JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2020, 60, 391−399

pubs.acs.org/jcim

To Improve Protein Sequence Profile Prediction through Image Captioning on Pairwise Residue Distance Map

Sheng Chen,[†] Zhe Sun,[†] Lihua Lin,[†] Zifeng Liu,[‡] Xun Liu,[‡] Yutian Chong,[‡] Yutong Lu,[†] Huiying Zhao,[*,§] and Yuedong Yang[*,†,∥]

[†]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China

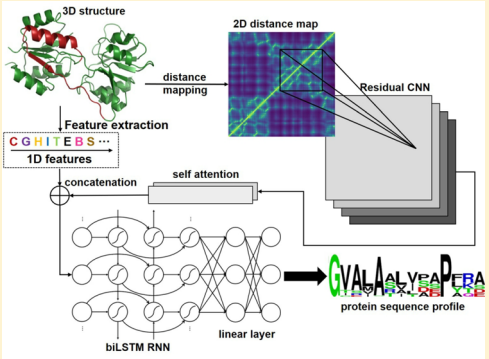
[‡]Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510000, China

[§]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China

[∥]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University) of the Ministry of Education, Guangzhou 510000, China

Supporting Information

ABSTRACT: Protein sequence profile prediction aims to generate multiple sequences from structural information to advance the protein design. Protein sequence profile can be computationally predicted by energy-based or fragment-based methods. By integrating these methods with neural networks, our previous method, SPIN2, has achieved a sequence recovery rate of 34%. However, SPIN2 employed only one-dimensional (1D) structural properties that are not sufficient to represent three-dimensional (3D) structures. In this study, we represented 3D structures by 2D maps of pairwise residue distances and developed a new method (SPROF) to predict protein sequence profiles based on an image captioning learning frame. To our best knowledge, this is the first method to employ a 2D distance map for predicting protein properties. SPROF achieved 39.8% in sequence recovery of residues on the independent test set, representing a 5.2% improvement over SPIN2. We also found the sequence recovery increased with the number of their neighbored residues in 3D structural space, indicating that our method can effectively learn long-range information from the 2D distance map. Thus, such network architecture using a 2D distance map is expected to be useful for other 3D structure-based applications, such as binding site prediction, protein function prediction, and protein interaction prediction. The online server and the source code is available at http://biomed.nscc-gz.cn and https://github.com/biomed-AI/SPROF, respectively.



1. INTRODUCTION

Computational protein design attempts to design a protein sequence that will fold into a predefined structure to perform a desired function. The motivation of studies in this area is not only to supplement, modify, or improve the function of wild-type proteins but also to improve our fundamental comprehension of the relationship between protein sequences, structures, and functions. The past three decades have witnessed significant progress in de novo protein design.[1] More recently, by using the Rosetta package, Silva et al. designed potent and selective mimics of anticancer drugs IL-2 and IL-15.[2] Such advances have shown the potential to design novel proteins for diagnostic, therapeutic, and industrial purposes. While significant progress has been made, existing protein design approaches have low success rates.[3] This has led to efforts on building a library of designed sequences or sequence profiles (sequences randomly generated by specific probabilities of 20 standard amino acids at each site) for guiding experimental screening or directed evolution.[4−7]

Typically, protein sequences or sequence profiles can be generated by applying mutations on a random sequence iteratively to minimize its folding free energy with a proper optimization algorithm.[8−12] However, the search of global minima is not guaranteed since it is an NP-hard combinatorial optimization problem.[13] To explore the possibility of more computationally efficient protein design methods, Dai et al. proposed a fragment-based method by searching structurally similar fragments from known protein structures.[14,15] For a given target protein structure, the sequence profile obtained from structurally similar fragments shows high similarity to its sequence. This fragment-based method is of high computational efficiency but has a lack of information on nonlocal residue interactions (close in three-dimensional (3D) structure but not in sequence). Li et al. employed a knowledge-based scoring function to compute residue specific energy values according to 3D structures, and integrated them with the profiles derived from fragments into neural networks.[16] The developed SPIN method by training neural network with the local (e.g., fragment-derived) and nonlocal (e.g., energy-based)

Received: May 28, 2019
Published: December 4, 2019

Downloaded via GRIFFITH UNIV on April 26, 2020 at 01:53:02 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© 2019 American Chemical Society

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

features achieved a sequence recovery of 30%. Later, the sequence recovery was improved to 33% by using a deep learning method.[17] At the same time, SPIN2,[18] an updated version of SPIN, was also developed by utilizing a deep learning network with additional features, slightly improving the sequence recovery to 34.4%. However, all these prediction methods utilized only one-dimensional (1D) structural properties that are not sufficient to represent 3D structures.

In order to make a full use of protein 3D structural information, a few studies attempted to input the whole 3D structural information into a 3D-covolutional neural network (3D-CNN) for different biological problems, such as protein−ligand scoring prediction,[19] protein-binding site prediction,[20] side chain conformation prediction,[21] and quality assessment of protein folds.[22] However, it remains challenging to train an accurate 3D-CNN network from the large number of redundant variables involved in the highly sparse 3D matrix with the limited number of 3D structures deposited in the protein data bank (PDB).

On the other hand, it was well-known that a 3D structure can be alternatively represented by the 2D contact map, which simply shows whether distance of each residue pair is below a threshold (usually 8 Å). For example, Skolnick et al. stated that their algorithm was able to successfully fold a small protein even with a small portion of inter-residue contacts.[23] Many recent reports showed that predicted contact map could even produce high-quality 3D protein structures.[24] Moreover, the 2D contact map is an image that can be efficiently modeled by modern deep learning techniques, such as ResNet[25] in the contact map prediction task,[26,27] and the prediction from 2D contact maps to sequence profiles is similar to the image captioning problem.[28]

There exists a few differences with traditional image captioning tasks. First, classical image captioning tasks take only a single 2D image as input, while our method's inputs include both 2D distance maps and 1D structural features. Second, in image caption scenarios, images are often preprocessed to a fixed size, but our distance maps should not be resized because each pixel represents exactly one residue pair, and residues far in the sequence might be neighbored in 3D space. Third, the target output of image captioning task is a sentence whose length is irrelevant with input, while our input distance map is of size $L \times L$, where L is equal to the length of our target output ($L \times 20$).

