JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

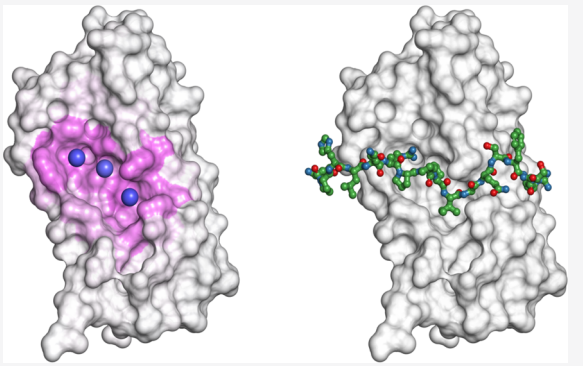
Protein−Peptide Binding Site Detection Using 3D Convolutional Neural Networks

Igor Kozlovskii and Petr Popov*

Cite This: https://doi.org/10.1021/acs.jcim.1c00475 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Peptides and peptide-based molecules represent a promising therapeutic modality targeting intracellular protein−protein interactions, potentially combining the beneficial properties of biologics and small-molecule drugs. Protein−peptide complexes occupy a unique niche of interaction interfaces with respect to protein−protein and protein−small molecule complexes. Protein−peptide binding site identification resembles image object detection, a field that had been revolutionalized with computer vision techniques. We present a new protein−peptide binding site detection method called BiteNet$_{Pp}$ by harnessing the power of 3D convolutional neural network. Our method employs a tensor-based representation of spatial protein structures, which is fed to 3D convolutional neural network, resulting in probability scores and coordinates of the binding "hot spots" in the input structures. We used the domain adaptation technique to fine-tune model trained on protein−small molecule complexes using a manually curated set of protein−peptide structures. BiteNet$_{Pp}$ consistently outperforms existing state-of-the-art methods in the independent test benchmark. It takes less than a second to analyze a single-protein structure, making BiteNet$_{Pp}$ suitable for the large-scale analysis of protein−peptide binding sites.



INTRODUCTION

Protein−peptide interactions are prevalent within a cell that play a vital role by mediating or regulating up to 40% of all cellular processes, such as signal transduction or endogenous regulation.[1] New protein−peptide interactions are regularly deposited to the Protein Data Bank,[2] and accumulated structural information revealed that protein−peptide interactions require relatively small binding sites that vary in geometrical and physicochemical properties.[3] Protein−peptide binding sites are of great interest for pharmacology, as they occur in many interaction networks.[4] Furthermore, protein−peptide binding sites can be exploited to design synthetic peptides to modulate protein−protein interactions (PPIs),[5] representing another important class of pharmacological targets.[6] For a long time, PPIs were considered as intractable targets. On the one hand, while used for targeting extracellular PPIs, large biologics cannot reach intracellular PPIs due to the limitations in crossing the cellular membrane. On the other hand, traditional small-molecule scaffolds may diffuse through the membrane but are not always suitable for diverse and shallow PPI interfaces.[7] PPI interfaces possess distinctive features, including large contact area (ca. 1500−3000 Å^2 for PPI vs ca. 300−1000 Å^2 for protein−small molecule interaction[8]) and lack of pronounced deep binding sites, typically observed for small molecules (~270 Å^3 in volume[9]). Notably, PPI interfaces often contain a few small binding pockets (~100 Å^3 in volume[10]) that are essential for the binding affinity.[11]

Peptides and peptide-based molecules occupy a unique niche of chemical space with respect to small molecules (molecular weight < 0.5 kDa) and biologics (molecular weight > 150 kDa). They represent a promising therapeutic modality targeting intracellular PPIs, potentially combining the beneficial properties of biologics (e.g., low toxicity, high specificity, and affinity) and small molecules (e.g., permeability).[7] Structure-based design of therapeutic peptide modalities requires knowledge of the corresponding target's binding site. Discovery of novel protein−peptide binding sites potentially expands "druggable" genome, which, in turn, opens new opportunities for pharmacology. While experimental identification of binding sites is resource-consuming, powerful computational methods could provide faster and cheaper alternatives. Developed computational approaches vary in methodology to predict protein−peptide binding sites and can be roughly categorized into knowledge-based,

Received: April 28, 2021

fragment-based, feature-based, and end-to-end methods. The knowledge-based methods search for templates to the target protein within a collection of known protein−peptide complexes; putative binding sites are then predicted from the best aligned and clustered templates.[12,13] Fragment-based methods search for "hot spots" in the target protein structure by probing sample molecules (chemical fragments or amino acids) and scoring them by interaction energy.[14,15] Feature-based methods utilize either sequence (e.g., position-specific scoring matrices, hidden Markov model profile features, etc.) or structural (e.g., solvent-accessible surface area, secondary structure type, etc.) descriptors fed into a machine learning model to score each amino acid for belonging to a binding site.[16−21] End-to-end methods operate directly with protein sequences[22,23] or structures[24] by taking advantage of deep learning methods capable of learning features during training. With a continuously growing amount of structural data, it becomes possible to develop more robust methods using end-to-end deep learning approaches that work directly with spatial structures of protein complexes. Particularly structure-based methods are suitable for protein conformation analysis, which is useful in cryptic binding site discovery, protein−peptide docking,[25,26] and studying the dynamic behavior of the binding sites.

Here, we present a deep learning model for identifying protein−peptide binding sites by considering protein structures as 3D images subject to the object detection. We represent a peptide-binding site as a set of "hot spots" in the protein structure contributing most to the protein−peptide interaction. Using the nonredundant set of protein−peptide complexes retrieved from Protein Data Bank (PDB), we trained a 3D convolutional neural network (CNN) initialized with protein−ligand binding site prediction model weights.[27] The obtained model, dubbed BiteNet$_{Pp}$, outputs the binding "hot spots" as a set of 3D coordinates along with the neighboring amino acid residues, given the input protein structure. To the best of our knowledge, this is the first time the domain adaptation technique is used to fine-tune a model trained on protein−small molecule complexes toward protein−peptide complexes. Our model outperforms state-of-the-art methods reaching the area under the ROC curve (ROC AUC) of 0.91 and the Matthew correlation coefficient (MCC) of 0.49 on a commonly used protein−peptide benchmark.[17] BiteNet$_{Pp}$ is fast enough for large-scale analysis of protein structures taking less than a second to analyze a single-protein structure.

