Peptide-binding specificity prediction using fine-tuned protein structure prediction networks

Amir Motmaen[a,b,c] (ID), Justas Dauparas[a,b] (ID), Minkyung Baek[a,b] (ID), Mohamad H. Abedi[a,b,d] (ID), David Baker[a,b,d,1] (ID), and Philip Bradley[a,b,e,1] (ID)

Peptide-binding proteins play key roles in biology, and predicting their binding specificity is a long-standing challenge. While considerable protein structural information is available, the most successful current methods use sequence information alone, in part because it has been a challenge to model the subtle structural changes accompanying sequence substitutions. Protein structure prediction networks such as AlphaFold model sequence-structure relationships very accurately, and we reasoned that if it were possible to specifically train such networks on binding data, more generalizable models could be created. We show that placing a classifier on top of the AlphaFold network and fine-tuning the combined network parameters for both classification and structure prediction accuracy leads to a model with strong generalizable performance on a wide range of Class I and Class II peptide-MHC interactions that approaches the overall performance of the state-of-the-art NetMHCpan sequence-based method. The peptide-MHC optimized model shows excellent performance in distinguishing binding and non-binding peptides to SH3 and PDZ domains. This ability to generalize well beyond the training set far exceeds that of sequence-only models and should be particularly powerful for systems where less experimental data are available.

AlphaFold | fine-tuning | binding specificity prediction | peptide-MHC interactions | structure modeling networks

Sequence-based methods utilize large sets of experimentally validated binding and non-binding peptides to assemble position specific weight matrices or more sophisticated neural networks with several layers to discriminate binder from non-binder peptides (1–7). Methods such as NetMHCpan are the current state of the art to address key biological challenges like major histocompatibility complex (MHC)-peptide-binding specificity which is central to the adaptive immune system (T cell surveillance, differentiation, etc.), since they can readily optimize parameters over large sets of binding and non-binding peptides. However, sequence-based methods are limited by their inability to incorporate detailed structural information, and as a result, they have reduced generalizability, particularly in cases where there is less training data. While structure-based methods have shown promise to fill this "gap", they have been limited by their inability to accurately predict protein and peptide backbone changes which can affect both affinity and specificity, and more importantly, they lack a way to optimize many model parameters on the large amounts of peptide-binding data that are often available (8).

Results

We reasoned that the recent advances in protein structure prediction could help overcome both limitations: that of structure-based methods in utilizing large amounts of known binding data, and sequence-based methods, in using structural information. AlphaFold (9) and RoseTTAFold (10) predict highly accurate structures (11) and structure quality confidence metrics that have been used to distinguish pairs of proteins which bind from those that don't with some success (12, 13). However, while these methods can be readily trained with structural data, in their current form it is not straightforward to train on binding data.

We set out to extend these networks to enable simultaneous training on structure and binding data. Because of the importance of the peptide-MHC interaction to adaptive immunity, and the very large datasets available for training and testing, we began by focusing on this binding interaction. We first explored how to provide available sequence and structure data as inputs to AlphaFold to obtain the most accurate peptide-MHC structure models (*Methods*) and obtained best performance using single sequence (rather than multiple sequence alignment (MSA) information) and peptide-MHC structure templates as inputs with the query peptide positionally aligned to the template peptides. This approach modeled peptide-MHC structures with a median peptide RMSD of 0.8 Å on backbone and 1.8 Å on all atoms for Class I (*SI Appendix*, Fig. S1 *A and B*) and 0.8 Å on backbone and 1.2 Å on

Significance

Peptide-binding proteins carry out a variety of biological functions in cells and predicting their binding specificity could significantly improve our understanding of molecular pathways. Deep neural networks have achieved high structure prediction accuracy, but are not trained to predict binding specificity. Here we describe an approach to extending such networks to jointly predict protein structure and binding specificity. We incorporate AlphaFold into this approach and fine-tune its parameters on peptide-MHC Class I and II structural and binding data. The fine-tuned model approaches state-of-the-art classification accuracy on peptide-MHC specificity prediction and generalizes to other peptide-binding systems such as the PDZ and SH3 domains.

Author affiliations: [a]Department of Biochemistry, University of Washington, Seattle, WA 98195; [b]Institute for Protein Design, University of Washington, Seattle, WA 98195; [c]Bioengineering Graduate Program, University of Washington, Seattle, WA 98195; [d]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195; and [e]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA 98109

all atoms for Class II (*SI Appendix,* Fig. S1 *C and D*) complexes. These levels of accuracy are consistent with other recent work predicting peptide-MHC structures (8, 14–17).

We found previously in studies of designed proteins bound to targets (18) that both the AlphaFold residue–residue accuracy estimate, predicted aligned error (PAE), and the per-residue accuracy estimate predicted local-distance difference test (pLDDT) are able to partially discriminate binders from non-binders. We carried out predictions for sets of binding and non-binding peptides from (1, 5, 19) and found that both the PAE, calculated between MHC and peptide, and the per-residue accuracy estimate pLDDT averaged over the peptide residues, provided some discrimination of binders from non-binders (*SI Appendix,* Fig. S2). However, the discrimination between binders and non-binders was considerably poorer than NetMHCpan and AlphaFold tended to dock non-binding peptides in the MHC-peptide-binding groove (Fig. 2*A*).

