JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 2033−2042

pubs.acs.org/jcim

Detecting Proline and Non-Proline Cis Isomers in Protein Structures from Sequences Using Deep Residual Ensemble Learning

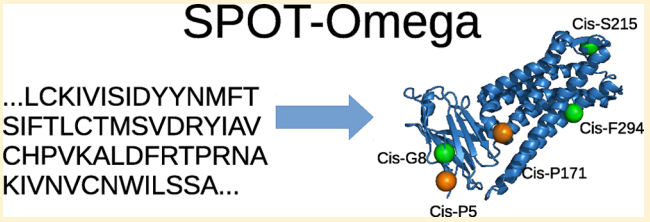
Jaswinder Singh,[†] Jack Hanson,[†] Rhys Heffernan,[†] Kuldip Paliwal,[†] Yuedong Yang,[‡,¶] and Yaoqi Zhou*[,‡]

[†]Signal Processing Laboratory, Griffith University, Brisbane, QLD 4122, Australia

[‡]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia

[¶]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong 510006, China

ABSTRACT: It has been long established that cis conformations of amino acid residues play many biologically important roles despite their rare occurrence in protein structure. Because of this rarity, few methods have been developed for predicting cis isomers from protein sequences, most of which are based on outdated datasets and lack the means for independent testing. In this work, using a database of >10000 high-resolution protein structures, we update the statistics of cis isomers and develop a sequence-based prediction technique using an ensemble of residual convolutional and long short-term memory bidirectional recurrent neural networks that allow learning from the whole protein sequence. We show that ensembling eight neural network models yields maximum Matthews correlation coefficient values of approximately 0.35 for *cis*-Pro isomers and 0.1 for *cis*-nonPro residues. The method should be useful for prioritizing functionally important residues in cis isomers for experimental validations and improving the sampling of rare protein conformations for ab initio protein structure prediction.



INTRODUCTION

Protein backbone structures can be characterized by three dihedral angles representing rotations around the N−C_α bond (ϕ), the C_α−C bond (ψ), and the C−N bond between two residues (ω). Unlike ϕ and ψ, which vary from −180° to 180°, the resonance in the C−N amide bond results in partial double-bond characteristics and causes atoms $C_\alpha(i)$, $C(i)$, $O(i)$, $N(i+1)$, $H(i+1)$, and $C_\alpha(i+1)$ to be approximately in the same plane. As a result, ω is bound to either approximately 180° in the trans conformation or approximately 0° in the cis conformation. Because of steric restraints, the trans conformation is energetically more favorable than the cis conformation by 2.5−2.6 kcal/mol for most residues. However, this energetic difference is only 0.5 kcal/mol for proline residues because of the cyclic side chain. As a result, while only 0.03% of Xaa−NonPro (where Xaa refers to any amino acid residue type) bonds are in the cis conformation, 5.2% of Xaa−Pro bonds in the Protein Data Bank (PDB) are cis isomers.[1] For these rare cases, the relative inherent instability of the cis isomer compared with the trans isomer has to be overcome by its nonlocal interactions with other residues that are close in three-dimensional structure but far away in sequence position. This extra evolutionary effort must occur for a structurally/functionally important reason.

Indeed, rare cis conformations are often located at functionally important regions[2] such as active sites[3] and binding interfaces[4,5] and are more conserved than trans conformations.[6] Cis−trans isomerization has been found to serve as molecular switches,[7] channel gaters,[8] protein stabilizers,[9] and expression regulators.[10]

The biological importance of thermodynamically stable cis conformations makes it important to identify residues with this rare conformation. Relying completely on experimental techniques such as NMR spectroscopy and X-ray crystallography is impractical because the structures of a massive number of proteins are unknown. Thus, it is highly desirable to employ computational methods to predict residues in cis conformations prior to experimental validations.

Unlike the prediction of protein secondary structure and backbone angles ϕ, ψ, θ, and τ,[11] only a few methods have been developed for detecting cis−trans isomerization of residues, many of which predict only for imide (Xaa−Pro) bonds and not amide (Xaa−nonPro) bonds in a protein. The first method for *cis*-proline prediction was developed by Frömmel and Preissner[12] using a vector of physicochemical properties for six neighboring residues, obtaining a recovery rate of 75% for 235 known *cis*-prolyl residues without false positives. Analysis of X-ray structures revealed residue type, secondary structure, and solvent accessibility preferences in the neighborhood of cis peptide bonds.[13] Later, Wang et al.[14] employed a one-hot encoding representation of the amino acid sequence in a 20-residue window as input to a support vector machine (SVM)[15] to achieve about 70% accuracy for an

Received: July 6, 2018
Published: August 17, 2018

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

independent test set of 1159 *cis*- and 5080 *trans*-prolyl samples. Song et al.[16] showed the importance of evolutionary sequence profile and predicted secondary structure in 21-residue windows and obtained an accuracy of 71.5% for 5-fold cross-validation of 2424 nonhomologous proteins. Exarchos et al.[17] analyzed patterns surrounding *cis*-nonPro peptide bonds and found that structural similarity has the most discriminative power in resolving ω angles. More recently, several methods for predicting *cis*-prolyl conformations were developed by using intelligent voting of multiple SVM models with evolutionary information as input and achieved a reported accuracy of >80%.[18,19] However, their results were based on a 1:1 ratio of cis and trans samples without independent tests. To the best of our knowledge, none of the computational servers for all of these methods are available except for CISPEPpred.[16] Our evaluation reveals its low precision (5%) and sensitivity (12%).

Several methods for predicting cis conformations for all 20 residue types have also been developed. Pahlke et al.[20] developed a rule-based technique (secondary structure and amino acid propensities) for all cis residues. Exarchos et al.[21] employed multiple SVM models with selected features from evolutionary profiles, secondary structure, solvent accessibility, and physicochemical properties in an 11-residue window and achieved 70% accuracy, 75% sensitivity, and 71% precision on a fully balanced data set. The actual performance in a real-world application (severely unbalanced data) is unknown, as no server is available for making an independent assessment.

