PepNN: a deep attention model for the identification of peptide binding sites

Osama Abdin[1], Han Wen[2], Philip M. Kim[1,2,3,*]

1. Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada
2. Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada
3. Department of Computer Science, University of Toronto, Toronto, ON M5S 3E1, Canada

* To whom correspondence should be addressed: pi@kimlab.org

Abstract

Protein-peptide interactions play a fundamental role in facilitating many cellular processes, but remain underexplored experimentally and difficult to model computationally. Here, we introduce PepNN-Struct and PepNN-Seq, structure and sequence-based approaches for the prediction of peptide binding sites on a protein given the sequence of a peptide ligand. The models make use of a novel reciprocal attention module that is able to better reflect biochemical realities of peptides undergoing conformational changes upon binding. To compensate for the scarcity of peptide-protein complex structural information, we make use of available protein-protein complex and protein sequence information through a series of transfer learning steps. PepNN-Struct achieves state-of-the-art performance on the task of identifying peptide binding sites, with a ROC AUC of 0.893 and an MCC of 0.483 on an independent test set. Beyond prediction of binding sites on proteins with a known peptide ligand, we also show that the developed models make reasonable agnostic predictions, allowing for the identification of novel peptide binding proteins.

Introduction

Interactions between proteins and peptides are critical for a variety of biological processes. A large fraction of protein-protein interactions are mediated by the binding of intracellular peptide recognition modules (PRMs) to linear segments in other proteins (Tompa *et al*, 2014). Moreover, peptide ligands binding to extracellular receptors have important functions (Krumm & Grisshammer, 2015). In total, it is estimated that there are roughly 10^4 human proteins that contain at least one PRM (Cunningham *et al*, 2020) and that there are over 10^6 peptide motifs encoded in the human proteome (Tompa *et al*, 2014). Disruption of these interactions and their regulation can consequently result in disease; for instance, many proteins with PRMs harbor oncogenic mutations (Yang *et al*, 2015). It has also been shown that viral proteins encode peptidic motifs that can potentially be used to hijack host machinery during infection (Hagai *et al*, 2014).

In the absence of ample experimental data including solved structures, gaining molecular insight into these interactions and their associated disease states is contingent on the ability to model peptide binding computationally. This has been a difficult problem that has traditionally been approached with peptide-protein docking (Ciemny *et al*, 2018). One widely used peptide docking tool is FlexPepDock, a Rosetta protocol that refines coarse-grain peptide-protein conformations by sampling from the degrees of freedom within a peptide (Raveh *et al*, 2010). In general, benchmarking studies have shown that peptide docking approaches often fail to accurately identify the native complex conformation (London *et al*, 2012; Agrawal *et al*, 2019; Weng *et al*, 2020), indicating that this problem remains unsolves; current approaches are limited by the high flexibility of peptides as well the inherent error of scoring heuristics (Ciemny *et al*, 2018). Machine learning approaches provide potential alternatives to docking, as they can sidestep the issue of explicit enumeration of conformational space and can learn scoring metrics directly from the data.

Different machine learning approaches have been applied to the preliminary problem of predicting the binding sites of peptides with a varying amount of success (Johansson-Åkhe *et al*, 2019; Zhao *et al*, 2018; Taherzadeh *et al*, 2016, 2018; Wardah *et al*, 2020; Iqbal & Hoque, 2018). Deep learning approaches have resulted in large improvements in many area, including in the domains of protein and structural biology (Senior *et al*, 2020). However, no such model has been developed for the identification of peptide binding sites; one hurdle has been the paucity of available data, as deep learning models usually require large training data sets.

Here, we sought to develop a novel deep learning architecture to improve upon existing approaches. In particular, we sought to exploit available protein-protein complex information, thereby adding an order of magnitude more training data. The "hot segment" paradigm of protein-protein interaction suggests that the interaction between two proteins can be mediated by a linear segment in one protein that contributes to the majority of the interface energy (London *et al*, 2010). Complexes of protein fragments with receptors thus represent a natural source of data for model pre-training. In addition, the idea of pre-training contextualized language models has recently been adapted to protein biology for the purpose of generating meaningful representations of protein sequences (Elnaggar *et al*, 2020; Rao *et al*, 2019). The success of these approaches provides an opportunity to develop a strictly sequence based peptide binding site predictor.

In this study, we integrate the use of contextualized-language models, available protein-protein complex data, and a task-specific attention-based architecture, to develop parallel models for both structure and sequence-based peptide binding site prediction: PepNN-Struct and

PepNN-Seq. Comparison to existing approaches reveals that our models perform better in most cases. We also show that the developed models can make reasonable peptide agnostic predictions, allowing for their use for the identification of novel peptide binding sites.

Results

Parallel models for structure and sequence-based peptide binding site prediction

We sought to develop a network that takes as input a representation of a protein as well as a peptide sequence, and outputs residue-wise scores representing the confidence that a particular residue is part of a peptide binding site (Fig 1A-B). The PepNN architecture is based in part on the Transformer, a model that makes use of repeated multi-head attention modules to efficiently learn long-range dependencies in sequence inputs (Vaswani *et al*, 2017). The Transformer architecture has also been adapted to graph inputs (Ingraham *et al*, 2019); graph convolutions have been shown to be effective for protein design (Strokach *et al*, 2020). PepNN-Struct makes use of these graph attention layers to learn from context within an input protein (Fig 1A). PepNN-Seq, on the other hand, generates predictions based solely on the input protein and peptide sequences (Fig 1B).

PepNN differs from conventional Transformers in that it does not follow an encoder-decoder architecture. This is based on the fact encoding the peptide sequence independently would implicitly assume that all information about the peptide is contained within its sequence. This assumption is not concordant with the fact that many disordered regions undergo conformational changes upon protein binding (Mohan *et al*, 2006). A peptide's sequence is thus insufficient by itself to determine its conformation in a particular system. As an alternative, we introduced multi-head reciprocal attention layers, a novel attention-based module that simultaneously updates the peptide and protein embeddings while ensuring that the unnormalized attention values from protein to peptide residues are equal to the unnormalized attention values in the other direction. This ensures that the protein residues involved in binding have influence on the peptide residues and vice versa, better reflecting the physical reality of the peptide-protein binding process. The exact model hyperparameters were determined using random search (see Methods) and we compared the performance of the model to a graph Transformer with the same hyperparameters on the preliminary task of identifying the binding sites of protein fragments. We found that the reciprocal attention variant outperforms the graph Transformer in its capacity to accurately identify fragment binding sites (Fig S1).