We reasoned that the relatively poor discrimination compared to NetMHCpan reflected the lack of training of AlphaFold on negative binding examples generally and on peptide-MHC binding data in particular. To overcome this limitation, we set out to implement a hybrid structurally informed classification approach for fine-tuning AlphaFold that incorporates both binder and non-binder peptide-MHC pairs. To enable direct training on binding data, we added a simple logistic regression layer on top of AlphaFold which converts the peptide-MHC PAE values into a normalized binder/non-binder score. We experimented with different training regimens for optimizing the combined model parameters, and obtained the best results by first fitting the logistic regression parameters, and then subsequently fine-tuning all of the AlphaFold parameters (Fig. 1 and *SI Appendix,* Fig. S3). To maintain structure prediction accuracy, we included peptide-MHC structure prediction examples during training. For cases where the structure was not known, we followed a "distillation" procedure (9) in which AlphaFold models of the complex were used as the ground truth conformation, but with the loss calculated only over the MHC to allow the peptide conformation to adjust during training. Leveraging the wealth of data available on peptide-MHC interactions, we assembled a diverse training set consisting of 10,340 peptide-MHC examples, 203 structurally characterized and 5,102 modeled peptide-MHC binder examples, and 5,035 non-binder examples, distributed across 68 Class I alleles and 39 Class II allele pairs

(*Methods*). Training the combined structure prediction-classification model on this set produced consistent drops in the combined loss function (*SI Appendix,* Fig. S4), suggesting that the model might be learning to better discriminate binder from non-binder peptides on the basis of structural information. The fine-tuned model developed a tendency to exclude non-binder peptides completely out of the MHC's peptide-binding groove (Fig. 2*A*).

We assessed prediction performance on an independent test set containing 2,402 binder and 2,717 non-binder peptide-MHC pairs distributed across 32 Class I and 26 Class II alleles (Fig. 2). This set contained no peptides with fewer than two mismatches to any peptide in the training set (*SI Appendix,* Fig. S5). During training, the Class I examples were restricted to 9-mer peptides, but in the test set we included Class I peptide-MHC pairs containing peptides of lengths 8, 9, and 10 residues to better assess generalizability of our fine-tuning approach. The combined structure prediction-classification model performs significantly better on this test set than AlphaFold with default parameters (Fig. 2 *B* and *C*). A direct comparison to the state-of-the-art sequence-based predictor NetMHCpan on classical human leukocyte antigen (HLA) alleles is challenging since the vast majority of available binding data is already contained in its training set. Despite our test examples being contained within NetMHCpan's training data, our fine-tuned model performed competitively with NetMHCpan, particularly on Class II, achieving an area under the receiver operating characteristic curve (AUROC) of 0.97 for Class I (Fig. 2*B*) and 0.93 for Class II (Fig. 2*C*) indicating excellent classification. Beyond classification, our model also has reasonable performance on peptide-MHC binding affinity, with a Pearson correlation coefficient of 0.79 for 9-residue peptides binding to HLA-A*02:01 (compared to 0.57 for AlphaFold with default parameters and 0.78 for NetMHCpan; *SI Appendix,* Fig. S6). Our combined prediction-classification approach appears to learn the biochemical space of peptide-MHC binding efficiently rather than relying solely on sequence patterns observed during training and retains good peptide structure prediction performance (*SI Appendix,* Figs. S7 and S8).

Divergence from NetMHCpan on 10-mer Peptides. Our structure-based approach to peptide-MHC binding prediction is an orthogonal and hence potentially complementary method to
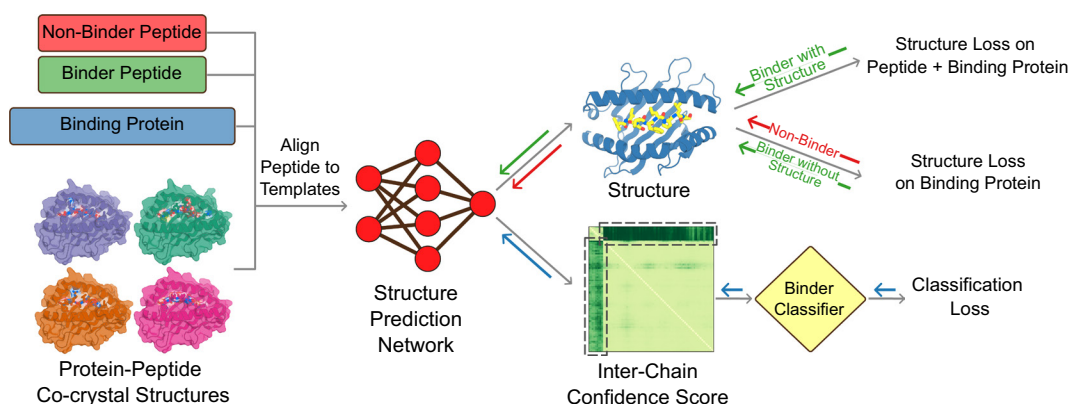