One interesting observation is the lack of neural networks and other modern machine learning technologies to predict residue cis conformations. In contrast, the prediction of secondary structure or local backbone angles other than ω has been improved greatly by neural-network-based techniques.[11] This could perhaps be due to the hitherto limited and unbalanced data in the training and test sets, particularly for non-prolyl residues. Nevertheless, the effectiveness of deep feature abstraction and the propagation of long-range nonlocal interactions throughout the whole protein sequence have been shown to contribute to significant improvements in several bioinformatics areas, such as intrinsically disordered regions in proteins,[22] protein contact map prediction,[23,24] microRNA secondary structure,[25,26] and, most relevant to this work, prediction of protein secondary structure,[27,28] including backbone torsion angles and other local structural properties.[29−32] This has been achieved through the use of deep long short-term memory bidirectional recurrent neural networks (LSTM-BRNNs),[33−35] and convolutional neural networks (CNNs).[36]

LSTM-BRNNs are designed to be effective in propagating long-range dependencies throughout an entire sequence. This ability is important in protein analysis, where interacting structural neighbors may be distantly separated in the protein primary sequence. CNNs also have this potential, but because of their typically small kernel sizes, they lack the network depth necessary to propagate information throughout particularly long sequences. Residual connections in ResNets (residual CNNs)[37,38] have overcome this problem by allowing CNNs to have a greatly increased depth by minimizing the vanishing gradient problem in neural network training. Our previous work showed that these residually-connected networks, like those in ResNets and residual LSTMs (ResLSTMs),[39] can be combined to provide state-of-the-art results in protein contact map prediction.[23] Convolutional and recurrent architectures are particularly useful in protein sequence learning because of their ability to accept variable-length inputs, as the base module can be "unfolded" for each sequence position. Each sequence residue is treated dependently on its surrounding sequence context.

In this paper, we first update the statistics of cis isomers using more than 10 000 nonredundant high-resolution protein structures. We have found that 4.6% of proline and 0.14% of non-proline residues are in cis isomers. The former resembles the values presented in Jabs et al.[1] from their small data set, whereas the latter is about 5 times higher. Moreover, we have found that more than 50% of proteins with 190 residues or longer have at least one cis isomer. Here we applied an ensemble of selected ResNet and ResLSTM network topologies to the problem of protein ω angle prediction to exploit the benefits of capturing both long- and short-range dependencies in sparse data. This model, named SPOT-Omega, achieves highly precise prediction for both imide and amide bonds without being trained specifically for either case. This performance is also obtained without training on an artificially balanced data set and thus is a more useful guide for real-world applications. This work confirms the effectiveness of capturing long-range nonlocal dependencies in local protein structure analysis despite the sparsity of the class labels. Independent tests on publicly available datasets also set a new benchmark for future ω angle prediction analysis. SPOT-Omega is available as a web server and as a stand-alone program along with training and test sets at http://sparks-lab.org.

METHODOLOGY

The Machine Learning Approach. We conducted several tests on different models to find the best approach for predicting ω angles. On the basis of results from our previous paper,[23] we form an ensemble of high-performing ResNets, ResLSTMs, and hybrid ResLSTM/ResNets for ω angle prediction.

The ResNet segments in our models consist of consecutive residual blocks, as illustrated in Figure 1A. Each residual block in our model follows the preactivation ResNet architecture shown in He et al.,[38] with each convolution using a one-dimensional (1D) kernel of size 3 with a varying number of filters and an exponential linear unit (ELU) activation

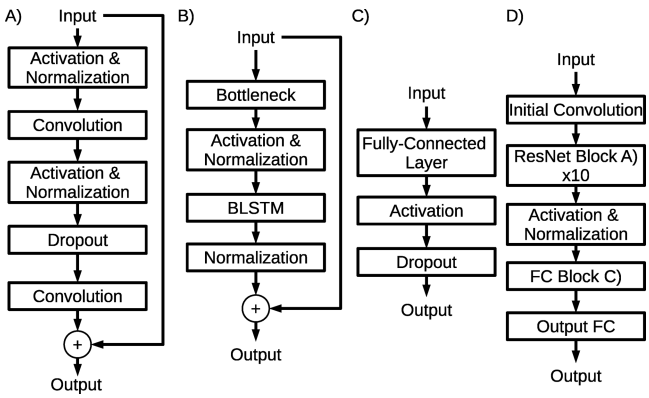


Figure 1. (A−C) Layout of three network blocks employed for building our deep learning models: (A) ResNet, (B) ResLSTM, and (C) fully connected (FC) architecture. (D) Example of a neural network model (model 1 from Table 1) using the ResNet and FC blocks in (A) and (C), respectively.

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

Table 1. Network Parameters for the Individual Networks Used in the Ensemble Model

		no. of blocks			no. of block neurons			
model	layout	ResNet	ResLSTM	FC	ResNet	ResLSTM	FC	bottleneck
0	ResNet	15	–	1	64	–	256	–
1	ResNet	10	–	1	64	–	512	–
2	ResLSTM	–	2	1	–	128	256	no
3	ResNet	10	–	1	128	–	256	–
4	ResLSTM/ResNet	10	2	1	64	128	256	yes
5	ResLSTM/ResNet	5	2	1	64	128	256	yes
6	ResLSTM/ResNet	15	2	1	64	128	256	no
7	ResNet/ResLSTM	15	2	1	64	128	256	no

function.[40] Because this model employs the preactivation architecture, a convolution operation is applied prior to the first residual block, and the output of the last residual block is activated and then normalized.

The ResLSTM segments consist of the residual blocks shown in Figure 1B. Each block contains one bidirectional LSTM (BLSTM) layer with 128 one-cell memory blocks for both the forward and backward directions concatenated together to produce 256 inputs in the succeeding layer. In some of our models, it was found that a bottleneck layer increased the performance of our ResLSTM blocks, as shown in Figure 1B. This bottleneck connection consisted of a 1×1 convolution operation with ELU activation and layer normalization.

All of the trained networks conclude with hidden fully connected (FC) layers and an output FC layer. The hidden FC layers have ELU activation, are regularized using dropout ($p(\mathbf{d}) = 0.5$),[41] and consist of varying numbers of neurons and one bias neuron. The output layer, on the other hand, has no dropout, three output nodes, a bias neuron, and softmax activation. The three outputs are used to represent the predicted probabilities that a residue will be in a trans, *cis*-Pro, or *cis*-nonPro conformation. We can combine our outputs to obtain the probability that a residue will be in a cis conformation by summing the two cis output nodes, as it is only useful to separate amide and imide bonds in training. In our work, we found it beneficial to provide separate thresholds for amide and imide bonds in order to give a binary output label for each input residue.

We have developed a number of models based on various combinations of ResNet, ResLSTM, and FC segments. The network parameters (numbers of segment blocks and neurons in each block) were obtained by extensive testing to find the most effective architectures as determined by their performance on a validation set. Specifically, we trained models based on all combinations of ResNet and/or ResLSTM blocks, 5−20 ResNet blocks with 64−128 filters, varying FC neuron depth, and varying ResLSTM layer size with and without bottleneck layers. The layouts of the eight final chosen models (three ResNets, one ResLSTM model, and four hybrid ResNet/ResLSTM models with varying numbers of ResNet, ResLSTM, and FC blocks and sizes) are elaborated in Table 1. As an illustration, the network architecture of model 1 from Table 1 is shown in Figure 1D.

The results of the eight models are combined to minimize generalization errors made on the data by each of our individual predictors.[42] As the cis conformation labels are few and far between, we examined several methods of combining each predictor's outputs and found similar performance. Thus, the mean over all of the outputs is employed as the final model.