Inspired by the image captioning tasks, we have designed a novel network architecture coupling bidirectional long short-term memory (BiLSTM) with self-attentional 2D-convolution neural networks (self-attentional CNN) to predict protein sequence profile, namely SPROF method. The deep neural network can process both 1D structural properties and a 2D distance map reflecting the continuous distances between residue pairs. To our best knowledge, this is the first study to utilize a 2D distance map for structure-based prediction of protein properties. The SPROF method achieved a sequence recovery rate of 39.8% on the independent test set, which is significantly higher than 34.6% achieved by the SPIN2 method trained from only 1D structural features. Further analysis indicated that the improvement was mostly contributed by residues most contacted with other residues, suggesting that the inclusion of 2D distance map can efficiently capture long-range contacted information. Therefore, such network architecture to utilize 2D distance map is expected to be useful for other 3D structure-based applications such as binding site prediction, protein function prediction, and protein interaction.

2. MATERIALS AND METHODS

2.1. Data Sets. Since a training deep learning network requires a large number of training samples, we employed the data set curated in 2017, as used in our previous study.[26] The data set consists of 12,450 nonredundant chains with resolution <2.5 Å, *R*-factor <1.0, sequence length ≥30, and sequence identity ≤25% from the cullpdb Web site. Among them, 11,200 chains deposited before June, 2015 were selected as a training set, and the remaining 1250 were used as an independent test set.

From this data set, we removed long chains with lengths ≥600 because the required memory for learning on these long chains is over the 12GB memory limitation of our used graphics processing unit (GPU) Nvidia GTX 1080 Ti. Finally, we kept a data set of 7134 chains for training and 922 chains for the test, namely TR7134 and TS922, respectively. To verify the robustness of our model, we also collected all single chains with structures deposited in the protein data bank from TBM-hard targets in CASP13. The data set consists of 22 chains, namely CASP13-TBM-hard.

2.2. Features Extraction. Our input features include both 1D structural features and 2D distance maps. The 1D structural features include 150 features that are similar to those used in SPIN2.[18] For completeness, we make a brief introduction on the 1D features.

1D Structural Features. The 1D structural features can be divided into four feature groups: the secondary structures (8), cosine and sine values of backbone angles φ, ψ, θ, ω, and τ (10), local fragment-derived profiles (20), and the global energy features (112), namely GF_SS, GF_AG, GF_FRAG, and GF_ENERGY, respectively. GF_GS are one-hot DSSP codes for eight-state protein secondary structures (C, G, H, I, T, E, B, S). GF_AG are sine and cosine values of five backbone angles φ, ψ, ω, θ and τ at each given position, where φ, ψ, and ω are three main-chain dihedral angles rotated along $N-C_\alpha$, $C_\alpha-C$, and C_i-N_{i+1} bonds, respectively, τ is a dihedral angle based on four neighboring C_α atoms $C_{\alpha i-1}-C_{\alpha i}-C_{\alpha i+1}-C_{\alpha i+2}$, and θ is an angle intervening $C_{\alpha i-1}-C_{\alpha i}-C_{\alpha i+1}$. GF_FRAG are the probabilities of 20 standard residue types at each position estimated from structurally similar fragments.[15] GF_ENERGY are the interaction energies of 20 standard residue types at a selected position with the rest of the backbone positions occupied by the alanine residue. The energies are computed by using the DFIRE statistical scoring function[29] based on preferred backbone states-dependent side-chain conformations as defined in the bbdep rotamer library.[30] If one residue type has >6 rotamers, only 6 most frequent rotamers were chosen. We also used the lowest energies among all rotameric states for each residue type. Finally, this generated a total of 112 (=(6 + 1) × 13 + (3 + 1) × 4 + (2 + 1) × 1 + 1 × 2) energy features. Different from SPIN2, we did not utilize distances between atoms within the same residue or belonging to neighboring residues since they might include residual information in the force field during the determination of protein 3D structures according to experimental data.

2D Distance Maps. In addition to the 1D structural features used by SPIN2, we derived an input feature of 2D distance matrix S (namely distance map) with its elements:

