Received: 6 November 2017 | Revised: 12 February 2018 | Accepted: 27 February 2018

DOI: 10.1002/prot.25489

RESEARCH ARTICLE

WILEY PROTEINS

SPIN2: Predicting sequence profiles from protein structures using deep neural networks

James O'Connell[1] | Zhixiu Li[2,3] | Jack Hanson[1] | Rhys Heffernan[1] | James Lyons[1] | Kuldip Paliwal[1] | Abdollah Dehzangi[1,4] | Yuedong Yang[2,5] | Yaoqi Zhou[2]

[1]Signal Processing Laboratory, Griffith University, Nathan, Australia

[2]Institute for Glycomics, Griffith University, Gold Coast, Australia

[3]Translational Genomics Group, Queensland University of Technology Translational Research Institute, Brisbane, Australia

[4]Department of Computer Science, Morgan State University, Baltimore, Maryland

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

Correspondence
Yaoqi Zhou, Institute for Glycomics, Griffith University, Gold Coast, Australia.
Email: yaoqi.zhou@griffith.edu.au

Funding information
Australian Research Council, Award/Grant number: LE150100161; National Health and Medical Research Council of Australia, Award/Grant numbers: 1059775, 1083450; National Natural Science Foundation of China, Award/Grant number: 61271378; Microsoft Azure for Research Award

Abstract

Designing protein sequences that can fold into a given structure is a well-known inverse protein-folding problem. One important characteristic to attain for a protein design program is the ability to recover wild-type sequences given their native backbone structures. The highest average sequence identity accuracy achieved by current protein-design programs in this problem is around 30%, achieved by our previous system, SPIN. SPIN is a program that predicts sequences compatible with a provided structure using a neural network with fragment-based local and energy-based nonlocal profiles. Our new model, SPIN2, uses a deep neural network and additional structural features to improve on SPIN. SPIN2 achieves over 34% in sequence recovery in 10-fold cross-validation and independent tests, a 4% improvement over the previous version. The sequence profiles generated from SPIN2 are expected to be useful for improving existing fold recognition and protein design techniques. SPIN2 is available at http://sparks-lab.org.

KEYWORDS
bioinformatics, deep learning, fold recognition, neural networks, structure prediction

1 | INTRODUCTION

Proteins consist of any combination of 20 amino acid residues, which provides a vast sequence space. For a protein sequence of 100 amino acids, the number of possible sequences is astronomically large: 20^{100} or 2×10^{130}, a number larger than the number of atoms in the universe. The majority of these artificial sequences do not have well-defined structures, including many biologically active proteins (intrinsically disordered or unstructured proteins).[1] For proteins with structures, Anfinsen's dogma states that their folded shapes are determined through their primary sequence.[2] Thus, it must also be possible to analyze this inversely, in that a protein's sequence can be deduced by analyzing its native structure. From this we can begin to explore the possibility of tailor-making proteins to fit a desired structure and function.

Many protein design programs have been designed based on this dogma. A typical program starts from a target backbone structure and a random sequence.[3–7] Random mutations to the sequence are coupled with a global energy minimization technique to search for sequences with the lowest energy. More sophisticated methods account for backbone flexibility in protein design.[8–10] Nevertheless, key to the success of any of these models is an accurate energy function.[11] However, existing energy functions for protein design are not yet sufficiently accurate,[11,12] although recent progress has been made.[13,14]

The effect of an insufficient energy function can be seen in the accuracies of current systems. One important measure for the quality

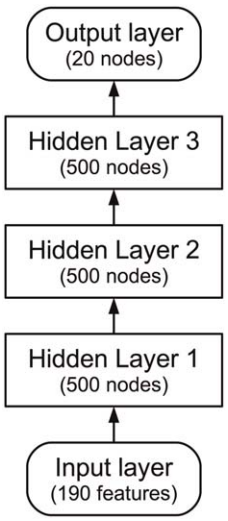

FIGURE 1 The general architecture of the deep neural network used in this project. Stacked auto-encoders begin encoding the first layer, then the first two layers and so on

of a design program is its ability to recover wild-type sequences given their native backbone structures. The average sequence recovery achieved by the current top-performing protein-design programs is only around 30%.

