ACS OMEGA

Open Access

PepCARES: A Comprehensive Advanced Refinement and Evaluation System for Peptide Design and Affinity Screening

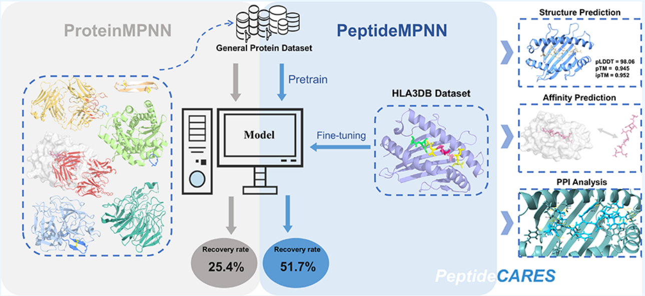
Wen Xu,[†] Zhipeng Wu,[†] Chengyun Zhang, Cheng Zhu, and Hongliang Duan*

Cite This: *ACS Omega* 2024, 9, 46429−46438

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Peptides are crucial in vaccine research, and their remarkable specificity and efficacy make them a promising potential drug class. However, designing and screening these peptides computationally is challenging. Here, we present the comprehensive advanced refinement and evaluation system (PepCARES), a program utilizing our novel model called PeptideMPNN and score evaluation for peptide design and affinity screening. PeptideMPNN, built on ProteinMPNN with transfer learning, significantly enhances sequence recovery (by 26.26%) and reduces perplexity (by 0.536) in a sequence generation task. We designed peptides targeting two HLA alleles and, using MHCfovea and PDBePISA, identified candidates with high potential. From 20 designed peptides, 14 and 7 peptides were selected, respectively. Our research provides a method for designing and screening peptides, making an important step toward the development of peptide-based vaccines.



■ INTRODUCTION

Peptides play a key role in various processes and have been utilized as complementary therapeutic agents to antibodies and small molecules.[1−3] They have the potential to traverse cell membranes to access intracellular targets, thereby targeting disease-related sites that may be inaccessible to antibodies or small molecules. Consequently, peptides represent a promising class of therapeutic agents.

Over the past decade, more than 20 T cell-based vaccines have been proposed for clinical development.[4] Compared to traditional antibody vaccines, these peptide vaccines display better experimental synthesis and immune cell response sensitivity.[5,6] To further accelerate the development of such potential vaccines for various diseases, many computational approaches have been proposed to design/find peptides for enhanced peptide−protein interaction (PPI).[5,7,8] Among these, traditional physics-based methods constitute a notable category. These methods rely on intricate scoring or energy functions and try to find low-energy conformations through energy minimization and mutations for designing plausible peptides. Taking Rosetta as an example, based on the initial structure's conformation and energy, this approach identifies plausible conformations by determining the global minimum of the energy function (representing the lowest energy state),[9−11] thereby producing corresponding peptide sequences. This program can also be applied to systematically investigate the design principles for macrocycle peptides with membrane permeability or oral bioavailability.[12] However, such methods face practical limitations, primarily due to the inherent inaccuracies in the energy functions and the necessity for specialized structural biology expertise.

With the rapid advancement of deep learning, research attention has been focused on deep learning-based methods. These methods can model protein backbone structures at the atomic level and then generate residue sequences that tend to fold to the reference structure. Several methods based on different neural network architectures rapidly emerged: CNN-based methods such as SPROF, DenseCPD, and ProDCoNN; GNN-based methods such as ProteinMPNN and PiFold; and transformer-based methods such as ABACUS-R and ProDesignLE.[13−19]

ProteinMPNN stands as a pioneering deep learning model that designs protein sequences to match the input protein backbones.[20] However, its lack of peptide structure may result in unsatisfactory generality for modeling peptide structures and sequences. In addition, the limited accessibility of experimental data also hinders the ability of this model. To address this issue, we introduced transfer learning as an effective solution. This strategy utilizes existing knowledge to help models better understand related tasks, thereby improving their efficiency and effectiveness.[21,22] In this study, we utilized pHLA crystal

Received: August 20, 2024
Revised: October 27, 2024
Accepted: November 4, 2024
Published: November 11, 2024

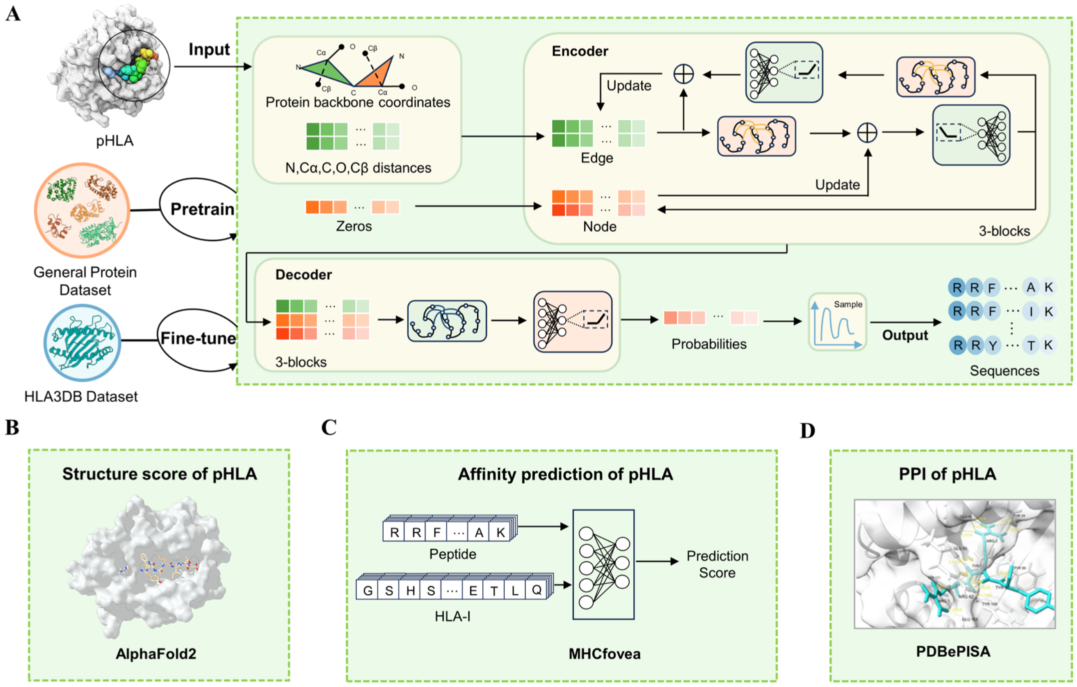


Figure 1. PepCARES framework for designing and screening peptides. (A) PeptideMPNN-designed peptide sequences based on the 3D coordinates of the input backbones. (B) Structure scores of pHLA using AlphaFold2. (C) Affinity prediction of pHLA using MHCfovea. (D) Peptide−protein interactions of pHLA using PDBePISA.

structures from the HLA3DB data set to fine-tune ProteinMPNN (Figure 1A).[23] In theory, transfer learning can serve as a tailored solution to improve the performance of this model in the peptide design task.