Each network was trained in TensorFlow version 1.4[43] using the Adam optimization algorithm[44] with a weight of 10 placed on the corresponding output node when computing the network cost to account for the class disparity. A higher class weight, more relative to the actual class disparity, led to degraded precision from the network. Each activated output in the residual blocks (i.e., all but the FC blocks) is normalized by layer normalization.[45] Training the model in TensorFlow and other similar libraries enables training to take place on our Nvidia GTX TITAN X graphics processing unit (GPU), which has been shown to speed up network training time by up to a factor of 20 for neural networks.[46] The total training time for a purely ResNet model was 1 min/epoch over our training set for a batch size of 50 sequences, whereas a pure ResLSTM model took 8−10 min/epoch depending on the network parameters.

Datasets. Here we have employed the same training and testing datasets as in our previous work.[23,32] In brief, we obtained 12 450 nonredundant proteins from the cullpdb website in February 2017 using the following criteria: resolution < 2.5 Å, R-factor < 1.0, and sequence identity cutoff < 25% according to BlastClust.[47] This gave us a database of 12 450 proteins, which we split into sets Train (10 200 proteins), TestI (1000 proteins), and TestII (1250 proteins). Set TestII was made of the set of proteins deposited after June 2015, and the Train and TestI sets were randomly divided from the remaining pool of proteins.

Here we consider a protein to be in the trans state when its measured ω is in the range of $[150°, 210°)$ and in the cis state when ω is in the range of $[-30°, 30°)$. Other residues that are measured outside of these ranges are ignored during back-propagation and in analysis but not during the feed-forward stage.

Input Features. Our method utilizes two evolutionary profiles, physicochemical representations of the amino acid residues, and the outputs of predicted 1D structural properties as its inputs. Each protein's PSSM profile was generated by three iterations of PSI-Blast against the NCBI's Non-Redundant (NR) database[47] updated in 2017. The HHM profile was generated by HHBlits version 2.0.15 by searching the 20% nonredundant Uniprot2016 database[48] using the default search parameters of HHBlits.[49] We also employ the outputs of our previous 1D structural property predictor SPIDER3,[31] consisting of the predicted values of the three secondary structure probabilities (for a residue to be in the helix, coil, or strand conformation), one relative ASA, eight sines/cosines of the θ, τ, ϕ, and ψ angles, and two HSEα up and down. To avoid training on data seen during the training of SPIDER3, we obtain the cross-validation outputs of the third iteration of SPIDER3 for proteins in our Train set that

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

Table 2. Cis and Trans Conformation Counts for Each Amino Acid Residue in Each of Our Datasets

ω conf.	P	non-P	A	C	D	E	F	G	H	I	K
cis	6185	3892	304	57	322	263	116	583	183	119	207
trans	127387	2804798	236770	36462	173845	199852	121320	205814	72718	169044	167998
% cis	4.855	0.139	0.128	0.156	0.185	0.132	0.096	0.283	0.252	0.070	0.123
ω conf.	L	M	N	Q	R	S	T	V	W	Y	
cis	226	97	228	159	150	342	226	168	41	101	
trans	279255	63126	126612	111060	152233	177974	158832	204250	41756	105877	
% cis	0.081	0.154	0.180	0.143	0.099	0.192	0.142	0.082	0.098	0.095	

Table 3. Cis Conformation Counts and Percentages for Each Secondary Structure State and Surface Exposures for Amide and Imide Bonds

	helix	coil	sheet	core	surface
cis-Pro	12 (0.19%)	5968 (96.82%)	184 (2.99%)	2344 (38.26%)	3783 (61.74%)
cis-nonPro	18 (0.53%)	3240 (96.09%)	114 (3.38%)	869 (29.96%)	2032 (70.04%)

were used in training of SPIDER3. All of the input data were standardized to have zero mean and unit variance using the means and variances of the training set before being input into the network.

Performance Evaluation. Because of the severe class disparity in this problem (~5% for *cis*-imide and 0.1% for *cis*-amide bonds), it is important to have skew-independent metrics, which can represent the accuracy of cis prediction levels against an overwhelming number of trans labels. One commonly used pair of performance measures are the sensitivity (fraction of correctly predicted positives, $\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$, where TP is the number of true positive predictions and FN is the number of false negative predictions) and the specificity (fraction of correctly predicted negatives, $\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$, where TN is the number of true negative predictions and FP is the number of false positive predictions). However, the dominance of negative cases (trans conformations) makes it easier to maximize the specificity by simply predicting all of the residues to be in trans states, which gives an overall accuracy of 99.6%. Thus, a number of performance measures are employed. In addition to the area under the receiver operating characteristic (ROC) curve (plot of Se vs 1 − Sp) ($\mathrm{AUC_{ROC}}$), we calculate the Matthews correlation coefficient (MCC),[50]

$$\mathrm{MCC} = \frac{(\mathrm{TP}+\mathrm{TN})(\mathrm{FP}+\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

which is a measure of the correlation between the expected and obtained class labels, and generate the precision−recall curve, where the precision is the fraction of correct positive predictions out of all positive predictions ($\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$). The precision is a more practically useful guide for experimentalists to estimate the minimum number of predicted cis conformers to be experimentally tested to yield some positive outcomes. This leads to another single-valued metric, the area under the precision−recall curve ($\mathrm{AUC_{PR}}$), which provides a more informative analysis for unbalanced class prediction.[51] For completeness, we also calculated the overall accuracy, $\mathrm{Q2} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}$. We should note that a naïve single-state prediction would lead to a Q2 of 0.948 for proline and 0.998 for non-proline.

Because of the large propensity for cis conformations in imide bonds compared to amide bonds, we separately analyze the performance of our model for proline only and non-proline only. This is achieved by masking the outputs and labels corresponding to the amino acid residue for each sequence. Thus, we obtain separate thresholds for amide bonds and imide bonds.

There is only one method for which a functioning server is available to compare to our model. CISPEPpred[16] offers comparisons of sequence- and profile-based methods for proline residues and is available at http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/cispep/.

RESULTS

Statistical Analysis of Cis Conformations. Analyzing the database of 12 450 proteins reveals that 4.6% of proline and 0.14% of non-proline residues are cis isomers. The former value is similar to a previous estimate of 5.2% while the latter is about 5 times higher than the value of 0.03% for Xaa-nonPro residues reported by Jabs et al.[1] To confirm the statistics, we examined a high-resolution subset of the 12 450 proteins (2929 proteins at 1.6 Å) and found that the fractions of *cis*-Pro and *cis*-nonPro isomers are 5.2% and 0.14%, respectively. Thus, our obtained *cis*-nonPro isomer percentage is not caused by the resolution of protein structures. On the other hand, a slight increase in the fraction of *cis*-Pro isomers in high-resolution structures (5.2% vs 4.6%) suggests a possible underestimation of *cis*-Pro isomers for low-resolution structures. For the training set and two test sets, the fractions of proline residues in cis isomers are 4.8%, 4.8%, and 5.1%, respectively. For non-proline residues, they are 0.139%, 0.134%, and 0.140%, respectively. These fractions are consistent with each other despite the fact that the two test sets are substantially smaller than the training set, indicating reasonably representative test sets.

