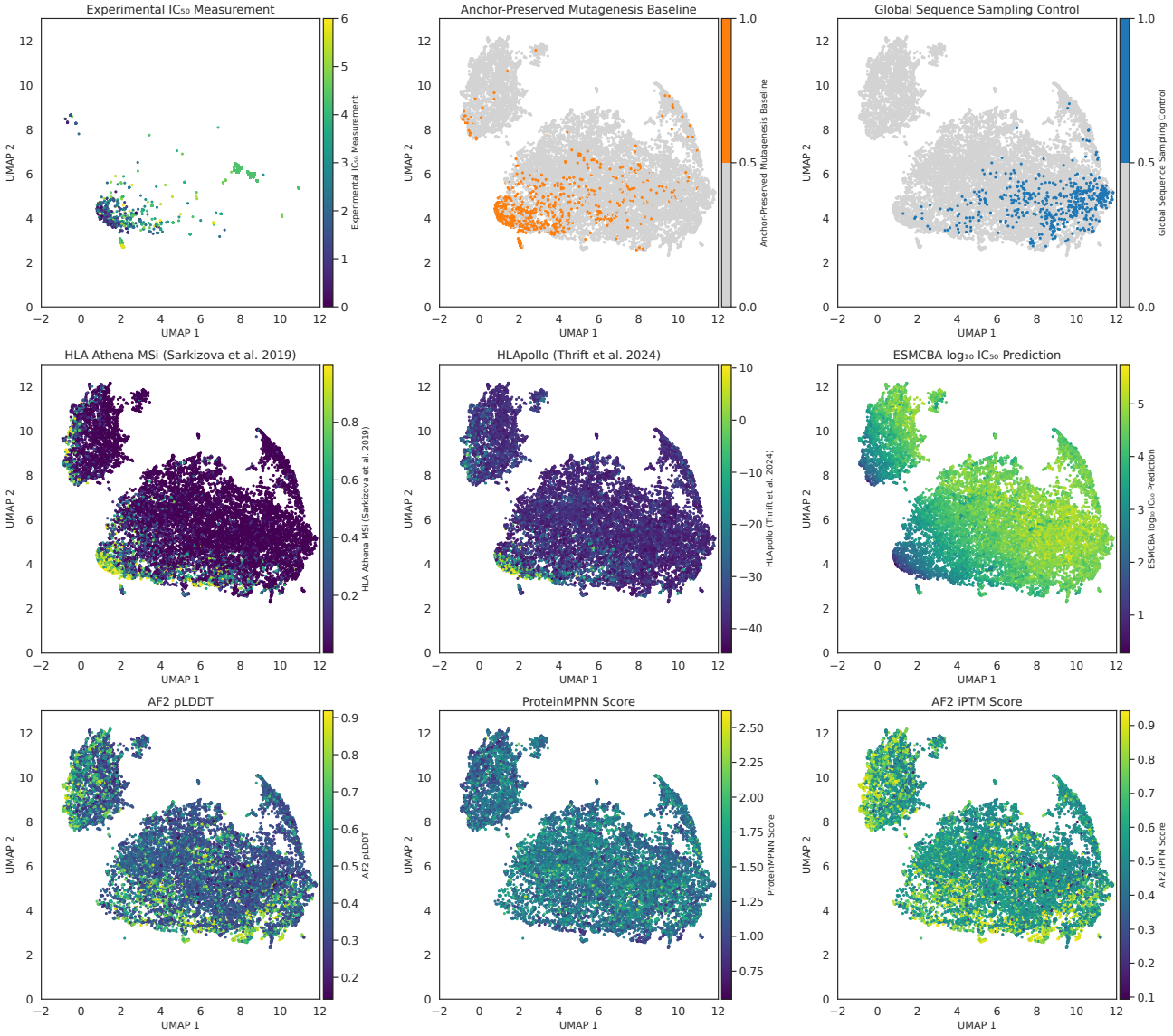

Figure S13. UMAP of the embeddings of ESMCBA for HLA-A*02:01.