Combined with the protein design model ProteinMPNN and transfer learning strategy, we proposed a novel model called PeptideMPNN for peptide design. To demonstrate the feasibility and efficiency of this model, we adopt it to generate peptide vaccines targeting human leukocyte antigen (HLA) molecules (see Supplementary Section S1 for more detailed information). HLAs play a crucial role in the T cell-mediated adaptive immune response and are universally involved in disease immunity across diverse populations.[24] Such molecules are expressed in all nucleated cells, and their central role in a wide range of clinical situations, from infectious diseases to cancer immunotherapy, makes them ideal targets for related drug development (see Supplementary Section S1).[4] The majority of peptides targeting HLA molecules fall within the 8−10 amino acid range, which optimally fits into the binding groove of HLA molecules, thus facilitating stable and effective interactions.[25] Namely, such short peptides are the focus of this study. There are two major classes of HLAs: HLA class I (HLA-I) and HLA class II (HLA-II). Each class is expressed on different types of immune cells and performs specific recognition functions (see Supplementary Section S1 for more detailed information). We selected HLA-I molecules (hereafter referred to as HLA) as research objects due to their critical involvement in presenting peptides to CD8$^+$ T cells, which recognize the peptide-HLA (pHLA) complexes and effectively eliminate infected cells (Supplementary Figure 1).[26,27]

Our model was compared to another model in various aspects of peptide design, including HLA allele types and peptide lengths. Moreover, the peptides designed by PeptideMPNN underwent additional AlphaFold2-based analysis (Figure 1B). Notably, PeptideMPNN not only excelled in the sequence recovery rate but also demonstrated remarkable trends for achieving native-like peptide folding. In addition, we specifically designed peptides targeting two HLA alleles and further explored their binding affinity. We employed the pHLA affinity prediction model, MHCfovea, along with native peptides, to predict the binding probability of the designed peptides (Figure 1C). Consequently, those peptides with high binding probabilities were selected for further PDBePISA-based peptide−protein interaction analysis with their corresponding targets (Figure 1D). In this assessment, we conducted comprehensive evaluations including Gibbs free energy (ΔG) calculations, hydrogen bonding analysis, and surface electrostatic potential analysis to explore the potential of the designed peptides.

In summary, this study presents the comprehensive advanced refinement and evaluation system (PepCARES) for designing peptide sequences and selecting peptides with powerful potential against targets of interest. This versatile framework can be readily applied to peptide design tasks, thereby accelerating the progression of pertinent drug studies.

■ MATERIALS AND METHODS

Data Set. In this study, we used a large protein data set from the Protein Data Bank (PDB) for pretraining.[28] The PDB database provides enriched structural information about proteins, nucleic acids, and other biomolecules investigated by techniques such as X-ray crystallography or cryoelectron microscopy. Dauparas et al. screened the PDB database and used the mmseqs2 clustering tool to group them with a 30% sequence identity cutoff, yielding 25,361 clusters.[16] These clusters were then randomly assigned to the training, validation, and test sets, comprising 23,358, 1464, and 1539 clusters, respectively.

In the fine-tuning phase, we utilized the HLA3DB database constructed by Gupta et al. as our data source.[23] The database

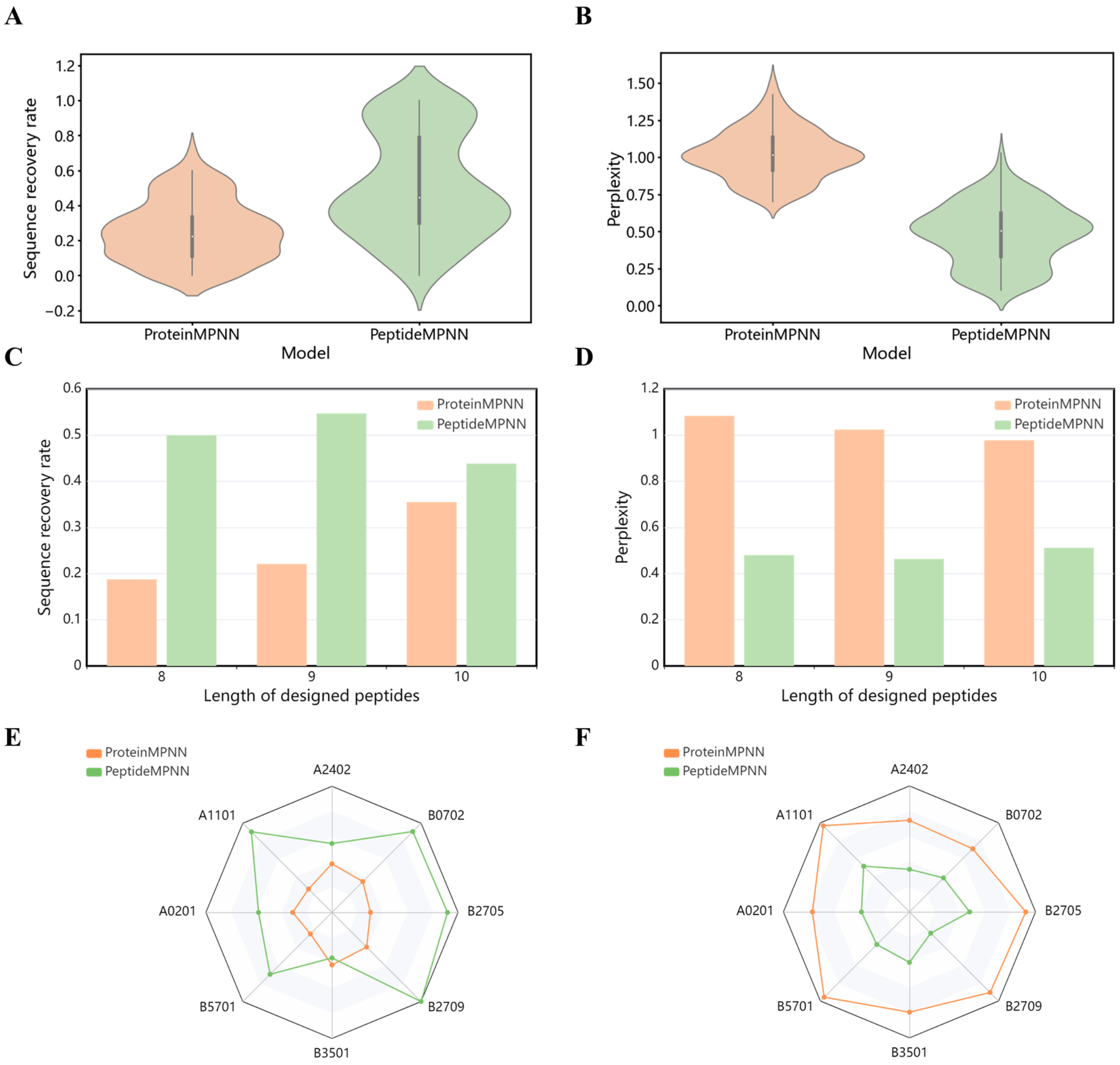


Figure 2. Comparison between PeptideMPNN (green) and ProteinMPNN (orange) in the peptide design task. (A,B) Sequence recovery rate and perplexity comparisons between ProteinMPNN and PeptideMPNN. (C,D) Comparison between ProteinMPNN and PeptideMPNN in designing peptides with different lengths. (E,F) Comparison between ProteinMPNN and PeptideMPNN in designing peptides against different HLA allele proteins.

contained 556 complexes with high structural resolution under 3.0 Å from PDB. All of them are peptide-HLA complexes with peptide lengths ranging from 8 to 10. We divided each genotype into training, validation, and test sets with a 8:1:1 ratio. To minimize bias from insufficient data, entries with genotypes fewer than 10 were used only for training or validation and were excluded from the test set. Ultimately, the data set was divided into 438, 67, and 51 entries for training, validation, and testing, respectively.