Energy functions for protein design are usually a composite of single-body, two-body, and/or multi-body energetic terms. A single-body (profile-based) method can obtain structure-compatible sequences directly from structurally homologous fragments.[15] This method by Zhou and Zhou[15] was subsequently improved using neural-networks trained on local fragment-derived and nonlocal energy-based profiles for SPIN.[16] The sequence recovery increased from 24% for the fragment-based technique to 30% for SPIN. Here, we develop SPIN2 by employing additional features and a deep-learning neural network. The average sequence recovery reaches >34%, comparable or better than existing top-performing protein design programs.

2 | MATERIALS AND METHODS

2.1 | Datasets

We used the same benchmark that was established for developing the SPIN program.[16] This data set consists of 2032 proteins with resolutions better than 3.0 Å, and pairwise sequence identity <30%. The set was further randomly divided into a training set (1532 proteins) and a test set (500 proteins) labeled TR1532 and TS500, respectively. Structural fragments for generating fragment-derived sequence profiles are from a completely independent template library of 2282 proteins (TL2282) with <30% sequence identity to TR1532 and TS500.

2.2 | Previously-employed features

SPIN employed only three sets of features: backbone torsion angles, local fragment-derived profiles, and global energy-based features. Backbone torsion angles are the rotational angles (ϕ and ψ angles) in N—Cα and Cα—C bonds, respectively. The local fragment-derived sequence profiles[12] were obtained by comparing 5-residue structural segments from i to $i+4$, where $i \in \{1, 2, \ldots, L-4\}$ in the target structure of length L, to 5-residue fragments in template library TL2282. The sequences of the most structurally similar fragments (measured in Root Mean Squared Deviation) were utilized to calculate the probability of an amino acid residue at each sequence position and establish a sequence profile. For each sequence position, this profile has a twenty substitution probabilities corresponding to the 20 residue types.

The global energy-based features, on the other hand, are interaction energies between a residue type j at position i, its side-chain rotamer, k, and the rest of the backbone positions occupied by the alanine residue. The lowest energies for all rotameric states and the energies of the six most frequent rotameric states for each residue type were employed as nonlocal energy-based features. Here, the bbdep02 rotamer library,[17] and a knowledge-based energy function based on the distance-scaled finite-ideal gas reference state were employed.[18,19]

2.3 | New local features employed

In addition to torsion angles ϕ and ψ, we have incorporated two other backbone angles θ and ι. Local angle θ is based on neighboring Cα atoms ($C\alpha_{i-1}$—$C\alpha_i$—$C\alpha_{i+1}$) and the N_i—$C\alpha_i$—C_i bond angle, while dihedral angle ι is based on four neighboring Cα atoms ($C\alpha_{i-1}$—$C\alpha_i$—$C\alpha_{i+1}$—$C\alpha_{i+2}$). Unlike in SPIN, all angles were represented by their sine and cosine—both of which are required to map all possible angles to a unique representation. Sine and cosine features also allow identical or nearby angles that are numerically distant when represented in degrees or radians, such as $-\pi/\pi$ or $-179°/179°$, to map to similar locations in feature space. Pairwise atomic distances within a single residue and between the two neighboring residues (excluding covalently bonded atoms) were also utilized.

2.4 | New nonlocal features employed

We calculated contact numbers, defined as the number of neighboring alpha carbons, between 5 and 20 Å in 1 Å steps from a given Cα atom. This feature provides information about the local density at specific positions and measures how deeply a position is buried. The distance-dependent contact number at each distance bin was obtained, providing 16 features. In addition, we also obtained two distances within a single residue (intra-residue distances) of N—O and Cα—O, and 30 inter-residue atomic distances between the four main-chain atoms of

TABLE 1 Performance comparison between SPIN and SPIN2 in terms of percent of sequence recovery

Method	TR1532[a]	TS500[b]	Top 2 match[c]
SPIN	30.7%	30.3%	43.8%
SPIN2	34.4%	34.4%	49.1%

[a]10-fold cross-validation.
[b]Independent test.
[c]Match of wild-type sequence to one of the top-2 predicted residue types for TS500.

TABLE 2 Contribution of each feature group to SPIN2 in term of percent of sequence recovery

Feature Excluded	TR1532[a]	TS500[b]
None	34.40%	34.43%
Energy-based	30.30%	30.06%
Inter-residue distance	32.41%	32.39%
Fragment-based	33.25%	33.28%
Angles	33.94%	33.95%
Intra-residue distance	34.16%	34.16%
Contact number	34.24%	34.35%

[a]10-fold cross-validation.
[b]Independent test.

the current and two nearest neighboring residues (eg, N_{i-1}—N_i, N_{i-1}—$C\alpha_i$, N_{i-1}—O_i, N_{i-1}—C_i, N_i—N_{i+1}, N_i—$C\alpha_{i+1}$, N_i—O_{i+1}, N_i—C_{i+1}).

This results in a total of 54 new features: 6 Cα-based angles, 16 contact numbers, 2 intra-residue distances, and 30 neighboring residue distances. When combined with the 136 previously employed features (2 sine and cosine sets of 2 backbone torsions, 20 fragment-derived features, and 112 energy-based features), it gives us a total of 190 features into our network. These features were utilized without a window as we found that additional neighboring information did not further improve our results with the classifier used.

2.5 | Deep neural network

This study employed a Deep Neural Network (DNN). DNN's are fully-connected artificial neural networks with three or more hidden layers. In this study, we employed a network with three hidden layers with 500 sigmoided nodes each (Figure 1). The output layer had 20 softmax nodes representing the 20 types of amino acid residues. The network was developed in Matlab using the DeepLearnToolbox.[20]

Using several hidden layers enables our model to create several layers of data abstraction (or representation) necessary to effectively model complex underlying relationships in the input data. This layer depth is what enables our model to truly utilize deep learning, one of the most powerful learning tools in the literature.[21] However, with each additional hidden layer, the computational complexity of training larger and larger models requires much larger amounts of input data. Thus, three hidden layers have been empirically chosen as good compromise between high-level feature abstraction and limited input data.

In this study, neuron weights in the DNN were initialized using stacked sparse auto-encoders. An auto-encoder is a neural network pretraining method which trains a layer of hidden units to replicate the input, in order to obtain an effective initialization of the neural network weights. We also employed a sparsity penalty to prevent the network learning of the identity function.[22] Stacked auto-encoders simply train each hidden layer in the network using individual auto-encoders, with each auto-encoder using all of the pretrained weights of the previous layers' auto-encoders. Following initialization of all hidden layers, back-propagation was used to fine-tune the network from the training set TR1532.

In addition to initialization of the network weights, L2 regularization was employed to prevent overfitting of the training data and to improve generalization of the network. Increasing the depth of a neural network often leads to overfitting, so regularization is key to obtaining a general and meaningful output for unseen input data. When employed in iterative training methods, L2 regularization has the effect of proportionally reducing each weight toward zero at every update. It is thus commonly referred to as a shrinking factor.

2.6 | Evaluation

To study the effectiveness and generality of our proposed technique, this study adopted two evaluation methods. First, 10-fold cross-validation was used for training. We applied 10-fold cross-validation by randomly dividing the TR1532 train set into ten different subsets and using nine of these subsets as training data and the remaining subset in testing. This was repeated ten times for a given model, such that all ten subsets were used exactly once for testing. The combined results formed the final training accuracy. These cross-validation scores were then used to select a final model, which was subsequently trained on the entirety of the TR1532 set and tested on the independent test set (TS500). The robustness of this method in avoiding overfitting was confirmed by the similarity between the cross-validation and independent test set scores.

3 | RESULTS AND DISCUSSION

Table 1 compares the performance of SPIN2 (this work) and the previous SPIN method. There is an overall 4% improvement from SPIN to SPIN2 in sequence recovery of wild-type sequences for both 10-fold cross-validation and the independent test set. The performance increase between SPIN and SPIN2 for TS500 obtains a P values of $<1 \times 10^{10}$ based on a confidence interval of 99%, rejecting the null