In analyzing the residue-wise constitution of cis isomers, it was found that different residues behaved differently. In the 12 450 protein set, glycine has the highest fraction of cis isomers (0.28%) and isoleucine has the lowest (0.08%), as shown in Table 2. We formed a correlation analysis between the fraction of cis isomers of 19 residue types and >500 physicochemical properties collected in AAindex (http://www.genome.jp/aaindex/).[52] We found that the highest correlation

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

coefficients (CCs) are 0.784 and 0.770 for the weights for a coil at window positions −1 and 0, respectively.[53] This link between the coil propensity and the cis isomer of a residue suggests that cis isomers are likely to be located in coil regions. Indeed, as shown in Table 3, we found that more than 96% of both *cis*-Pro and *cis*-nonPro isomers are in coil regions, whereas only 3% and 0.2−0.5% are in sheet and helix conformations, respectively. If we define surface residues as those residues with more than 25% solvent-accessible surface area, the majority of cis isomers are at the surface (61.7% for *cis*-Pro and 70.0% for *cis*-nonPro). Similar trends but with significantly less data were observed previously.[13]

To further explore possible patterns in neighboring residues surrounding cis isomers, we employed WebLogo plots[54] for sequence and secondary structure distributions around the cis isomers. Figures 2 and 3 show the distributions of five

Figure 2. WebLogo plot for the surrounding amino acid content for all *cis*-Pro isomers in the data, with a window size of ±5 residues.

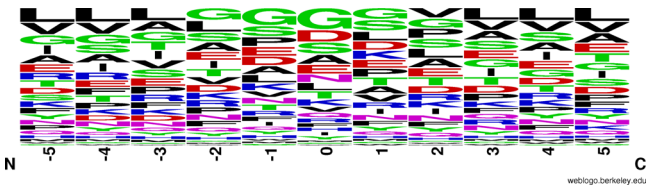


Figure 3. WebLogo plot for the surrounding amino acid content for all *cis*-nonPro isomers in the data, with a window size of ±5 residues.

neighboring residues for *cis*-Pro and *cis*-nonPro isomers, respectively. No clear amino acid patterns can be found for either case. On the other hand, the WebLogo plot of secondary structure around cis isomers (Figure 4) shows the dominant

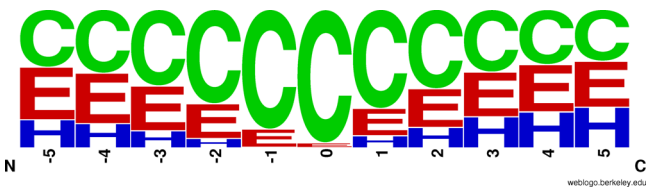


Figure 4. WebLogo plot for the surrounding secondary structure elements (coil (C), helix (H), and strand (E)) for all cis isomers in the data, with a window size of ±5 residues.

coil at the center with a gradual increase in the residues in strand or helical states away from the center. These results suggest that it is nearly impossible to locate cis isomers by relying on single-sequence and secondary structure information alone.

To get statistics of how many proteins have cis isomers, Figure 5 shows a cumulative distribution of proteins with at least one cis isomer as a function of protein chain length. The figure demonstrates that more than 50% of proteins longer than ∼190 residues and close to 100% of proteins with more than 600 residues have at least one cis isomer. This confirms

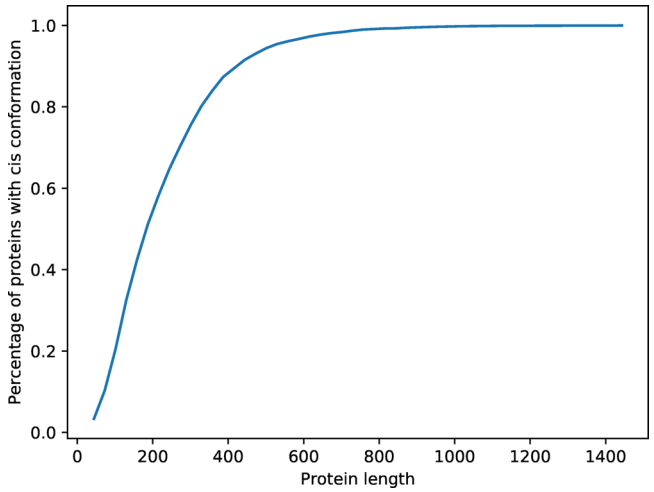


Figure 5. Cumulative distribution of proteins with a cis conformation.

the importance of detecting functionally important cis isomers in protein structures.

Overall Performance of SPOT-Omega for *cis*-Pro Prediction. Table 4 shows the mean results of 10-fold cross-validation on the Train set for each of the eight individual models used in the ensemble and for the ensemble model itself for *cis*-Pro prediction. For the individual models, the AUC values range from 0.829 to 0.843, the maximized MCC values from 0.30 to 0.32, and the sensitivities from 16% to 21% at a nearly constant precision of 50% according to a preset threshold. Similar performance among individual deep learning models indicates similar ability to learn from the same data without overtraining. The consensus prediction by the simple average yields the best prediction with an AUC of 0.858, maximized MCC value of 0.35, and sensitivity of 24%. The low standard deviations across 10 folds for all of the metrics indicate the stability of performance across each testing fold. The ensemble model outperforms each individual model in AUC (0.8575 vs 0.8428 for the best single model) and MCC (0.350 vs 0.325 for the best single model). The difference in AUC is statistically significant, with $P \leq 3 \times 10^{-3}$.[55]

Table 5 compares the performances of a baseline model, a single-state model, and the final ensemble model for predicting *cis*-Pro isomers in two independent test sets. The baseline model was obtained by a random classifier. The single-state model simply assigns all residues as trans isomers. For *cis*-Pro residues, the baseline model achieves a random AUC (0.5), MCC value (0.0), and Q2 (50%), as expected. The naive single-state model obtains a Q2 of 95.4%. For the ensemble model, the AUC values are 0.852 and 0.860 for the TestI and TestII sets, respectively. The corresponding MCC values are 0.352 and 0.360, respectively. The ROC curves of the ensemble model for the TestI and TestII sets are compared in Figure 6. A slightly better performance for the TestII set is observed, but the difference between the two ROC curves is statistically insignificant ($P \leq 0.29$). Here, for practical purposes, the threshold T used to generate the sensitivity, specificity, precision, and Q2 metrics is chosen so that the precision on the TestI set is at least 50% for each model. The MCC value is obtained from a second threshold, taken from the value that maximizes the MCC value on the TestI set. With the threshold determined by the TestI set, the ensemble model achieves a precision of 57% and sensitivity of 25% for the