PeptideMPNN. In this study, the popular protein design model ProteinMPNN was used as the baseline model. ProteinMPNN is based on the structured transformer model proposed by Ingraham et al.,[29] using the structural features characterized by N, C_α, C, O, and virtual C_β atoms of the protein skeleton. In addition, order-agnostic decoding was implemented in this model for the fixed-target ligand generation task. Such decoding skips the fixed regions but includes them in the sequence context for the remaining positions, which can effectively infer structures with unknown regions.

We combined transfer learning with the ProteinMPNN model and proposed a model called PeptideMPNN. During training, we fine-tuned the ProteinMPNN model based on the HLA3DB database, and all training samples were processed in each epoch. We set the initial learning rate to 5×10^{-4} and used the Adam optimizer to adjust the learning rate. To ensure effective model adaptation during training, the learning rate would automatically decrease by a factor of 10 if the validation loss does not improve within 10 consecutive epochs. After 400 epochs, the training process stopped.

AlphaFold2. In this study, we used the AlphaFold2 model to filter pHLA complex structures for subsequent tasks. This structure prediction model was proposed by Google's DeepMind team in 2021, making a significant impact on biomedical research and drug development in recent years.[30] In our work, we implemented a locally installed version of AlphaFold2. This model was optimized to support the parallel execution of multiple structure prediction tasks on multiple GPUs. Based on the input sequences, AlphaFold2 can predict corresponding structures and provide confidence scores such as the predicted local distance difference test (pLDDT), the predicted template modeling (pTM), and the interface predicted template modeling (ipTM).

MHCfovea. To predict the binding affinity of peptides targeting HLA molecules, we implemented the MHCfovea model. MHCfovea, an ensemble of convolutional neural network (CNN) models, can predict the binding probability of peptides targeting the HLA allele, making it a suitable tool for assessing the quality of our designed peptides.[31] In this study, we used PeptideMPNN to design 20 peptides against the HLA-A*02:01 and the HLA-B*27:05 targets, respectively, and then applied MHCfovea to assess them. By comparing them to the native peptides, we selected designed peptides with higher binding potentials.

RESULTS AND DISCUSSION

Performance Comparison between PeptideMPNN and ProteinMPNN. In this study, we developed the peptide design model PeptideMPNN by fine-tuning the protein design model ProteinMPNN on the HLA3DB structural data set (detailed information can be found in the Materials and Methods). Based on the pHLA complex backbones, PeptideMPNN can generate peptide sequences that potentially align with the provided backbones. To conduct a thorough assessment of PeptideMPNN's performance, we utilized ProteinMPNN as our baseline model for comparison.

We evaluated the performance of the models using two key metrics: the sequence recovery rate and perplexity. The sequence recovery rate assesses the model's ability to reconstruct natural sequences, where a high recovery rate signifies that the model has learned the structure-induced sequence constraints. On the other hand, perplexity quantifies the certainty surrounding the native amino acid residues, with a lower perplexity score indicating a more concentrated probability distribution.[32]

In the general test set, PeptideMPNN demonstrated an impressive sequence recovery rate of 51.70%, significantly surpassing the rate achieved by ProteinMPNN, which was 25.44% (Figure 2A). This result underscores PeptideMPNN's efficacy in capturing the sequence characteristics of pHLA complexes. Recognizing that a high sequence recovery rate alone may not comprehensively demonstrate performance differences, we delved deeper by comparing the two models using perplexity. PeptideMPNN exhibited superior performance with a perplexity score of 0.486, markedly lower than ProteinMPNN's score of 1.022 (Figure 2B). This further highlights the superiority of our PeptideMPNN.

In addition, we evaluated the ability of the model to design peptides with different lengths. It is noteworthy that the anchor residues of HLA molecules exhibit positional preferences that are influenced by the peptide length, thereby affecting the interaction within peptide-HLA complexes.[25] For peptides with lengths of 8, 9, and 10 amino acids, PeptideMPNN achieved sequence recovery rates of 50.00, 54.71, and 43.83%, respectively. In contrast, ProteinMPNN's recovery rates for these same peptide lengths were significantly lower, at 18.75, 22.10, and 35.50% (Figure 2C). These findings further demonstrate PeptideMPNN's superiority in designing peptides of different lengths compared to ProteinMPNN.

In addition, we noted that PeptideMPNN performed best in designing 9-amino acid peptides. It may be attributed to the fact that such peptides can optimally fit within the HLA binding groove due to their small lengths.[33] This length allows the peptides to interact with key anchor residues at both the N- and C-termini, thereby stabilizing the peptide-HLA complexes. Furthermore, the perplexities of PeptideMPNN in designing peptides with different lengths were in the range from 0.463 to 0.512, significantly lower than those of ProteinMPNN (Figure 2D).

Moreover, we explored the performance of models in designing peptides against different HLA alleles. HLA polymorphism and allele-specific sequence motifs play a crucial role in peptide binding. HLA genes exhibit a wide variety of alleles at different loci, requiring that the amino acids of polypeptide anchor residues are specifically complementary to those in the binding grooves of specific HLA alleles (Supplementary Figure 2).[34] Due to limited experimental data, we only focused on the analysis of representative eight targets of HLA-A and HLA-B alleles. When targeting three common HLA-A alleles, PeptideMPNN demonstrated sequence recovery rates of 46.58, 72.33, and 43.71%, respectively, which were significantly higher than ProteinMPNN's rates of 24.93, 21.00, and 30.85% (Figure 2E). These results underscore the superior affinity of peptides designed by PeptideMPNN toward these HLA-A alleles. PeptideMPNN also performed better than ProteinMPNN when against HLA-B alleles, although the recovery rates of PeptideMPNN were lower for the HLA-B*35:01 allele. This could be attributed to the fewer associated samples available in the test set. Additionally, we compared the perplexities of both models for this task (Figure 2F). PeptideMPNN consistently outperformed ProteinMPNN against all eight representative HLA groups, further validating the effectiveness of PeptideMPNN in terms of diversity and specificity.

In summary, PeptideMPNN exhibited superior performance compared to ProteinMPNN in multiple tasks, validating our fine-tuning strategy's ability to not only accurately recover specific residues within particular structures but also effectively capture the intricate mapping relationship between the tertiary structure and the primary sequence. This approach significantly enhanced sequence recovery rates and minimized perplexity, thereby demonstrating PeptideMPNN's extensive applicability and efficacy in the realm of peptide design.

Structure Assessment for Designed Peptides by Using AlphaFold2. After confirming that PeptideMPNN could design peptides with high accuracy, we further explored whether the designed sequences could accurately fold into the desired structures and form stable complexes with targets. We used AlphaFold2 to predict the structures of the designed sequences. The predicted structures were then evaluated by three confidence metrics: pLDDT, pTM, and ipTM. These metrics were universally adopted in related work. Especially in the works of the Baker group, these metrics proved helpful in finding binders with high affinity.[35,36] However, predicting biological activity remains a formidable challenge for current methods. Although AlphaFold2-based validation offers researchers valuable insight for assessing the biological activity of designed peptides, the correlations between these structural metrics and biological activity are unclear. However, until now, these metrics are still general and effective ways to assess the designed peptides and proteins.