METHODS

Training and Test Sets. For rigorous comparison with the other protein−peptide binding site prediction methods, we used the TR1154 training set and the TS125 test benchmark[17] collected from the BioLip database.[28] TR1154 and TS125 are composed of 1154 and 125 protein−peptide complexes, respectively, with peptides of at most 30 amino acid residues and with at least three protein amino acid residues within 3.5 Å from a peptide. TR1154 and TS125 were constructed by clustering the BioLip database with the sequence identity threshold of 30% and retrieving a single structure from each cluster such that no similar proteins are shared between the training and the test sets. For raw structures we observed irrelevant peptides, missing atoms, amino acid residues, and protein chains; to remove potential artifacts related to the low-quality structures, we restored missing fragments and refined the structure using the ICM-Pro suite (molsoft.com). More specifically, we added protein chains present in the crystallographical structure but absent in the dataset; then, we modeled atoms and residues missing in the crystallographical structure; finally, we refined the protein structures by optimizing side chains of histidine, asparagine, glutamine, and cysteine amino acid residues. For training, we performed five-fold cross-validation with a random split on TR1154 and tested the models on TS125. Closer examination, however, revealed similar structures shared between TR1154 and TS125. Indeed, there are 50 (28) proteins in TR1154 (TS125) that have >30% sequence similarity with at least one protein from TS125 (TR1154) (see Table S1). Moreover, we found identical protein sequences shared between TR1154 and TS125 (for example, PDB IDs: 3QPZ vs 1MW4, 2HO2 vs 2OEI, 1K9Q vs 1JMQ). Here, to compute the sequence similarity, we aligned the PDB protein sequences using Biopython[29] and normalized the alignment score by the alignment's length, where the alignment score comprises +1 for the identical amino acid residues, 0 for the different ones, −1 for the opening gaps, and 0 for the extending gaps. Therefore, the models trained on TR1154 likely demonstrate biased metrics on TS125. For these reasons, in addition to TR1154, we composed a different training set BN$_{Pp}$956.

To compose the BN$_{Pp}$956 training set, we first retrieved high-resolution (<3 Å) crystallographical protein structures from PDB. Then, we kept only structures containing a protein chain with at least 50 amino acid residues and a peptide of 5−20 amino acid residues such that there are at least 20 protein heavy atoms within 4 Å from any peptide heavy atom. Additionally, we disregarded structures with more than eight protein chains or exceeding 200 Å along the first principal axis. We considered only structures with the relevant binding sites that have at least one "hot spot" (see the Metrics section for the "hot spot" definition). Structures with the relevant binding sites, where a nonpeptide small molecule is close to a peptide (distance between any heavy atom pair is smaller than 4 Å), were discarded. Finally, we filtered out structures with "hot spots" observed in the crystallographic symmetry mates to ensure that the binding site is formed mainly by the asymmetric unit. In the next step, we restored missing atoms and short loops (at most 10 amino acid residues), refined protein structures, and remodeled non-standard amino acid residues back to their standard analogues using ICM-Pro. We further discarded structures with refinement errors and structures for which the refinement changed the size of binding site interface by more than two atoms. Overall, this procedure yielded a dataset of 4556 protein−peptide structures (BN$_{Pp}$4556).

For rigorous comparison with the other methods, we further removed protein structures similar to the TS125 test benchmark as it follows. We calculated pairwise sequence and structural similarities between the obtained dataset and TS125. Here, we calculated the sequence similarity as the alignment score normalized by the average length of two sequences. Note that this is more strict normalization resulting in higher sequence similarity values, compared to the normalizing by the length of the alignment. The structural similarity was calculated as the sum of TM scores computed with the TM-Align suite,[30] divided by the total length of two sequences. Then, we removed structures with sequence similarity >0.7 or structure similarity >0.8 with any complex from TS125, resulting in a dataset of 4089 complexes without proteins similar to TS125. To remove redundancy, we

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. BiteNet$_{Pp}$ Neural Network Model Architecture

index	input shape	output shape	layer	kernel	stride	filters
1	64 × 64 × 64 × 11	64 × 64 × 64 × 32	Conv3D	3	1	32
2	64 × 64 × 64 × 32	32 × 32 × 32 × 32	Conv3Dpool	3	2	32
3	32 × 32 × 32 × 32	32 × 32 × 32 × 32	Conv3D	3	1	32
4	32 × 32 × 32 × 32	32 × 32 × 32 × 32	Conv3D	3	1	32
5	32 × 32 × 32 × 32	16 × 16 × 16 × 32	Conv3Dpool	3	2	32
6	16 × 16 × 16 × 32	16 × 16 × 16 × 64	Conv3D	3	1	64
7	16 × 16 × 16 × 64	16 × 16 × 16 × 64	Conv3D	3	1	64
8	16 × 16 × 16 × 64	8 × 8 × 8 × 64	Conv3Dpool	3	2	64
9	8 × 8 × 8 × 64	8 × 8 × 8 × 128	Conv3D	3	1	128
10	8 × 8 × 8 × 128	8 × 8 × 8 × 4	Conv3D	3	1	4

clustered complexes by the sequence similarity threshold of 0.9 and kept only a single random structure from each cluster. Finally, we clustered the obtained set of protein−peptide complexes by the structural similarity threshold of 0.5 such that any two structures from different clusters share at most 0.5 structural similarity. To perform five-fold cross-validation, we randomly split the obtained clusters forming five folds of 131, 138, 392, 167, and 128 protein−peptide structures (the largest fold of size 392 corresponds to the largest cluster). It is important to note that there are only 16 pairs of structures in BN$_{Pp}$956 and TS125 that have sequence similarity values higher than 30% (see Table S2); hence, our filtering criteria are more strict than that used in the previous studies. For fair evaluation of the fine-tuned model, we applied the same procedures to the protein−ligand dataset to re-train BiteNet[27] (see Table S3).

Model Training. We adapted the BiteNet's 3D CNN architecture for model training.[27] Briefly, a protein structure is represented as the 4D tensor, where the first three dimensions correspond to the x, y, and z dimensions, and the fourth dimension corresponds to 11 channels that store atomic densities for different atom types: sulfur, amide nitrogen, aromatic nitrogen, guanidinium nitrogen, ammonium nitrogen, carbonyl oxygen, hydroxyl oxygen, carboxyl oxygen, sp^2 carbon, aromatic carbon, and sp^3 carbon. The 4D tensor is split into the 4D subtensors of fixed size (64 × 64 × 64 × 11 corresponding to 64 × 64 × 64 = 262 144 voxels of 1 Å^3, that store 11 channels), which are fed into the 3D CNN comprising 10 3D convolutional layers. A single subtensor, in turn, is represented as 512 nonoverlapping cells of 8 × 8 × 8 voxels with 11 channels. The model outputs four values, namely, the probability score and x, y, and z coordinates, for putative "hot spots" within each cell constituting a single subtensor. Therefore, the model predicts 512 putative "hot spots" with the probability scores varying from 0.0 (no "hot spot") to 1.0 (confident "hot spot"). The non-max-suppression is applied to get the final predictions of the "hot spots" coordinates. More specifically, once the top-score prediction is selected, all of the other predictions within 4 Å are removed; then, the second top-score prediction is selected and so on, until all lower-score predictions within 4 Å from a higher-score prediction will be removed. Note that, as "hot spots" may be located close to each other, we lowered the distance threshold for the non-max-suppression procedure to 4 Å, compared to 8 Å in the original BiteNet approach. We initialized model weights with the BiteNet ones, thus fine-tuning model to the protein−peptide binding sites. The detailed architecture of the model is listed in Table 1. Note that all, except for the last one, convolutional layers are followed by the batch normalization layer and the rectified linear unit (ReLU) activation function. During training, we used implicit data augmentation by a random rotation of proteins in each epoch. The loss function to minimize consists of three terms: (i) cross-entropy for the probability score, (ii) mean squared error for cells with the true "hot spots", and (iii) the L_2 regularization term