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

Table 4. Performance of 10-Fold Cross-Validation on Our Training Set for *cis*-Pro Prediction, with Standard Deviations in Parentheses, by Individual Neural Network Models and the Ensemble

predictor	AUC_{ROC}	AUC_{PR}	max MCC	Se	Sp	Pr	Q2
model 0	0.8371 (0.02)	0.3019 (0.04)	0.3213 (0.03)	0.1848 (0.06)	0.9912 (0.00)	0.5044 (0.00)	0.9540 (0.00)
model 1	0.8428 (0.01)	0.3156 (0.02)	0.3245 (0.02)	0.2110 (0.03)	0.9900 (0.00)	0.5039 (0.00)	0.9541 (0.00)
model 2	0.8320 (0.01)	0.2895 (0.03)	0.3077 (0.03)	0.1772 (0.05)	0.9917 (0.00)	0.5063 (0.00)	0.9541 (0.00)
model 3	0.8331 (0.03)	0.2948 (0.06)	0.3094 (0.05)	0.1813 (0.07)	0.9915 (0.00)	0.5006 (0.02)	0.9541 (0.00)
model 4	0.8401 (0.01)	0.3035 (0.02)	0.3168 (0.02)	0.1930 (0.04)	0.9908 (0.00)	0.5032 (0.00)	0.9540 (0.00)
model 5	0.8380 (0.01)	0.3018 (0.03)	0.3164 (0.03)	0.1895 (0.04)	0.9910 (0.00)	0.5044 (0.00)	0.9541 (0.00)
model 6	0.8379 (0.01)	0.3037 (0.03)	0.3134 (0.03)	0.1876 (0.04)	0.9912 (0.00)	0.5059 (0.01)	0.9541 (0.00)
model 7	0.8289 (0.01)	0.2804 (0.03)	0.3004 (0.02)	0.1592 (0.06)	0.9925 (0.00)	0.5056 (0.00)	0.9540 (0.00)
ensemble	0.8575 (0.01)	0.3450 (0.02)	0.3503 (0.02)	0.2384 (0.05)	0.9887 (0.00)	0.5030 (0.00)	0.9540 (0.00)

Table 5. Performance of the Ensemble Models in *cis*-Pro Prediction for Two Independent Test Sets

test set	predictor	AUC_{ROC}	AUC_{PR}	max MCC	*T*	Se	Sp	Pr	Q2
I	random	0.5000	0.048	0.0000	0.500	0.5000	0.5000	0.048	0.5000
	single-state	–	–	–	–	0.0000	1.0000	0.0000	0.9544
	ensemble	0.8523	0.3390	0.3523	0.802	0.1992	0.9907	0.5052	0.9546
II	ensemble	0.8603	0.3778	0.3603	0.802	0.2489	0.9905	0.5710	0.9546
	CISPEP sequence	–	–	0.0072	–	0.1230	0.8877	0.0519	0.8513
	CISPEP profile	–	–	0.0070	–	0.1230	0.8875	0.0518	0.8511

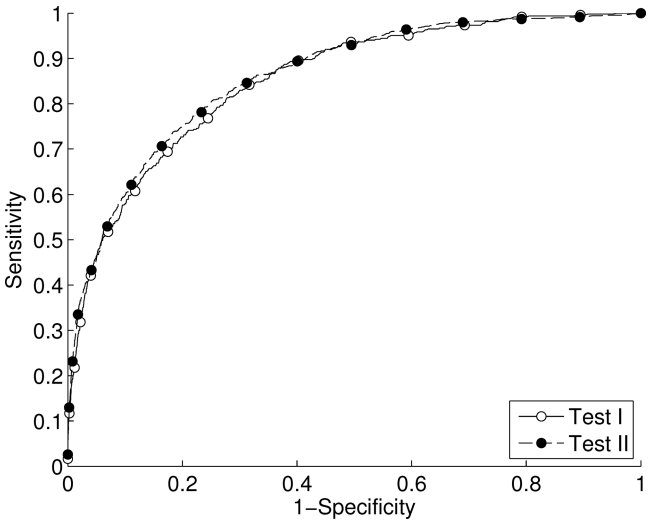


Figure 6. Receiver operating characteristic (ROC) curves for *cis*-Pro predictions by the ensemble model on the two test sets (I and II) as labeled.

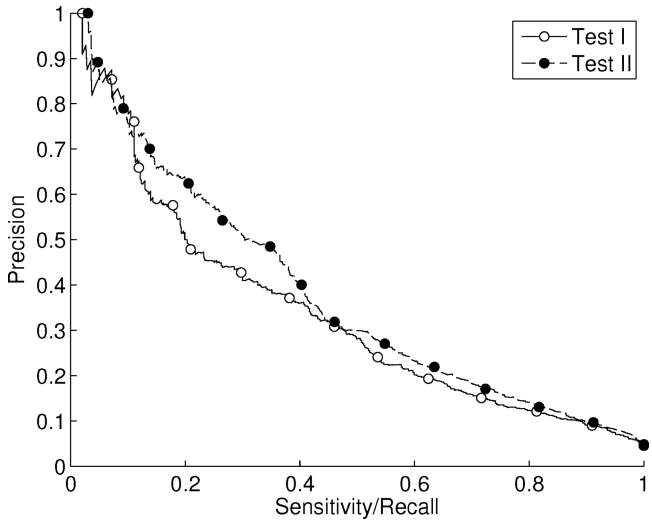


Figure 7. Precision−recall curves for the *cis*-Pro predictions from the ensemble model on the two test sets (I and II) as labeled.

TestII set. The consistent performance of the ensemble model across 10 folds and two test sets confirms the robustness of the models trained.

The performance of another predictor, CISPEPpred, is also shown in Table 5. CISPEPpred's single-sequence and profile-based predictions for the TestII set (proteins deposited after June 2015) both show a near-zero MCC value with low sensitivity (12%) and low precision (5%), indicating that the method is not generalizable to detect cis isomers in recently deposited protein structures.

Precision−recall curves of the ensemble model are shown in Figure 7 for the two test sets. The figure shows that a precision of nearly 100% is possible at the cost of a minimal sensitivity. The choice of threshold at 50% precision with about 20% coverage of all *cis*-Pro isomers (from the TestI set) is a compromise to balance the requirements of precision and sensitivity.

Overall Performance of SPOT-Omega for *cis*-NonPro Prediction. Table 6 shows the mean results of 10-fold cross-validation on the Train set for *cis*-nonPro prediction. For the individual network models, the AUC values range from 0.828 to 0.843, maximized MCCs from 0.09 to 0.12, and sensitivities from 2.1% to 4.2% for precisions ranging from 29% to 31% according to a preset threshold (set to achieve a precision of ∼30%). It was more difficult to predict at a consistent precision for *cis*-nonPro because of the high difficulty of separating the rare *cis*-nonPro events. The consensus prediction by the simple average yields the best prediction with an AUC of 0.857, MCC of 0.13, and sensitivity of 4.6% for a precision of 31%. The low standard deviations across 10 folds for all metrics indicate the stability of the performance across each testing fold. The ensemble model outperforms each individual model in AUC (0.857 vs 0.843 for the best single model) and MCC (0.133 vs 0.125 for the best single

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

Table 6. Performance of 10-Fold Cross-Validation on Our Training Set for *cis*-NonPro Prediction, with Standard Deviations in Parentheses, by Individual Neural Network Models and the Ensemble

predictor	AUC_{ROC}	AUC_{PR}	max MCC	Se	Sp	Pr	Q2
model 0	0.8344 (0.02)	0.0374 (0.01)	0.1245 (0.03)	0.0420 (0.02)	0.9999 (0.00)	0.2925 (0.05)	0.9985 (0.00)
model 1	0.8412 (0.01)	0.0369 (0.01)	0.1241 (0.02)	0.0371 (0.02)	0.9999 (0.00)	0.3058 (0.00)	0.9985 (0.00)
model 2	0.8328 (0.01)	0.0305 (0.01)	0.1122 (0.02)	0.0316 (0.01)	0.9999 (0.00)	0.3111 (0.01)	0.9986 (0.00)
model 3	0.8313 (0.03)	0.0342 (0.01)	0.1152 (0.04)	0.0359 (0.02)	0.9999 (0.00)	0.2898 (0.05)	0.9985 (0.00)
model 4	0.8404 (0.01)	0.0343 (0.01)	0.1167 (0.03)	0.0378 (0.02)	0.9999 (0.00)	0.3067 (0.01)	0.9985 (0.00)
model 5	0.8389 (0.02)	0.0337 (0.01)	0.1169 (0.02)	0.0338 (0.01)	0.9999 (0.00)	0.3074 (0.01)	0.9986 (0.00)
model 6	0.8432 (0.01)	0.0333 (0.01)	0.1161 (0.03)	0.0371 (0.02)	0.9999 (0.00)	0.3026 (0.01)	0.9985 (0.00)
model 7	0.8281 (0.02)	0.0230 (0.01)	0.0933 (0.02)	0.0211 (0.01)	0.9999 (0.00)	0.2862 (0.05)	0.9986 (0.00)
ensemble	0.8566 (0.01)	0.0418 (0.01)	0.1331 (0.02)	0.0467 (0.02)	0.9999 (0.00)	0.3070 (0.00)	0.9985 (0.00)

Table 7. Performance of the Ensemble Models in *cis*-NonPro Prediction for Two Independent Test Sets

test set	predictor	AUC_{ROC}	AUC_{PR}	max MCC	T	Se	Sp	Pr	Q2
I	random	0.5000	0.0014	0.0000	0.500	0.5000	0.5000	0.0014	0.5000
	single-state	–	–	–	–	0.0000	1.0000	0.0000	0.9987
	ensemble	0.8754	0.0516	0.1417	0.634	0.0471	0.9999	0.3111	0.9986
II	ensemble	0.8766	0.0470	0.0943	0.634	0.0142	1.0000	0.4615	0.9986

model). The difference between the AUC values is statistically significant with $P \leq 1.4 \times 10^{-2}$.[55]

Table 7 compares the performances of a baseline model, a single-state model, and the final ensemble method for predicting *cis*-nonPro isomers in the two test sets. The baseline model achieves a random AUC (0.5), MCC (0.0), and Q2 (50%), as expected. The naïve single-state model obtains a Q2 of 99.87%. For the ensemble model, the AUC values are 0.875 and 0.8766 for the two test sets, respectively. The corresponding maximized MCC values are 0.14 and 0.09, respectively. The ROC curves of the ensemble model for the two test sets are compared in Figure 8. The MCC value is

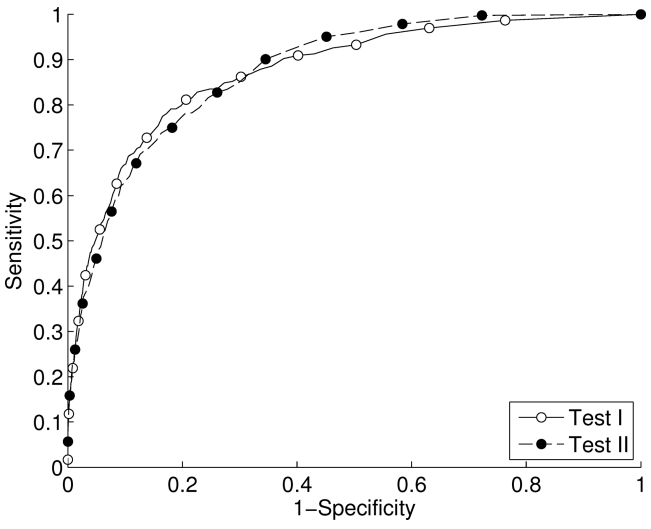


Figure 8. As in Figure 6 but for the *cis*-nonPro predictions.

somewhat lower in TestII (0.09) than in TestI (0.14) because the threshold T used to maximize the MCC on the TestI set was applied to the TestII set. For the same reason, a much higher precision but a lower sensitivity was observed in TestII compared with TestI, as the threshold from TestI at ∼30% precision was employed. The overall difference in performance (in terms of AUC) between TestI and TestII is statistically insignificant ($P \leq 0.44$). The precision–recall curve of the ensemble method for *cis*-nonPro isomers (Figure 9) is unable to achieve 100% precision at extremely low sensitivity,

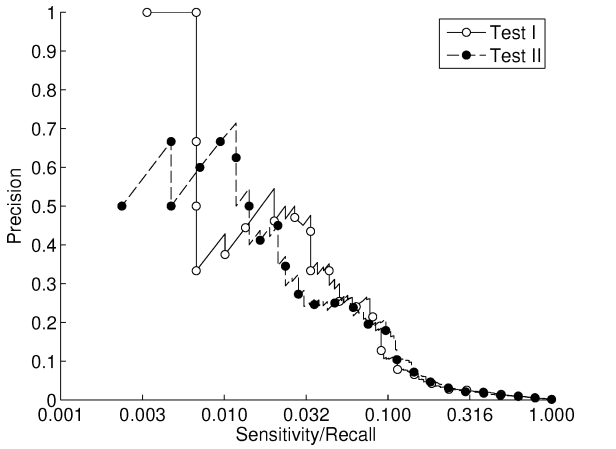


Figure 9. As in Figure 7 but for the *cis*-nonPro predictions. The *x* axis is in a log scale for clarity.

highlighting the challenge of resolving the rare *cis*-nonPro isomers from an overwhelming number of *trans*-nonPro isomers. Nevertheless, the result for the TestI set is nearly the same as that for the TestII set, confirming that the model obtained can be applied to unseen proteins with similar precision.

Importance of Feature Types. To get a sense of the contribution of individual feature groups to the performance of cis isomer prediction, we separate the features into sequence profiles generated from PSI-Blast (PSSM), profiles from HHblits (HHM), physicochemical properties (PHYS), and predicted 1D structural properties from SPIDER3 (SPD3). The results of these feature tests are shown in Table 8. Interestingly, removing HHM has the least impact on the AUC for predicting *cis*-Pro isomers but the highest impact for *cis*-nonPro isomers. HHM is also the best single feature group for predicting *cis*-nonPro but not for *cis*-Pro isomers, suggesting the inherent difference in the formation of *cis*-Pro and *cis*-nonPro isomers.

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033–2042

Table 8. Contributions to the AUC Metric by Different Feature Groups as a Single Feature or When Removed from the Full Feature Set on TestII (Model 1 Architecture Used)

	experiment	all features	PHYS	PSSM	HHM	SPD3
cis-Pro	singular	–	0.680	0.782	0.752	0.786
	removed	0.841	0.832	0.836	0.834	0.822
cis-nonPro	singular	–	0.685	0.778	0.821	0.795
	removed	0.858	0.846	0.837	0.815	0.840

DISCUSSION

We have developed a new method for an understudied problem: predicting cis isomers from protein sequences. Our large-scale statistics from over 10 000 proteins indicate that the fraction of *cis*-nonPro isomers (0.14%) is 5 times larger than a previous estimation. More importantly, more than 50% of proteins with sequence lengths of 190 residues or longer have cis isomers. We showed that by using an ensemble of high-performing ResNets and ResLSTM networks we can achieve MCCs of about 0.35 for predicting *cis*-Pro isomers and about 0.1 for *cis*-nonPro isomers. The performance is consistent among 10-fold cross-validation and two independent test sets. For *cis*-Pro isomers, our results, a sensitivity of 25% and a precision of 57%, are substantially better than those of the only available server, which gave a sensitivity of 12% and a precision of 5%. No method for *cis*-nonPro isomers is available for comparison.

One interesting observation is that the fraction of *cis*-nonPro isomers is strongly correlated with the propensity for coil residues, with a correlation coefficient of 0.78, consistent with the fact that the majority of cis isomers (>96%) are located in coil regions. It is a possibility that the less structured backbone makes a possible cis conformation sterically more tolerable.

Our SPOT-Omega method provides reasonably accurate prediction of *cis*-Pro isomers. According to the precision−recall curve (Figure 7), we can achieve >80% precision at about 10% coverage of all true *cis*-Pro isomers. In other words, the confidence is very high if the predicted probability is greater than 0.914. One can also have a high success rate of 30% at about 50% coverage with a lower threshold. Thus, employing the predicted *cis*-Pro probability to prioritize potential *cis*-prolines for experimental studies would allow for a significant reduction in cost for experimental validation.

Compared with *cis*-Pro isomers, it is significantly more challenging to predict *cis*-nonPro isomers. The maximum MCC value is about 0.14 for *cis*-nonPro isomers on TestI, compared with 0.35 for *cis*-Pro isomers. Moreover, at 50% confidence (precision), one can achieve only 1.4% coverage of all true *cis*-nonPro isomers, compared with 31% for *cis*-Pro isomers. This is largely because the number of positive instances (only a few hundred per residue type) is too small to allow the neural networks to learn the relation between the sequence and cis conformations. Nevertheless, the robust performance across different datasets indicates that it is possible to develop a reliable predictive model even for an extremely imbalanced data set (1.4:1000) without over- or undersampling techniques commonly used in other machine learning methods such as support vector machines.[56,57] Such data set balancing techniques often lead to poor performance in real-world applications, as the real data are deeply imbalanced. Nevertheless, the method could be further improved in the future if more *cis*-nonPro cases are collected.

In summary, this work represents the first reliable neural network model for predicting cis isomers, which are biologically important. The method will be useful for assisting not only function prediction but also protein structure prediction. This is the case because a 180° change in one ω angle will lead to a completely different orientation in the protein backbone structure and significantly affect the overall fold of the entire protein structure in protein folding simulations. Thus, the coupling of ω angle prediction with the prediction of other backbone torsion angles (ϕ and ψ) from a method such as SPIDER3 should be directly useful in sampling rare but important protein conformations. The method is available as a server at http://sparks-lab.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yaoqi.zhou@griffith.edu.au.

ORCID

Jack Hanson: 0000-0001-6956-6748

Yaoqi Zhou: 0000-0002-9958-5699

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DP180102060 to Y.Z. and K. P.) and in part by National Health and Medical Research Council of Australia (1121629 to Y.Z.). We also acknowledge the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

REFERENCES

(1) Jabs, A.; Weiss, M. S.; Hilgenfeld, R. Non-Proline Cis Peptide Bonds in Proteins 1. *J. Mol. Biol.* **1999**, *286*, 291−304.

(2) Craveur, P.; Joseph, A. P.; Poulain, P.; de Brevern, A. G.; Rebehmed, J. Cis−Trans Isomerization of Omega Dihedrals in Proteins. *Amino Acids* **2013**, *45*, 279−289.

(3) Weiss, M. S.; Metzner, H. J.; Hilgenfeld, R. Two Non-Proline Cis Peptide Bonds May Be Important for Factor XIII Function. *FEBS Lett.* **1998**, *423*, 291−296.

(4) Koo, B.-K.; Park, C.-J.; Fernandez, C. F.; Chim, N.; Ding, Y.; Chanfreau, G.; Feigon, J. Structure of H/ACA RNP Protein Nhp2p Reveals Cis/trans Isomerization of a Conserved Proline at the RNA and Nop10 Binding Interface. *J. Mol. Biol.* **2011**, *411*, 927−942.

(5) van Aalten, D. M.; Komander, D.; Synstad, B.; Gåseidnes, S.; Peter, M. G.; Eijsink, V. G. Structural Insights into the Catalytic Mechanism of a Family 18 Exo-Chitinase. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 8979−8984.

(6) Lorenzen, S.; Peters, B.; Goede, A.; Preissner, R.; Frömmel, C. Conservation of Cis Prolyl Bonds in Proteins During Evolution. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 589−595.

(7) Sarkar, P.; Saleh, T.; Tzeng, S.-R.; Birge, R. B.; Kalodimos, C. G. Structural Basis for Regulation of the Crk Signaling Protein by a Proline Switch. *Nat. Chem. Biol.* **2011**, *7*, 51.

(8) Lummis, S. C.; Beene, D. L.; Lee, L. W.; Lester, H. A.; Broadhurst, R. W.; Dougherty, D. A. Cis−Trans Isomerization at a Proline Opens the Pore of a Neurotransmitter-Gated Ion Channel. *Nature* **2005**, *438*, 248.