Fig. 1. Extending structure prediction networks for binder classification. For fine-tuning the combined model to distinguish binding from non-binding peptides, inputs (*Left*) are sets of known binding (green) and non-binding (red) sequences, the sequence of the target protein(s), and the available co-crystal structures which provide templates for modeling the complex. The peptide sequence is positionally aligned to the peptides in the co-crystal structures, and the structure of the complex is predicted with AlphaFold. A logistic regression classifier converts the output PAE values into a normalized binder/non-binder score and prediction. During training, the combined model parameters were optimized to minimize the loss terms in the final column, and model weights were updated through gradient backpropagation as indicated by the solid arrows. The classification loss (cross-entropy) is computed on all training examples; the structure loss is computed over the entire complex for binding peptides with co-crystal structures and over the binding protein alone for binding peptides without co-crystal structures and for non-binding peptides.
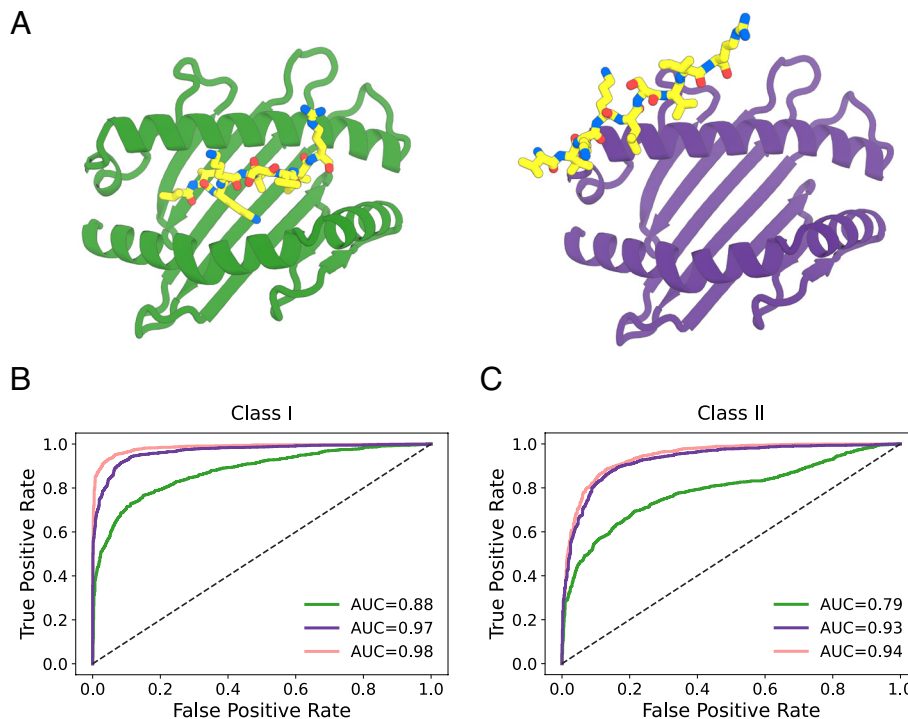
Fig. 2. Fine-tuning improves pMHC-peptide binder classification to state-of-the-art levels. (*A*) Structure prediction for a non-interacting peptide-MHC pair using AlphaFold before (*Left*) and after (*Right*) fine-tuning on joint structure prediction and binder classification. The peptide is incorrectly predicted to bind the MHC in the first case but not the second. (*B* and *C*) Binder/non-binder peptide classification ROC curves on test sets (*Methods*): (*B*) MHC Class I, (*C*) MHC Class II. AlphaFold (green), combined structure prediction-classification model (purple), and NetMHCpan (pink).

purely sequence-based tools such as NetMHCpan. As an illustration, we noticed a striking divergence between structural predictions and NetMHCpan predictions in a scan of the Hepatitis C virus (HCV) genome polyprotein for 10-mer peptides predicted to bind to the HLA-A*02:01 allele. Surprisingly, the two top-ranked NetMHCpan peptides (AKLMPQLPGI and ADLMGYIPLV) both had charged amino acids at the first anchor position (peptide position 2, a position which strongly prefers hydrophobic amino acids in known HLA-A*02:01 binders; Fig. 3*B*). By contrast, these two peptides are not predicted to bind to HLA-A*02:01 using our joint structure and binding prediction model (Fig. 3*A*). When predicting affinities for Class I peptides of length greater than 9, NetMHCpan aligns target peptides to a 9-mer binding model, allowing consecutive positions to "gap out" to match the 9-mer model length. For both of these peptides, peptide position 2 was not aligned to the 9-mer model, making position 3 the anchor position for the purposes of scoring. This likely explains the highly favorable scores assigned to these peptides, as position 3 is a leucine (a strongly preferred anchor amino acid) in both. To assess more broadly the relationship between the NetMHCpan gap position and the consistency with our predictions, we assigned a rank error equal to the absolute difference in ranks assigned by NetMHCpan and our structural approach to each 10-mer peptide window in the target and grouped these rank error values based on the NetMHCpan gap position. As shown in Fig. 3*C*, the positions with highest mean rank error are the two canonical HLA-A*02:01 anchor positions (2 and 10, for a 10-mer peptide). These results suggest that there may be differences in the reliability of NetMHCpan predictions for peptides of length greater than 9, depending on the location of the gap position(s), and more generally, that our combined structure prediction-classification model may be more robust than purely sequence-based models to the inevitable biological variations in peptide length and other properties.

Combined Structure Prediction-Classification Model Extends to Other Protein-Peptide Systems. To evaluate the generalizability of our combined structure prediction-classification model beyond the peptide-MHC binding data it was trained on, we assessed performance on two additional systems: peptide recognition by PDZ domains, which bind C-terminal peptides, and peptide recognition by SH3 domains, which recognize proline-rich peptides. The PDZ dataset was taken from ref. 20 and consists of 17 PDZ domains with an experimentally determined co-complex structure and peptide binders from phage-display selection experiments (21) (*SI Appendix*, Table S1). The SH3 dataset consists of 19 SH3 domains with an experimentally determined co-complex structure and peptide binders from phage-display available in the PRM-DB database (http://www.prm-db.org/) (22) (*SI Appendix*, Table S2). For comparison with experiments, we performed an "in silico selection" experiment in which 20,000 random peptides were modeled and ranked by our structure prediction-classification model (Fig. 4*A*). Position weight matrices (PWMs) were constructed from the top-ranked binding sequences, and these PWMs were compared with PWMs constructed from the experimentally selected binders using two PWM divergence measures proposed by Smith and Kortemme (20). The fraction of top-ranked sequences used to build the PWM (f_{PWM}) is a free parameter in this approach; we varied this parameter from 10^{-4} (i.e., only two sequences used to build the PWM) to one (all sequences used) and evaluated the accuracy of the resulting PWMs. The prediction errors, plotted as a function of f_{PWM}, are shown in Fig. 4*B* for our fine-tuned parameters (purple lines) and for AlphaFold with default parameters (green lines). From these error curves, we can see that there appears to be an optimal value for f_{PWM} somewhere around 0.01: using too few top-ranked sequences likely introduces noise into the PWMs, while using too many of the randomly sampled sequences reduces the degree of enrichment of preferred
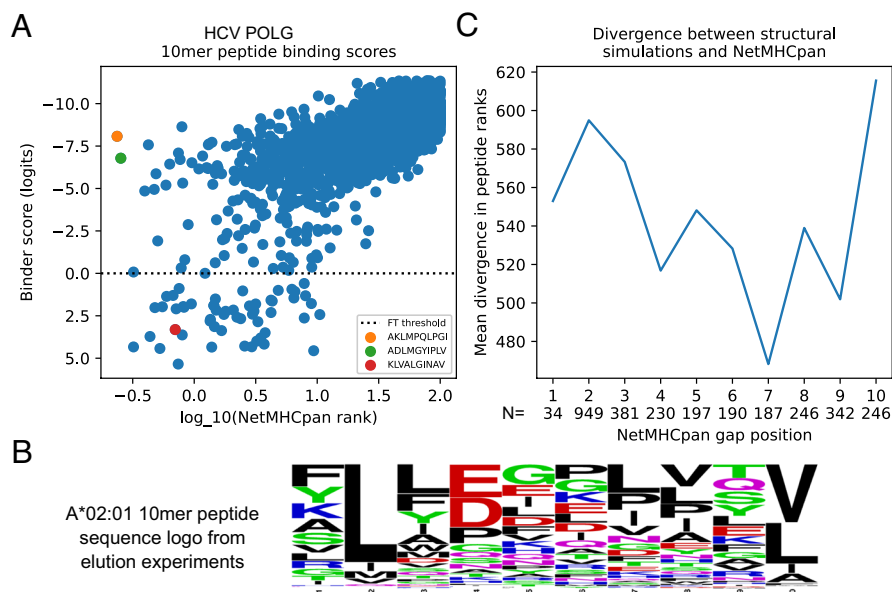
Fig. 3. Prediction of 10-mer peptide binding highlights differences between sequence- and structure-based approaches. (*A*) Scatterplot of NetMHCpan rank score (*x*-axis, $\log_{10}$-transformed) versus combined structure prediction-classification model binder score (*y*-axis, reversed so that predicted strong binders for both methods are in the lower left) reveals strong overall correlation ($R^2 = 0.49$) as well as peptides with divergent predictions (upper left and lower right corners). The top two NetMHCpan predicted peptides (orange and green markers) have charged amino acids at the first anchor position (peptide position 2). A known immunodominant T cell epitope (red marker) scores well by both methods. (*B*) Sequence logo built from 10-mer peptides eluted from A*02:01 (5) shows strong preference for hydrophobic amino acids at the anchor positions. (*C*) Average disagreement in peptide rank between NetMHCpan and the structural model (*y*-axis) is highest when the NetMHCpan gap position is located at one of the two anchor positions (2 and 10).

amino acids. Notably, fine-tuning AlphaFold's parameters on peptide-MHC binding data does not degrade performance on these other, structurally distinct, classes of protein–peptide interactions; the fine-tuned parameters perform better than the default parameters (see *SI Appendix*, Fig. S11 for a comparison to the sequence-based machine learning method HSM (23) on these two systems).

Discussion

The outstanding performance of AlphaFold on protein structure prediction and NetMHCpan on peptide-MHC binding specificity prediction illustrates the power of networks trained on very large datasets; the Protein Data Bank (PDB) in the first case, experimental peptide-MHC binding data in the second. Here, we show how to train a single network on both structural and binding data simultaneously and demonstrate that this enables considerable generalization of binding specificity prediction beyond the available training sets. While sequence-only networks such as NetMHCpan can achieve outstanding performance within their target domains, combined sequence and structure models can be more robust to biological variation and limited training data, and they are able to generalize to new biological systems. Moving forward, there is much to explore in network architectures for combined structure and specificity prediction; our combination of a simple classification module (logistic regression) on top of the AlphaFold network provides a starting benchmark for comparison. One notable limitation of our framework is the dichotomous treatment of binding affinity: Peptides are either "binders" or "non-binders", which neglects variation in binding affinity that can have important biological consequences. Although the network trained on binder/non-binder classification does have some ability to predict binding strength (*SI Appendix*, Fig. S6), it may be possible to incorporate binding affinity data into the training process and thereby produce a more accurate model.

Materials and Methods

Peptide-MHC Dataset. Structures of peptide-MHC complexes were downloaded from the RCSB Protein Databank (24) with an inclusion cutoff of 2021-08-05. Peptide-MHC binding data were extracted from the training sets generously provided by the NetMHCpan developers on their website (https://services.health-tech.dtu.dk/service.php?NetMHCpan-4.1). The training set consisted of 10,340 peptide-MHC examples, 203 structurally characterized and 5,102 modeled peptide-MHC binder examples, and 5,035 non-binder examples, distributed across 68 Class I alleles and 39 Class II allele pairs. Twenty-three mouse peptide-MHC structures were included in the training set; the remainder of the training and testing data was restricted to human alleles. We assessed prediction performance on an independent test set containing 2,402 binder and 2,717 non-binder peptide-MHC pairs distributed across 32 Class I and 26 Class II alleles. This set contained no peptides with fewer than two mismatches to any peptide with the same binder/non-binder label in the training set (*SI Appendix*, Fig. S5). For Class I, only nine-residue peptides were used for training; for testing, we included 21 8-residue and 475 ten-residue peptides. Class II peptides were modeled as the core nine-residue window together with one amino acid of context on either side (11 residues total). During inference, longer Class II peptides were broken into overlapping 11-residue windows for modeling and the best scoring window was used to make the binding prediction. During training, Class II peptides were represented by their best scoring 11-residue window; thus each peptide, regardless of length, mapped to a single training example. All examples used for training and testing are provided in the GitHub repository associated with this manuscript.

Peptide-MHC Template Search and Query-to-Template Mapping. The amino acid sequence of the α1 and α2 domains for MHC Class I and the α1 and β1 domains for MHC Class II were considered as the MHC query sequence and were pairwise aligned to their corresponding domains across all entries in the PDB template set. The pairwise alignment was done with the Needleman–Wunsch local alignment algorithm using the BLOSUM62 substitution score matrix while setting gap open penalty to −11 and gap extend penalty to −1. The Biopython package (25) was used to run the alignment algorithm. The alignments were used to generate query-to-template mapping strings that map each query residue to the corresponding residue in the template crystal structure. To generate query-to-template mapping for the peptide, sequence identity was ignored and
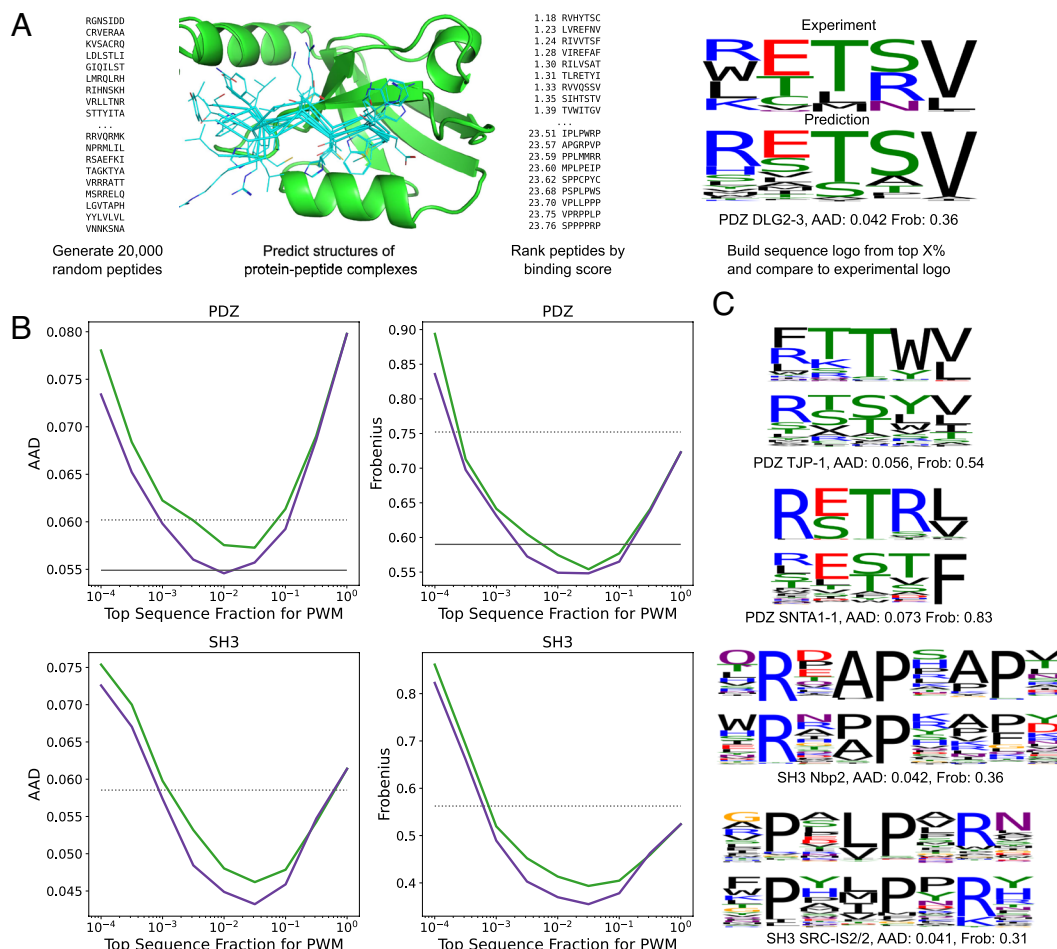
Fig. 4. Accurate prediction of PDZ-peptide and SH3-peptide interactions using the combined structure prediction-classification network. (*A*) Prediction of position-weight matrices by random sequence generation followed by network evaluation. (*B*) Combined network fine-tuned on peptide-MHC interactions outperforms AlphaFold with default parameters in recapitulating experimental PWMs. AAD: average absolute difference in amino acid frequency; Frobenius: average Frobenius distance between PWM columns (lower is better for both; see *Methods*). AlphaFold (green) and fine-tuned AlphaFold (purple). The solid black lines in the PDZ panels indicate the performance of the structure-based method from ref. 20, which was optimized on PDZ domains. Dashed black lines indicate the average AAD and Frobenius scores of PWMs constructed from the sequences of the template PDB peptides used for modeling. (*C*) Examples of predicted and experimental sequence logos (see *SI Appendix*, Figs. S9 and S10 for a complete set of sequence logos).

mappings were generated based on position along the peptide sequence. For Class I molecules, the query peptide was mapped to the template one to one if the length of the two peptides was equal, and if not, the first three and the last three residues of the query were mapped to the corresponding residues of the template. For Class II queries, the core nine-mer residues of the peptide and their immediate neighboring residues (11-mer residues in total) were mapped to the corresponding residues in the templates. If the 9-mer core was not determined in the query, each possible consecutive 11-residue window along the peptide was used as the query peptide to model the complex with all possible positions for the 9-mer core and the best model was selected by picking the one with the lowest inter-chain PAE among the windows. Finally, the MHC and peptide mapping strings were combined by putting the peptide either in the N terminus or the C terminus of the MHC mapping string.

AlphaFold Modeling of Peptide-MHC Pairs. Query peptide sequence was appended to the N terminus or C terminus of the query MHC sequence. At chain break junctions, the residue indices were shifted by 200. Single sequence and templates were used as inputs, and AlphaFold's model_1, model_1_ptm, model_2, or model_2_ptm weights were used for inference. The crystal structures of the top four hits from the template search based on the MHC alignment scores were used to construct the template features in AlphaFold. Both PDB outputs of the predictions and predicted AlphaFold metrics were saved for analysis purposes.

Hybrid AlphaFold Fine-Tuning on Structure and Binding Data. Combined structure prediction-classification approach for fine-tuning AlphaFold was performed in two stages (*SI Appendix*, Fig. S3). First, the binder model parameters

were fitted using logistic regression, while the AlphaFold parameters were kept fixed at their starting values. The AlphaFold MHC-peptide inter-PAE scores for the training set, together with their binder/non-binder labels, were provided to the LogisticRegression class from the Scikit-learn (26) package linear_model. This produced slope and intercept values that defined a linear mapping between inter-PAE values and logit values. We noted that these fitted parameters differed between MHC Class I and MHC Class II: A model fitted on Class I peptides had a higher midpoint PAE (8.03) than a model fitted on Class II peptides (4.34). This difference likely reflects structural differences in the peptide-binding mode, with Class I peptides bulging out of the MHC pocket, anchored at the ends, while Class II peptides follow a direct and fully extended path with termini protruding at the ends of the pocket. To account for these differences, and allow for future applications in which even more disparate binding data might be combined, we allowed the logistic regression midpoint parameter to vary as a function of a predefined class value (here I or II) for each training example.