$$\begin{aligned}\text{loss} = &\sum_{i=1}^{N_{\text{cells}}} (-s_i \log(\hat{s}_i) - (1-s_i)\log(1-\hat{s}_i)) + + \lambda_{\text{coord}} \\ &\sum_{i=1}^{N_{\text{cells}}} s_i((x_i-\hat{x}_i)^2 + (y_i-\hat{y}_i)^2 + (z_i-\hat{z}_i)) + + \lambda L_2 \end{aligned} \quad (1)$$

where N_{cells} is the number of cells in the single cubic grid; s_i and $\hat{s}_i$ are the true (0 or 1) and predicted probability scores for the cell, respectively; x_i, y_i, z_i and $\hat{x}_i$, $\hat{y}_i$, $\hat{z}_i$ are the true and predicted coordinates for i-th cell, respectively; L_2 is the euclidean norm of the model's weights; and the coefficients λ_{coord} and γ were set to 5 and 1×10^{-5}, respectively, as in the previous study.[27] In the inference mode, we averaged results obtained for 50 replicas of the input structure obtained by rotation about 10 different axes corresponding to the centroids of the icosahedron facets[31] by $\pi/3$, $2\pi/3$, π, $4\pi/3$, and $5\pi/3$ angles.

Metrics. To determine "hot spots" in protein−peptide structures, we used the following procedure (see Figure 1). We represented a peptide as a set of overlapping segments of one, two, three, or four amino acid residues. For each segment, we calculated "hotness" as the size of protein interface, that is, the number of protein heavy atoms within 4 Å. We kept segments with "hotness" ≥15. Next, we formed all possible combinations of not overlapping segments. We sorted the combinations by the total number of segments and the total "hotness" and selected the top combination. Finally, we defined the "hot spots" as the geometric centers of each segment's protein interface from the top combination.

We defined true positives (TP) as the predictions with coordinates within $d_{\text{threshold}}$ = 4 Å from any "hot spot", and false positives (FP) otherwise. False negatives (FN) correspond to the number of "hot spots" with no correct prediction. Note that true negatives (TNs) cannot be defined in a robust way because there is an infinite number of points around the protein structure; hence, the number of TNs would strongly depend on the point selection procedure. The precision and recall metrics are calculated as follows

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

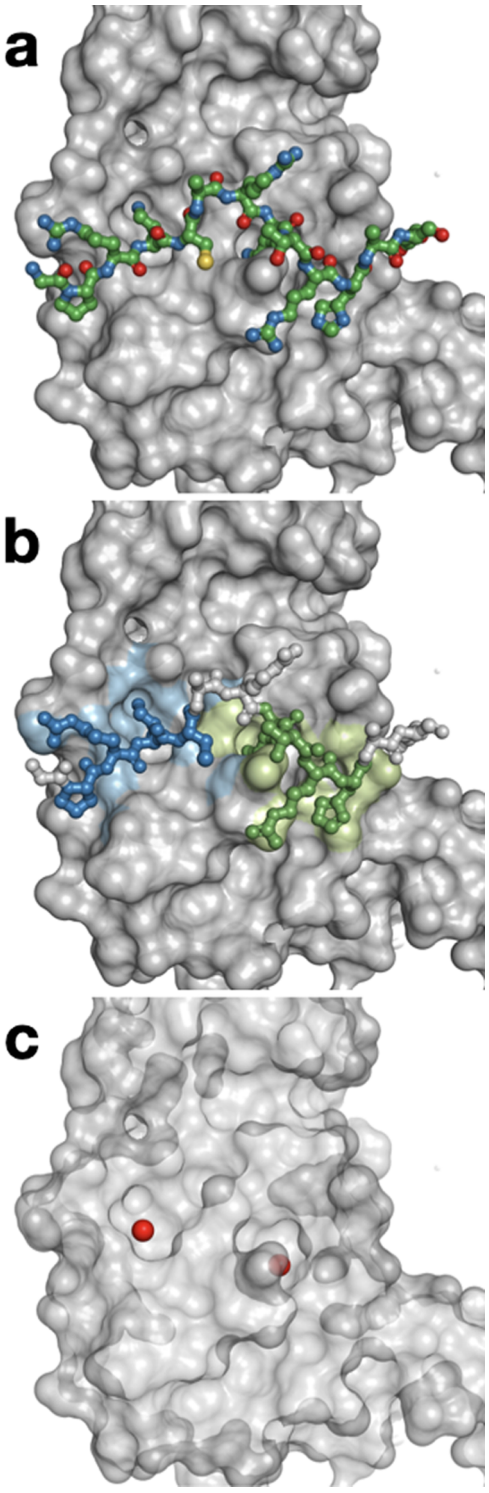


Figure 1. Example of "hot spot" assignment for a protein−peptide binding site. (a) Protein−peptide interaction interface. Protein and peptide are shown with surface and sticks, respectively. (b) Peptide split into two nonoverlapping segments of four amino acid residues with "hotness" of 20 and 21 and colored green and blue, respectively. The corresponding protein interfaces are colored with light green and light blue, respectively. (c) The "hot spots" corresponding to each protein interface are shown with red spheres.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

As the target performance metric for training, we used the average precision (AP), that is, the area below the precision−recall curve. AP is a commonly used metric for the object detection problem in computer vision also useful for the binding site detection problem.[27] Most of the existing approaches, however, assess performance by the Area Under the Receiver Operating Curve (ROC AUC) and Matthews Correlation Coefficient (MCC) metrics calculated for the amino acid residues classified as either belonging to the protein−peptide binding site or not. The former is computed as the area under the true-positive rate vs false-positive rate curve, and the latter is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

Therefore, we used the binary residue labels for the TS125 set[19,21] and found that these labels correspond to the amino acid residues within 3.8 Å from a peptide (see Table S4). To convert BiteNet$_{Pp}$ predictions to the residue-based probability scores, we used the following procedure. First, we score each atom a within the distance threshold d from prediction p obtained with the non-max suppression of 1 Å and probability score threshold of 0.001 as

$$s_a = \max_{p, \|\mathbf{r}_p - \mathbf{r}_a\| \le d} s_p \times e^{-\|\mathbf{r}_p - \mathbf{r}_a\|^2 / 2r_{\text{norm}}} \tag{5}$$

where s_p is the probability score of the prediction p; $\mathbf{r}_p$ and $\mathbf{r}_a$ are the coordinates of p and a, respectively; and r_{norm} is the normalization coefficient. In case there are several predictions within d from a, the maximum s_a is taken. Then, the amino acid residue score is calculated as the maximum of its heavy atom scores. Finally, we defined amino acid residues with scores higher than score threshold s_r as belonging to the binding site. We determined the optimal parameters d = 5 Å, r_{norm} = 5, and s_r = 0.2 that provide the best averaged AP and MCC metrics on the cross-validation.

We would like to note that in terms of residue-based binary classification, TS125 is a highly imbalanced dataset with <0.06 balance ratio (1719 positive and 29 151 negative examples). Therefore, accuracy and specificity are likely unrepresentative metrics, as they weight all errors equally; and for precision and recall, one can obtain almost perfect values by predicting all negative or positive. ROC AUC takes into account score ranking; however, it is also affected by high class imbalance. MCC is a more suitable metric for imbalanced datasets but operates with binary labels. The average precision metric (AP; PR AUC) operates with scores and is suitable for imbalanced sets with a low positive-to-negative example ratio. Therefore, we also considered the residue-based AP as the performance metric.