Transfer learning results in large improvements in model performance

We used transfer learning in two ways to improve model performance. The first was to pretrain the model on a large protein fragment-protein complex dataset before fine-tuning with a smaller dataset of peptide-protein complexes (Fig 1C). To generate the fragment dataset, we scanned all protein-protein complex interfaces in the PDB with the PeptiDerive Rosetta protocol (Sedan *et al*, 2016) to identify protein fragments of length 5-25 amino acids that contribute to a large portion of the complex interface energy (Fig S2). These fragment-protein complexes were filtered based on their estimated interface energy as well as the buried surface area to ensure that they had binding properties that were reasonably close to that of peptide-protein complexes. The second application of transfer learning was the use a pre-trained contextualized language model, ProtBert (Elnaggar *et al*, 2020), to embed protein sequences. These high dimensional, information-rich, embeddings were used as input to PepNN-Seq (Fig 1B).

To evaluate the impact of transfer learning on model performance, we trained PepNN-Struct and PepNN-Seq using different procedures. Pre-training PepNN-Struct resulted in significant improvement over models trained on only the fragment or peptide complex dataset, both in terms of over all binding residue prediction, and in terms of prediction for individual proteins (Fig 2A, B). Model predictions on the Bro domain of HD-PTP demonstrate this difference in performance, as only the pre-trained variant of the model correctly predicts the peptide binding site (Fig 2C). This example furthermore illustrates that the pre-training step helps bring the parameters closer to an optimum for general peptide binding site prediction, rather than improving performance solely on examples that match patterns seen in the fragment-complex dataset.

Embedding protein sequences with ProtBert resulted in large performance improvements over learned embedding parameters for PepNN-Seq (Fig 2D, E). Interestingly, pretraining on the fragment complexes did not have a large impact on PepNN-Seq performance (Fig 2B, D). This may suggest that pre-training on the fragment complexes allows PepNN-Struct to learn reasonable protein embeddings while the use of a pre-trained contextualized language model is sufficient for the generation of reasonable embeddings in the case of PepNN-Seq.

PepNN achieves state-of-the-art performance on peptide binding site prediction

We initially evaluated the developed models on an independent test set derived from the peptide complex dataset. Unsurprisingly, we found that PepNN-Struct outperforms PepNN-Seq (Table 1). We additionally ran the sequence-based PBRpredict-Suite model on this test dataset (Iqbal & Hoque, 2018). All three variants of this model performed worse than PepNN on this dataset (Table 1) and notably, the observed performance was drastically lower than the performance reported in the original publication. This could potentially be due to the fact a smoothing approach was used to annotate binding sites in the PBRpredict-Suite study (Iqbal & Hoque, 2018), while binding site residues annotations were made based only on distance to peptide residues in this study.

Most other existing approaches lack programmatic access and a portion rely on alignments to reference datasets that overlap with the test set. We hence used values reported in the literature for comparison. To ensure an unbiased comparison, the model was re-trained on the training datasets used in different studies prior to comparison on their test sets. In all cases, PepNN-Struct largely outperforms existing approaches in terms of ROC AUC (Table 1). In most cases, PepNN-Seq also outperforms existing approaches by this metric. PepNN does, however, perform worse in terms of MCC in a couple of cases, suggesting that there exist thresholds at which the models do not perform was well as the PepBind approach, despite having more robust performance at different prediction thresholds. It is worth noting that the training datasets used in other studies were substantially smaller and thus training on them resulted in lower performance of our models overall (Table 1). This was both due to the fact that the datasets used in other studies are relatively outdated and that a larger portion of the available data was used for testing.

Peptide-agnostic prediction allows the identification of putative novel peptide binding proteins

To quantify the extent to which the model relies on information from the protein when making predictions, we tested the ability of PepNN-Struct and PepNN-Seq to predict peptide binding sites using random length poly-glycine peptides as input sequences. While the models did perform better when given the native peptide sequence than with a poly-glycine sequence (p-value < 2.2e-16 for both PepNN-Struct and PepNN-Seq, DeLong test), there was only a small overall decrease in the ROC AUC when a poly-glycine was given (Fig 3A, B). Comparing the probabilities that the model assigns to different residues shows that in both the case of PepNN-Struct and PepNN-Seq, providing the native peptide increases the model’s confidence when predicting binding residues (Fig S3). Providing the native peptide sequence is thus more important for reducing false negatives. Overall, these results suggest that while providing a known peptide can increase model accuracy, the model can make reasonable peptide-agnostic predictions and could potentially be used to identify novel peptide binders.

To quantify the model's confidence that a protein is a peptide-binding module, we generated a score that takes into account the binding probabilities that the model assigns the residues in the protein, as well as the percentage of residues that the model predicts are binding residues with high confidence. To compute this score, a Gaussian distribution was fit to the distribution of binding residue percentages in each protein from the training dataset (Fig S4A). The resulting score was the weighted average of the top *n* residue probabilities and the likelihood that a binding site would be composed of those *n* residues based on the aforementioned Gaussian. For each protein, *n* was chosen to maximize the score. As done in a previous study (Johansson-Åkhe *et al*, 2019), the weight assigned to each component of the score was chosen to maximize the correlation between the MCC of the prediction for each protein in the validation dataset, and its score (Fig S4B,C). This was motivated by the fact that the confidence of the model should correlate with its correctness.

We used the models to predict binding sites for domains in every unique chain in the PDB not within 30% homology of a sequence in the training dataset and domains in every sequence in the reference human proteome from UniProt (Consortium, 2018), not within 30% homology of a sequence in the training dataset. Domains were extracted by assigning PFAM (Finn *et al*, 2013) annotations using InterProScan (Jones *et al*, 2014) (Table S1, S2). To assess the capacity of the models to discriminate between peptide binding modules and other domains, we compared the distribution of scores for canonical PRMs to that of other proteins. Previously defined modular protein domains (Jadwin *et al*, 2012), and peptide binding domains

(Cunningham *et al*, 2020) were considered canonical PRMs. In both the case of the PDB and the human proteome, the distribution of scores for canonical PRMs was higher than the background distribution (Fig 3C, D).

In total, PepNN-Struct assigns 39 623 domains in the PDB a score higher than the mean PRM score and PepNN-Seq assigns 10 332 domains in the human proteome a score higher than the mean PRM score. Analysis of the distribution of scores for different domains reveals that many DNA binding domains, including different transcription factors and DNA modifying enzymes, were assigned low scores on average by PepNN (Table S3, S4). This indicates that PepNN has the capacity to discriminate between different types of binding sites. There are, nonetheless, some nucleic acid binding domains with high scores (Table S3, S4) suggesting that there are false positives and that downstream computational and experimental work is required to validate putative peptide binding sties.