(9) Truckses, D. M.; Prehoda, K. E.; Miller, S. C.; Markley, J. L.; Somoza, J. R. Coupling Between Trans/Cis Proline Isomerization and Protein Stability in Staphylococcal Nuclease. *Protein Sci.* **1996**, *5*, 1907−1916.

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

(10) Nelson, C. J.; Santos-Rosa, H.; Kouzarides, T. Proline Isomerization of Histone H3 Regulates Lysine Methylation and Gene Expression. *Cell* **2006**, *126*, 905−916.
(11) Yang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch? *Briefings Bioinf.* **2018**, *19*, 482−494.
(12) Frömmel, C.; Preissner, R. Prediction of Prolyl Residues in Cis-Conformation in Protein Structures on the Basis of the Amino Acid Sequence. *FEBS Lett.* **1990**, *277*, 159−163.
(13) Pal, D.; Chakrabarti, P. Cis Peptide Bonds in Proteins: Residues Involved, Their Conformations, Interactions and Locations1. *J. Mol. Biol.* **1999**, *294*, 271−288.
(14) Wang, M.-L.; Li, W.-J.; Xu, W.-B. Support Vector Machines for Prediction of Peptidyl Prolyl Cis/Trans Isomerization. *J. Pept. Res.* **2004**, *63*, 23−28.
(15) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
(16) Song, J.; Burrage, K.; Yuan, Z.; Huber, T. Prediction of Cis/Trans Isomerization in Proteins Using PSI-BLAST Profiles and Secondary Structure Information. *BMC Bioinf.* **2006**, *7*, 124.
(17) Exarchos, K. P.; Exarchos, T. P.; Papaloukas, C.; Troganis, A. N.; Fotiadis, D. I. Detection Of Discriminative Sequence Patterns in the Neighborhood of Proline Cis Peptide Bonds and Their Functional Annotation. *BMC Bioinf.* **2009**, *10*, 113.
(18) Al-Jarrah, O. Y.; Yoo, P. D.; Taha, K.; Muhaidat, S.; Shami, A.; Zaki, N. Randomized Subspace Learning For Proline Cis-Trans Isomerization Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2015**, *12*, 763−769.
(19) Yoo, P. D.; Muhaidat, S.; Taha, K.; Bentahar, J.; Shami, A. Intelligent Consensus Modeling for Proline Cis-Trans Isomerization Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11*, 26−32.
(20) Pahlke, D.; Leitner, D.; Wiedemann, U.; Labudde, D. COPS-Cis/Trans Peptide Bond Conformation Prediction of Amino Acids on the Basis of Secondary Structure Information. *Bioinformatics* **2005**, *21*, 685−686.
(21) Exarchos, K. P.; Papaloukas, C.; Exarchos, T. P.; Troganis, A. N.; Fotiadis, D. I. Prediction of Cis/Trans Isomerization Using Feature Selection and Support Vector Machines. *J. Biomed. Inf.* **2009**, *42*, 140−149.
(22) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving Protein Disorder Prediction by Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. *Bioinformatics* **2017**, *33*, 685−694.
(23) 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**, DOI: 10.1093/bioinformatics/bty481.
(24) 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*, e1005324.
(25) Lee, B.; Baek, J.; Park, S.; Yoon, S. deepTarget: End-to-End Learning Framework for microRNA Target Prediction Using Deep Recurrent Neural Networks. *Proc. ACM BCH10.* **2016**, 434−442.
(26) Park, S.; Min, S.; Choi, H.; Yoon, S. deepMiRGene: Deep Neural Network based Precursor microRNA Prediction. 2016, arXiv:1605.00017 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1605.00017.
(27) Sønderby, S. K.; Winther, O. Protein Secondary Structure Prediction with Long Short Term Memory Networks. 2014, arXiv:1412.7828 [q-bio.QM]. arXiv.org e-Print archive. https://arxiv.org/abs/1412.7828.
(28) Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, *6*, 18962.
(29) Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y. Improving Prediction of Secondary Structure, Local Backbone Angles, and Solvent Accessible Surface Area of Proteins by Iterative Deep Learning. *Sci. Rep.* **2015**, *5*, 11476.
(30) Heffernan, R.; Dehzangi, A.; Lyons, J.; Paliwal, K.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y.; Yang, Y. Highly Accurate Sequence-Based Prediction of Half-Sphere Exposures of Amino Acid Residues in Proteins. *Bioinformatics* **2016**, *32*, 843−849.
(31) 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. *Bioinformatics* **2017**, *33*, 2842−2849.
(32) Heffernan, R.; Paliwal, K.; Yang, Y.; Zhou, Y. Single-Sequence-Based Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility, Half-Sphere Exposure, and Contact Number by Long Short-Term Memory Bidirectional Recurrent Neural Networks. *J. Comput. Chem.* **2018**, in press.
(33) Schuster, M.; Paliwal, K. K. Bidirectional Recurrent Neural Networks. *IEEE T Signal Proces* **1997**, *45*, 2673−2681.
(34) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **1997**, *9*, 1735−1780.
(35) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. *Nature* **1986**, *323*, 533−536.
(36) LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput* **1989**, *1*, 541−551.
(37) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proc. IEEE CVPR.* **2016**, 770−778.
(38) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. *Proc. CV* **2016**, *9908*, 630−645.
(39) Kim, J.; El-Khamy, M.; Lee, J. Residual LSTM: Design of a Deep Recurrent Architecture for Distant Speech Recognition. 2017, arXiv:1701.03360 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1701.03360.
(40) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 2015, arXiv:1511.07289 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1511.07289.
(41) 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.
(42) Hansen, L. K.; Salamon, P. Neural Network Ensembles. *IEEE T Pattern Anal* **1990**, *12*, 993−1001.
(43) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Józefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F. B.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016, arXiv:1603.04467 [cs.DC]. arXiv.org e-Print archive. https://arxiv.org/abs/1603.04467.
(44) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv:1412.6980 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980.
(45) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer Normalization. 2016, arXiv:1607.06450 [stat.ML]. arXiv.org e-Print archive. https://arxiv.org/abs/1607.06450.
(46) Oh, K.-S.; Jung, K. GPU Implementation of Neural Networks. *Pattern Recognition* **2004**, *37*, 1311−1314.
(47) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(48) UniProt Consortium. Reorganizing the Protein Space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042

(49) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM−HMM Alignment. *Nat. Methods* **2012**, *9*, 173−175.
(50) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.
(51) Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. *Proc. ML23* **2006**, 233−240.
(52) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2007**, *36*, D202−D205.
(53) Qian, N.; Sejnowski, T. J. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.* **1988**, *202*, 865−884.
(54) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188−1190.
(55) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29−36.
(56) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.
(57) Japkowicz, N.; Stephen, S. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* **2002**, *6*, 429−449.

DOI: 10.1021/acs.jcim.8b00442
J. Chem. Inf. Model. 2018, 58, 2033−2042