RESULTS AND DISCUSSION

BiteNet$_{Pp}$. We trained BiteNet$_{Pp}$ on curated protein structures using 3D CNN architecture, proven to be top-performing for protein−small molecule binding site detection[27] (see the Methods section), and Figure 2 illustrates the BiteNet$_{Pp}$ workflow. In the first step, the input protein structure is voxelized, resulting in the fixed-size 3D images (64 × 64 × 64) representing 64 Å^3 spatial cube, and each voxel (1 Å^3) contains 11 channels corresponding to the atomic

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

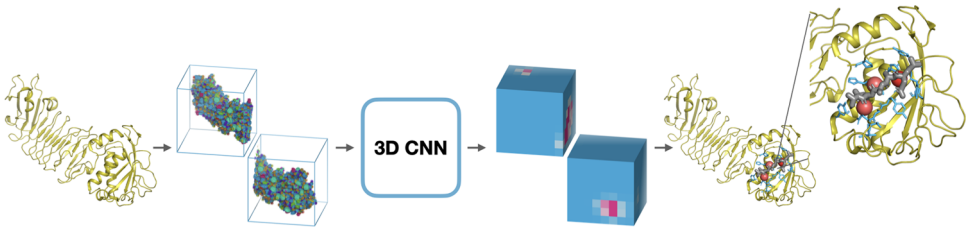


Figure 2. Illustration of the BiteNet$_{Pp}$ workflow. The input protein structure is voxelized into fixed-size 4D tensors. The tensors are fed to 3D CNN that outputs probability scores and coordinates of putative "hot spots" for regions of 8 Å × 8 Å × 8 Å. Finally, the postprocessing procedure keeps only the most relevant predictions. Protein structure, protein interface, and predicted "hot spots" are shown with gold cartoon, blue sticks, and red spheres, respectively. Voxels are colored with respect to the different channels with nonzero values.

densities of a certain type. In the second step, the 3D images are fed into 3D CNN to output tensors of size 8 × 8 × 8 × 4, where the first three dimensions correspond to the cell coordinates relatively to the 3D image (region of 8 × 8 × 8 voxels), and the four scalars of the last dimension correspond to the probability score of a "hot spot" being in the cell and its 3D coordinates. Finally, in the third step, the obtained tensors are processed to output the most relevant "hot spot" predictions. Overall, the input to BiteNet$_{Pp}$ is the spatial structure of a protein and the output is the scored centers of the predicted "hot spots", along with the list of scored amino acid residues associated with each center.

While protein−peptide and protein−small molecule binding sites likely differ in size and geometry, the interaction patterns share the same physical principles. To employ protein−small molecule interaction patterns captured by BiteNet, we used its weights as the initial weights for the BiteNet$_{Pp}$ training. We performed five-fold cross-validations on the TR1154 and BN$_{Pp}$956 sets with and without fine tuning, and Figure 3 shows the obtained results. Indeed, we observed that such fine tuning improves the averaged AP from 0.290 to 0.326 and from 0.337 to 0.374 in cross-validation on TR1154 and BN$_{Pp}$956, respectively (see Table S5). Notably, the original BiteNet performs poorly on the protein−peptide structures with an AP of 0.121 on training set, emphasizing the difference between the protein−small molecule and protein−peptide binding sites. Overall, the models trained on TR1154 and BN$_{Pp}$956 showed similar performance and the hyperparameter search did not reveal significant improvement in any cross-validation. Note that BiteNet$_{Pp}$'s output is the 3D coordinates of "hot spots" along with the probability scores; however, most of the methods for protein−peptide binding site prediction operate with the binary classification of binding and nonbinding amino acid residues. To convert BiteNet$_{Pp}$'s predictions into the residue-based predictions, we used cross-validation to find optimal parameters for the score assignment of each amino acid residue (see the Metrics section). We want to emphasize that as the target metric, we used the AP metric predicted for the "hot spot" coordinates and not for amino acid residues. The two final BiteNet$_{Pp}$ models were trained for 300 epochs on the whole TR1154 and BN$_{Pp}$956 sets.

Comparison with Other Methods. When BiteNet$_{Pp}$ was applied to the TS125 test benchmark, we observed superior performance in terms of the ROC AUC and MCC metrics, compared to all state-of-the-art methods (see Figure 4 and Table S6). The model trained on TR1154 demonstrated slightly better metrics than the model trained on BN$_{Pp}$956 (0.891 and 0.457 vs 0.881 and 0.440 for ROC AUC and MCC,

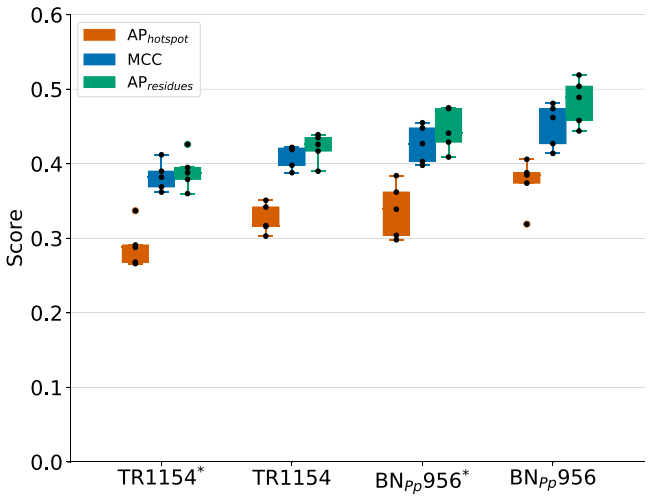


Figure 3. Cross-validation performance metrics for BiteNet$_{Pp}$ models trained on TR1154 set and BN$_{Pp}$956 without fine-tuning (marked with asterisk) and with fine-tuning. The mean and standard deviation values were calculated for AP$_{hotspot}$ (red), MCC (blue), and AP$_{residues}$ (green) on five validation sets. Individual measurements are shown with black dots.

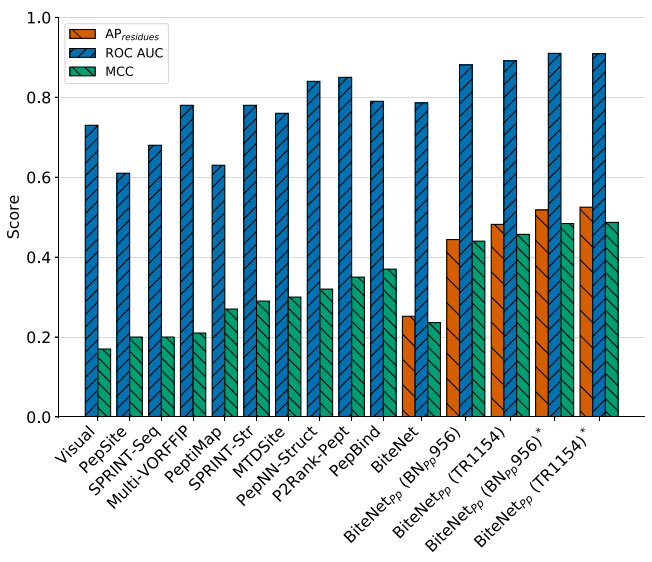


Figure 4. AP$_{residues}$ (red), ROC AUC (blue), and MCC (green) performance metrics on the TS125 benchmark for the state-of-the-art methods and BiteNet$_{Pp}$ models. BiteNet$_{Pp}$(BN$_{Pp}$956)* and BiteNet$_{Pp}$(TR1154)* correspond to the performance metrics calculated on the TS125$_{curated}$ benchmark. For the other methods, the metric values were retrieved from the corresponding papers.