One domain identified by PepNN-Struct is the sterile alpha motif (SAM) domain of the Deleted-in-liver cancer 1 (DLC1) protein (Table S1). This domain was recently shown to be a peptide binding module (Joshi *et al*, 2020), demonstrating the capacity of the model to identify novel peptide binders. Another interesting hit identified using PepNN-Struct is the ORF7a accessory protein from the SARS-Cov-2 virus (Table S1). The model predicts that this protein has a peptide binding site located between two beta-sheets at the N-terminal end of the protein (Fig 4A). Validating this peptide binding site involves identifying a binding peptide and showing that the residues that comprise the binding site are necessary for the interaction. The ORF7a homolog from SARS-Cov has been shown to bind the ectodomain of the human BST-2 protein (Taylor *et al*, 2015). BST-2 binds and tethers viral particles to the cell membrane, thereby preventing viral exit (Taylor *et al*, 2015). It was shown that by binding BST-2, ORF7a prevents its glycosylation and thus reduces its ability to inhibit viral exit (Taylor *et al*, 2015). Given the fact that BST-2 forms a coiled-coil structure, it is possible that a linear segment along one of its helices binds to ORF7a at the predicted peptide-binding pocket.

As a preliminary, unbiased, test of this prediction, we performed global docking of BST-2 onto ORF7a using the ClusPro webserver (Kozakov *et al*, 2017; Vajda *et al*, 2017). In seven of the top ten poses, BST-2 was found to interact with ORF7a at the predicted binding site. In four of these poses, the N70 residue on BST-2, a known glycosylation site (Wollscheid *et al*, 2009), was completely buried. To validate these docking results, those four systems were subject to short, 50 ns, MD simulations. ORF7a was stably bound to BST-2 in one of the four systems. To better evaluate this putative binding conformation at a longer time scale, a truncated system was built and it was subjected to three simulations of at least 200 ns. ORF7a remained bound to BST-2 throughout the different trajectories (Fig 4B), and hydrogen bond analysis showed that several charged/polar sidechains at the interface contribute to the majority of the binding affinity (Fig 4C).

Application of PepNN to epitope prediction

The binding of antibodies to their target antigens is largely facilitated by a set of variable segments known as complementarity-determining regions (CDRs). It has been shown that synthetic peptides derived from the sequences of these CDRs can bind the target antigen of the antibody from which they were derived (Williams *et al*, 1991, 1988; Taub *et al*, 1989). We thus re-trained PepNN to predict the binding sites of different CDRs given an antigen structure. The estimated interface energy of peptide-protein complexes is greater than that of CDR-protein complexes (Fig S5). The pre-training dataset was consequently remade with less stringent thresholds (see Methods). We also ensured that fragments forming helix or strand secondary structures were filtered from the pre-training dataset. We trained the model to predict binding sites given H1, H2, and H3 loops. To generate a full epitope prediction, we assigned each residue the maximum score from the three models. Overall, the observed performance was worse than that on peptide binding site prediction (Fig S6A). Nevertheless, the model makes reasonable predictions on numerous test antigens (Fig S6B).

Discussion

We have developed parallel structure and sequence-based models for the prediction of peptide binding sites. These models, PepNN-Struct and PepNN-Seq, make use of a novel attention-based deep learning module that is integrated with transfer learning to compensate for the scarcity of peptide-protein complex data. Comparison to existing approaches shows that PepNN achieves state-of-the-art on the task of identifying peptide binding sites. In addition, unlike previously developed approaches, PepNN does not rely on structural or sequence alignments and is thus not dependent on the presence of structural data for homologs. Given the success of this approach, PepNN can be incorporated into local docking pipelines in order to facilitate the generation of protein-peptide complex models, a necessary step in delineating the molecular mechanisms underlying many cellular processes.

We furthermore demonstrated that PepNN can make accurate peptide-agnostic predictions. This observation is concordant with recent work that has suggested that a protein's surface contains the majority of information regarding its capacity for biomolecular interactions (Gainza *et al*, 2020). Other approaches, trained on negative binding data, are better suited than PepNN to discriminate between identified binding peptides (Lei *et al*, 2020; Cunningham *et al*, 2020). By contrast, PepNN can uniquely be used to score proteins lacking a known peptide ligand to predict their ability to bind peptides. Running this procedure on all proteins in the PDB and the reference human proteome revealed a number of putative novel peptide recognition modules, suggesting that a large portion of the space of PRMs has yet to be characterized. As a demonstration of the model's capacity to identify novel peptide binders, we performed MD simulations on putative ORF7a/BST-2 complexes, suggesting that the former protein can potentially bind a linear fragment of BST-2 at a predicted peptide binding site. The observation that PepNN can make predictions in the absence of a known peptide binder can also be used to discern regions of proteins that can be readily targeted by peptides. PepNN predictions can thus be used to inform the application of high-throughput experimental approaches to different proteins for the purpose of identifying therapeutic peptides.

Materials and Methods

Datasets

A dataset of protein-peptide complexes was generated by filtering complexes in the PDB. Crystal structures with a resolution of at least 2.5 Å that contain a chain of at least 50 amino acids in complex with a chain of 25 or less amino acids were considered putative peptide-protein complexes. Using FreeSASA (Mitternacht, 2016), complexes with a buried surface area of less than 400 Å^2 were filtered out, leaving 3046 complexes. The sequences of the receptors in the remaining complexes were clustered at a 30% identity threshold using PSI-CD-HIT (Fu *et al*, 2012), and the resulting clusters were divided into training, validation, and test sets at proportions of 80%, 10% and 10% respectively. The test set contains 305 examples and is referred to as TS305.

A similar process was used to generate a dataset of protein fragment-protein complexes. Using the PeptiDerive Rosetta protocol (Sedan *et al*, 2016), the PDB was scanned for protein fragments of length 5-25 amino acids with a high predicted interface energy when in complex with another chain of at least 50 amino acids. Complexes were filtered out based on the distribution of predicted interface energies from the dataset of real protein-peptide complexes. Only complexes with an interface score less than one standard deviation above the mean of the peptide-protein complex distribution were maintained. The complexes were also filtered by buried surface area. Complexes with less than 400 Å^2 were once again filtered out. The final dataset contained 406 365 complexes. For data splitting, complexes were again clustered at 30% identity. In both datasets, binding residues were defined as those residues in the protein receptor with a heavy atom within 6 Å^2 from a heavy atom in the interacting chain.

In addition to TS305, the models were also tested on benchmark datasets compiled in other studies. This includes the test dataset used to evaluate the Interpep approach (Johansson-Åkhe *et al*, 2019) (TS251), the test dataset used to evaluate the PepBind approach (Zhao *et al*, 2018) (TS639), and the test dataset used to evaluate SPRINT-Str (Taherzadeh *et al*, 2017) (TS125).

Complexes from the non-redundant SAbDab dataset were used for training the model to predict epitopes (Dunbar *et al*, 2014). CDRs were defined using the PyIgClassify dataset (Adolf-Bryfogle *et al*, 2015). Complexes where a particular CDR was not in contact with the antigen were filtered out when training the model to predict the binding site of that CDR. Antigen sequences were clustered at 30% identity before splitting the dataset. For pre-training on fragment-protein complexes, less stringent thresholds of a Rosetta interface score of -10 and a buried surface area of 250 Å^2 were used for filtering. In addition, secondary structure annotations were assigned to each fragment in the dataset using the MDTraj software (McGibbon *et al*, 2015), and any fragment with more than two residues in the helix or strand classes were filtered out. The resulting dataset contained 684 912 entries.

Input representation

In the case of PepNN-Struct, input protein structures are encoded using a previously described graph representation (Ingraham *et al*, 2019), with the exception that additional node features are added to encode the side chain conformation at each residue. In this representation, a local coordinate system is defined at each residue based on the relative position of the Cα to the other backbone atoms (Ingraham *et al*, 2019). The edges between residues encode information about the distance between the resides, the relative direction from one Cα to another, a quaternion representation of the rotation matrix between the local coordinate systems, and an embedding of the relative positions of the residues in the protein sequence (Ingraham *et al*, 2019). The nodes include a one-hot representation of the amino acid identity and the torsional backbone angles (Ingraham *et al*, 2019).

To encode information about the side-chain conformation, the centroid of the heavy side chain atoms at each residue is calculated. The direction of the atom centroid from the Cα is represented using a unit vector based on the defined local coordinate system. The distance is encoded using a radial basis function, similar to the encoding used for inter-residue distances in the aforementioned graph representation (Ingraham *et al*, 2019). A one-hot encoding is used to represent protein and peptide sequence information. The pre-trained contextualized language model, ProtBert (Elnaggar *et al*, 2020), is used to embed the protein sequence in PepNN-Seq.

Model architecture

The developed architecture takes inspiration the original Transformer architecture (Vaswani *et al*, 2017), as well the Structured Transformer, developed for the design of proteins with a designated input structure (Ingraham *et al*, 2019). Like these models, the PepNN architecture consists of repeating attention and feed forward layers (Fig 1A). PepNN differs from conventional Transformers, however, in that does not follow an encoder-decoder attention architecture and it makes use of multi-head reciprocal attention. This is a novel attention-based module that shares some conceptual similarity to a layer that was recently used for salient object detection (Xia *et al*, 2019). Conventional scaled dot attention, mapping queries, represented by matrix Q, and key-value pairs, represented by matrices K and V, to attention values takes the following form (Vaswani *et al*, 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In reciprocal attention modules, protein residue embeddings are projected to a query matrix, $Q \in \mathbb{R}^{\mathrm{n}\times\mathrm{d_k}}$ and a value matrix, $\mathrm{V_{prot}} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d_v}}$, where n is the number of protein residues. Similarily, the peptide residue embeddings are projected a key matrix, $K \in \mathbb{R}^{m\times d_k}$, and a value matrix, $V_{pep} \in \mathbb{R}^{m\times d_v}$, where m is the number of peptide residues. The resulting attention values are as follows:

$$\text{Attention}_{prot}(Q, K, V_{pep}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V_{pep}$$

$$\text{Attention}_{pep}(Q, K, V_{prot}) = \text{softmax}\left(\frac{KQ^T}{\sqrt{d_k}}\right)V_{prot}$$

Projecting the residue encodings multiple times and concatenating the resulting attention values allows extension to multiple heads, as described previously (Vaswani *et al*, 2017). The overall model architecture includes alternating self-attention and reciprocal attention layers, with a final set of layers to project the protein residue embedding down to a residue-wise probability score (Fig 1A). For the purpose of regularization, dropout layers were included after each attention layer.

Model hyperparameters were tuned using random search to optimize the cross-entropy loss on the fragment complex validation dataset. Specifically, eight hyperparameters were tuned; d_{model} (the model embedding dimension), d_i (the dimension of the hidden layer in the feed forward layers), d_k, d_v, the dropout percentage, the number of repititions of the reciprocal attention module, the number of heads in each attention layer, and the learning rate. In total, 100 random hyperparameter trials were attempted. d_{model} was set to 64, d_i was set to 64, d_k was set to 64, $d_v$was set to 128, dropout percentage was set to 0.2, the number of repetitions of the reciprocal attention module was set to 6, and each multi-head attention layer was composed of 6 heads.

Training

Training was done using an Adam optimizer with a learning rate of 1e-4. A weighted cross-entropy loss was optimized to take into account the fact that the training dataset is skewed towards non-binding residues. In both the pre-training step with the fragment complex dataset and the training with the peptide complex dataset, early stopping was done based on the validation loss. Training was at most 500 000 iterations during the pre-training step and the at most 25 000 iterations during the fine-tuning step.

Scoring potential novel peptide binding sites

Peptide-agnostic prediction of proteins in the human proteome and the PDB was performed by providing the model with a protein sequence/structure and a poly-glycine sequence of length 10 as the peptide. When computing scores using PepNN-Struct, the weight given to the top n residue probabilities was 0.97. When computing scores using PepNN-Seq, this weight was set to 0.99. Pairwise comparisons were done with the distributions of every PFAM domain to remaining domains with a Wilcoxon rank-sum test and multiple testing correction was done using the Benjamini-Hochberg procedure.

Statistical tests

Wilcoxon signed-rank and rank-sum tests were done using the SciPy python library (Virtanen *et al*, 2020). Multiple testing correction was done using the statsmodels python package (Seabold & Perktold, 2010). The DeLong test was done using the pROC R package (Robin *et al*, 2011).

Protein-protein docking and molecular dynamics simulations on ORF7a/BST-2

The structure of the SARS-CoV-2 ORF7a encoded accessory protein (PDB ID 6W37) and mouse BST-2/Tetherin Ectodomain (PDB ID 3NI0 (Swiecki *et al*, 2011)) were used as input structures for the ClusPro webserver (Kozakov *et al*, 2017; Vajda *et al*, 2017). The top 10 results, ranked by binding affinity, were retrieved for further analysis. The ClusPro docking poses of the ORF7a/BST-2 complex were directly used as input to the Charmm-gui webserver (Brooks *et al*, 2009; Jo *et al*, 2008; Lee *et al*, 2016) to set up MD systems. The systems have a size of approximately 1803 Å^3 and a total of ~570,000 atoms. To speed up the simulation, a truncated system was also created. Amino acids after residue 100 in BST-2 were removed, resulting in a system of size ~1003 Å^3 and approximately 91,300 atoms. The energy minimization and MD simulations were performed with the GROMACS program (Pronk *et al*, 2013) version 2019.3 GPU using the CHARMM36 force field (Klauda *et al*, 2010; Huang & MacKerell Jr, 2013) and TIP3P water model (Jorgensen *et al*, 1983).

Code and data availability

The datasets used in this study and the code to run PepNN are available at https://gitlab.com/oabdin/pepnn.

References

Agrawal P, Singh H, Srivastava HK, Singh S, Kishore G & Raghava GPS (2019) Benchmarking of different molecular docking methods for protein-peptide docking. *BMC Bioinformatics* 19: 426

Brooks BR, Brooks III CL, Mackerell Jr. AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, *et al* (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30: 1545–1614

Ciemny M, Kurcinski M, Kamel K, Kolinski A, Alam N, Schueler-Furman O & Kmiecik S (2018) Protein–peptide docking: opportunities and challenges. *Drug Discov Today* doi:10.1016/j.drudis.2018.05.006 [PREPRINT]

Consortium TU (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47: D506–D515

Cunningham JM, Koytiger G, Sorger PK & AlQuraishi M (2020) Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat Methods* 17: 175–183

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, *et al* (2020) ProtTrans: Towards Cracking the Language of Life{\textquoteright}s Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv*

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al* (2013) Pfam: the protein families database. *Nucleic Acids Res* 42: D222–D230

Fu L, Niu B, Zhu Z, Wu S & Li W (2012) CD-HIT: Accelerated for clustering the nextgeneration sequencing data. *Bioinformatics*

Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM & Correia BE (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 17

Hagai T, Azia A, Babu MM & Andino R (2014) Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. *Cell Rep* 7: 1729–1739

Huang J & MacKerell Jr AD (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 34: 2135–2145

Ingraham J, Garg VK, Barzilay R & Jaakkola T (2019) Generative models for graph-based protein design. In *Deep Generative Models for Highly Structured Data, DGS@ICLR 2019 Workshop*

Iqbal S & Hoque MT (2018) PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence. *Bioinformatics* 34: 3289–3299

Jadwin JA, Ogiue-Ikeda M & Machida K (2012) The application of modular protein domains in proteomics. *FEBS Lett* 586: 2586–2596

Jo S, Kim T, Iyer VG & Im W (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *J Comput Chem* 29: 1859–1865

Johansson-Åkhe I, Mirabello C & Wallner B (2019) Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Sci Rep* 9

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al* (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics*

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW & Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79: 926–935

Joshi R, Qin L, Cao X, Zhong S, Voss C, Min W & Li SSC (2020) {D}{L}{C}1 {S}{A}{M} domain-binding peptides inhibit cancer cell growth and migration by inactivating {R}ho{A}. *J Biol Chem* 295: 645–656

Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD & Pastor RW (2010) Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J Phys Chem B* 114: 7830–7843

Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, Beglov D & Vajda S (2017) The ClusPro web server for protein--protein docking. *Nat Protoc* 12: 255–278

Krumm BE & Grisshammer R (2015) Peptide ligand recognition by G protein-coupled receptors. *Front Pharmacol* 6: 48

Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, Wei S, Buckner J, Jeong JC, Qi Y, *et al* (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* 12: 405–413

Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, Zhao D & Zeng J (2020) CAMP: a Convolutional Attention-based Neural Network for Multifaceted Peptide-protein Interaction Prediction. *bioRxiv*

London N, Raveh B, Movshovitz-Attias D & Schueler-Furman O (2010) Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? *Proteins Struct Funct Bioinforma*

London N, Raveh B & Schueler-Furman O (2012) Modeling Peptide--Protein Interactions. In

Homology Modeling: Methods and Protocols, Orry AJW & Abagyan R (eds) pp 375–398. Totowa, NJ: Humana Press

Mitternacht S (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK & Uversky VN (2006) Analysis of Molecular Recognition Features (MoRFs). *J Mol Biol* 362

Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, *et al* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29: 845–854

Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny JF, Abbeel P & Song YS (2019) Evaluating Protein Transfer Learning with {TAPE}. *CoRR* abs/1906.0

Raveh B, London N & Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins Struct Funct Bioinforma* 78: 2029–2040

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C & Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12: 77

Seabold S & Perktold J (2010) Statsmodels: econometric and statistical modeling with Python. In *9th Python in Science Conference*

Sedan Y, Marcu O, Lyskov S & Schueler-Furman O (2016) Peptiderive server: derive peptide inhibitors from protein-protein interactions. *Nucleic Acids Res*

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, *et al* (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577: 706–710

Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A & Kim PM (2020) Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst* 11: 402-411.e4

Swiecki M, Scheaffer SM, Allaire M, Fremont DH, Colonna M & Brett TJ (2011) Structural and biophysical analysis of BST-2/tetherin ectodomains reveals an evolutionary conserved design to inhibit virus release. *J Biol Chem* 286: 2987–2997

Taherzadeh G, Yang Y, Zhang T, Liew AW-C & Zhou Y (2016) Sequence-based prediction of protein–peptide binding sites using support vector machine. *J Comput Chem* 37: 1223–1229

Taherzadeh G, Zhou Y, Liew AW-C & Yang Y (2017) Structure-based prediction of protein–peptide binding regions using Random Forest. *Bioinformatics* 34: 477–484

Taherzadeh G, Zhou Y, Liew AWC & Yang Y (2018) Structure-based prediction of protein-peptide binding regions using Random Forest. *Bioinformatics* 34

Taylor JK, Coleman CM, Postel S, Sisk JM, Bernbaum JG, Venkataraman T, Sundberg EJ & Frieman MB (2015) Severe Acute Respiratory Syndrome Coronavirus ORF7a Inhibits Bone Marrow Stromal Antigen 2 Virion Tethering through a Novel Mechanism of Glycosylation Interference. *J Virol* 89: 11820–11833

Tompa P, Davey NE, Gibson TJ & Babu MM (2014) A Million peptide motifs for the molecular biologist. *Mol Cell* doi:10.1016/j.molcel.2014.05.032 [PREPRINT]

Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR & Kozakov D (2017) New additions to the ClusPro server motivated by CAPRI. *Proteins* 85: 435–444

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł & Polosukhin I (2017) Attention is all you need. In *Advances in Neural Information Processing Systems*

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17

Wardah W, Dehzangi A, Taherzadeh G, Rashid MA, Khan MGM, Tsunoda T & Sharma A (2020) Predicting protein-peptide binding sites with a deep convolutional neural network. *J Theor Biol* 496

Weng G, Gao J, Wang Z, Wang E, Hu X, Yao X, Cao D & Hou T (2020) Comprehensive Evaluation of Fourteen Docking Programs on Protein–Peptide Complexes. *J Chem Theory Comput* 16: 3959–3969

Wollscheid B, Bausch-Fluck D, Henderson C, O'Brien R, Bibel M, Schiess R, Aebersold R & Watts JD (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* 27

Xia C, Li J, Su J & Tian Y (2019) Exploring Reciprocal Attention for Salient Object Detection by Cooperative Learning. [PREPRINT]

Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M & Roth FP (2015) Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLOS Comput Biol* 11: 1–30

Zhao Z, Peng Z & Yang J (2018) Improving Sequence-Based Prediction of Protein-Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method. *J Chem Inf Model* 58

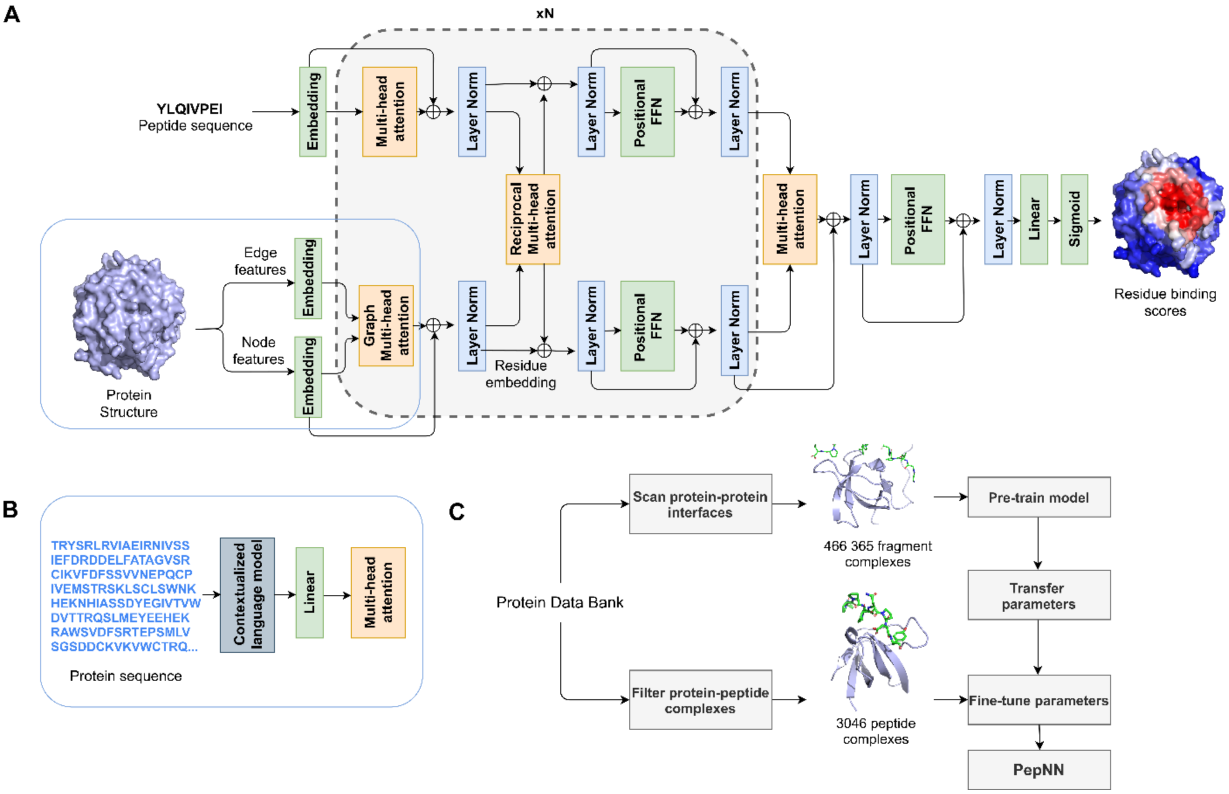


Figure 1: Model architecture and training procedure. A) Attention layers are indicated with orange, normalization layers are indicated with blue and simple transformation layers are indicated with green. B) Input layers for PepNN-Seq. C) Transfer learning pipeline used for model training.

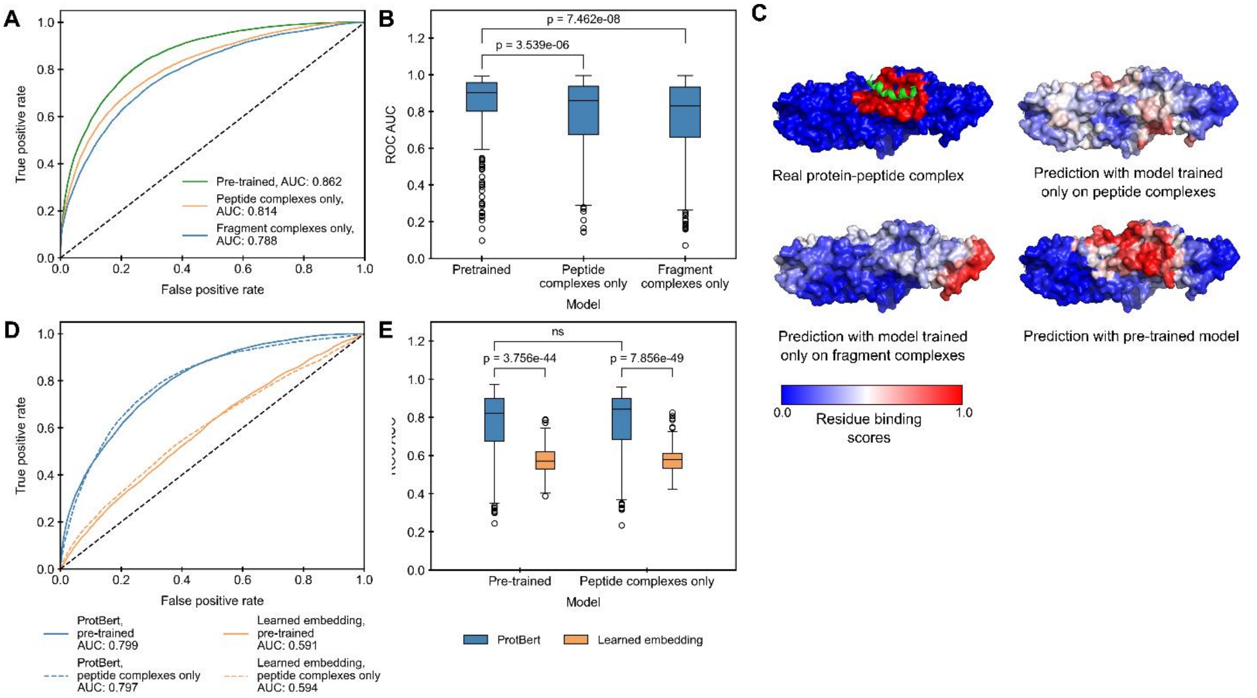


Figure 2: Impact of transfer learning on model performance on the peptide complex validation dataset. A) ROC curves on all residues in the dataset using predictions from PepNN-Struct trained on different datasets. B) Comparison of the distribution of ROC AUCs on different input proteins using predictions from PepNN-Struct with different training procedures and sequence embeddings (Wilcoxon signed-rank test). C) Predictions of the binding site of the Bro domain of HD-PTP (PDB code 5CRV) using PepNN-Struct trained on different datasets. D) ROC curves on all residues in the dataset using predictions from the sequence model with different training procedures and sequence embeddings. E) Comparison of the distribution of ROC AUCs on different input proteins using predictions from PepNN-Seq trained on different datasets (Wilcoxon signed-rank test).