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

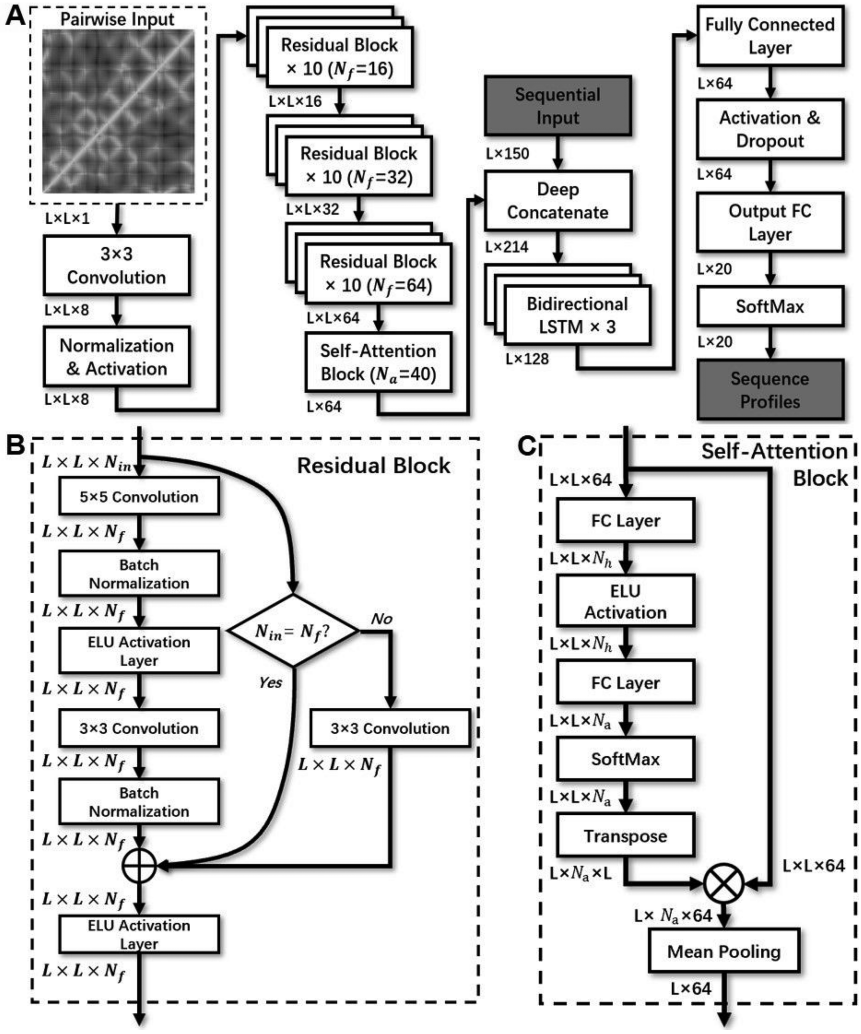


Figure 1. The neural network layout of SPROF with (A) illustrating the overall network architecture of SPROF, where L is the length of protein sequence. (B) Details of the residual block, where N_{in} is the number of input layers, and N_f represents the number of kernels in each convolution layer. (C) The structure of the multihead self-attention block, where N_a represents the number of parallel attention layers (chosen as 40 by fine-tuning), and N_h is the number of hidden states of fully connected (FC) layers (chosen as 50).

$$s_{ij} = \frac{2}{1 + \frac{\max(d_0, d_{ij})}{d_0}} \quad (1)$$

where d_{ij} is the distance between C_α atoms of residues i and j, and d_0 was set 4.0 Å, as also used in definition of the SP-score.[31] This conversion of distance ensures a score ranging from 0 to 1, with a good discrimination for distances between 4 and 8 Å. We did not use exactly the same formula as SP-score (d_{ij}^2 used in SP-score) since it produced slightly worse results (results not shown).

2.3. Deep Learning Method. The deep neural network (DNN) used in our previous work SPIN2[18] is a fully connected neural network, which is proved to be efficient in encoding 1D sequential information, but not powerful for processing a 2D distance map. The 2D distance map can be viewed as a special image, which is used to produce a protein sequence profile as the caption of the image. Inspired by the image captioning learning architecture, we have designed a deep learning network coupling RNN and CNN to extract features from 1D and 2D features, respectively. As shown in Figure 1A, a self-attentional ultradeep residual convolutional neural network (ResNet-CNN)[25] encoded the 2D distance map into a vector representation, which was then concatenated with our 1D structural features and fed into an RNN module to generate a protein sequence profile.

CNN Module. CNN has demonstrated superior performance in image tasks because of its implementation of shift, scale, and distortion invariance through local receptive fields, shared weights, and subsampling. Though the representation depth of CNN is beneficial for the classification,[32] the vanishing gradient problem has become a major obstacle to increasing the depth of CNN. In 2015, He et al. proposed ResNet, an ultradeep residual neural network to solve the vanishing gradient problem by employing shortcut connection between outputs of a convolution layer and its previous layer.[25] ResNet has been widely used in the task of protein contact map prediction.[26,27] The ResNet used in our model differs in the fact that we employed a self-attention mechanism[33] to convert the output-size of ResNet module from $L \times L \times N_f$ to $L \times N_f$, where L is the length of protein sequence and N_f represents the kernels amount for the last convolution layer.

RNN Module. The features from the CNN module and 1D structural features were concatenated together and fed into a bidirectional long short-term memory recurrent neural net-

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

work (LSTM-BRNN or BiLSTM) to generate the protein sequence profile. Unlike standard feed forward neural networks, RNN retains a state that can represent information from an arbitrarily long context window.[34] However, traditional RNNs have no ability to learn long-range dependencies as a result of gradient vanishing problems. To overcome this problem, Hochreiter and Schmidhuber proposed the LSTM technique[35] using carefully designed nodes with recurrent edges of fixed unit weight.[36] Later, RNN with bidirectional LSTM to exploit both preceding and following dependencies was proposed and has been proved to outperform unidirectional ones in framewise phoneme classification.[37] Currently, LSTM-BRNN has been widely used in many bioinformatics studies.[38,39]

2.4. Neural Networks Implementation Details. In SPROF, 2D distance map ($L \times L$ with L as the protein length) was encoded by the self-attentional ResNet into sequential tensors ($L \times 64$), and was then concatenated with 1D structural features ($L \times 150$). The concatenated features ($L \times 214$) were fed into a bidirectional LSTM to generate a decoded tensor of $L \times 128$. Finally, a series of fully connected (FC) layers and activation layers conducted nonlinear-transformations on the output of bidirectional LSTM to obtain a prediction result of ($L \times 20$), which represented possibilities of 20 amino acid types on each sequence position.

Our self-attentional ResNet module is composed of a series of residual blocks (Figure 1B) and a self-attention block (Figure 1C).

Residual Block. Our residual block employed an exponential linear unit (ELU) as activation layer instead of a rectified linear unit (ReLU) used by typical residual blocks. The ELU activation function was shown to be more effective than standard ReLU in learning of the ResNet.[40] Furthermore, before each activation layer, regularization was applied to the network through the use of batch normalization.[41] Considering the limitation of the used GPU memory size (12GB), we employed 30 ResNet blocks ($30 \times 2 = 60$ convolution layers) in our final model. It should be noted that a smaller window size with more convolution layers is beneficial for the performance of CNN.[32] Thus, the kernel size of two convolution layers in each residual block was chosen as 5×5 or 3×3, respectively.

Self-Attention Block. The self-attention block shown in Figure 1C converts feature size from $L \times L \times 64$ into $L \times 64$. Mathematically, self-attention mechanism is implemented by a fully connected neural network, SoftMax, and matrix dot-multiplication. The SoftMax is applied on the second dimension, making elements in each slice sum up to 1. Biologically, the i-th slice represents the i-th residue, and the j-th element in this slice represents the importance of the j-th residue on the i-th residue. The attention mechanism was designed to catch the most affecting residues for prediction of each given residue. As shown in Figure S1A, the multihead attention map highlights the most important residue pairs, which are often spatially close (shown in Figure S1B).

Handling Variable Length Inputs. Different from general image tasks that often preprocess images to the same size, protein sequence profile prediction has to handle proteins of variable sizes. Therefore, we had to design a CNN that could process inputs of variable sizes and ensure the output to have a size equaling to its size. Finally, our neural networks do not have pooling layers as CNN networks often do, and the output of the last residual block remains the same value (L) of width and height.

Bidirectional LSTM. The input of our bidirectional LSTM is in size of $L \times 214$. The bidirectional LSTM module consists of three layers. In each layer, there are two independent LSTM representing two directions, respectively. Our LSTM cells consist of 64 one-cell memory blocks, culminating in 128 hidden states for each bidirectional LSTM layer.

Linear Layers. Our linear layers are fully connected. The first FC layer consists of 64 nodes plus a bias node with an ELU activation. The FC output layer has 20 output neurons and a sigmoid activation to convert the output into a likelihood of each amino acid type at each position ($L \times 20$).

Tools. We trained our model in the framework of Facebook's PyTorch library (v0.4.0), which enables us to accelerate the model training on an Nvidia GeForce GTX 1080 GPU. It has been shown that the use of a GPU for training a neural network can speed up by a factor up to 20.[42]

Optimization Algorithm and Dropout. Our model was trained with cross entropy as the loss function and ADAM algorithm for optimization.[43] ADAM optimization algorithm is generally considered to be robust for the selection of hyperparameters and converges more quickly than the traditionally used stochastic gradient descent (SGD). We used a learning rate of 0.0005 in this study. Furthermore, a 50% dropout rate was adopted at the output of the fully connected layer during training to reduce overfitting.[44]

Hyperparameters Tuning by the Cross Validation. The architecture and hyperparameters were optimized by the 5-fold cross validation, where the training set was randomly divided into five different subsets. Each time four of these subsets were used to train a model, and the left one was used for the test. This process was repeated for five times so that all five subsets were tested exactly once, and the average accuracy over five tests was used for the overall performance. With the hyperparameters achieving the best performance, the final model was trained on the whole training set and tested on the independent test set.

Evaluation. We evaluated the performance by the native sequence recovery rate that is the percentage of residues that were correctly predicted. A residue was considered to be correctly predicted if the wild-type residue type has the highest value in the predicted profile for 20 residue type at the position. As the evaluation metric based on one-to-one mapping is strict, we employed the other evaluation metric called positively matched rate. For this metric, we considered a prediction is correct if the predicted and actual amino acids have a positive value in the BLOSUM62 matrix.[45]

In addition, we also evaluated the performance of different types of residues, we calculated precision and recall for residue R as

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

and

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

,where TP is the number of correctly predicted residues for type R, FP is the number of residues wrongly predicted as R, and FN is the number of incorrectly predicted residues of wild-type R.

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

3. RESULTS

3.1. Model Selection and Feature Importance. Table 1 illustrates the performance of SPROF and its variants with different network architectures. SPROF achieved sequence recovery of 39.9% and 39.8% on the 5-fold cross validation (CV) and independent tests, respectively. The consistent results indicate the robustness of the SPROF method. An exclusion of the self-attention block (SPROF-noAtt) caused a decrease of 0.9% in the CV and 1.2% in the independent tests. The removal of RNN module and 1D structural features (SPROF-CNN) decreased the sequence recovery to 36.1% in the CV. The slightly higher recovery rate (36.3%) in the test set than the CV should result from random fluctuations. The greatest drop was from SPROF-RNN that excluded the CNN module and 2D distance map. This caused a 6% and 6.6% drop of native sequence recoveries in the training set and independent test, respectively. The results demonstrate the benefits of utilizing the distance map features and image captioning method.

Table 1. Native Sequence Recovery Rates Achieved by the SPROF and Its Variants on 5-Fold Cross Validation and Independent Test

model	TR7134 (cross validation)	TS922 (test)
SPROF[a]	39.9%	39.8%
SPROF-noAtt[b]	39.0%	38.6%
SPROF-CNN[c]	36.1%	36.3%
SPROF-RNN[d]	33.9%	33.2%

[a]The best performed model, details shown in Figure 1. [b]SPROF without self-attention block. [c]Using the CNN module without bidirectional LSTM module or input of 1D structural features. [d]Using the RNN module without self-attentional ResNet module or input of 2D distance maps.

It is of interest to see which type of features made the greatest contribution in the prediction. We excluded each type of features one-by-one to obtain five different feature sets for model training and then compared the performance of each model. Table 2 shows the sequence recoveries of these five models on the independent test set. As expected, 2D distance map features contributed the most in the sequence recovery (contributing 6.6% on independent test), followed by energy-based features (1.8%) that made the highest contribution in the SPIN2. The exclusion of fragment-based features made overall sequence recovery 0.7% lower, and the exclusion of secondary structure features or backbone torsion angles features also marginally decreased the overall sequence recovery (0.3% and 0.2%, respectively). These results highlight the importance of distance map in our prediction model, which inspire us to employ distance map features on other 3D structure-based applications in future.

Table 2. Comparison of Sequence Recoveries after Excluding One Feature Group From SPROF

feature excluded	TR7134 (cross validation)	TS922 (test)
SPROF	39.9%	39.8%
distance map	33.9%	33.2%
GF_ENERGY	37.8%	38.0%
GF_FRAG	38.7%	39.1%
GF_SS	39.6%	39.5%
GF_AG	39.7%	39.6%

3.2. Comparison with Other Methods. We further made a direct comparison with SPIN2 on the test set TS922 and CASP13-TBM-hard. As shown in Table 3, there is about 5% consistent improvement from SPIN2 to SPROF for the native sequence recovery rates on both the TS922 and CASP13-TBM-hard data sets. When evaluated by the positively matched rate, both methods have higher absolute values, and our method has a consistent improvement of around 5% over SPIN2. The performances of CASP13-TBM-hard targets by both methods are detailed in Table S1. Since Wang's method[17] is not available online, we cannot make a direct comparison. According to published results, SPIN2 and Wang's method should be close because the sequence recovery of SPIN2 is over 4% higher than SPIN's, while Wang's method is about 3% higher than SPIN's.

Table 3. Performance Comparison of SPIN2 and SPROF on TS922 and CASP13-TBM-hard Datasets

	performance evaluation			
	sequence recovery rate[a]		positively matched rate[b]	
method	TS922	CASP13-TBM-hard	TS922	CASP13-TBM-hard
SPROF	39.8%	39.2%	59.9%	57.2%
SPIN2	34.6%	34.6%	54.3%	52.1%

[a]The proportion of matched residues between prediction and target. [b]The proportion of residues in the prediction sequence that have positive values in BLOSUM62 with residues in the target sequence.

We compared the performance of SPROF, SPROF-CNN, SPROF-RNN, and SPIN2 for proteins chains with different lengths on TS922. As shown in Figure 2A, SPROF consistently outperformed SPIN2 in all intervals and SPROF-CNN model is somewhere in between. SPROF-RNN model is less accurate than SPIN2, likely because SPROF-RNN model excluded partial features employed by SPIN2. A direct comparison of the sequence recovery rates (Figure 2B) suggests that SPROF is significantly better than SPIN2 (P-value $<10^{-99}$) according to the pairwise t test, where SPROF outperformed SPIN2 on 815 protein chains, worse on 76 chains, and tied on the remained 31 chains.

For a given residue type, we compared the recall and precision score of SPROF and SPIN2 on TS922, as shown in Figure 2C,D, respectively. SPROF outperformed SPIN2 in 15 (75% of 20) amino acids types for recall and 17 (85% of 20) for precision.

We noticed that in Figure 2B, there are a few chains with high sequence recoveries (~50−60%) and a few with very low ones (~10−20%). Table S2 lists chains with 10 highest and 10 lowest sequence recovery rates. Chains of low recovery rates are mostly short chains and thus do not have a well-formed hydrophobic core. On the other hand, the 10 chains of highest sequence recovery rates show higher proportions of buried residues than the average.

To explore why SPROF outperformed SPIN2, we plotted the prediction accuracy of residues as a function of their contact number for different methods. The contact number was defined as the number of neighboring C_α atoms no farther than 13 Å from a given C_α atom. As shown in Figure 3A, SPROF and SPROF-CNN show an increase of prediction accuracy with the addition of neighbored residues. By

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

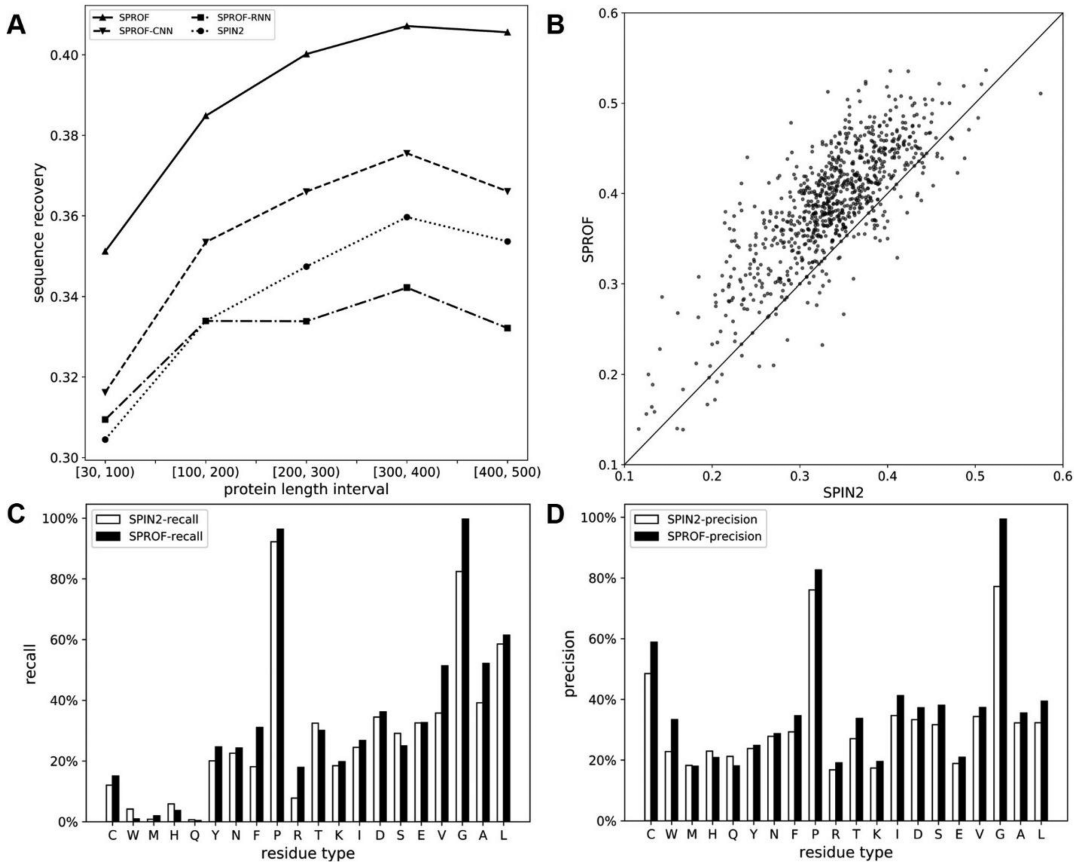


Figure 2. (A) The average sequence recovery rates of protein chains in different length intervals by four methods. (B) The sequence recovery for each chain in TS922 by SPROF and SPIN2. (C) The recall and (D) precision for different amino acids residue types by SPIN2 and SPROF over TS922.

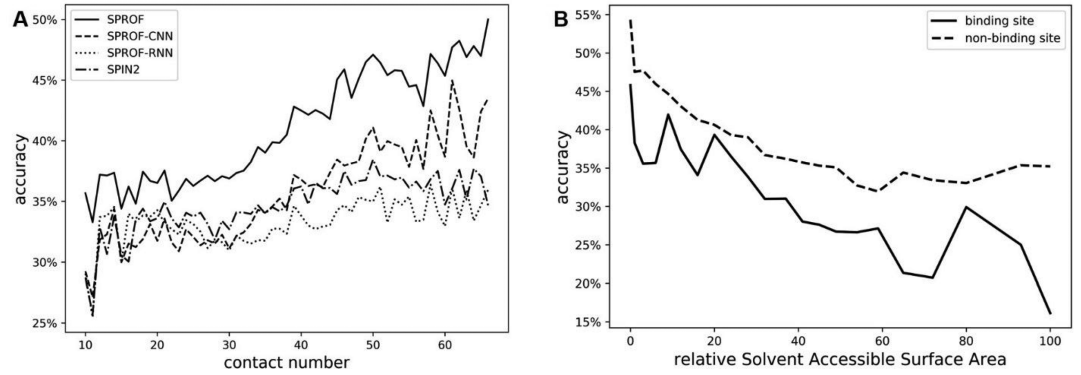


Figure 3. (A) The accuracy for residues as a function of their contact numbers for SPROF and SPIN2. (B) The line plot of prediction accuracy for residues in different rASA intervals and binding or nonbinding site on 357 chains (overlap with BioLip) of TS922.

comparison, SPROF-RNN and SPIN2 without using a 2D distance map show close to flat performances for all residues. This comparison indicated that the inclusion of 2D distance map helped the model to capture information on residues contacted in 3D structure.

We also compared the prediction accuracies between binding and nonbinding sites. By mapping the protein chains of TS922 to those defined in BioLip,[46] we generated a data set of 357 chains data set. As shown in Figure 3B, the prediction accuracies of residues decrease with the relative solvent surface area, and the accuracy of binding residues is consistently lower than that of nonbinding residues. This is as expected because buried residues maintain 3D spatial structures, and binding residues are evolved mainly for protein function, and not necessary to be optimized for 3D structure.

3.3. Case Study. To illustrate our method, we chose the precorrin-6A reductase cobK (pdb ID: 5c4n chain D) for comparison of methods. This protein chain contains 8 helical and 12 β sheet fragments, in total 244 amino acids. For a clear look of the predicted sequence profile, we plotted sequence logos for fragments of residue index 75-104, the red part in the Figure 4A. SPROF and SPIN2 achieved accuracies of 60% and 26.7% for this fragment, respectively. As shown in the Figure 4C,D, SPROF has made a correct prediction for 11 amino acids (red amino acids in Figure 4E) that are not correctly predicted by SPIN2. A deep look indicates most of the amino

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

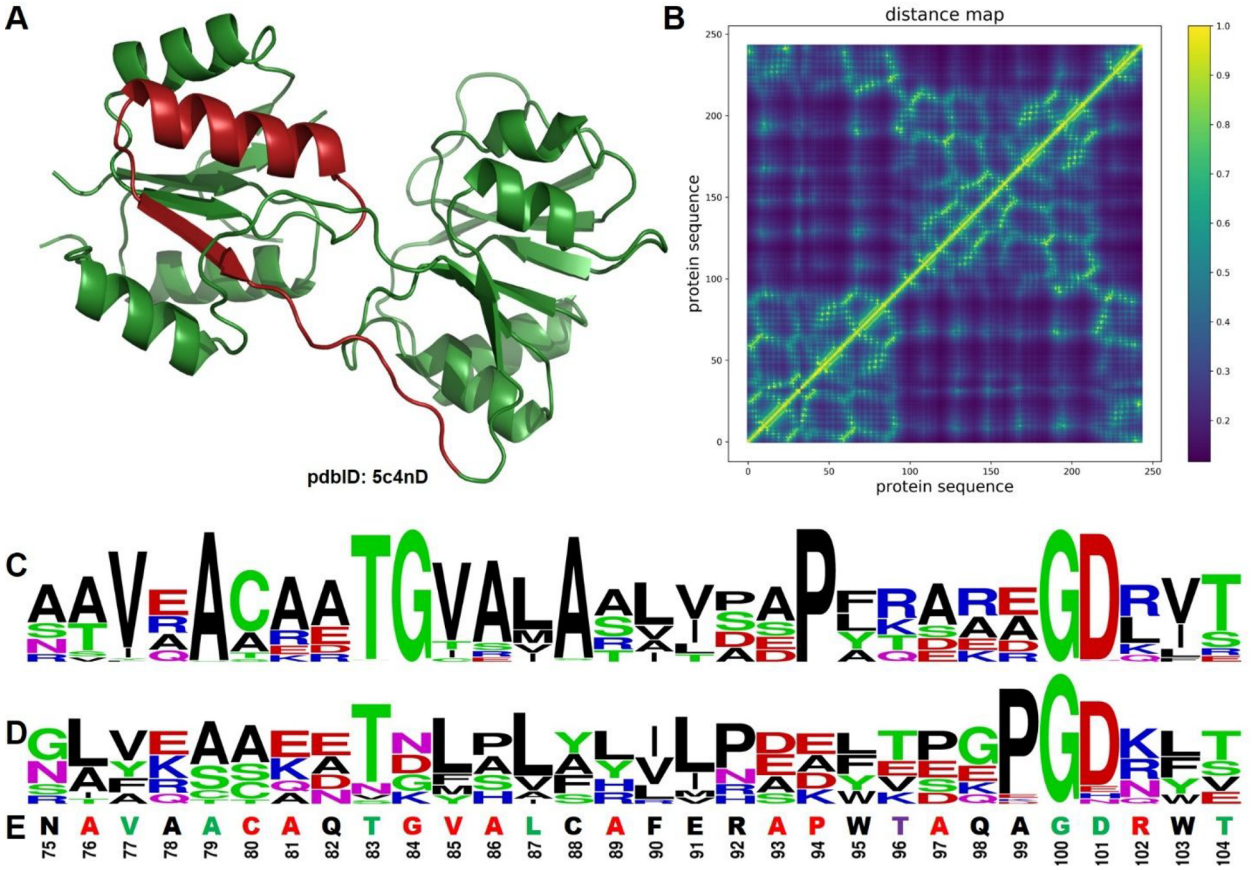


Figure 4. (A) 3D structure, (B) distance map, and sequence logo generated by (C) SPROF and (D) SPIN2 for the precorrin-6A reductase cobK (pdb ID: 5c4nD). For a clear look, only fragment 75-104 (red in the 3D structure) was shown in the sequence logo generated by SPIN2 and SPROF. The wild-type sequence and indexes were provided in (E) with red, purple, green, and black for correct prediction of amino acids by SPORF only, SPIN2 only, both methods, and none, respectively.

acids in the list are hydrophobic (6 alanine and 1 valine). This result is consistent with our expectation because our method is better for predicting most contacted residues that are frequently hydrophobic amino acids. SPROF only misses one prediction (no. 96) that is correctly predicted by SPIN2. On this position, the native amino acid (threonine) ranked the third by our prediction.

4. CONCLUSIONS

This study highlights the power of applying an image captioning method on a 2D distance map for protein sequence profile prediction. We proposed a protein sequence profile prediction method SPROF which combined recurrent neural network, convolution neural network, and attention mechanism. SPROF has improved the native sequence recovery from 34.6% (previous method SPIN2) to 39.8% on our independent test set. The improvement is consistent regardless of protein lengths, test sets (cross validation and independent test), evaluation metrics (sequence recovery or positively matched rate), or types of amino acids (in precision or recall score). We also trained a model by using only 1D structural features, which is significantly lower than SPROF with inclusion of 2D distance map. This is reasonable because distance maps are capable of encoding the 3D structural information on proteins. Inspired by an image captioning method, SPROF is capable of extracting these 3D structural information and thus obtains higher accuracy for sequence prediction. More importantly, this study has provided a new architecture to advance structure-based predictions, such as predictions of protein interaction,[47] binding site,[48] and protein function.[49] For example, we have recently combined this architecture for prediction of protein−drug interaction.[50] Moreover, there is a significant progress to make a prediction of contact map from sequence by combining deep learning and genomic big data,[26,51] which provides a potential way to employ a predicted 2D contact map to substitute or concatenate with the one-hot encoding for sequential information in tasks like protein function prediction.[52]

The generated sequence profiles could be integrated into software like Rosetta to further improve the result of protein design, as reported in previous study.[15] To validate the ability of such designed sequences to fold into the respective scaffold, either computational[53,54] or experimental[55] validation could be applied in the future. Moreover, the generated sequence profiles have been proven beneficial for improving existing fold recognition technique studies,[14,56] so our improved prediction of the sequence profile might advance the structure prediction. In addition, the significant difference of the native residue recovery between binding and nonbinding residues may be helpful for discriminating functional residues.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00438.

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

A pair of figures indicating the capability of self-attention mechanism to catch the most important residue pairs. A table comparing the performance of SPIN2 and SPROF on CASP13-TBM-hard test set. A table listing chains with the top10 highest or lowest sequence recovery rates in the TS922 test set (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: yangyd25@mail.sysu.edu.cn.
*E-mail: zhaohy8@mail.sysu.edu.cn.

ORCID

Sheng Chen: 0000-0003-1428-6778
Huiying Zhao: 0000-0001-9429-3016
Yuedong Yang: 0000-0002-6782-2813

Author Contributions

C.S., Z.H., and Y.Y. designed the method. C.S. and S.Z. developed and implemented methods and produced results. C.S., S.Z., L.L, Z.H., and Y. Y. wrote the manuscript. All authors read and approved the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work has been supported in part by the National Key R&D Program of China (2018ZX10301402), National Natural Science Foundation of China (U1611261, 61772566, and 81801132), Guangdong Frontier & Key Tech Innovation Program (2018B010109006, 2019B020228001), and Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

REFERENCES

(1) Liu, H.; Chen, Q. Computational Protein Design for Given Backbone: Recent Progresses in General Method-Related Aspects. *Curr. Opin. Struct. Biol.* **2016**, *39*, 89−95.
(2) Silva, D. A.; Yu, S.; Ulge, U. Y.; Spangler, J. B.; Jude, K. M.; Labao-Almeida, C.; Ali, L. R.; Quijano-Rubio, A.; Ruterbusch, M.; Leung, I.; Biary, T.; Crowley, S. J.; Marcos, E.; Walkey, C. D.; Weitzner, B. D.; Pardo-Avila, F.; Castellanos, J.; Carter, L.; Stewart, L.; Riddell, S. R.; Pepper, M.; Bernardes, G. J. L.; Dougan, M.; Garcia, K. C.; Baker, D. De Novo Design of Potent and Selective Mimics of Il-2 and Il-15. *Nature* **2019**, *565*, 186−191.
(3) Li, Z.; Yang, Y.; Zhan, J.; Dai, L.; Zhou, Y. Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. *Annu. Rev. Biophys.* **2013**, *42*, 315−335.
(4) Treynor, T. P.; Vizcarra, C. L.; Nedelcu, D.; Mayo, S. L. Computationally Designed Libraries of Fluorescent Proteins Evaluated by Preservation and Diversity of Function. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 48−53.
(5) Guntas, G.; Purbeck, C.; Kuhlman, B. Engineering a Protein-Protein Interface Using a Computationally Designed Library. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19296−19301.
(6) Allen, B. D.; Nisthal, A.; Mayo, S. L. Experimental Library Screening Demonstrates the Successful Application of Computational Protein Design to Large Structural Ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19838−19843.
(7) Hayes, R. J.; Bentzien, J.; Ary, M. L.; Hwang, M. Y.; Jacinto, J. M.; Vielmetter, J.; Kundu, A.; Dahiyat, B. I. Combining Computational and Experimental Screening for Rapid Optimization of Protein Properties. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 15926−15931.
(8) Dahiyat, B. I.; Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **1997**, *278*, 82−87.
(9) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**, *332*, 449−460.
(10) Lippow, S. M.; Tidor, B. Progress in Computational Protein Design. *Curr. Opin. Biotechnol.* **2007**, *18*, 305−311.
(11) Liu, Y.; Kuhlman, B. Rosettadesign Server for Protein Design. *Nucleic Acids Res.* **2006**, *34*, W235−8.
(12) Regan, L.; DeGrado, W. F. Characterization of a Helical Protein Designed from First Principles. *Science* **1988**, *241*, 976−978.
(13) Pierce, N. A.; Winfree, E. Protein Design Is Np-Hard. *Protein Eng., Des. Sel.* **2002**, *15*, 779−782.
(14) Zhou, H.; Zhou, Y. Fold Recognition by Combining Sequence Profiles Derived from Evolution and from Depth-Dependent Structural Alignment of Fragments. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 321−328.
(15) Dai, L.; Yang, Y.; Kim, H. R.; Zhou, Y. Improving Computational Protein Design by Using Structure-Derived Sequence Profile. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 2338−2348.
(16) Li, Z.; Yang, Y.; Faraggi, E.; Zhan, J.; Zhou, Y. Direct Prediction of Profiles of Sequences Compatible with a Protein Structure by Neural Networks with Fragment-Based Local and Energy-Based Nonlocal Profiles. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 2565−73.
(17) Wang, J.; Cao, H.; Zhang, J. Z.; Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **2018**, *8*, 6349.
(18) O'Connell, J.; Li, Z.; Hanson, J.; Heffernan, R.; Lyons, J.; Paliwal, K.; Dehzangi, A.; Yang, Y.; Zhou, Y. Spin2: Predicting Sequence Profiles from Protein Structures Using Deep Neural Networks. *Proteins: Struct., Funct., Genet.* **2018**, *86*, 629−633.
(19) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.
(20) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A.; De Fabritiis, G. Deepsite: Protein-Binding Site Predictor Using 3d-Convolutional Neural Networks. *Bioinformatics* **2017**, *33*, 3036−3042.
(21) Liu, K.; Sun, X.; Ma, J.; Zhou, Z.; Dong, Q.; Peng, S.; Wu, J.; Tan, S.; Blobel, G.; Fan, J. Prediction of Amino Acid Side Chain Conformation Using a Deep Neural Network. *arXiv preprint arXiv:1707.083812017*https://arxiv.org/abs/1707.08381 (Accessed on July 26, 2017).
(22) Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep Convolutional Networks for Quality Assessment of Protein Folds. *Bioinformatics* **2018**, *34*, 4046−4053.
(23) Skolnick, J.; Kolinski, A.; Ortiz, A. R. Monsster: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *J. Mol. Biol.* **1997**, *265*, 217−241.
(24) Adhikari, B.; Cheng, J. Confold2: Improved Contact-Driven Ab Initio Protein Structure Modeling. *BMC Bioinf.* **2018**, *19*, 22.
(25) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *In Proc. IEEE Conf. Comput. Vis. Pattern Recog.* **2016**, *2016*, 770−778.
(26) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* **2018**, *34*, 4039−4045.
(27) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13*, No. e1005324.
(28) Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. *In Proc. IEEE Conf. Comput. Vis. Pattern Recog.* **2015**, *2015*, 3156−3164.
(29) Yang, Y.; Zhou, Y. Ab Initio Folding of Terminal Segments with Secondary Structures Reveals the Fine Difference between Two Closely Related All-Atom Statistical Energy Functions. *Protein Sci.* **2008**, *17*, 1212.
(30) Dunbrack, R. L., Jr; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci.* **1997**, *6*, 1661−1681.
(31) Yang, Y.; Zhan, J.; Zhao, H.; Zhou, Y. A New Size-Independent Score for Pairwise Protein Structure Alignment and Its Application to

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399

Structure Classification and Nucleic-Acid Binding Prediction. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 2080−2088.
(32) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* 2014, https://arxiv.org/abs/1409.1556 (Accessed on April 10, 2015).
(33) Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-Attentive Sentence Embedding. *arXiv preprint arXiv:1703.03130* 2017, https://arxiv.org/abs/1703.03130 (Accessed on March 9, 2017).
(34) Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. Proceedings from the *Eleventh Annual Conference of the International Speech Communication Association, 2010*, Makuhari, Chiba, Japan, September 26−30, 2010; International Speech Communication Association: Baixas, France, 2010.
(35) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.
(36) Lipton, Z. C.; Berkowitz, J.; Elkan, C. A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv preprint arXiv:1506.00019* 2015, https://arxiv.org/abs/1506.00019 (Accessed on Oct 17, 2015).
(37) Graves, A.; Schmidhuber, J. Framewise Phoneme Classification with Bidirectional Lstm and Other Neural Network Architectures. *Neural Netw.* **2005**, *18*, 602−610.
(38) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing Non-Local Interactions by Long Short-Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers and Solvent Accessibility. *Bioinformatics* **2017**, *33*, 2842−2849.
(39) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving Protein Disorder Prediction by Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. *Bioinformatics* **2016**, *33*, 685−692.
(40) Shah, A.; Kadam, E.; Shah, H.; Shinde, S.; Shingade, S., Deep Residual Networks with Exponential Linear Unit. *arXiv preprint arXiv:1604.04112* 2016, https://arxiv.org/abs/1604.04112 (Accessed on Oct 5, 2016).
(41) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167* 2015, https://arxiv.org/abs/1502.03167 (Accessed on March 2, 2015).
(42) Oh, K.-S.; Jung, K. Gpu Implementation of Neural Networks. *Pattern Recogn.* **2004**, *37*, 1311−1314.
(43) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* 2014, https://arxiv.org/abs/1412.6980 (Accessed on Jan 30, 2017).
(44) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(45) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.
(46) 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.
(47) Litfin, T.; Zhou, Y.; Yang, Y. Spot-Ligand 2: Improving Structure-Based Virtual Screening by Binding-Homology Search on an Expanded Structural Template Library. *Bioinformatics* **2017**, *33*, 1238−1240.
(48) 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.
(49) Zhao, H.; Yang, Y.; Zhou, Y. Highly Accurate and High-Resolution Function Prediction of Rna Binding Proteins by Fold Recognition and Binding Affinity Prediction. *RNA Biol.* **2011**, *8*, 988−996.
(50) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting Drug Protein Interaction Using Quasi-Visual Question Answering System.-*bioRxiv preprint bioRxiv* 2019, https://www.biorxiv.org/content/10.1101/588178v1 (Accessed on March 25, 2019).
(51) Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D. Protein Structure Determination Using Metagenome Sequence Data. *Science* **2017**, *355*, 294−298.
(52) Kulmanov, M.; Hoehndorf, R. Deepgoplus: Improved Protein Function Prediction from Sequence. *Bioinformatics* **2019**, 615260.
(53) Greener, J. G.; Moffat, L.; Jones, D. T. Design of Metalloproteins and Novel Protein Folds Using Variational Autoencoders. *Sci. Rep.* **2018**, *8*, 16189.
(54) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19−25.
(55) Xiong, P.; Hu, X.; Huang, B.; Zhang, J.; Chen, Q.; Liu, H., Increasing the Efficiency and Accuracy of the Abacus Protein Sequence Design Method. *Bioinformatics* **2019**. DOI: 10.1093/bioinformatics/btz515
(56) Schmidt am Busch, M.; Mignon, D.; Simonson, T. Computational Protein Design as a Tool for Fold Recognition. *Proteins: Struct., Funct., Genet.* **2009**, *77*, 139−158.

DOI: 10.1021/acs.jcim.9b00438
J. Chem. Inf. Model. 2020, 60, 391−399