Sequence and structure based deep learning models for the identification of peptide binding sites

Osama Abdin[*]
University of Toronto
Toronto ON, M5S3E1
osama.abdin@kimlab.org

Han Wen[†]
University of Toronto
Toronto ON, M5S3E1
han.wen@kimlab.org

Philip M. Kim[*, †, ‡]
University of Toronto
Toronto ON, M5S3E1
pi@kimlab.org

Abstract

Protein-peptide interactions play a fundamental role in facilitating many cellular processes. 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. These models make use of a novel reciprocal attention module that simultaneously updates peptide and protein embeddings while enforcing symmetry in the attention values, thereby better reflecting biochemical realities of peptides undergoing conformational changes upon binding. To compensate for the scarcity of peptide-protein complex structural information, we used transfer learning in two ways; pre-training on a large dataset derived from protein-protein complexes, and using a pre-trained contextualized language model to embed protein sequences. On an independent test set, PepNN-Struct achieved an area under the ROC curve (ROC AUC) of 0.893 and a Matthews correlation coefficient (MCC) of 0.483. The ROC AUC and MCC of PepNN-Seq on the same dataset were 0.859 and 0.401 respectively. The models were furthermore tested on benchmark datasets from recent studies and PepNN-Struct resulted in up to a 9.3% increase in ROC AUC relative to the best performing existing approaches. Beyond prediction of binding sites on proteins with a known peptide ligand, we also showed that the developed models can make reasonable peptide-agnostic predictions, allowing for the identification of novel peptide binding proteins. One identified putative novel peptide binding module is the ORF7a accessory protein from Sars-Cov-2. Molecular dynamics simulations suggest that a linear segment of the Bone marrow stromal antigen 2 (BST-2) human protein can indeed stably bind to the predicted binding site.

1 Introduction

Interactions between proteins and peptides are critical for a variety of biological processes. These interactions are especially prevalent in signal transduction pathways and include the binding of peptide ligands to extracellular receptors [1], as well as the binding of intracellular peptide recognition

[*]Department of Molecular Genetics
[†]Donnelly Centre for Cellular and Biomolecular Research
[‡]Department of Computer Science

Machine Learning for Structural Biology Workshop, NeurIPS 2020, Vancouver, Canada.

modules (PRMs) to linear segments in other proteins [2]. Modifying these interactions and their regulation consequently has implications for disease. Many proteins with PRMs encode sites of oncongenic mutations [3]. It has also been shown that viral proteins encode peptidic motifs that can potentially be used to hijack host machinery during infection [4].

In the absence of experimentally solved structures, gaining molecular insight into these interactions is contingent on the ability to model peptide binding computationally. Traditionally, this has been done using peptide-protein docking, which involves sampling from the conformational space of a protein and an interacting peptide, and evaluating each conformation using a function based on geometry or electrostatics [5]. 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 [6]. In general, benchmarking studies have shown that peptide docking approaches often fail to accurately identify the native complex conformation [7, 8, 9]. These approaches are limited by the high flexibility of peptides as well the inherent error of scoring heuristics [5]. Machine learning approaches have significant potential as alternatives to docking, as they can sidestep the issue of explicit enumeration of conformational space and can learn scoring metrics directly from the data.

Both random forest models and support vector machines (SVMs) have been applied with some success to the preliminary problem of predicting the binding sites of peptides [10, 11, 12, 13]. While contemporary deep learning approaches have resulted in large improvements to multiple problems in the domains of protein and structural biology, larger models have had limited success on this task [14], likely due to the the paucity of available structural protein-peptide complex data. Based on this, we sought to incorporate transfer learning with a modern deep learning architecture to improve upon existing approaches. In particular we sought to exploit the large amount of available protein-protein complex information. 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 [15]. Complexes of protein fragments with receptors thus represent a natural source of data for model pre-training.

Recently, the idea of pre-training contextualized language models has been adapted to protein biology for the purpose of generating meaningful representations of protein sequences [16, 17]. The success of these approaches provides an opportunity to develop a strictly sequence based peptide binding site predictor. By integrating the use of contextualized-language models, available protein-protein complex data, and a task-specific attention based architecture, we developed parallel models for both structure and sequence based peptide binding site prediction; PepNN-Struct and PepNN-Seq. Comparison to existing approaches revealed that our models perform better in most cases. The developed models can make reasonable peptide agnostic predictions, allowing for their use for the identification of novel peptide binding sites.