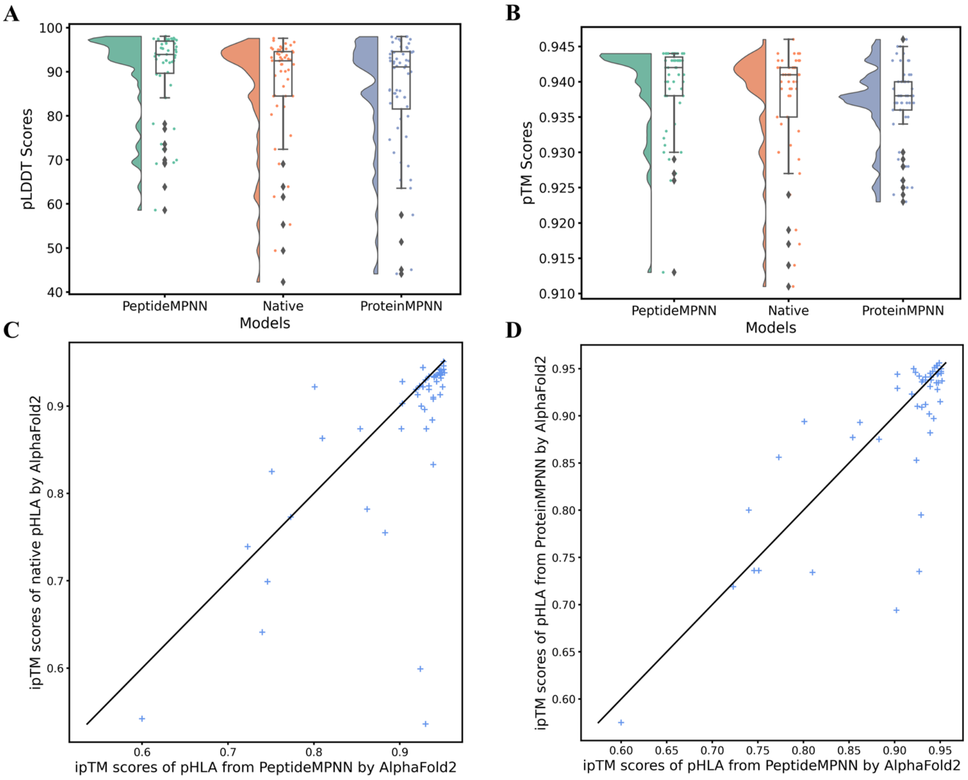


Figure 3. Structure confidence evaluation by AlphaFold2 on native and designed pHLA complexes by PeptideMPNN and ProteinMPNN. (A) pLDDT score comparison between native and designed peptides. (B) pTM score comparison between native and designed pHLA complexes. (C) ipTM score comparison between native and designed pHLA complexes by PeptideMPNN. (D) ipTM score comparison between designed pHLA complexes by PeptideMPNN and ProteinMPNN.

Table 1. Scores Provided by AlphaFold2 and MHCfovea for the Peptides Generated against the HLA-A*02:01

pHLA	pLDDT	pTM	ipTM	RMSD_Cα (Å)	affinity prediction score
HLA-A*02:01-P1	97.38	0.944	0.950	0.474	0.981
HLA-A*02:01-P2	96.76	0.944	0.948	0.138	0.980
HLA-A*02:01-P3	97.59	0.944	0.952	0.271	0.955
HLA-A*02:01-P4	94.81	0.945	0.948	0.293	0.984
HLA-A*02:01-P5	98.06	0.945	0.952	1.191	0.924
HLA-A*02:01-P6	95.99	0.944	0.948	0.472	0.981
HLA-A*02:01-P7	98.05	0.943	0.946	0.693	0.994
HLA-A*02:01-P8	97.02	0.941	0.930	0.450	0.986
HLA-A*02:01-P9	93.89	0.946	0.953	0.306	0.991
HLA-A*02:01-P10	97.70	0.944	0.950	0.307	0.948
HLA-A*02:01-P11	97.10	0.945	0.952	0.447	0.986
HLA-A*02:01-P12	97.26	0.943	0.944	0.420	0.915
HLA-A*02:01-P13	97.81	0.944	0.952	0.482	0.981
HLA-A*02:01-P14	97.38	0.942	0.935	0.077	0.976
HLA-A*02:01-P15	95.00	0.944	0.951	0.323	0.165
HLA-A*02:01-P16	97.26	0.943	0.945	0.176	0.817
HLA-A*02:01-P17	96.36	0.944	0.943	0.435	0.893
HLA-A*02:01-P18	97.49	0.945	0.952	0.304	0.949
HLA-A*02:01-P19	97.26	0.944	0.943	0.353	0.926
HLA-A*02:01-P20	97.47	0.944	0.952	0.283	0.954
7UR1					0.927

The peptides designed by PeptideMPNN gained notably high-confidence AlphaFold structural scores (Figure 3A). Specifically, the pLDDT scores of these peptides surpassed those of the peptides designed by ProteinMPNN and the native peptides. This suggests that PeptideMPNN possesses the capability to design peptides that fold into the desired 3D structures with greater accuracy and, in certain instances, even form more stable structural scaffolds than the native sequences. Regarding the accuracy of the protein complex structure, the pTM score serves as a reliable metric.[37] Since the HLA targets have been fixed in both the predicted and native sequences, all sequence-to-structure mappings generally achieved high pTM

Table 2. Scores Provided by AlphaFold2 and MHCfovea for the Peptides Generated against the HLA-B*27:05

pHLA	pLDDT	pTM	iPTM	RMSD_Cα (Å)	affinity prediction score
HLA-B*27:05-P1	93.77	0.946	0.931	0.32	0.993
HLA-B*27:05-P2	93.74	0.945	0.928	0.875	0.985
HLA-B*27:05-P3	95.38	0.947	0.939	0.213	0.990
HLA-B*27:05-P4	93.77	0.946	0.931	0.324	0.993
HLA-B*27:05-P5	94.71	0.946	0.933	0.432	0.998
HLA-B*27:05-P6	94.71	0.946	0.933	0.432	0.998
HLA-B*27:05-P7	94.82	0.946	0.936	1.308	0.996
HLA-B*27:05-P8	93.04	0.945	0.923	0.302	0.998
HLA-B*27:05-P9	91.52	0.944	0.919	0.327	0.995
HLA-B*27:05-P10	94.93	0.944	0.935	1.173	0.995
HLA-B*27:05-P11	90.66	0.943	0.913	1.243	0.995
HLA-B*27:05-P12	95.38	0.946	0.937	0.198	0.999
HLA-B*27:05-P13	89.85	0.943	0.905	1.258	0.999
HLA-B*27:05-P14	96.76	0.946	0.947	0.277	0.997
HLA-B*27:05-P15	93.98	0.946	0.932	0.246	0.990
HLA-B*27:05-P16	91.99	0.945	0.920	0.267	0.998
HLA-B*27:05-P17	95.90	0.946	0.940	0.974	0.993
HLA-B*27:05-P18	92.60	0.943	0.926	0.882	0.997
HLA-B*27:05-P19	88.74	0.943	0.905	0.306	0.994
HLA-B*27:05-P20	92.07	0.944	0.922	0.325	0.997
1JGE					0.993

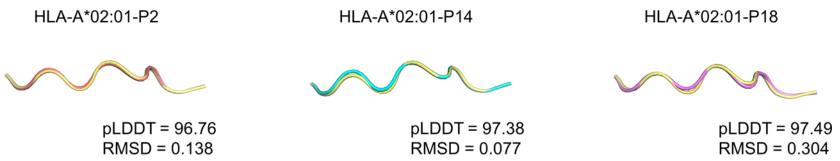


Figure 4. Confidence metric comparisons between the native peptides (yellow) and designed peptides with high-confidence metrics against the HLA-A*02:01 allele.

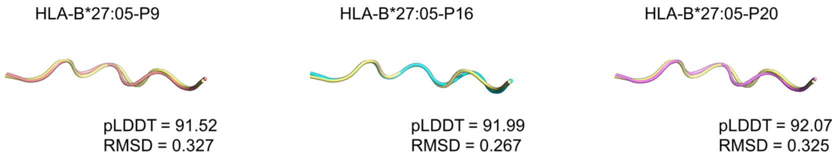


Figure 5. Confidence metric comparisons between the native (yellow) and designed peptides with high-confidence metrics against the HLA-B*27:05 allele.

scores. However, peptides designed by PeptideMPNN attained the highest pTM scores when compared to those designed by ProteinMPNN and the native peptide sequences (Figure 3B). These findings further validate PeptideMPNN's superiority in designing peptides with stable structural configurations.

In addition, we also compared their ipTM scores (more detailed information can be found in Supplementary Table 1). This metric can reflect the quality of the interaction interface of the predicted protein complexes and the binding potential of designed peptides against protein targets.[37] Notably, O'Reilly et al. confirmed that higher ipTM scores indicated a higher likelihood of target−ligand interactions, too.[38] Furthermore, complexes with ipTM scores exceeding 0.85 were demonstrated to exhibit reliable binding. The majority of the peptides designed in our study achieved higher ipTM scores, with PeptideMPNN surpassing ProteinMPNN in performance (Figure 3C,D), suggesting that the sequences optimized by PeptideMPNN are more likely to form effective bindings with HLA targets. Overall, the AlphaFold-based assessment demonstrates that PeptideMPNN is a suitable tool for designing peptides targeting HLA molecules.

Binding Potential of Designed Binders against the HLA-A*02:01 and the HLA-B*27:05 Targets. As mentioned in the article above, the specificity of peptides binding to HLA molecules is influenced by allosteric constraints. To explore whether PeptideMPNN could design peptides with enhanced binding capabilities against different HLA alleles, we applied our design strategy to two specific targets: the widely prevalent HLA-A*02:01 allele and the specific genotype-encoded HLA-B*27:05 allele. The HLA-A*02:01 allele is the most common allele across multiple ethnicities worldwide and is associated with multiple immune response mechanisms.[39,40] The HLA-B*27:05 allele, a not common subtype of HLA-B27, is predominantly found in Caucasians and is strongly associated with ankylosing spondylitis (AS).[41] Validation that involved these two targets provided an effective assessment of our design strategy.

PeptideMPNN designed 20 peptide sequences for each target using known pHLA complex structures (7U1R and 1JGE) as templates.[42,43] To validate the structural feasibility of these sequences, the designed pHLA complexes were input into AlphaFold2 for structure prediction. AlphaFold2 provided high-confidence structural predictions for peptides designed by PeptideMPNN against the HLA-A*02:01 and the HLA-B*27:05 alleles. Specifically, all pTM and ipTM scores exceeded 0.9, and most pLDDT scores also exceeded 90 (Tables 1 and 2), indicating that the designed peptides exhibited minimal deviation in spatial structure folding and binding position compared to the native peptides. In addition, RMSDs of the designed peptides targeting the HLA-A*02:01 allele ranged from 0.077 to 1.191, while those targeting the HLA-B*27:05 allele ranged from 0.198 to 1.308. Figures 4 and 5 show the structural alignments between several designed peptides and native peptides. There were great structural similarities between the designed peptides and the native peptides. Namely, designed peptides were likely to interact effectively with the targets.

We employed the deep learning-based framework MHCfovea to predict the binding potential of designed peptides.[31] MHCfovea is a collection of multiple CNN models that use HLA allele sequences and peptide sequences as inputs to predict binding probabilities (detailed information can be found in the Materials and Methods). The majority of the designed peptides exhibited a high probability of binding, with more than half of the designed peptides having higher predicted binding probabilities than the native peptides (Tables 1 and 2). To explore the interactions between the designed peptides and their target proteins, we selected those with higher affinity prediction scores than the native peptides. We then investigated interactions within these selected complexes by using the PPI analysis tool PDBePISA.[44] Against the HLA-A*02:01 target, all 14 designed peptides exhibited ΔG compared to the native peptides, and their interaction surface areas were similar (Table 3). Against the HLA-B*27:05 target, all the designed peptides had larger interaction surface areas, and seven designed peptides had lower ΔG than the natives (Table 4). The numbers of hydrogen bonds and salt bridges also increased, which confirms that PeptideMPNN was capable of designing peptides with enhanced binding potential against various HLA alleles. Although we have made interaction observations based on several tools, experimental validation is still the most effective and reliable means of verification. We strongly recommend conducting necessary in-laboratory experiments for further verification, when resources permit.

Table 3. Interaction Analysis for Designed Peptides That Exhibited Affinity Prediction Scores Higher than Those of the Native Peptides when Targeted against the HLA-A*02:01 Allele

pHLA	interface area (Å^2)	ΔG (kcal/mol)	# HB	# SB
HLA-A*02:01-P1	899.8	−10.5	10	4
HLA-A*02:01-P2	884.6	−10.4	12	3
HLA-A*02:01-P3	881.1	−10.4	11	1
HLA-A*02:01-P4	911.1	−9.8	10	2
HLA-A*02:01-P6	902.5	−10.6	14	4
HLA-A*02:01-P7	852.2	−9.2	12	4
HLA-A*02:01-P8	936.5	−11.3	12	5
HLA-A*02:01-P9	894.5	−9.3	11	4
HLA-A*02:01-P10	896.3	−10.6	8	2
HLA-A*02:01-P11	936.5	−11.3	12	5
HLA-A*02:01-P13	900.3	−10.5	11	5
HLA-A*02:01-P14	927.2	−9.6	11	3
HLA-A*02:01-P18	912.1	−12.3	11	5
HLA-A*02:01-P20	894.1	−11.1	10	3
7UR1	916.3	−6.7	18	6

Table 4. Interaction Analysis for Designed Peptides That Exhibited Affinity Prediction Scores Higher than Those of the Native Peptides when Targeted against the HLA-B*27:05 Allele

pHLA	interface area (Å^2)	ΔG (kcal/mol)	# HB	# SB
HLA-B*27:05-P5	960.2	−3.0	17	11
HLA-B*27:05-P6	961.2	−3.0	16	10
HLA-B*27:05-P7	984.5	−6.1	13	10
HLA-B*27:05-P8	1055.5	−8.0	14	10
HLA-B*27:05-P9	1019.0	−7.1	14	11
HLA-B*27:05-P10	980.8	−5.1	17	11
HLA-B*27:05-P11	1000.6	−4.0	17	11
HLA-B*27:05-P12	953.1	−5.4	15	10
HLA-B*27:05-P13	975.4	−4.2	16	11
HLA-B*27:05-P14	1078.4	−6.2	18	11
HLA-B*27:05-P16	969.1	−7.9	14	12
HLA-B*27:05-P18	937.5	−3.2	17	10
HLA-B*27:05-P19	931.3	−4.4	17	15
HLA-B*27:05-P20	1012.7	−8.4	13	8
1JGE	815.5	−5.2	17	9

To explore the potential of the designed peptides, we conducted a visual analysis of the PPI for the 14 candidate peptides against the HLA-A*02:01 target and seven candidate peptides against the HLA-B*27:05 target. We presented the top three designed peptides with the lowest ΔG here, and the other designed peptides are shown in Supplementary Figures 3 and 4. These candidate peptides were compared to the native peptides to assess their binding potential by calculating hydrogen bonds at the active sites. Seven hydrogen bonds were formed between the native peptide and the HLA-A*02:01 target. Through sequence optimization by PeptideMPNN, the designed candidate peptides not only retained these hydrogen bonds but also established additional residue connections (Figure 6A). This enhancement in binding affinity likely accounted for the reduced ΔG values observed for the candidate peptides, in comparison to the native peptide. Furthermore, the shorter hydrogen bond distances between the candidate peptides and their target receptor underscored their stronger intermolecular interactions. For the HLA-B*27:05 target, the designed peptides also formed more and shorter hydrogen bonds compared to the native peptide (Figure 6B). Additionally, we investigated the surface electrostatic potential of the peptides. Differences in the surface electrostatic potential can lead to variations in electrostatic interactions, thus affecting ligand−target binding.[45] The analysis indicates a high consistency of the electrostatic potential between the candidate peptides and the native peptides (Figure 7). The surface electrostatic potential maps of the other candidate peptides are shown in Supplementary Figures 5 and 6. Collectively, these analyses demonstrate that 14 candidate peptides against the HLA-A*02:01 target and seven candidate peptides against the HLA-B*27:05 target had strong binding potential, validating the effectiveness of our design strategy.

■ CONCLUSIONS

In this study, we combined the ProteinMPNN architecture and transfer learning strategy to build a model called Pepti-

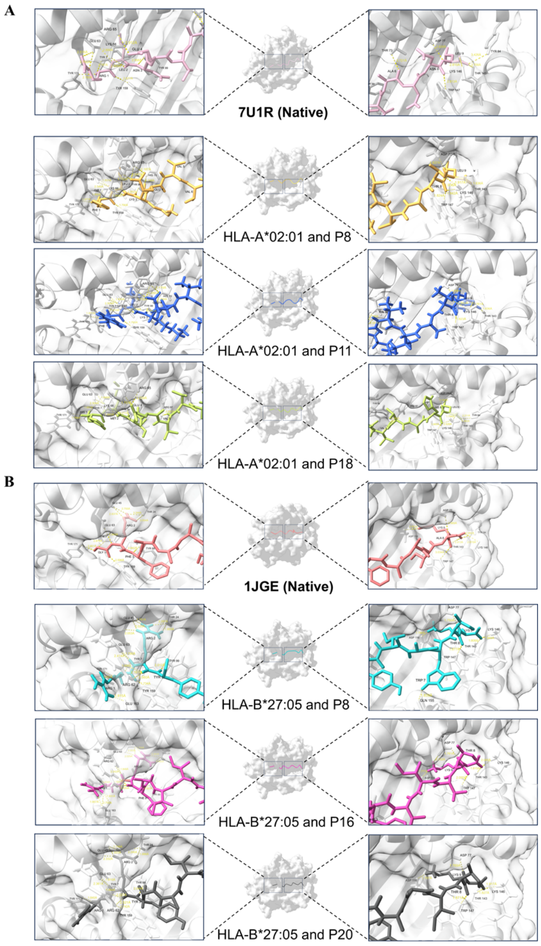


Figure 6. Interaction visualization analysis of candidate peptides. (A) Three candidate peptides against the HLA-A*02:01 allele with the lowest Δ*G* and the native peptide. (B) Three candidate peptides against the HLA-B*27:05 allele with the lowest Δ*G* and the native peptide.

deMPNN for peptide sequence design. This model showed a significant improvement in the sequence recovery rate of 26.26% and a reduction in the perplexity of 0.536 compared to the baseline model. It also performed well in peptide design involving different peptide lengths and HLA alleles. However, while PeptideMPNN has demonstrated superior performance in designing short peptides (8−10 residues) compared to ProteinMPNN, its scalability to longer peptides requires further exploration. Moreover, we used AlphaFold2 to evaluate the structure confidence of the designed peptides, which further demonstrated PeptideMPNN's superiority in peptide design.

Furthermore, we used PeptideMPNN to design peptides against the representative allele HLA-A*02:01 and the specific allele HLA-B*27:05. Through affinity prediction and interaction interface evaluation, 14 peptides against the HLA-A*02:01 allele and seven peptides against the HLA-B*27:05 allele were selected, respectively. These candidates not only outperformed the native peptides in binding probability but also exhibited lower Gibbs free energy.

Additionally, we may adapt PeptideMPNN to accommodate unnatural amino acids and peptide-like molecules in future research, considering their importance in related drug development. Moreover, we plan to implement the powerful model AlphaFold3 in our study for improved accuracy and rationalization in the future.

In conclusion, the PepCARES framework enables the rapid and efficient peptide design and can effectively identify peptides with higher affinity. This approach can accelerate the development of peptide vaccines and provide a new opportunity to improve the targeting of immune responses through peptide design. Moreover, the generalized design principles established by our model based on HLA epitopes may be transferred to other targets within the immune system.

■ ASSOCIATED CONTENT

Data Availability Statement

The data set of general protein structures used for pretraining in this study is available at https://github.com/dauparas/ProteinMPNN. The data set of pHLA complex structures used for fine-tuning in this study is available at https://hla3db.

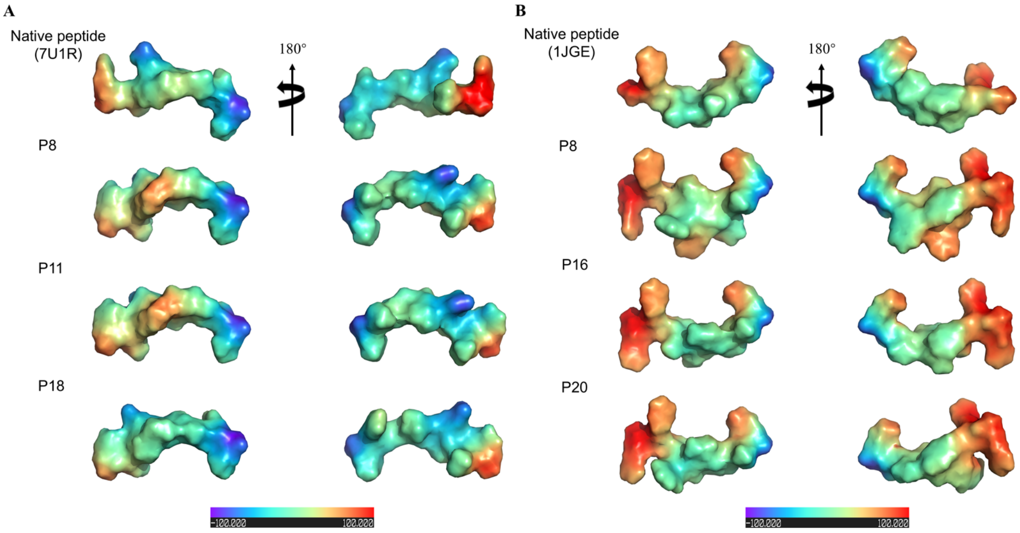


Figure 7. Surface electrostatic potential maps of candidate peptides. (A) Three candidate peptides with the lowest Δ*G* and the native peptide against the HLA-A*02:01 target. (B) Three candidate peptides with the lowest Δ*G* and the native peptide against the HLA-B*27:05 target.

research.chop.edu. The codes of PeptideMPNN in this study are available at https://github.com/xw09/PeptideMPNN.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c07682.