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

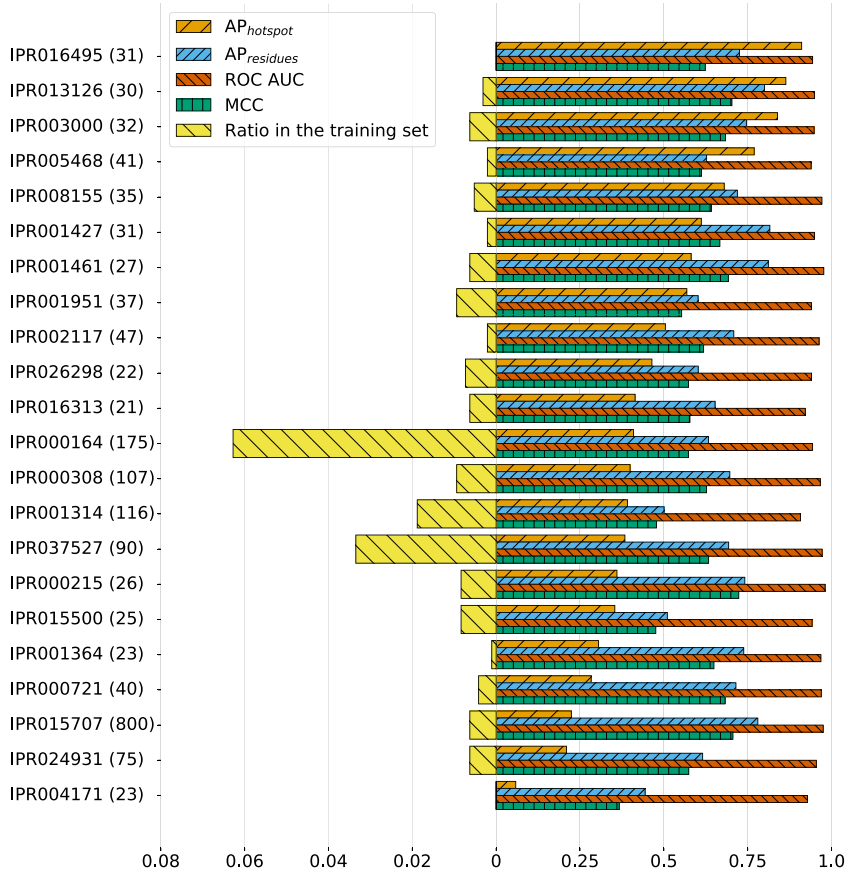


Figure 5. BiteNet$_{Pp}$'s performance for the most represented on protein families from the BN$_{Pp}$4556 curated dataset of 4556 protein structures. Number of protein structures is alongside InterPro protein family identifier. The ratio of protein structures in BN$_{Pp}$4556 is shown with orange bars. The AP$_{hotspot}$, AP$_{residues}$, ROC AUC, and MCC metrics are shown with orange, blue, red, and green bars, respectively.

respectively); however, TS125 contains 28 and 15 proteins that share ≥30% sequence similarity with TR1154 and BN$_{Pp}$956, respectively (see the Training and Test Sets section). Nonetheless, we did not observe any drop in the performance when we excluded similar proteins from TS125. Namely, the residue-based ROC AUC and MCC are 0.893 and 0.454 (0.878 and 0.451) for the model trained on TR1154 (BN$_{Pp}$956) and tested on the subset of 97 (110) complexes from TS125. A closer examination of the similar complexes revealed that there is one structure from BN$_{Pp}$956 (PDB ID: 6G5K) corresponding to the subdomain of the structure from TS125 (PDB ID: 2NP0). One can calculate sequence similarity using normalization by the minimal length of two sequences to avoid such cases, rather than the average length; however, we did not observe any model's improvement or deterioration when applied this type of normalization.

These observations emphasize the importance of careful train-validation-test dataset preparation. Unfortunately, currently available structural data is not enough to compose "ideal" train-validation-test splits. Indeed, using sequence similarity of 30% and structural similarity of 50%, corresponding to the unrelated protein folds, resulted in a single cluster on our dataset. Fortunately, due to the progress in macromolecular structure determination, the number of deposited protein structures grows exponentially; thus, one should expect richer datasets in the nearest future. We also want to point out that widely used ROC AUC and MCC metrics are calculated solely based on the amino acid residue classification. This, in turn, introduces bias to the performance with respect to the protein structures because larger binding sites impact more on the metrics. Therefore, alternative metrics are needed for a more robust estimation of model performance. For example, one computed the MCC values for each binding site and used averaged MCC across the dataset as the performance metric,[12,13] where a binding site is predicted correctly, if at least 50% of its residues correspond to the true binding site. Here, we also computed the residue-based AP; unfortunately, it was not possible to calculate it for the other methods for comparison.

Another source of potential bias lies in considering buried residues—such residues unlikely participate in binding and can be trivial negative examples to predict, thus yielding overestimated metrics. We calculated the relative solvent-accessible surface area (RSASA) (using FreeSASA[32]) for each residue in TS125 and observed that ∼30% (9164 out of 29 151) of nonbinding residues have RSASA < 0.1, while binding residues with RSASA < 0.1 constitute only ∼12.5% (215 out of 1719) (see Figure S1 and the Supporting Information). We want to emphasize that BiteNet$_{Pp}$ is robust to such a bias since it was trained to optimize the "hot spot"-based AP, rather than the residue-based classification metrics. However, the mapping from the "hot spot" to the residue predictions could be prone to such a bias. To ensure that this is not the case, we further excluded amino acid residues with RSASA varying from 0.05 to 0.5 and recalculated the residue-based AP, ROC AUC, and MCC values. We did not observe inferior performance metrics, indicating that there is no bias (see Figure S2). Unfortunately, we could not calculate these types of results for the other