The binder model parameters were then fixed at their optimal values, while the AlphaFold parameters were fine-tuned in the context of a hybrid structure and binder loss function. The loss for a single training example was equal to the softmax cross-entropy of the binder/non-binder prediction for that example plus the AlphaFold structure loss, which was multiplied by a weight of 1.0, if the ground truth structure for that example was an experimentally determined structure, or 0.25, if the ground truth structure was an AlphaFold predicted structure. Predicted structural models were used as the ground truth conformations for the non-binder peptides and for binder peptides without solved structures; for these examples, the AlphaFold structure loss was computed only over the MHC to

allow the peptide conformation to adjust during fine-tuning. For the parameters evaluated here, we chose a fairly conservative stopping point after two epochs of training (when each training example had been seen twice) to avoid excessive drift in the AlphaFold model. A Python script that performs parameter fine-tuning, together with command line parameters and example inputs, is provided in the GitHub repository associated with this manuscript. Full inputs for fine-tuning, including the predicted AlphaFold models used as ground truth structures, are provided in the GitHub repository associated with this manuscript.

Evaluation of Classification Performance on Peptide-MHC. Pairs of peptide-MHC from the dataset under evaluation were modeled with AlphaFold or Fine-Tuned AlphaFold parameters. Inter-chain PAE terms corresponding to the peptide-MHC interactions were averaged. Mean inter-chain PAE values and binder non-binder class labels were used to construct Receiver Operating Characteristic (ROC) curves and the area under the curve values were used to compare classification performance of different models. To compare our prediction results to NetMHCpan, we used NetMHCpan-4.1 and NetMHCIIpan-4.1 (1, 2). These software tools were installed on our server and executed through the command line.

AlphaFold Modeling of PDZ and SH3 Domains. The PDZ domain dataset (*SI Appendix*, Table S1) was taken from ref. 20 and consisted of 17 human PDZ domains with experimentally determined structures. Binder peptides for the 17 PDZ domains were downloaded from supplemental data for ref. 21 (https://bad-erlab.org/Data/PDZ). Experimental amino acid frequency matrices (PWMs) were constructed from the PDZ binder peptides, with clone frequency weighting. For the AlphaFold simulations, 20,000 random peptide sequences of length equal to the peptide in the experimental structure were generated using NNK codon frequencies to match the amino acid bias in the phage-display libraries. The experimental structure listed in the template column of *SI Appendix*, Table S1 was used as the sole template, with the random peptide sequences aligned to the template peptide and single-sequence MSA information.

The SH3 domain dataset (*SI Appendix*, Table S2) consisted of 19 SH3 domains with experimentally determined structures extracted from the Database of Peptide Recognition Modules (http://prm-db.org/) (22). Experimental PWMs were downloaded from the PRM-DB. SH3 domains can bind peptides in two orientations, denoted Class I and Class II, which have opposite chain orientations: Class I peptides often match a +XXPXXP sequence motif, where "+" denotes a positively charged amino acid and X is any amino acid; Class II peptides often match a PXXPX+ motif. SH3-peptide PWMs from PRM-DB were annotated as Class I or Class II by choosing the class whose sequence motif had the highest PWM frequency (averaged over the three motif positions). Five of the SH3 domains had multiple PWMs in the PRM-DB, one of which was assigned as Class I and one as Class II; these domains were modeled twice, once in each orientation. For AlphaFold modeling, the native PDB structure listed in *SI Appendix*, Table S2 ("SH3 template" column) was used as the template for the SH3 domain. Four peptide-SH3 structures with peptides in the desired orientation (i.e., Class I or Class II) were chosen as "Peptide templates" based on SH3 domain sequence identity (*SI Appendix*, Table S2, Peptide templates column). The peptides in these structures were transformed into the reference frame of the SH3 domain template by structural superimposition to create hybrid template models for AlphaFold. Multiple structural alignment was used to identify the core motif positions (+XXPXXP or PXXPX+) in each template peptide. The peptide sequence modeled in the AlphaFold runs consisted of the core motif together with one residue on either side (nine residues for Class I and eight residues for Class II).

For comparison with experimental PWMs, predicted PWMs were constructed from the top-ranked peptide sequences. Peptides were ranked by protein–peptide

inter-PAE: The sum of the residue–residue PAE scores for all (protein, peptide) and (peptide, protein) residue pairs, where PAE is AlphaFold's "predicted aligned error" accuracy measure. The experimental PWMs were derived from phage-display experiments with random peptide libraries of size 10^9 and greater, whereas the predicted PWMs were based on 20,000 modeled peptides. To account for this differential and better match the entropy of the amino acid frequency distributions, we squared the predicted amino acid frequencies and renormalized them to sum to 1. This had the effect of increasing the information content of the predicted PWMs without changing the order of amino acid preference. The exponent of 2 can be partly rationalized by the approximate twofold differential in log search-space size between predictions and experiments. Following ref. 20, predicted and experimental PWMs for PDZ domains were compared over the last five C-terminal peptide positions. Predicted and experimental PWMs for SH3 domains were compared over the core 7 (for class I) or 6 (for class II) positions of the SH3 motif together with the immediately adjacent positions, if those positions were present in the experimental PWM. Two measures of PWM column divergence were used to assess predictions: average absolute difference (AAD) and the Frobenius metric. The AAD for a single PWM position equals the sum of the absolute frequency deviations for all amino acids, divided by 20; AAD ranges from 0.0 (perfect agreement) to 0.1 (maximal divergence). The Frobenius metric for a single PWM position equals the square root of the sum of the squared frequency deviations; it ranges from 0.0 (perfect agreement) to the square root of 2 (maximal divergence). The AAD and Frobenius values in Fig. 4 were averaged over all PWM columns.

HCV Experiment. We scanned the sequence of the Hepatitis C virus (HCV) genome polyprotein (3,011 amino acids; Uniprot ID P27958) to find potential 10-mer peptide epitopes presented by HLA-A*02:01. We ran NetMHCpan with FASTA format input and default parameters. NetMHCpan compares 10-mer peptides to its internal 9-mer binding model by dropping one gap position from the peptide:model alignment. We parsed the location of the gap position from the output columns ("Of", "Gp", and "Gl") and confirmed by matching to the sequence in the "Core" column. To assess divergence between our structure-based predictions and NetMHCpan, raw scores from each method were sorted and converted to rank scores. Peptides were grouped by the NetMHCpan-assigned gap position, and the mean absolute difference in rank score was computed for each gap position.

1. B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, 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 Res.* **48**, W449–W454 (2020).
2. B. Reynisson *et al.*, Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J. Proteome Res.* **19**, 2304–2315 (2020).
3. J. Jin *et al.*, Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. *Proteins* **89**, 866–883 (2021).
4. G. Venkatesh, A. Grover, G. Srinivasaraghavan, S. Rao, MHCAttnNet: Predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. *Bioinformatics* **36**, i399–i406 (2020).
5. S. Sarkizova *et al.*, A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
6. H. Zeng, D. K. Gifford, DeepLigand: Accurate prediction of MHC class I ligands using peptide embedding. *Bioinformatics* **35**, i278–i283 (2019).
7. D. Gfeller *et al.*, Predictions of immunogenicity reveal potent SARS-CoV-2 CD8+ T-cell epitopes. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.05.23.492800 (Accessed 26 March 2022).
8. M. A. S. Perez, M. A. Cuendet, U. F. Röhrig, O. Michielin, V. Zoete, Structural prediction of peptide-MHC binding modes. *Methods Mol. Biol.* **2405**, 245–282 (2022).
9. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
11. J. Jumper *et al.*, Applying and improving AlphaFold at CASP14. *Proteins* **89**, 1711–1721 (2021).

12. I. R. Humphreys *et al.*, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
13. L. Chang, A. Perez, AlphaFold encodes the principles to identify high affinity peptide binders. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.03.18.484931 (Accessed 26 March 2022).
14. M.-S. Park, S. Y. Park, K. R. Miller, E. J. Collins, H. Y. Lee, Accurate structure prediction of peptide-MHC complexes for identifying highly immunogenic antigens. *Mol. Immunol.* **56**, 81–90 (2013).
15. J. M. Khan, S. Ranganathan, pDOCK: A new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res.* **6** (suppl. 1), S2 (2010).
16. T. Liu *et al.*, Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *J. Chem. Inf. Model.* **54**, 2233–2242 (2014).
17. D. A. Antunes, D. Devaurs, M. Moll, G. Lizée, L. E. Kavraki, General prediction of peptide-MHC binding modes using incremental docking: A proof of concept. *Sci. Rep.* **8**, 4327 (2018).
18. N. Bennett *et al.*, Improving de novo protein binder design with deep learning. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.06.15.495993 (Accessed 26 March 2022).
19. J. G. Abelin *et al.*, Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326 (2017).
20. C. A. Smith, T. Kortemme, Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.* **402**, 460–474 (2010).
21. R. Tonikian *et al.*, A specificity map for the PDZ domain family. *PLoS Biol.* **6**, e239 (2008).
22. J. Teyra *et al.*, Large-scale survey and database of high affinity ligands for peptide recognition modules. *Mol. Syst. Biol.* **16**, e9310 (2020).
23. J. M. Cunningham, G. Koytiger, P. K. Sorger, M. AlQuraishi, Biophysical prediction of protein-peptide interactions and signaling networks using machine learning. *Nat. Methods* **17**, 175–183 (2020).
24. H. M. Berman *et al.*, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
25. P. J. A. Cock *et al.*, Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
26. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12** 2825–2830 (2011).