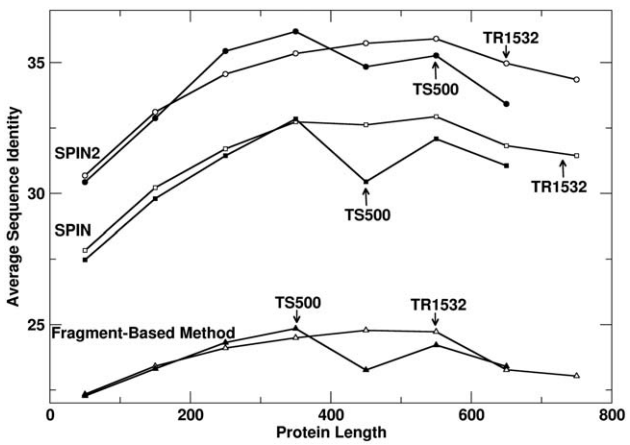

FIGURE 2 Average sequence identity between predicted and wild-type sequences as a function of protein length (10-fold cross-validation on TR1532, open symbols, and independent test on TS500, filled symbols) by SPIN and SPIN2

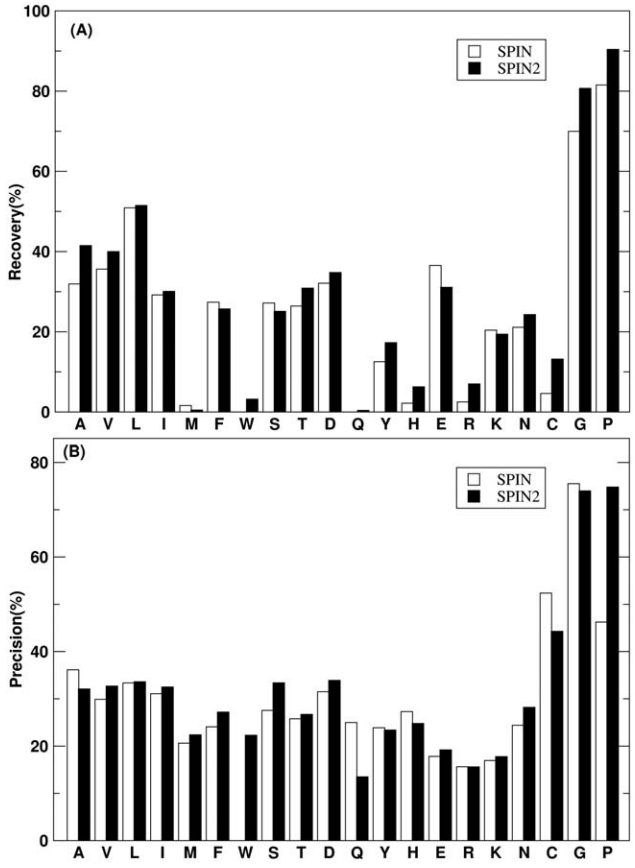

FIGURE 3 A, Recovery rate for each amino acid residue type by SPIN and SPIN2. B, Precision for each amino acid residue type by SPIN and SPIN2

hypothesis.[23] The consistency between 10-fold cross-validation and the independent test set confirms the robustness of the method. For the top-two match (matching wild-type sequence to one of the top-two predicted residue types) there is a 5.3% improvement.

To determine the source of the improvements, we investigated the effectiveness of each of the employed feature groups on the achieved results. Table 2 examines the importance of each feature group after excluding them one-by-one. Energy-based profiles developed for SPIN made the highest contribution, followed by local distances between the main atoms of the current and neighboring residues (inter-residue distances), then fragment-based profiles, and finally contact numbers. Contact numbers did contribute the least to the overall accuracy, but still improved them marginally. Interestingly, local distances contributed more information than angles in determining the sequences of the proteins.

Figure 2 compares the fragment-based method SPIN and the updated SPIN2 by examining sequence recovery as a function of protein sequence length. SPIN2 consistently improves over SPIN at all protein lengths. Moreover, 10-fold cross-validation (TR1532) and the independent test set (TS500) yielded consistent results, indicating the robustness of SPIN2 for different datasets.

Figure 3 compares recovery and precision for individual amino acids. For all amino acid types, SPIN2 improves over SPIN in either precision (65% of amino acids) or recovery (75% of amino acids). We note that glycine and proline have the highest recovery rates and precision, compared to all other residue types. This is understandable because glycine does not have side-chain and it can visit backbone angles that are not reachable by other amino acids due to steric constraints. On the other hand, the cyclic structure of proline side-chain limits the backbone ϕ angle at about -65 degree.[24] After excluding glycine and proline, SPIN2 continues to improve over SPIN in either precision (67% of 18 amino acids) or recovery (72% of 18 amino acids).

The results of the feature-omission analysis produced no other apparent residue-specific patterns except that the accuracy of proline residues is most affected by removing inter-residue features, the accuracy of cystine residues are impacted by intra-residue, rotomer-based, and residue angle features, and finally, the accuracy of tryptophan is greatly boosted by having fragment-based features. As is expected, these are some of the residues that show the greatest improvements over SPIN in Figure 3.

4 | CONCLUSIONS

This article illustrates the power of deep learning, and it's applicability in the field of protein sequence design. SPIN2 has pushed the limit of sequence recovery for profile-based techniques to >34% through the use of deep learning and an extended feature set. SPIN2's improvements over our previous predictor have been illustrated in both our datasets, and for the majority of amino acids. The more accurate structure-derived sequence profiles produced by SPIN2 will be useful for further enhancing fold recognition[15,25] and improving protein design.[12]

As deep learning methods allow for a high level of feature abstraction, adding new features allows the network to explore the deep hidden relationships between our features. The features added to our network from SPIN2 all contribute meaningfully to the accuracies of our predictions, which may not have been the case for a smaller network incapable of modeling deep patterns in the data.

Finally, without the need for energy minimization, SPIN2 is highly efficient in predicting sequence profiles; only a few minutes are required on an Intel(R) Xeon(R) CPU E5-2670 0 at 2.60 GHz (an average over a few small proteins of <100 residues).

ORCID

Jack Hanson http://orcid.org/0000-0001-6956-6748
Yaoqi Zhou http://orcid.org/0000-0002-9958-5699

REFERENCES

[1] Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol.* 2008; 18(6):756–764.

[2] Anfinsen CB. Principles that govern the folding of protein chains. *Science.* 1973;181(4096):223–230.

[3] Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. *Science.* 1997;278(5335):82–87.

[4] 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(2):449–460.

[5] Lippow SM, Tidor B. Progress in computational protein design. *Curr Opin Biotechnol*. 2007;18(4):305–311.

[6] Liu Y, Kuhlman B. Rosettadesign server for protein design. *Nucleic Acids Res*. 2006;34(Web Server):W235–W238.

[7] Regan L, DeGrado WF. Characterization of a helical protein designed from first principles. *Science*. 1988;241(4868):976–978.

[8] Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. *Science*. 1998;282(5393):1462–1467.

[9] Mandell DJ, Kortemme T. Backbone flexibility in computational protein design. *Curr Opin Biotechnol*. 2009;20(4):420–428.

[10] Murphy GS, Mills JL, Miley MJ, Machius M, Szyperski T, Kuhlman B. Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure*. 2012;20(6):1086–1096.

[11] Li Z, Yang Y, Zhan J, Dai L, Zhou Y. Energy functions in de novo protein design: current challenges and future prospects. 2013;42:315–335.

[12] Dai L, Yang Y, Kim HR, Zhou Y. Improving computational protein design by using structure-derived sequence profile. *Proteins*. 2010;78(10):2338–2348.

[13] Rocklin GJ, Chidyausiku TM, Goreshnik I, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*. 2017;357(6347):168–175.

[14] Xiong P, Wang M, Zhou X, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat Commun*. 2014;5:5330.

[15] Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*. 2005;58(2):321–328.

[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*. 2014;82(10):2565–2573.

[17] Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*. 1997;6(8):1661–1681.

[18] Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins*. 2008;72(2):793–803.

[19] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 2002;11(11):2714–2726.

[20] Palm RB. *Prediction as a Candidate for Learning Deep Hierarchical Models of Data*. Master's Thesis. Technical University of Denmark; 2012.

[21] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.

[22] Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: International Conference on Advances in Neural Information Processing Systems (Long Beach, CA, USA); 2007:153–160.

[23] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.

[24] Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Proteins*. 1992;12(4):345–364.

[25] Schmidt Am Busch M, Mignon D, Simonson T. Computational protein design as a tool for fold recognition. *Proteins*. 2009;77(1):139–158.

How to cite this article: O'Connell J, Li Z, Hanson J, et al. SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins*. 2018;86:629–633. https://doi.org/10.1002/prot.25489