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

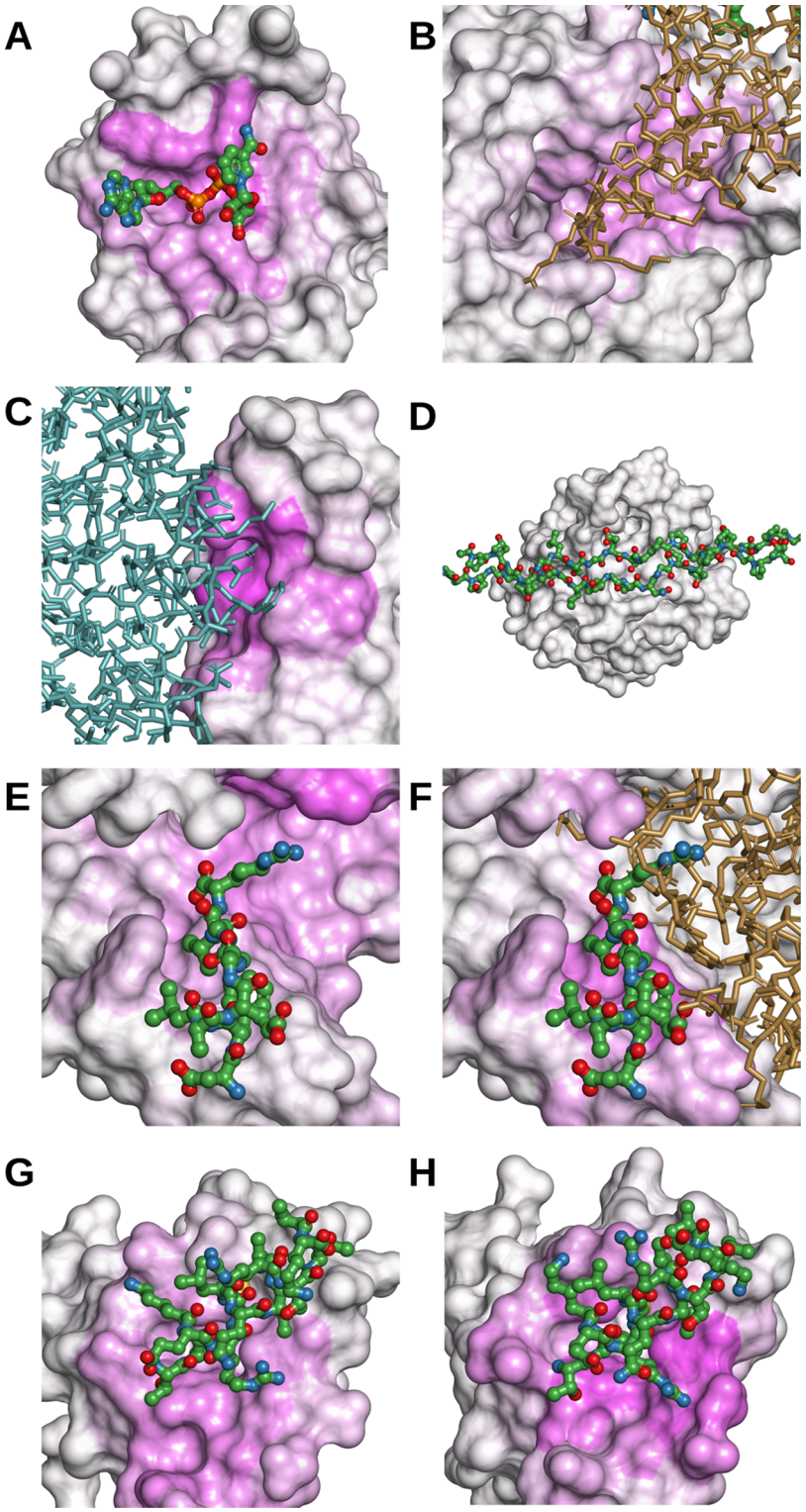


Figure 6. Some examples of false-positive and false-negative predictions for TS125. The target proteins are represented with white surfaces, and the amino acid residues are colored with magenta according to the probability score. Small molecules and peptides are shown with ball and sticks. Missing in TS125 protein chains are shown with brown and teal sticks for the asymmetric unit and symmetry mate, respectively. (A−C) BiteNet$_{Pp}$ model predicts binding interface corresponding to the nonpeptide molecule: (A) NAD small molecule (PDB ID: 4HANA); (B) protein chain in the asymmetric unit (PDB ID: 4CC9); and (C) protein chain in the symmetry mate (PDB ID: 2AST). (D) False-negative predictions corresponding to a peptide of more than 20 amino acid residues (PDB ID: 2M32; 23 residues). (E, F) Predictions for a complex, where a peptide binding interface is formed with two protein chains: (E) a single-protein chain as in TS125 and (F) missing chain added (PDB ID: 1BBR). (G, H) Predictions for an NMR complex for (G) the first NMR model and (H) the top-scored model (H) (1HV2; model index 14).

methods, but for PepNN,[24] that shows a similar trend (see Figure S2).

By construction, TS125 contains a single-protein chain and a peptide; however, in the corresponding PDB entries, there

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

might be other protein chains that participate in binding interface formation. TS125 also contains NMR structures with several conformations that slightly differ in binding interfaces, thus affecting the models' performance. To assess the effect of such peculiarities on the performance of BiteNet$_{Pp}$, we restored missing protein chains, refined the obtained complexes, applied BiteNet$_{Pp}$, and calculated the performance metrics (PDB IDs: 1QRJ, 3LYV, 4HFX did not pass the refinement procedure and were disregarded). In the case of NMR structures, we considered a single conformation corresponding to the best residue-based AP score. We observed improvement in the residue-based AP, ROC AUC, and MCC from 0.482, 0.891, and 0.457 to 0.525, 0.909, and 0.487 for BiteNet$_{Pp}$ model trained on TR1154 and from 0.444, 0.881, and 0.440 to 0.518, 0.910, and 0.484 for model trained on BN$_{Pp}$956, respectively (see Table S7 for more details).

To demonstrate the predictive power of BiteNet$_{Pp}$ for different protein structures, we assigned the InterPro[33] family identifier to each protein in the BN$_{Pp}$4556 dataset and considered protein families counting at least 20 protein structures. Figure 5 shows the AP, AUC, and MCC metrics calculated for each protein family for BiteNet$_{Pp}$, as well as the ratio of structures from this family presented in the BN$_{Pp}$956 training set. Similarly to the entire training set, the "hot spot"-based AP averaged over the protein families is 0.48, and we did not observe any significant correlation between the ratio of protein structures presented in the training set and the performance metrics.

Error Analysis. We examined protein complexes, on which BiteNet$_{Pp}$ produced incorrect predictions and observed several ambiguous cases among false-positive and false-negative predictions.