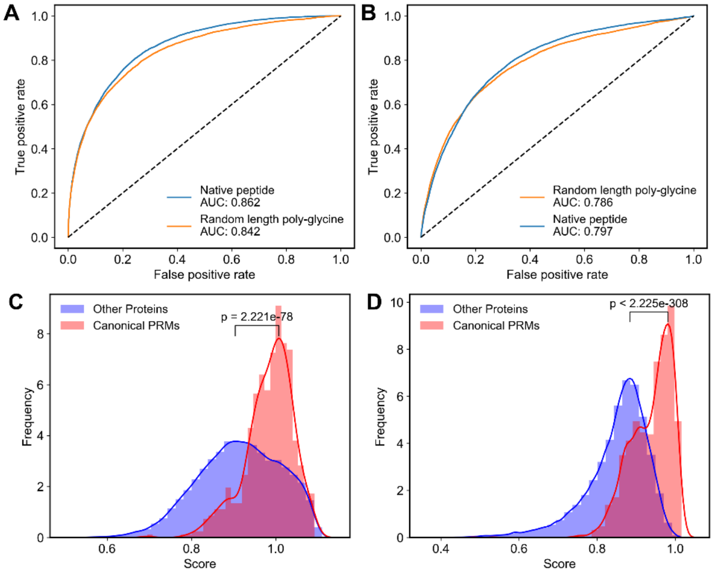


Figure 3: Peptide-agnostic binding site prediction using PepNN-Struct and PepNN-Seq. A) ROC curves on the validation dataset using PepNN-Struct with different input peptide sequences. B) ROC curves on the validation dataset using PepNN-Seq with different input peptide sequences. C) Scores assigned by PepNN-Struct to different domains in the PDB (Wilcoxon rank-sum test). D) Scores assigned by the PepNN-Seq to different domains in the reference human proteome (Wilcoxon rank-sum test).

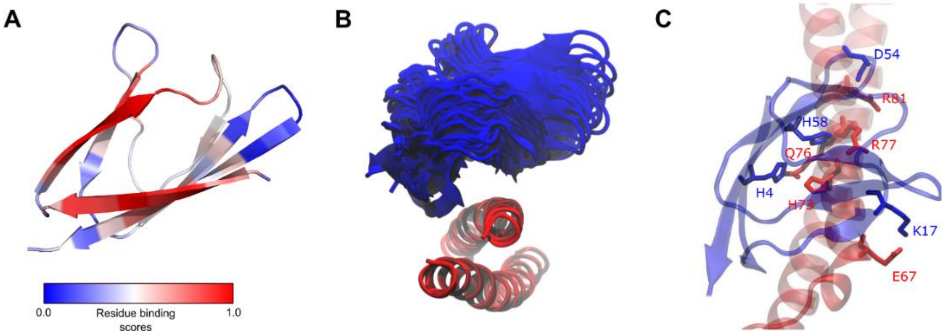
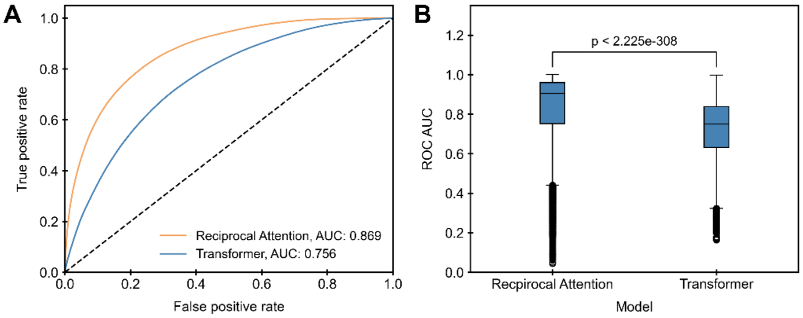


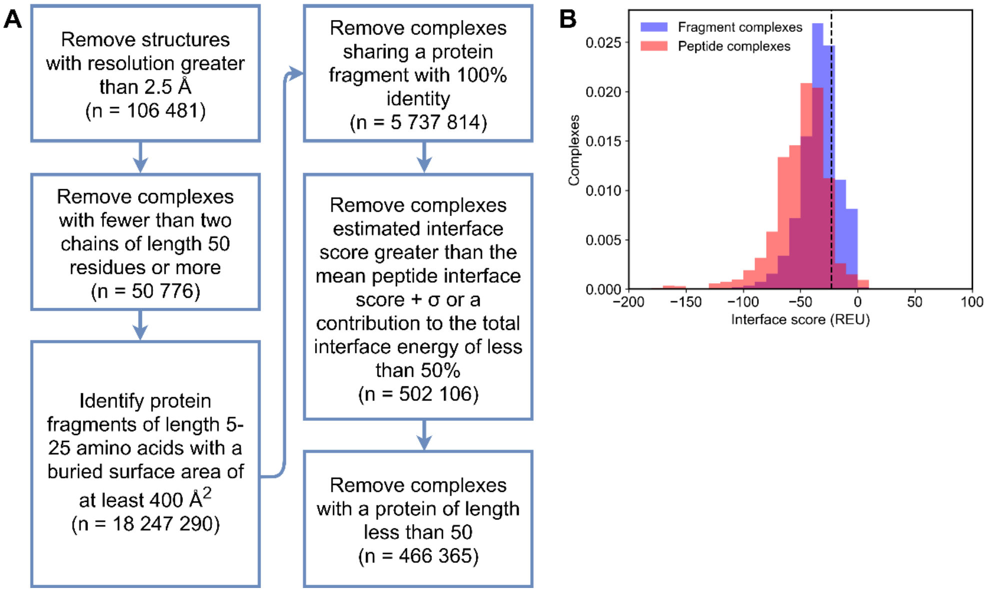
Figure 4: A) ORF7a peptide binding site prediction. B) Ensemble plot of putative ORF7a/BST-2 complex from a 300 ns MD simulation. C) Hydrogen bonds between residues at the BST/ORF7a interface in the predicted complex