> Introduction about the HLA-I molecules including the mechanism and nomenclature of HLA; detailed information about the parameters of the PeptideMPNN model; ipTM score comparison between native and designed pHLA complexes by PeptideMPNN and ProteinMPNN; visualization analysis of candidate peptides targeting the HLA-A*02:01 and the HLA-B*27:05 allele with lower ΔG than the native one (DOCX)

AUTHOR INFORMATION

Corresponding Author

Hongliang Duan − *Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China;* orcid.org/0000-0002-9194-0115; Email: hduan@mpu.edu.mo

Authors

Wen Xu − *College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China;* orcid.org/0009-0007-2656-0988

Zhipeng Wu − *College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China;* orcid.org/0000-0003-3535-1081

Chengyun Zhang − *AI Department, Shanghai Highslab Therapeutics, Inc., Shanghai 201203, China*

Cheng Zhu − *College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c07682

Author Contributions

[†]W. Xu and Z. Wu contributed equally. Z. Wu and W. Xu conceived the basic idea. Z. Wu and W. Xu constructed the training data, built the model, and performed computational analysis. All authors wrote the first draft of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This project was supported by the Natural Science Foundation of Zhejiang Province (no. LD22H300004) and the Science and Technology Development Fund, Macao (no. 0177/2023/RIA3).

REFERENCES

(1) Dietrich, U.; Dürr, R.; Koch, J. Peptides as drugs: from screening to application. *Curr. Pharm. Biotechnol.* **2013**, *14* (5), 501−512.

(2) Lau, J. L.; Dunn, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorg. Med. Chem.* **2018**, *26* (10), 2700−2707.

(3) Henninot, A.; Collins, J. C.; Nuss, J. M. The current state of peptide drug discovery: Back to the future? *J. Med. Chem.* **2018**, *61* (4), 1382−1414.

(4) Muraduzzaman, A. K. M.; Illing, P. T.; Mifsud, N. A.; Purcell, A. W. Understanding the role of HLA class I molecules in the immune response to influenza infection and rational design of a peptide-based vaccine. *Viruses* **2022**, *14* (11), 2578.

(5) Purcell, A. W.; McCluskey, J.; Rossjohn, J. More than one reason to rethink the use of peptides in vaccine design. *Nat. Rev. Drug Discovery* **2007**, *6* (5), 404−414.

(6) Slingluff, C. L. The present and future of peptide vaccines for cancer: Single or multiple, long or short, alone or in combination? *Cancer J. Sudbury Mass* **2011**, *17* (5), 343−350.

(7) Govindarajan, K. R.; Kangueane, P.; Tan, T. W.; Ranganathan, S. MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinforma. Oxf. Engl.* **2003**, *19* (2), 309−310.

(8) Koşaloğlu-Yalçın, Z.; Lanka, M.; Frentzen, A.; Logandha Ramamoorthy Premlal, A.; Sidney, J.; Vaughan, K.; Greenbaum, J.; Robbins, P.; Gartner, J.; Sette, A.; Peters, B. Predicting T cell recognition of MHC class I restricted neoepitopes. *OncoImmunology* **2018**, *7* (11), No. e1492508.

(9) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein structure prediction using rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.

(10) Bradley, P.; Chivian, D.; Meiler, J.; Misura, K. M. S.; Rohl, C. A.; Schief, W. R.; Wedemeyer, W. J.; Schueler-Furman, O.; Murphy, P.; Schonbrun, J.; Strauss, C. E. M.; Baker, D. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Proteins Struct. Funct. Bioinforma.* **2003**, *53* (S6), 457−468.

(11) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P.-S.; Huber, T.; Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.; Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D.-A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, R. Y.-R.; Watkins, A.; Zimmerman, L.; Bonneau, R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **2020**, *17* (7), 665−680.

(12) Bhardwaj, G.; O'Connor, J.; Rettie, S.; Huang, Y. H.; Ramelot, T. A.; Mulligan, V. K.; Alpkilic, G. G.; Palmer, J.; Bera, A. K.; Bick, M. J.; Di Piazza, M.; Li, X.; Hosseinzadeh, P.; Craven, T. W.; Tejero, R.; Lauko, A.; Choi, R.; Glynn, C.; Dong, L.; Griffin, R.; van Voorhis, W. C.; Rodriguez, J.; Stewart, L.; Montelione, G. T.; Craik, D.; Baker, D. Accurate de novo design of membrane-traversing macrocycles. *Cell* **2022**, *185* (19), 3520−3532.

(13) Chen, S.; Sun, Z.; Lin, L.; Liu, Z.; Liu, X.; Chong, Y.; Lu, Y.; Zhao, H.; Yang, Y. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *J. Chem. Inf. Model.* **2020**, *60* (1), 391−399.

(14) Qi, Y.; Zhang, J. Z. H. DenseCPD: Improving the accuracy of neural-network-based computational protein sequence design with denseNet. *J. Chem. Inf. Model.* **2020**, *60* (3), 1245−1252.

(15) Zhang, Y.; Chen, Y.; Wang, C.; Lo, C.-C.; Liu, X.; Wu, W.; Zhang, J. ProDCoNN: Protein design using a convolutional neural network. *Proteins* **2020**, *88* (7), 819−829.

(16) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; De Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King,

N. P.; Baker, D. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **2022**, *378* (6615), 49–56.

(17) Gao, Z.; Tan, C.; Chacón, P.; Li, S. Z. PiFold: Toward effective and efficient protein inverse folding. *arXiv* April 13, **2023**. .

(18) Liu, Y.; Zhang, L.; Wang, W.; Zhu, M.; Wang, C.; Li, F.; Zhang, J.; Li, H.; Chen, Q.; Liu, H. Rotamer-free protein sequence design based on deep learning and self-consistency. *Nat. Comput. Sci.* **2022**, *2* (7), 451–462.

(19) Huang, B.; Fan, T.; Wang, K.; Zhang, H.; Yu, C.; Nie, S.; Qi, Y.; Zheng, W.-M.; Han, J.; Fan, Z.; Sun, S.; Ye, S.; Yang, H.; Bu, D. Accurate and efficient protein sequence design through learning concise local environment of residues. *Bioinforma. Oxf. Engl.* **2023**, *39* (3), btad122.