1.1 Related work

Like the approach developed in this study, the Interpep approach integrates machine learning (random forest classifiers) with available protein-protein complex information to improve peptide binding site prediction [10]. Unlike this approach, however, Interpep relies on explicit structural alignments with receptors from protein-protein complexes, inherently limiting the approach to complexes with structural homologs in the Protein databank (PDB). PepBind is another approach that uses sequence and structure based alignment to generate features as input to an SVM [11]. Other machine learning approaches have been developed for this task, but with considerably less success [13, 12, 14]. An approach has also been previously developed to elucidate the basis of peptide-protein binding at the molecular level by learning graph representations of inter-molecular amino acid dependencies [18]. While this study focused specifically on PDZ domains, the use of graphs to model inter-molecular dependencies is similar to the use of attention in this study to relate peptide and protein residues.

While more modern machine learning approaches have shown limited success on the task of identifying peptide binding sites, recent studies have applied such approaches to related tasks. A novel machine learning method was developed for the task of predicting peptide-protein interactions [19]. A deep learning approach has also been developed to extract information about potential biomolecular interactions from a protein's surface [20]. This approach was applied to the identification of sites of protein-protein interactions, as well as the interactions between proteins and small molecules [20]. Deep learning approaches using graph convolutional layers and attention modules have furthermore been used to rank protein-protein and protein-peptide docks [21, 22, 23].

2 Methods

2.1 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 [24], 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 [25], 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 will hereafter be referred to as TS305.

A similar process was used to generate a dataset of protein fragment-protein complexes. Using the PeptiDerive Rosetta protocol [26], 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 furthermore filtered by buried surface area. Complexes with less than 400 $Å^2$ were once again filtered out. The final dataset contained 406 365 complexes. In both datasets, binding residues were defined as those residues in the protein receptor with a heavy atom within 6 Å 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 [10] (TS251), the test dataset used to evaluate the PepBind approach [11] (TS639), and the test dataset used to evaluate SPRINT-Str [12] (TS125).

2.2 Input representation

In the case of PepNN-Struct, input protein structures are encoded using a previously described graph representation [27], 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 [27]. 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 [27]. The nodes include a one-hot representation of the amino acid identity and the torsional backbone angles [27].

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 [27]. A one-hot encoding is used to represent protein and peptide sequence information. The pre-trained contextualized language model, ProtBert [17], is used to embed the protein sequence in PepNN-Seq.

2.3 Model architecture

The developed architecture takes inspiration the original *Transformer* architecture [28], as well the *Structured Transformer*, developed for the design of proteins with a designated input structure [27]. The model uses multi-head attention layers developed in the former to encode peptide and protein sequence context, and graph attention modules developed in the latter to encode structural context (Figure 1A, B). The model introduced here differs, however, in the fact that it does not follow an encoder-decoder architecture. This is based on the fact that 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 [29]. A peptide's sequence is thus insufficient to determine its conformation in a particular system.

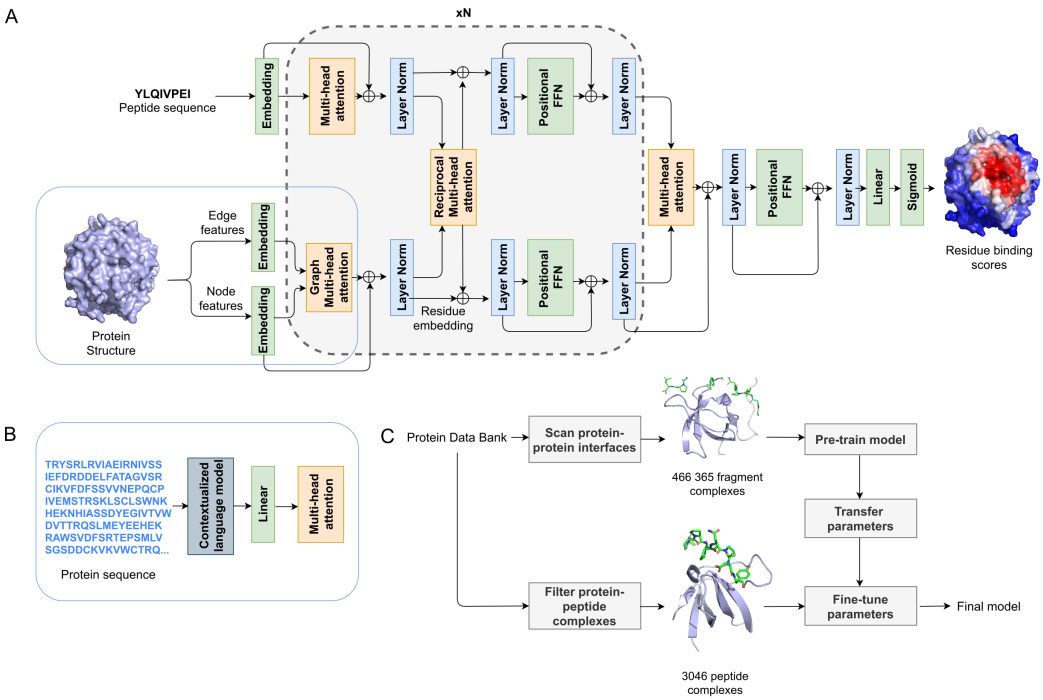

Figure 1: Model architecture and training procedure. A) Attention layers are indicated with orange, normalization layers are indicated with blue and simple transformation layers are indicated with green. B) Input layers for PepNN-Seq. C) Transfer learning pipeline used for model training.

As an alternative, we introduced multi-head reciprocal attention layers, a novel attention-based module with some similarity to a layer that was recently used for salient object detection [30]. This module 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. Consequently, symmetry is enforced in the updates of the protein and peptide embeddings. Scalar dot product attention, mapping queries, represented by matrix Q, and key-value pairs, represented by matricies K and V, to attention values typically takes the following form:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \text{ [28]}$$

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

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

Projecting the residue encodings multiple times and concatenating the resulting attention values allows extension to multiple heads, as described previously [28]. 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 (Figure 1A). For the purpose of regularization, dropout layers were included after each attention layer.

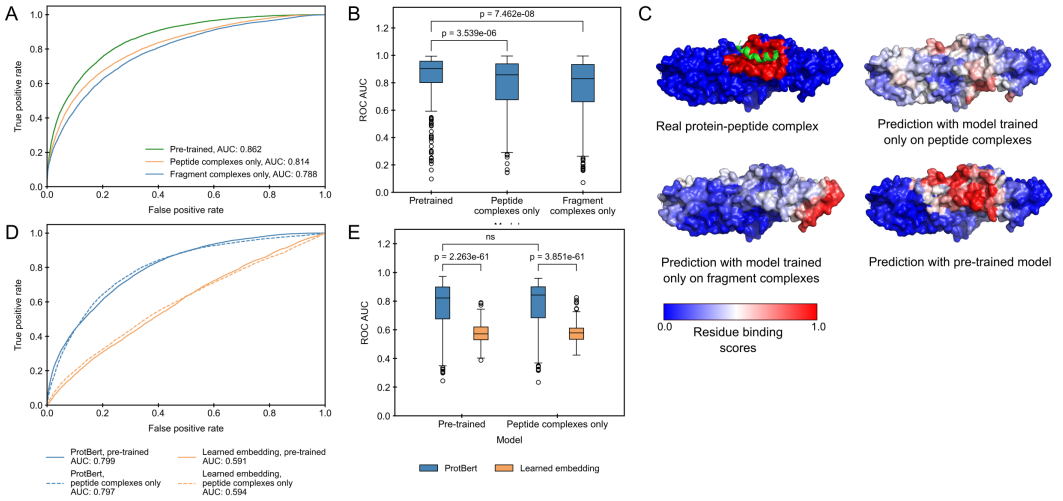

Figure 2: Impact of transfer learning on model performance on the peptide complex validation dataset. A) ROC curves on all residues in the dataset using predictions from PepNN-Struct trained on different datasets. B) Comparison of the distribution of ROC AUCs on different input proteins using predictions 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).

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.

2.4 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.

2.5 Scoring potential novel peptide binding sites

Peptide-agnostic prediction was performed by providing the model with a protein sequence/structure and a poly-glycine sequence of length 10 as the peptide. To quantify the model's confidence that a protein is a peptide-binding module, a score was generated 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. This score quantifies the likelihood that a protein contains a site that could reasonably bind a peptide ligand.

To compute this score, a Gaussian distribution was fit to the distribution of binding residue percentages in each protein from the training dataset. 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

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
TS251	251	PepNN-Struct	**0.817**	**0.370**
		PepNN-Seq	0.758	0.278
		Interpep [10]	0.793	—
TS639	640	PepNN-Struct	**0.838**	0.301
		PepNN-Seq	0.792	0.251
		PepBind [11]	0.767	**0.348**
TS125	640	PepNN-Struct	**0.841**	0.321
		PepNN-Seq	0.805	0.278
		PepBind [11]	0.793	**0.372**
	1156	SPRINT-Str [12]	0.780	0.290
	1199	SPRINT-Seq [13]	0.680	0.200
	1004	Visual [14]	0.730	0.170

based on the aforementioned Gaussian. For each protein, n was chosen to maximize the score. On the basis that the score should correlate with the correctness of model predictions, the weight assigned to each component of the score was chosen so that the correlation between the MCC of each protein in the validation dataset and its score was maximized.

2.6 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 [31]) were used as input structures for the ClusPro webserver [32, 33]. 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 [34, 35, 36] 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 [37] version 2019.3 GPU using the CHARMM36 force field [38, 39] and TIP3P water model [40].

3 Results

3.1 Transfer learning significantly improves 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 (Figure 1C). The second was to use a pre-trained contextualized language model, ProtBert [17], to embed protein sequences in PepNN-Seq (Figure 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 (Figure 3A, 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 (Figure 3C).

Embedding protein sequences with ProtBert resulted in large performance improvements over learned embedding parameters for PepNN-Seq (Figure 3D, E). Interestingly, pretraining on the fragment complexes did not have a large impact on PepNN-Seq performance (Figure 3B, D). This may suggest that pre-training on the fragment complexes allows PepNN-Struct to learn reasonable

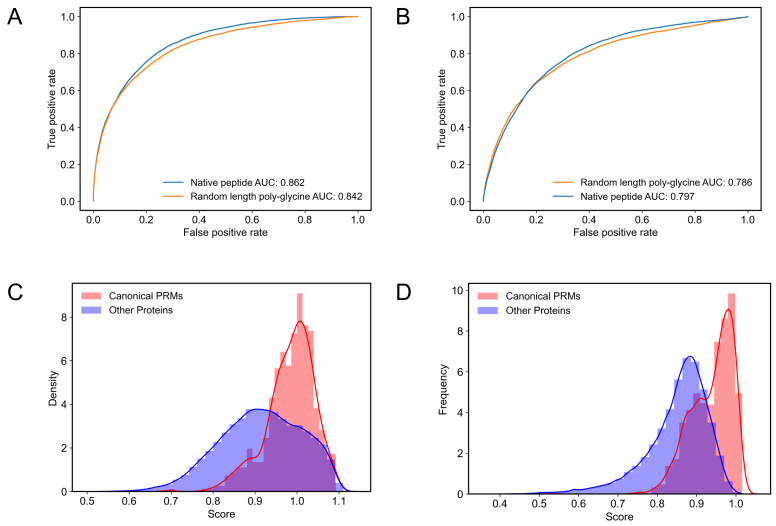

Figure 3: Peptide-agnostic binding site prediction using PepNN-Struct and PepNN-Seq. A) ROC curves on the validation dataset using PepNN-Struct with different input peptide sequences. B) ROC curves on the validation dataset using PepNN-Seq with different input peptide sequences. C) Scores assigned by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

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.

3.2 Comparision to existing approaches

We initially evaluated the developed models on the independent test set derived from the peptide complex dataset. Unsurprisingly, PepNN-Struct outperforms PepNN-Seq (Table 1). For a more unbiased comparison to existing approaches, we also re-trained the models on the datasets used in recently developed machine learning approaches 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. The models do, 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. 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 in these studies.

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

Recent work has suggested that a protein's surface contains the majority of information regarding its capacity for biomolecular interactions [20]. 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. In both the case of PepNN-Struct and PepNN-Seq, replacing the native peptide sequence with a poly-glycine peptide resulted in a slight decrease in performance (Figure 3A, B). This suggests that while providing a known peptide can increase model accuracy, the model can make reasonable peptide-agnostic predictions.

Based on this observation, 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 [41], not within 30% homology of a sequence in the training dataset. Scores were generated for the various domains as described in the

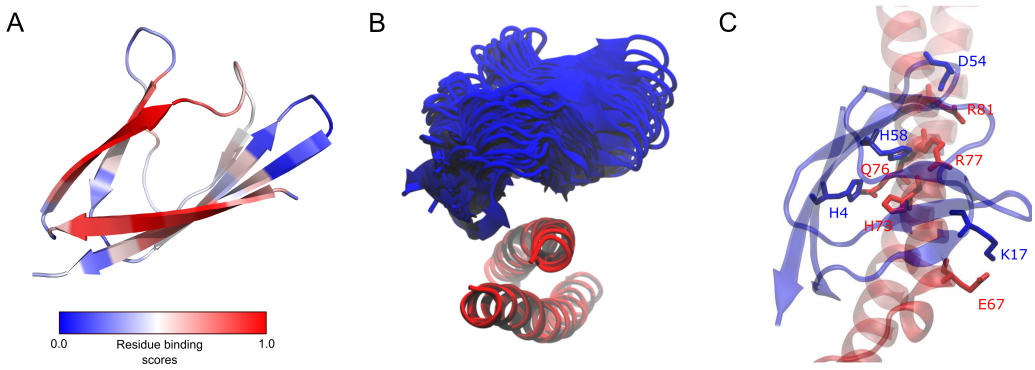

Figure 4: A) ORF7a peptide binding site prediction. B) Ensemble plot of putative ORF7a/BST-2 complex from a 300 ns MD simulation. C) Hydrogen bonds between residues at the BST/ORF7a interface in the predicted complex.

methods. Domains were extracted by assigning PFAM [42] annotations using InterProScan [43]. 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 [44], and peptide binding domains [19] were considered canonical PRMs.

In both the case of the PDB and the human proteome, the model generally assigns higher scores to canonical PRMs than other domains (Figure 3C, D). Nonetheless, there was overlap between the two distributions. Many of the non-canonical proteins that scored highly are other proteins with known peptide-binding capabilities, such as major histocompatibility complexes. 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. One domain identified by PepNN-Struct is the sterile alpha motif (SAM) domain of the Deleted-in-liver cancer 1 (DLC1) protein, which was assigned a very high score (1.095). This domain was recently shown to be a peptide binding module [45], demonstrating the capacity of the model to identify novel peptide binders.

3.4 The ORF7a protein is a potential peptide-binding module

Another interesting hit identifed using PepNN-Struct is the ORF7a accessory protein from the SARS-Cov-2 virus (score of 1.045). The model predicts that this protein has a peptide binding site located between two beta-sheets at the N-terminal end of the protein (Figure 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 [46]. BST-2 binds and tethers viral particles to the cell membrane, thereby preventing viral exit [46]. It was shown that by binding BST-2, ORF7a prevents its glycosylation and thus reduces its ability to inhibit viral exit [46]. 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 [32, 33]. 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 [47], 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 (Figure 4B), and hydrogen bond analysis showed that several charged/polar sidechains at the interface contribute to the majority of the binding affinity (Figure 4C).

4 Conclusions

We developed parallel structural and sequence based models for the prediction of peptide binding sites. This was done by developing of a novel attention based module as well as the use of transfer learning to compensate for the paucity of peptide-protein complex data. The developed model outperformed existing approaches in terms of ROC AUC on multiple test datasets. Unlike previously developed approaches, the model does not rely on structrual or sequence alignments and is thus also more versatile. The model is furthermore capable of making peptide-agnostic predictions. We showed that the model predictions can be turned into a score to quantify the model's confidence that a module can bind peptides. 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 model can thus be used to uncover new biology about proteins as well as identify sites on proteins that can potentially be targetted for therapeutic applications.

Broader Impact

The work presented here has potential applications in both exploratory research and therapeutic design. The ability to identify novel peptide binders can potentially be used to discover new biology mediated by peptide interactions. The model can also be incorporated into local docking pipelines to improve the generation of protein-peptide complex models. This can be used to gain an understanding of the moleular mechanisims underlying various cellular processes. The ability to accurately identify peptide binding sites in a peptide-agnostic matter can furthermore be used to discern regions in a protein that can be readily be targetted by peptides. Model predictions can thus be used to inform the application of experimental approaches such as phage dispay to different proteins for the potential identification of therapeutic peptides.

References

[1] Brian E. Krumm and Reinhard Grisshammer. Peptide ligand recognition by g protein-coupled receptors. *Frontiers in Pharmacology*, 6:48, 2015. ISSN 1663-9812. doi: 10.3389/fphar.2015.00048. URL https://www.frontiersin.org/article/10.3389/fphar.2015.00048.

[2] P. Tompa, N. E. Davey, T. J. Gibson, and M. M. Babu. A million peptide motifs for the molecular biologist. *Mol Cell*, 55(2):161–169, Jul 2014.

[3] Fan Yang, Evangelia Petsalaki, Thomas Rolland, David E. Hill, Marc Vidal, and Frederick P. Roth. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLOS Computational Biology*, 11(3):1–30, 03 2015. doi: 10.1371/journal.pcbi.1004147. URL https://doi.org/10.1371/journal.pcbi.1004147.

[4] Tzachi Hagai, Ariel Azia, M. Madan Babu, and Raul Andino. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Reports*, 7(5):1729 – 1739, 2014. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2014.04.052. URL http://www.sciencedirect.com/science/article/pii/S2211124714003702.

[5] Maciej Ciemny, Mateusz Kurcinski, Karol Kamel, Andrzej Kolinski, Nawsad Alam, Ora Schueler-Furman, and Sebastian Kmiecik. Protein–peptide docking: opportunities and challenges. *Drug Discovery Today*, 23(8):1530 – 1537, 2018. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2018.05.006. URL http://www.sciencedirect.com/science/article/pii/S1359644617305937.

[6] Barak Raveh, Nir London, and Ora Schueler-Furman. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 78(9):2029–2040, 2010. doi: https://doi.org/10.1002/prot.22716. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22716.

[7] Nir London, Barak Raveh, and Ora Schueler-Furman. *Modeling Peptide–Protein Interactions*, pages 375–398. Humana Press, Totowa, NJ, 2012. ISBN 978-1-61779-588-6. doi: 10.1007/978-1-61779-588-6_17. URL https://doi.org/10.1007/978-1-61779-588-6_17.

[8] Piyush Agrawal, Harinder Singh, Hemant Kumar Srivastava, Sandeep Singh, Gaurav Kishore, and Gajendra P. S. Raghava. Benchmarking of different molecular docking methods for protein-peptide docking. *BMC Bioinformatics*, 19(13):426, Feb 2019. ISSN 1471-2105. doi: 10.1186/s12859-018-2449-y. URL https://doi.org/10.1186/s12859-018-2449-y.

[9] Gaoqi Weng, Junbo Gao, Zhe Wang, Ercheng Wang, Xueping Hu, Xiaojun Yao, Dongsheng Cao, and Tingjun Hou. Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. *Journal of Chemical Theory and Computation*, 16(6):3959–3969, 2020. doi: 10.1021/acs.jctc.9b01208. URL https://doi.org/10.1021/acs.jctc.9b01208. PMID: 32324992.

[10] Isak Johansson-Åkhe, Claudio Mirabello, and Björn Wallner. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Scientific Reports*, 9(1):4267, Mar 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-38498-7. URL https://doi.org/10.1038/s41598-019-38498-7.

[11] Zijuan Zhao, Zhenling Peng, and Jianyi Yang. Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *Journal of Chemical Information and Modeling*, 58(7):1459–1468, Jul 2018. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00019. URL https://doi.org/10.1021/acs.jcim.8b00019.

[12] Ghazaleh Taherzadeh, Yaoqi Zhou, Alan Wee-Chung Liew, and Yuedong Yang. Structure-based prediction of protein– peptide binding regions using Random Forest. *Bioinformatics*, 34(3): 477–484, 09 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx614. URL https://doi.org/10.1093/bioinformatics/btx614.

[13] Ghazaleh Taherzadeh, Yuedong Yang, Tuo Zhang, Alan Wee-Chung Liew, and Yaoqi Zhou. Sequence-based prediction of protein–peptide binding sites using support vector machine. *Journal of Computational Chemistry*, 37(13):1223–1229, 2016. doi: 10.1002/jcc.24314. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.24314.

[14] Wafaa Wardah, Abdollah Dehzangi, Ghazaleh Taherzadeh, Mahmood A. Rashid, M.G.M. Khan, Tatsuhiko Tsunoda, and Alok Sharma. Predicting protein-peptide binding sites with a deep convolutional neural network. *Journal of Theoretical Biology*, 496:110278, 2020. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2020.110278. URL http://www.sciencedirect.com/science/article/pii/S0022519320301338.

[15] Nir London, Barak Raveh, Dana Movshovitz-Attias, and Ora Schueler-Furman. Can self-inhibitory peptides be derived from the interfaces of globular protein–protein interactions? *Proteins: Structure, Function, and Bioinformatics*, 78(15):3140–3149, 2010. doi: 10.1002/prot.22785. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22785.

[16] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John F. Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. *CoRR*, abs/1906.08230, 2019. URL http://arxiv.org/abs/1906.08230.

[17] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020. doi: 10.1101/2020.07.12.199554. URL https://www.biorxiv.org/content/early/2020/07/12/2020.07.12.199554.

[18] Hetunandan Kamisetty, Bornika Ghosh, Christopher James Langmead, and Chris Bailey-Kellogg. Learning sequence determinants of protein:protein interaction specificity with sparse graphical models. *Journal of Computational Biology*, 22(6):474–486, 2015. doi: 10.1089/cmb.2014.0289. URL https://doi.org/10.1089/cmb.2014.0289. PMID: 25973864.

[19] Joseph M. Cunningham, Grigoriy Koytiger, Peter K. Sorger, and Mohammed AlQuraishi. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature Methods*, 17(2):175–183, Feb 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0687-1. URL https://doi.org/10.1038/s41592-019-0687-1.

[20] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, Feb 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0666-6. URL https://doi.org/10.1038/s41592-019-0666-6.

[21] Isak Johansson-Åkhe, Claudio Mirabello, and Björn Wallner. Interpeprank: Assessment of docked peptide conformations by a deep graph network. *bioRxiv*, 2020. doi: 10.1101/2020.09.07.285957. URL https://www.biorxiv.org/content/early/2020/09/08/2020.09.07.285957.

[22] Yue Cao and Yang Shen. Energy-based graph convolutional networks for scoring protein docking models, 2019.

[23] Stephan Eismann, Raphael J. L. Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, and Ron Dror. Hierarchical, rotation-equivariant neural networks to predict the structure of protein complexes, 2020.

[24] S Mitternacht. Freesasa: An open source c library for solvent accessible surface area calculations [version 1; peer review: 2 approved]. *F1000Research*, 5(189), 2016. doi: 10.12688/f1000research.7931.1.

[25] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

[26] Yuval Sedan, Orly Marcu, Sergey Lyskov, and Ora Schueler-Furman. Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Research*, 44(W1):W536–W541, 05 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw385. URL https://doi.org/10.1093/nar/gkw385.

[27] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 15820–15831. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9711-generative-models-for-graph-based-protein-design.pdf.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[29] Amrita Mohan, Christopher J. Oldfield, Predrag Radivojac, Vladimir Vacic, Marc S. Cortese, A. Keith Dunker, and Vladimir N. Uversky. Analysis of molecular recognition features (morfs). *Journal of Molecular Biology*, 362(5):1043 – 1059, 2006. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2006.07.087. URL http://www.sciencedirect.com/science/article/pii/S0022283606009831.

[30] Changqun Xia, Jia Li, Jinming Su, and Yonghong Tian. Exploring reciprocal attention for salient object detection by cooperative learning, 2019.

[31] Melissa Swiecki, Suzanne M. Scheaffer, Marc Allaire, Daved H. Fremont, Marco Colonna, and Tom J. Brett. Structural and biophysical analysis of bst-2/tetherin ectodomains reveals an evolutionary conserved design to inhibit virus release. *The Journal of biological chemistry*, 286(4):2987–2997, Jan 2011. ISSN 1083-351X. doi: 10.1074/jbc.M110.190538. URL https://pubmed.ncbi.nlm.nih.gov/21084286. 21084286[pmid].

[32] Sandor Vajda, Christine Yueh, Dmitri Beglov, Tanggis Bohnuud, Scott E. Mottarella, Bing Xia, David R. Hall, and Dima Kozakov. New additions to the cluspro server motivated by capri. *Proteins*, 85(3):435–444, Mar 2017. ISSN 1097-0134. doi: 10.1002/prot.25219. URL https://pubmed.ncbi.nlm.nih.gov/27936493. 27936493[pmid].

[33] Dima Kozakov, David R. Hall, Bing Xia, Kathryn A. Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The cluspro web server for protein–protein docking. *Nature Protocols*, 12(2):255–278, Feb 2017. ISSN 1750-2799. doi: 10.1038/nprot.2016.169. URL https://doi.org/10.1038/nprot.2016.169.

[34] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009. doi: 10.1002/jcc.21287. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21287.

[35] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. Charmm-gui: A web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29(11):1859–1865, 2008. doi: 10.1002/jcc.20945. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20945.

[36] Jumin Lee, Xi Cheng, Jason M. Swails, Min Sun Yeom, Peter K. Eastman, Justin A. Lemkul, Shuai Wei, Joshua Buckner, Jong Cheol Jeong, Yifei Qi, Sunhwan Jo, Vijay S. Pande, David A. Case, Charles L. Brooks, Alexander D. MacKerell, Jeffery B. Klauda, and Wonpil Im. Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *Journal of Chemical Theory and Computation*, 12 (1):405–413, 2016. doi: 10.1021/acs.jctc.5b00935. URL https://doi.org/10.1021/acs.jctc.5b00935. PMID: 26631602.

[37] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 02 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt055. URL https://doi.org/10.1093/bioinformatics/btt055.

[38] Jeffery B. Klauda, Richard M. Venable, J. Alfredo Freites, Joseph W. O'Connor, Douglas J. Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D. MacKerell, and Richard W. Pastor. Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *The Journal of Physical Chemistry B*, 114(23):7830–7843, Jun 2010. ISSN 1520-6106. doi: 10.1021/jp101759q. URL https://doi.org/10.1021/jp101759q.

[39] Jing Huang and Alexander D. MacKerell Jr. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of Computational Chemistry*, 34(25): 2135–2145, 2013. doi: 10.1002/jcc.23354. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23354.

[40] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. doi: 10.1063/1.445869. URL https://doi.org/10.1063/1.445869.

[41] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1049. URL https://doi.org/10.1093/nar/gky1049.

[42] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1223. URL https://doi.org/10.1093/nar/gkt1223.

[43] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 01 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu031. URL https://doi.org/10.1093/bioinformatics/btu031.

[44] Joshua A. Jadwin, Mari Ogiue-Ikeda, and Kazuya Machida. The application of modular protein domains in proteomics. *FEBS letters*, 586(17):2586–2596, Aug 2012. ISSN 1873-3468. doi: 10.1016/j.febslet.2012.04.019. URL https://pubmed.ncbi.nlm.nih.gov/22710164. 22710164[pmid].

[45] R. Joshi, L. Qin, X. Cao, S. Zhong, C. Voss, W. Min, and S. S. C. Li. DLC1 SAM domain-binding peptides inhibit cancer cell growth and migration by inactivating RhoA. *J Biol Chem*, 295(2):645–656, 01 2020.

[46] Justin K. Taylor, Christopher M. Coleman, Sandra Postel, Jeanne M. Sisk, John G. Bernbaum, Thiagarajan Venkataraman, Eric J. Sundberg, and Matthew B. Frieman. Severe acute respiratory syndrome coronavirus orf7a inhibits bone marrow stromal antigen 2 virion tethering through a novel mechanism of glycosylation interference. *Journal of Virology*, 89(23):11820–11833, 2015. ISSN 0022-538X. doi: 10.1128/JVI.02274-15. URL https://jvi.asm.org/content/89/23/11820.

[47] Bernd Wollscheid, Damaris Bausch-Fluck, Christine Henderson, Robert O'Brien, Miriam Bibel, Ralph Schiess, Ruedi Aebersold, and Julian D. Watts. Mass-spectrometric identification and relative quantification of n-linked cell surface glycoproteins. *Nature Biotechnology*, 27(4): 378–386, Apr 2009. ISSN 1546-1696. doi: 10.1038/nbt.1532. URL https://doi.org/10.1038/nbt.1532.