Table 1: Comparison of the developed model to existing approaches

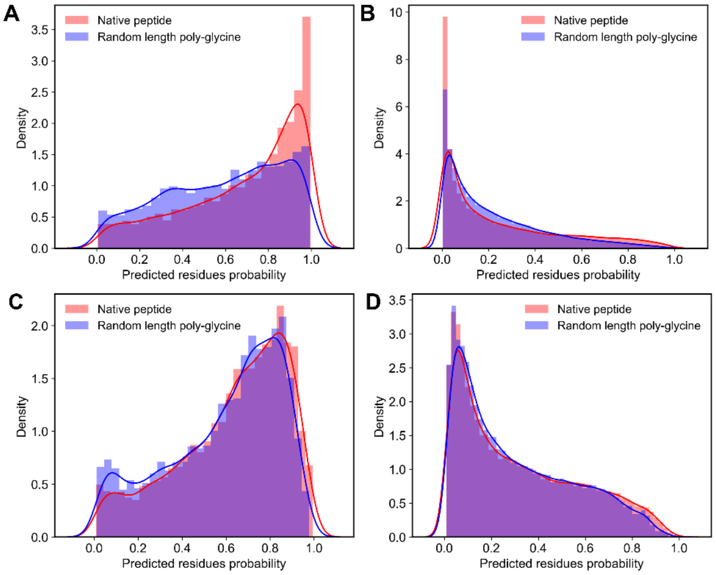
Test dataset	**Training dataset size**	**Model**	**ROC AUC**	**MCC**
TS305	2394	PepNN-Struct	**0.893**	**0.483**
		PepNN-Seq	0.859	0.401
	475	PBRpredict-flexible (Iqbal & Hoque, 2018)	0.653	0.139
		PBRpredict-moderate (Iqbal & Hoque, 2018)	0.620	0.127
		PBRpredict-strict (Iqbal & Hoque, 2018)	0.598	0.100
TS251	251	PepNN-Struct	**0.817**	**0.370**
		PepNN-Seq	0.758	0.278
		Interpep (Johansson-Åkhe *et al*, 2019)	**0.793**	---
TS639	640	PepNN-Struct	**0.838**	0.301
		PepNN-Seq	0.792	0.251
		PepBind (Zhao *et al*, 2018)	0.767	**0.348**
TS125	640	PepNN-Struct	**0.841**	0.321
		PepNN-Seq	0.805	0.278
		PepBind (Zhao *et al*, 2018)	0.793	**0.372**
	1156	SPRINT-Str (Taherzadeh *et al*, 2017)	0.780	0.290
	1199	SPRINT-Seq (Taherzadeh *et al*, 2016)	0.680	0.200
	1004	Visual (Wardah *et al*, 2020)	0.730	0.170



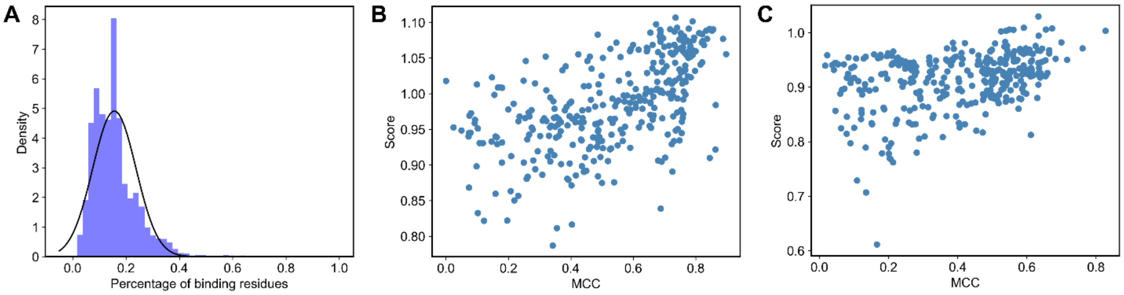
Supplementary Figure 1: Comparison of the performance of the developed model to a Transformer with the same hyperparameters on the fragment complex validation dataset. A) ROC curves on all residues in the dataset. B) Comparison of distribution of ROC AUCs on different input proteins (Wilcoxon signed-rank test).


A
Remove structures with resolution greater than 2.5 Å (n = 106 481)
Remove complexes with fewer than two chains of length 50 residues or more (n = 50 776)
Identify protein fragments of length 5-25 amino acids with a buried surface area of at least 400 Å^2 (n = 18 247 290)
Remove complexes sharing a protein fragment with 100% identity (n = 5 737 814)
Remove complexes estimated interface score greater than the mean peptide interface score + σ or a contribution to the total interface energy of less than 50% (n = 502 106)
Remove complexes with a protein of length less than 50 (n = 466 365)
B
Fragment complexes
Peptide complexes
Complexes
0.000
0.005
0.010
0.015
0.020
0.025
−200
−150
−100
−50
0
50
100
Interface score (REU)

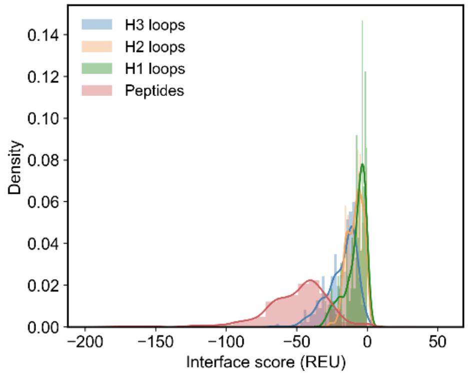
Supplementary Figure 2: A) Curation pipeline for generation of a protein fragment-protein dataset. B) Comparison of estimated interface distribution for the all fragment-protein complexes and the dataset of peptide-protein complexes. The dashed lines indicates the threshold used for filtering fragment-protein complexes.



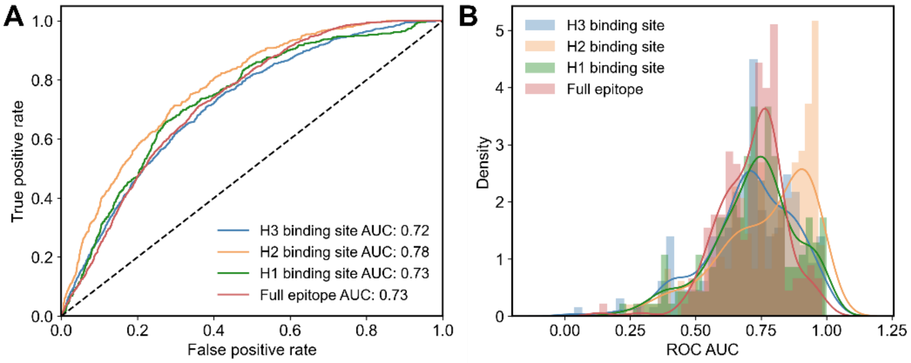
Supplementary Figure 3: Probabilities assigned by PepNN-Struct and PepNN-Seq to different models residues with and without the native peptide sequence. A) Probabilities assigned by PepNN-Struct to binding residues. B) Probabilities assigned by PepNN-Struct to non-binding residues. C) Probabilities assigned by PepNN-Seq to binding residues. D) Probabilities assigned by PepNN-Seq to non-binding residues.



Supplementary Figure 4: A) The percentage of binding residues in different examples in the training dataset. B) Relationship between scores assigned by PepNN-Struct and MCC of predictions. C) Relationship between scores assigned by PepNN-Seq and MCC of predictions.



Supplementary Figure 5: Comparison of estimated interface energies for peptide-protein complexes and CDR-protein complexes.



Supplementary Figure 6: Performance of PepNN on the task of epitope prediction. A) ROC curves on all residues in the test dataset. B) Distribution of ROC AUC for different input antigens from the test set.