For 20 out of 28 structures corresponding to the false-positive predictions, the predicted binding site was occupied not by a peptide, but a small molecule or a protein chain fragment presented in the raw structure (see Figure 6A,B). As for the false-negative predictions, for 9 out of 30 structures, the peptide binding interface is formed not by a single protein chain, but multiple protein chains missing in the original TS125 benchmark. Inclusion of missing protein chains into the benchmark improved the performance metrics of BiteNet$_{Pp}$ (see Figure 6E,F); calculated solely for these protein complexes, the residue-based AP, ROC AUC, and MCC increased from 0.162, 0.642, and −0.021 to 0.495, 0.894, and 0.314, respectively. For several complexes, such errors can be explained by a symmetry mate, rather than a co-crystallized molecule: in the case of a false-positive prediction, the binding interface is formed between the symmetry mate and asymmetric unit, and in the case of a false-negative prediction, the peptide interacts with asymmetric unit and the symmetry mate, forming the binding interface (see Figure 6C). To emphasize the importance of missing atoms, we calculated the number of protein atoms in the peptide-binding interface belonging to the missing protein chains and symmetry mates. Indeed, such a number varies from 0 to 141 (8.9 ± 18.8) and from 0 to 140 (9.3 ± 22.7) for missing protein chains; and from 0 to 79 (6.2 ± 11.4) and from 0 to 44 (7.3 ± 10.7) in the original TR1154 and TS125 sets, respectively. Note that in our training set, there are no missing protein chains, and we filtered out structures with a "hot spot" observed in a symmetry mate. Next, we would like to note that by construction, only the first NMR model is included in the TS125 benchmark; however, the other models may possess more representative binding interface. We observed that often BiteNet$_{Pp}$ achieved the best score not on the first NMR model (see Figure 6G,H), comparing the performance metrics for the best vs the first conformation yielding improvement in the residue-based AP, ROC AUC, and MCC from 0.271, 0.709, and 0.078 to 0.496, 0.822, and 0.115, respectively. Finally, 11 out of 30 false-negative predictions correspond to the binding sites with peptides disregarded in the BiteNet$_{Pp}$'s training set, as having more than 20 amino acid residues in size (see Figure 6D). We believe that the dataset issues described above should be taken into account in the next-generation predictive models, and the Supporting Information contains the refined protein structures used in this study. Overall, this error analysis demonstrates that BiteNet$_{Pp}$ is sensitive to protein conformations, and its predictive power is even higher within the applicability domain (see Table S8 for case-by-case performance in TS125).

It is important to note that one of the limitations of BiteNet$_{Pp}$ is rotation variance with respect to the input orientation of the protein structure. Here, we used data augmentation to increase the accuracy and to improve the robustness of the model. To demonstrate the importance of data augmentation, we generated 50 different orientations for each protein structure in TS125 and calculated per-protein performance metrics for the models trained with and without data augmentation on the BN$_{Pp}$956 training set. Indeed, the model trained with data augmentation demonstrates more accurate and robust predictions compared to the model trained with no data augmentation in terms of the mean standard deviation of the performance metrics. More specifically, the "hot spot"-based AP is higher (0.440 vs 0.382) and the mean standard deviation for the residue-based AP, ROC AUC, and MCC is lower (0.08, 0.05, and 0.10 vs 0.12, 0.09, and 0.14, respectively) for the model trained with data augmentation (see Figures S3−S5). We anticipate that development rotation-invariant and rotation-equivariant and neural network architectures[34] could further improve binding site detection methods.

CONCLUSIONS

Here, we presented BiteNet$_{Pp}$, a new protein−peptide binding site detection method, which utilizes a 3D convolutional neural network applied to the voxelized representation of protein structures. BiteNet$_{Pp}$ outputs coordinates of "hot spots" constituting protein−peptide binding site along with its probability scores. We used the domain adaptation technique to improve the performance of BiteNet$_{Pp}$, namely, we fine-tuned the original BiteNet model trained on protein−small molecule complexes on the curated protein−peptide dataset. The proposed method is fast enough for large-scale binding site detection campaigns, taking less than a second to analyze a single-protein structure. Our method outperforms the state-of-the-art approaches on the widely used TS125 benchmark, and for the first time, we achieved 0.49 and 0.91 milestones in terms of MCC and ROC AUC, respectively. BiteNet$_{Pp}$ is available at https://sites.skoltech.ru/imolecule/tools/bitenet/.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00475.

Similarity between proteins in training and test sets in terms of the sequence identity (normalized by the

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

alignment length); sizes of train and validation splits used in the cross-validation; number of amino acid residues classified as binding and nonbinding in TS125 with respect to the distance threshold; precision, recall, and $AP_{hotspot}$ metrics on cross-validation; "hot spot"-based and residue-based performance metrics for the cross-validation sets and the test benchmarks; distribution of the relative solvent-accessible surface area for the nonbinding (blue) and binding (orange) amino acid residues in TS125; residue-based AP (orange), ROC AUC (blue), and MCC (green) calculated for the amino acid residues with the relative solvent-accessible surface area higher than the threshold; and the distribution of the residue-based AP, ROC AUC, and MCC for 51 different orientations of each complex in TS125 (PDF)

Mapping between TS125 and $TS125_{curated}$ (CSV)

Coordinates of the target "hot spots" for $BN_{Pp}956$, TR1154, TS125, and $TS125_{refined}$ datasets (ZIP)

Binary labels for nonbinding and binding amino acid residues in $TS125_{refined}$ and all NMR models in TS125 (ZIP)

PDB files with the refined structures for $TS125_{refined}$ dataset (ZIP)

PDB files with the curated structures for $TS125_{curated}$ dataset (ZIP)

The "hot spot" predictions of the $BiteNet_{Pp}$ models trained with and without fine tuning and the original BiteNet model for the TS125 and $TS125_{refined}$ sets (ZIP)

The residue-based scores of the $BiteNet_{Pp}$ models trained with and without fine tuning and the original BiteNet model for the TS125, $TS125_{refined}$, and $TS125_{curated}$ sets (ZIP)

Relative solvent-accessible surface areas calculated for the amino acid residues in TS125, $TS125_{refined}$, and $TS125_{curated}$ datasets (ZIP)

Lists of PDB IDs used for the cross-validation folds for $BN_{Pp}956$ and TR1154 (ZIP)

AUTHOR INFORMATION

Corresponding Author

Petr Popov − *iMolecule, Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 121205, Russia;* orcid.org/0000-0003-3745-7154; Email: p.popov@skoltech.ru

Author

Igor Kozlovskii − *iMolecule, Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 121205, Russia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00475

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors used the Zhores supercomputer[35] to train $BiteNet_{Pp}$.

REFERENCES

(1) Petsalaki, E.; Russell, R. B. Peptide-mediated interactions in biological systems: new discoveries and applications. *Curr. Opin. Biotechnol.* **2008**, *19*, 344−350.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(3) Stanfield, R. L.; Wilson, I. A. Protein-peptide interactions. *Curr. Opin. Struct. Biol.* **1995**, *5*, 103−113.

(4) Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. Synthetic Therapeutic Peptides: Science and Market. *Drug Discovery Today* **2010**, *15*, 40−56.

(5) Díaz-Eufracio, B. I.; Naveja, J. J.; Medina-Franco, J. L. Protein-protein Interaction Modulators for Epigenetic Therapies. *Adv. Protein Chem. Struct. Biol.* **2018**, *110*, 65−84.

(6) Ruffner, H.; Bauer, A.; Bouwmeester, T. Human Protein-protein Interaction Networks and the Value for Drug Discovery. *Drug Discovery Today* **2007**, *12*, 709−716.

(7) Tsomaia, N. Peptide Therapeutics: Targeting the Undruggable Space. *Eur. J. Med. Chem.* **2015**, *94*, 459−470.

(8) Smith, M. C.; Gestwicki, J. E. Features of Protein-protein Interactions That Translate into Potent Inhibitors: Topology, Surface Area and Affinity. *Expert Rev. Mol. Med.* **2012**, *14*, No. e16.