(20) Bennett, N. R.; Coventry, B.; Goreshnik, I.; Huang, B.; Allen, A.; Vafeados, D.; Peng, Y. P.; Dauparas, J.; Baek, M.; Stewart, L.; DiMaio, F.; De Munck, S.; Savvides, S. N.; Baker, D. Improving de novo protein binder design with deep learning. *Nat. Commun.* **2023**, *14* (1), 2625.

(21) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3* (1), 9.

(22) Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **2019**, *20* (1), 723.

(23) Gupta, S.; Nerli, S.; Kutti Kandy, S.; Mersky, G. L.; Sgourakis, N. G. HLA3DB: comprehensive annotation of peptide/HLA complexes enables blind structure prediction of T cell epitopes. *Nat. Commun.* **2023**, *14* (1), 6349.

(24) Lu, T.; Zhang, Z.; Zhu, J.; Wang, Y.; Jiang, P.; Xiao, X.; Bernatchez, C.; Heymach, J. V.; Gibbons, D. L.; Wang, J.; Xu, L.; Reuben, A.; Wang, T. Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat. Mach. Intell.* **2021**, *3* (10), 864–875.

(25) Chu, Y.; Zhang, Y.; Wang, Q.; Zhang, L.; Wang, X.; Wang, Y.; Salahub, D. R.; Xu, Q.; Wang, J.; Jiang, X.; Xiong, Y.; Wei, D.-Q. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nat. Mach. Intell.* **2022**, *4* (3), 300–311.

(26) Jensen, P. E. Recent advances in antigen processing and presentation. *Nat. Immunol.* **2007**, *8* (10), 1041–1048.

(27) Nguyen, A. T.; Szeto, C.; Gras, S. The pockets guide to HLA class I molecules. *Biochem. Soc. Trans.* **2021**, *49* (5), 2319–2331.

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(29) Ingraham, J.; Garg, V. K.; Barzilay, R.; Jaakkola, T. Generative models for graph-based protein design. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019, pp 15820–15831.

(30) Bryant, P.; Pozzati, G.; Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **2022**, *13* (1), 1265.

(31) Lee, K.-H.; Chang, Y.-C.; Chen, T.-F.; Juan, H.-F.; Tsai, H.-K.; Chen, C.-Y. Connecting MHC-I-binding motifs with HLA alleles via deep learning. *Commun. Biol.* **2021**, *4* (1), 1–12.

(32) Zhang, X.; Yin, H.; Ling, F.; Zhan, J.; Zhou, Y. SPIN-CGNN: Improved fixed backbone protein design with contact map-based graph construction and contact graph neural network. *PLOS Comput. Biol.* **2023**, *19* (12), No. e1011330.

(33) Bassani-Sternberg, M.; Pletscher-Frankild, S.; Jensen, L. J.; Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein sbundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **2015**, *14* (3), 658–673.

(34) Shiina, T.; Hosomichi, K.; Inoko, H.; Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **2009**, *54* (1), 15–39.

(35) Cao, L.; Coventry, B.; Goreshnik, I.; Huang, B.; Sheffler, W.; Park, J. S.; Jude, K. M.; Marković, I.; Kadam, R. U.; Verschueren, K. H. G.; Verstraete, K.; Walsh, S. T. R.; Bennett, N.; Phal, A.; Yang, A.; Kozodoy, L.; DeWitt, M.; Picton, L.; Miller, L.; Strauch, E.-M.; DeBouver, N. D.; Pires, A.; Bera, A. K.; Halabiya, S.; Hammerson, B.; Yang, W.; Bernard, S.; Stanton, L.; Wilson, I. A.; Ruohola-Baker, H.; Schlessinger, J.; Lee, S.; Savvides, S. N.; Garcia, K. C.; Baker, D. Design of protein-binding proteins from the target Structure Alone. *Nature* **2022**, *605* (7910), 551–560.

(36) Zambaldi, V.; La, D.; Chu, A. E.; Patani, H.; Danson, A. E.; Frerix, T.; Schneider, R. G.; Saxton, D.; Thillaisundaram, A.; Moraes, I.; Lange, O.; Papa, E.; Stanton, G.; Martin, V.; Singh, S.; Wong, H.; Bates, R.; Kohl, S. A.; Abramson, J.; Senior, A. W.; Alguel, Y.; Wu, M. Y.; Aspalter, I. M.; Bentley, K.; Bauer, D. L. V.; Cherepanov, P.; Hassabis, D.; Kohli, P.; Fergus, R.; Wang, J. De novo design of high-affinity protein binders with AlphaProteo. *arXiv* 2024. .

(37) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **2004**, *57* (4), 702–710.

(38) O'Reilly, F. J.; Graziadei, A.; Forbrig, C.; Bremenkamp, R.; Charles, K.; Lenz, S.; Elfmann, C.; Fischer, L.; Stülke, J.; Rappsilber, J. Protein complexes in cells by AI-assisted structural proteomics. *Mol. Syst. Biol.* **2023**, *19* (4), No. e11544.

(39) Ellis, J. M.; Henson, V.; Slack, R.; Ng, J.; Hartzman, R. J.; Katovich Hurley, C. Frequencies of HLA-A2 alleles in five U.S. population groups: Predominance Of A*02011 and identification of HLA-A*0231. *Hum. Immunol.* **2000**, *61* (3), 334–340.

(40) Sant, S.; Quiñones-Parra, S. M.; Koutsakos, M.; Grant, E. J.; Loudovaris, T.; Mannering, S. I.; Crowe, J.; van de Sandt, C. E.; Rimmelzwaan, G. F.; Rossjohn, J.; Gras, S.; Loh, L.; Nguyen, T. H. O.; Kedzierska, K. HLA-B*27:05 alters immunodominance hierarchy of universal influenza-specific CD8$^+$ T cells. *PLoS Pathog.* **2020**, *16* (8), No. e1008714.

(41) Cauli, A.; Shaw, J.; Giles, J.; Hatano, H.; Rysnik, O.; Payeli, S.; McHugh, K.; Dessole, G.; Porru, G.; Desogus, E.; Fiedler, S.; Hölper, S.; Carette, A.; Blanco-Gelaz, M. A.; Vacca, A.; Piga, M.; Ibba, V.; Garau, P.; La Nasa, G.; López-Larrea, C.; Mathieu, A.; Renner, C.; Bowness, P.; Kollnberger, S. The arthritis-associated HLA-B*27:05 allele forms more cell Surface B27 dimer and free heavy chain ligands for KIR3DL2 than HLA-B*27:09. *Rheumatol. Oxf. Engl.* **2013**, *52* (11), 1952–1962.

(42) Bank, R. P. D. *RCSB PDB - 7U1R: SARS-CoV-2 Spike-derived peptide S1185–1193 K1191N mutant (RLNEVANNL) presented by HLA-A*02:01*. https://www.rcsb.org/structure/7U1R (accessed 2024–09–25).

(43) Hülsmeyer, M.; Hillig, R. C.; Volz, A.; Rühl, M.; Schröder, W.; Saenger, W.; Ziegler, A.; Uchanska-Ziegler, B. HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. *J. Biol. Chem.* **2002**, *277* (49), 47844–47853.

(44) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **2007**, *372* (3), 774–797.

(45) Gribenko, A. V.; Patel, M. M.; Liu, J.; McCallum, S. A.; Wang, C.; Makhatadze, G. I. Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (8), 2601–2606.