(9) Buchwald, P. Small-molecule Protein−protein Interaction Inhibitors: Therapeutic Potential in Light of Molecular Size, Chemical Space, and Ligand Binding Efficiency Considerations. *IUBMB Life* **2010**, *62*, 724−731.

(10) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting Druggable Binding Sites at the Protein−protein Interface. *Drug Discovery Today* **2009**, *14*, 155−161.

(11) Clackson, T.; Wells, J. A. A Hot Spot of Binding Energy in a Hormone-receptor Interface. *Science* **1995**, *267*, 383−386.

(12) Litfin, T.; Yang, Y.; Zhou, Y. Spot-peptide: Template-based Prediction of Peptide-binding Proteins and Peptide-binding Sites. *J. Chem. Inf. Model.* **2019**, *59*, 924−930.

(13) Johansson-Åkhe, I.; Mirabello, C.; Wallner, B. Predicting Protein-peptide Interaction Sites Using Distant Protein Complexes as Structural Templates. *Sci. Rep.* **2019**, *9*, No. 4267.

(14) Lavi, A.; Ngan, C. H.; Movshovitz-Attias, D.; Bohnuud, T.; Yueh, C.; Beglov, D.; Schueler-Furman, O.; Kozakov, D. Detection of Peptide-binding Sites on Protein Surfaces: The First Step Toward the Modeling and Targeting of Peptide-mediated Interactions. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 2096−2105.

(15) Yan, C.; Zou, X. Predicting Peptide Binding Sites on Protein Surfaces by Clustering Chemical Interactions. *J. Comput. Chem.* **2015**, *36*, 49−61.

(16) Segura, J.; Jones, P. F.; Fernandez-Fuentes, N. A Holistic in Silico Approach to Predict Functional Sites in Protein Structures. *Bioinformatics* **2012**, *28*, 1845−1850.

(17) Taherzadeh, G.; Yang, Y.; Zhang, T.; Liew, A. W.-C.; Zhou, Y. Sequence-based Prediction of Protein-peptide Binding Sites Using Support Vector Machine. *J. Comput. Chem.* **2016**, *37*, 1223−1229.

(18) Taherzadeh, G.; Zhou, Y.; Liew, A. W.-C.; Yang, Y. Structure-based Prediction of Protein-peptide Binding Regions Using Random Forest. *Bioinformatics* **2018**, *34*, 477−484.

(19) Zhao, Z.; Peng, Z.; Yang, J. Improving Sequence-based Prediction of Protein−peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method. *J. Chem. Inf. Model.* **2018**, *58*, 1459−1468.

(20) Krivák, R.; Jendele, L.; Hoksza, D. In *Peptide-Binding Site Prediction from Protein Structure via Points on the Solvent Accessible Surface*, Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018; pp 645−650.

(21) Wardah, W.; Dehzangi, A.; Taherzadeh, G.; Rashid, M. A.; Khan, M. G.; Tsunoda, T.; Sharma, A. Predicting Protein-peptide Binding Sites with a Deep Convolutional Neural Network. *J. Theor. Biol.* **2020**, No. 110278.

(22) Sun, Z.; Zheng, S.; Zhao, H.; Niu, Z.; Lu, Y.; Pan, Y.; Yang, Y. To Improve the Predictions of Binding Residues with Dna, Rna, Carbohydrate, and Peptide via Multiple-task Deep Neural Networks. *bioRxiv* 2020.
(23) Lei, Y.; Li, S.; Liu, Z.; Wan, F.; Tian, T.; Li, S.; Zhao, D.; Zeng, J. Camp: A Convolutional Attention-based Neural Network for Multifaceted Peptide-protein Interaction Prediction. *bioRxiv* 2020.
(24) Abdin, O.; Wen, H.; Kim, P. M. In *Sequence and Structure Based Deep Learning Models for the Identification of Peptide Binding Sites*, Machine Learning for Structural Biology Workshop at Conference on Neural Information Processing Systems (NeurIPS), 2020.
(25) Agrawal, P.; Singh, H.; Srivastava, H. K.; Singh, S.; Kishore, G.; Raghava, G. P. Benchmarking of Different Molecular Docking Methods for Protein-peptide Docking. *BMC Bioinf.* **2019**, *19*, 105−124.
(26) Weng, G.; Gao, J.; Wang, Z.; Wang, E.; Hu, X.; Yao, X.; Cao, D.; Hou, T. Comprehensive Evaluation of Fourteen Docking Programs on Protein−peptide Complexes. *J. Chem. Theory Comput.* **2020**, *16*, 3959−3969.
(27) Kozlovskii, I.; Popov, P. Spatiotemporal Identification of Druggable Binding Sites Using Deep Learning. *Commun. Biol.* **2020**, *3*, No. 618.
(28) Yang, J.; Roy, A.; Zhang, Y. Biolip: A Semi-manually Curated Database for Biologically Relevant Ligand−protein Interactions. *Nucleic Acids Res.* **2012**, *41*, D1096−D1103.
(29) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(30) Zhang, Y.; Skolnick, J. Tm-align: A Protein Structure Alignment Algorithm Based on the Tm-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(31) Popov, P.; Grudinin, S. Eurecon: Equidistant Uniform Rigid-body Ensemble Constructor. *J. Mol. Graphics Modell.* **2018**, *80*, 313−319.
(32) Mitternacht, S. Freesasa: An Open Source C Library for Solvent Accessible Surface Area Calculations. *F1000Research* **2016**, *5*, No. 189.
(33) Mitchell, A. L.; Attwood, T. K.; Babbitt, P. C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S. D.; Chang, H.-Y.; El-Gebali, S.; Fraser, M. I.; Gough, J.; Haft, D. R.; Huang, H.; Letunic, I.; Lopez, R.; Luciani, A.; Madeira, F.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Necci, M.; Nuka, G.; Orengo, C.; Pandurangan, A. P.; Paysan-Lafosse, T.; Pesseat, S.; Potter, S. C.; Qureshi, M. A.; Rawlings, N. D.; Redaschi, N.; Richardson, L. J.; Rivoire, C.; Salazar, G. A.; Sangrador-Vegas, A.; Sigrist, C. J.; Sillitoe, I.; Sutton, G. G.; Thanki, N.; Thomas, P. D.; Tosatto, S. C.; Yong, S.-Y.; Finn, R. D. Interpro in 2019: Improving Coverage, Classification and Access to Protein Sequence Annotations. *Nucleic Acids Res.* **2019**, *47*, D351−D360.
(34) Weiler, M.; Geiger, M.; Welling, M.; Boomsma, W.; Cohen, T. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. 2018, arXiv:1807.02547. arXiv.org e-Print archive. https://arxiv.org/abs/1807.02547 (accessed July 6, 2018).
(35) Zacharov, I.; Arslanov, R.; Gunin, M.; Stefonishin, D.; Bykov, A.; Pavlov, S.; Panarin, O.; Maliutin, A.; Rykovanov, S.; Fedorov, M. "Zhores"−Petaflops Supercomputer for Data-driven Modeling, Machine Learning and Artificial Intelligence Installed in Skolkovo Institute of Science and Technology. *Open Eng.* **2019**, *9*, 512−520.

https://doi.org/10.1021/acs.jcim.1c00475
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX