SOFTWARE TOOL ARTICLE

REVISED RamaNet: Computational *de novo* helical protein backbone design using a long short-term memory generative neural network [version 2; peer review: 2 not approved]

Previously titled: RamaNet: Computational *de novo* helical protein backbone design using a long short-term memory generative adversarial neural network

Sari Sabban ^{1,2}, Mikhail Markovsky³

¹Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Makka, Saudi Arabia

²Electrical and Computer Engineering Department, College of Engineering, Effat University, Jeddah, Makka, Saudi Arabia

³North Caucasian Federal Scientific Center of Horticulture, Viticulture, Wine-making, Krasnodar, Russian Federation

V2 First published: 27 Apr 2020, 9:298
<https://doi.org/10.12688/f1000research.22907.1>

Second version: 11 Sep 2020, 9:298
<https://doi.org/10.12688/f1000research.22907.2>

Latest published: 12 Oct 2020, 9:298
<https://doi.org/10.12688/f1000research.22907.3>

Abstract

The ability to perform *de novo* protein design will allow researchers to expand the variety of available proteins. By designing synthetic structures computationally, they can utilise more structures than those available in the Protein Data Bank, design structures that are not found in nature, or direct the design of proteins to acquire a specific desired structure. While some researchers attempt to design proteins from first physical and thermodynamic principals, we decided to attempt to test whether it is possible to perform *de novo* helical protein design of just the backbone statistically using machine learning by building a model that uses a long short-term memory (LSTM) architecture. The LSTM model used only the ϕ and ψ angles of each residue from an augmented dataset of only helical protein structures. Though the network's generated backbone structures were not perfect, they were idealised and evaluated post generation where the non-ideal structures were filtered out and the adequate structures kept. The results were successful in developing a logical, rigid, compact, helical protein backbone topology. This paper is a proof of concept that shows it is possible to generate a novel helical backbone topology using an LSTM neural network architecture using only the ϕ and ψ angles as features. The next step is to attempt to use these backbone topologies and sequence design them to form complete protein structures.

Open Peer Review

Approval Status

	1	2
version 3 (revision) 12 Oct 2020		
version 2 (revision) 11 Sep 2020		 view
version 1 27 Apr 2020	 view	

1. **Matias Valdenegro-Toro** , German Research Center for Artificial Intelligence, Bremen, Germany
2. **Jan Ludwiczak** , University of Warsaw, Warsaw, Poland
Stanislaw Dunin-Horkawicz, University of Warsaw, Warsaw, Poland

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

Generative Adversarial Network, De novo Protein Design, Machine Learning, Neural Network, Deep Learning, Helical Proteins, Protein backbone, Long Short-Term Memory

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

This article is included in the **INCF** gateway.

Corresponding author: Sari Sabban (sari.sabban@gmail.com)

Author roles: **Sabban S:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation; **Markovsky M:** Investigation, Methodology, Software, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 Sabban S and Markovsky M. This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Sabban S and Markovsky M. **RamaNet: Computational *de novo* helical protein backbone design using a long short-term memory generative neural network [version 2; peer review: 2 not approved]** F1000Research 2020, 9:298 <https://doi.org/10.12688/f1000research.22907.2>

First published: 27 Apr 2020, 9:298 <https://doi.org/10.12688/f1000research.22907.1>

REVISED Amendments from Version 1

After close inspection of the SenseGen neural network code (which was adopted from the referenced paper) it is clear that it is an LSTM network and not an LSTM-GAN, thus the paper was changed to reflect this detail. Figure 2 was changed along with all sections to describe an LSTM rather than a GAN architecture. The output results remain the same. The paper title was also updated. This should address the reviewer's comments about a GAN architecture.

We did train an MLP-GAN (results which were not as good as the LSTM), as per reviewer's recommendation these results were added as a comparison at the end of the results section (Figure 5).

Other minor corrections were also amended.

Any further responses from the reviewers can be found at the end of the article

Introduction

The concept of amino acid sequences folding into globular protein molecules allows for proteins' large functional diversity, mediating all the functional aspects of living organisms, thus winning themselves attention from biochemists for decades. The fusion of machine learning and deep learning with computational biology is accelerating research in both fields and bringing humanity closer to the setup of performing most biological research quickly, cheaply, and safely *in silico*, while only translating the very crucial aspects of it. Having access to a large database of protein crystal structures has led to the use of machine learning to design proteins computationally.

De novo protein design (i.e. from the beginning) is very well explained in this review¹. Proteins fold into a specific shape depending on the sequence of their amino acids, and of course shape dictates function. The driving forces that allow proteins to fold are the hydrogen bond interactions within the backbone and between the side chains, the Van der Waals forces, and principally the interaction of hydrophobic side chains within the core. The space of all possible sequences for all protein sizes is extremely large (as an example there are 20²⁰⁰ possibilities for a 200-residue protein). Thus, is it not surprising that natural proteins exist in clusters close to each other, which is logical since proteins would evolve away from a central functional protein to fold correctly and acquire new folds and functions, rather than go through the tedious ordeal of finding a totally new protein structure within the space of all possibilities. Thus, even though the Protein Data Bank adds about 10,000 new structures to its repository every year, most of these new structures are not unique folds.

The relationship between the sequence of a protein and its specific structure is understood, but we still lack a unified absolute solution to calculate one from the other. Hence why some research groups generated man-made protein designs by altering already existing natural proteins², since randomly finding a functionally folded protein from the space of all possible protein sequences is more or less statistically impossible. On the other hand, other researchers have attempted *de novo* protein

design by designing a topology from assembling short sequence peptide fragments taken from natural protein crystal structures^{3,4}, these fragments are calculated statistically depending on the secondary structures they are found in. Sometimes this fragment system is combined with first physical principals to model the loops between secondary structures to achieve a desired three-dimensional topology⁵. Others have used parametric equations to study and specify the desired protein geometry⁶⁻¹¹. These solutions employ an energy function, such as REF2015, that uses some fundamental physical theories, statistical mechanical models, and observations of protein structures to approximate the potential energy of a protein¹². Knowing the protein potential energy allows us to guide our search for the structure of a protein given its sequence (the structure resides at the global energy minima of that protein sequence) thus attempting to connect the sequence of a protein with its structure. The critical tool is the energy function, the higher its accuracy the higher our confidence in knowing the computed structure is the real natural structure. Thus, using the energy function to perform structure prediction (going from a known sequence to find the unknown three-dimensional structure) can also be used to perform fixed-backbone design (going from a known three-dimensional structure to find the sequence that folds it). This is where this paper comes in. Where as in *de novo* design neither backbone nor sequence is known, knowing one results in finding the other using the same energy function¹, and a good starting point is to design the backbone.

Other researchers have used machine learning for protein sequence design, employing the constraints ($C\alpha$ - $C\alpha$ distances) as the input features for the network and using a sliding window to read a sequence of residues, getting their types and constraints then predicting the next one giving the output prediction as an amino acid sequence¹³, this architecture reported an accuracy of 38.3% and performs what is called sequence design: designing a sequence for a backbone, so when the protein is synthesised it folds to that backbone. In fact, in the⁵ paper the protocol first generated an all-valine backbone, then sequence designed that backbone. In this paper, we want to computationally generate a backbone so it can be sequence designed using other protocols such as RosettaDesign^{4,14} or the protocol from¹³.

The protein's backbone can be folded using the ϕ and ψ angles, which are the angles between $C\alpha$ and N for ϕ and $C\alpha$ and C for ψ that primarily move the amino acid's backbone (not the side chains), and thus can be used as features to control the topology of a backbone. They were even one of the features used to fold proteins and predict their structures in AlphaFold¹⁵.

But the question is: how do we decide the ideal angles for a helix, the length of each helix, the number of helices, as well as the lengths and angles of the loops between the helices that will also result is a compact folded protein backbone. These numerous values can be solved statistically using neural networks. Especially that we want to use the structures in the PDB to forward design (rather than discover) new protein folds that are not a result of evolution.

The deep neural network architecture that we chose was a long short-term memory (LSTM) network¹⁶. The LSTM is usually used in natural language and data sequence processing, but in our model the LSTM was used to generate a sequence of ϕ and ψ angles. The model was constructed as a stack of LSTM layers, followed by fully connected layers, followed by a mixture density network (MDN)¹⁷ and worked by using random noise numbers (a latent) as input to build the values for the ϕ and ψ angles¹⁸.

Our effort in this paper was to use deep learning to learn the general fold of natural proteins, and using this generalising statistical concept to design novel protein backbone topologies, thus only getting the three dimensional backbone structure so it can be sequence designed using other protocols. Our research at this moment was a proof of concept and only concerned with getting a new and unique folded ideal helical protein backbone rather than a protein with a specific sequence, nor a function, nor a specific predetermined structure. Our system resulted in random yet compact helical backbone topologies only.

Methods

The following steps were used to generate the augmented training dataset, along with details of the neural network architecture and how the output was optimised then evaluated.

Data generation

The entire PDB database was downloaded on 28th June 2018 (~150,000 structures), and each entry was divided into its constituent chains resulting in individual separate structures (i.e: each PDB file had only a single chain). Each structure was analysed and chosen only if it contained the following criteria: contained only polypeptides, had a size between 80 and 150 amino acids without any breaks in the chain (a continuous polypeptide), a sum of residues that made up helices and sheets were larger than the sum of amino acids that made up loops, and the final structure having an Rg (radius of gyration) value of less than 15 88 Å. The chosen structures were then further filtered by a human to ensure only the desired structure concepts were selected, removing the structures that slipped through the initial computational filter. Furthermore, a diversity of structure folds were achieved rather than numerous repeats of the same fold (the haemoglobin fold was quite abundant). In previous attempts, a mixture of different structure classes were used, where some structures were only helices, some were only sheets, and the remaining were a mix of the two. However, that proved challenging in optimising the network, and as such a dataset made up of only helical protein structures was chosen for this initial proof of concept. The final dataset had 607 ideal helical structures. These structures were then cleaned (non-amino acid atoms were removed) in preparation to push them through the Rosetta version 3 modelling software that only takes in polypeptide molecules.

Data augmentation

These 607 structures were augmented using the Rosetta Fast-Relax protocol¹⁹. This protocol performs multiple cycles of packing and minimisation. In other words, the protocol performs small slight random angle moves on the backbone and side

chains in an attempt to find the lowest-scoring variant, but the random backbone angle moves is what we were after. Its originally intended function was to move a structure slightly to find the conformation of the backbone and side chains that corresponds to the lowest energy state as per the REF2015 energy function. Since the protocol performs random moves, a structure relaxed on two separate occasions will result in two molecules that look very similar with similar minimum energy scores, but technically have different ϕ and ψ angle values. This is the concept we used to augment our structures, and each structure was relaxed 500 times to give a final dataset size of 303,500 structures.

Feature extraction

Using only the ϕ and ψ angle values from a crystal structure it was possible to re-fold a structure back to its correct native fold, thus these angles were the only relevant features required to correctly fold a structure, Figure 1A details the range of angles in the un-augmented data. Each amino acid's ϕ and ψ angle values were extracted and tabulates as in Table 1. This was the dataset used to train the neural network.

The neural network

The model in Figure 2 was built using the SenseGen model as a template¹⁸ and consisted of a network constructed from an LSTM layer with 64 nodes, followed by two dense fully connected MLPs with 32 nodes for the first layer and 12 nodes for the second one, both employed a sigmoid activation function:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

Which was followed by an MDN layer employing an MDN activation function:

$$p(y|x) = \sum_{c=1}^C \alpha_c(x) \mathcal{D}(y|\lambda_{1,c}(x), \lambda_{2,c}(x), \dots)$$

c : the index of the corresponding mixture component. α : the mixing parameter. $\mathcal{D}$: the corresponding distribution to be mixed. λ : the parameters of the distribution $\mathcal{D}$, as we denote $\mathcal{D}$ to be a Gaussian distribution, λ_1 corresponds to the conditional mean and λ_2 to the conditional standard deviation. The training was done using the Adam optimiser²⁰, for each parameter ω^j :

$$v_t = \rho v_{t-1} + (1 - \rho) * g_t^2 \Delta \omega_t = -\frac{\eta}{\sqrt{v_t + \epsilon}} * g_t \omega_{t+1} = \omega_t + \Delta \omega_t$$

η : initial learning rate. v_t : exponential average of squares of gradients. g_t : gradient at time t along ω^j . The loss defined as the root mean squared difference between the sequence of inputs and the sequence of predictions:

$$\text{loss} = \sum_{t=1}^T (x_t - y_t)^2$$

y_t : output. x_t : next step sample $x_{t+1} = y_t$.

The network used random noise as a starting seed, this noise was generated by taking a single randomly distributed number

Figure 1. Ramachandran plots. 7A: Ramachandran plot of the dataset showing the ϕ and ψ angles of each amino acid for each structure. This is the unaugmented data of structures that are only made of helices. Green represents the angles on amino acids in loops, while red represents the angles of amino acids in helices. Some orange can be seen where the DSSP algorithm classified the amino acids as sheets (though there were none). One point to note; the angles here are represented between the range -180° to 180° as is conventional, while in the actual dataset the range was from 0° to 360° . 7B: The network's output ϕ and ψ angles for 25 structures after the relaxation step. The green dots represent the angles of amino acids within loops, and red within helices clustering around the same location as Figure 1A within the fourth quadrant as is desired for an α -helix that has ideal angles around $(-60^\circ, -45^\circ)$. These structures culminated to all 25 structures in Figure 4, and had an angle range for the helices $(-127.4^\circ < \phi < -44.7^\circ, -71.3^\circ < \psi < 30.6^\circ)$ not including the outliers. The purple dots represent the helices in the control structures, and the black dots the loops.

Table 1. The PS_Helix_500.csv dataset: The first five examples of the PS_Helix_500.csv dataset showing the PDB ID_chain augmentation number, residue 1 ϕ angle, residue 1 ψ angle, all the way to residue 150. The value 360.0° was used for the first missing angle while 0.0° was used to represent no residues.

	PDB_ID	phi_1	psi_1	phi_2	psi_2	phi_3	psi_3		phi_150	psi_150
1	1TQG_A_0293.pdb	360	98.8	207.4	163.8	298.1	313.6		0.0	0.0
2	1EZ3_A_0261.pdb	360	227.8	208.3	37	306.7	316.4		0.0	0.0
3	5IP0_E_0241.pdb	360	86.7	293.2	328.7	292.2	313.1		0.0	0.0
4	2P5T_G_0123.pdb	360	185.9	254.3	176.2	308.4	139.3		0.0	0.0
5	5EOH_A_0211.pdb	360	144.4	293.5	334.6	320.9	320.9		0.0	0.0

Figure 2. The *de novo* helical protein backbone design protocol. The full protocol showing the structure of the neural network model and its output. The model employing an LSTM network. The network's output is the generated ϕ and ψ angle sequences which were applied to a primary structure (a fixed 150 valine length straight structure generated by PyRosetta) that resulted in the development of the secondary structures but not a final compact structure due to suboptimal loop structures as a result of their variability in the dataset. To overcome this, the structure was relaxed to bring the secondary structure helices together. This did result in more compact structures but was not always ideal, thus a filter was used to filter out non-ideal structures and keep an ideal structure when generated.

between [0, 1) as the first predicted values, these values were then reshaped to the same shape of the last item of the predicted value resulting in a final shape of (batch_size, step_number, 1). The network predicted the main parameters of the new value (μ , σ , π) several times (according to the number_of_mixtures

value) and selected the single mixture randomly but according to the π value. It then predicted the next value according to the normal distribution using the μ and σ values. It added the final value to the prediction chain and then returned to step 2 until the predefined sequence length was obtained. The initial random

number was stripped from the returned sequence. Once the networks were constructed, the dataset was normalised and the training was done using the following parameters: the learning rate was 0.0003 with a batch size of 4 over 18,000 epochs.

Post-backbone topology generation processing and filtering

The output of the neural network was always a sequence of 150 ϕ and ψ angle value combination for a structure with 150 amino acids. A straight chain of 150 valines was computationally constructed using PyRosetta and used as a primary structure. Each amino acid in the primary structure had its angles changed according to the ϕ and ψ angle values, which ended up folding that primary structure resulting in secondary structures of helices and loops between them. The COOH end was trimmed, if it was a loop, until it reached the first amino acid that comprised a helix, thus variable structure sizes were generated. The structure ended up with helices and loops yet still with an open conformation. The generated structure was therefore relaxed using **PyRosetta version 4** FastRelax protocol to idealise the helices and compact the structure. Furthermore, not every prediction from the neural network resulted in an ideal structure even after the relax step, therefore we employed a filter to filter out structures we deemed not ideal. The filter discards structures that were less than 80 amino acids, had more residues in loops than in helices, had less than 20% residues making up its core, and had a maximum distance between Ca1 and any other Ca greater than 88 Å (the largest value in the dataset). PyRosetta was used²¹ since it was easier to integrate the code with the neural network's python script and combine the Rosetta engine with Biopython²² and DSSP^{23,24}.

Rosetta *de novo* protein design as a comparison

As a control we used the *de novo* protein design protocol using the Rosetta Script from the⁵ paper's supplementary material. The protocol was modified slightly to accommodate new updates in the Rosetta software suite, but maintained the talaris2014 energy function as in the original paper, and we used the protocol to design helical proteins. These proteins had to pass through several filters including a talaris2014 score of less than -150, a packing threshold of more than 0.50, and a secondary structure threshold of more than 0.90. We attempted to design proteins with 3, 4, and 5 helices to compare the backbone quality our neural network outputs to the output of the⁵ paper.

Implementation

This setup used the following packages: python 3.6.9, PyRosetta 4, Rosetta 3, Tensorflow 1.13.1, BioPython 1.76, and DSSP 3.01 and was run on GNU/Linux Ubuntu 19.10. Further information on running the setup can be found on this [GitHub](#) repository which includes an extensive README file. To train the neural network a 3GB GPU and 12GB of RAM is recommended, while to run the trained neural network and generate a backbone structure an Intel i7-2620M 2.70GHz CPU and 3GB of RAM is recommended.

Operation

As Detailed in [Figure 2](#), executing the trained network will generate a set of random numbers that are pushed through the

network which (using the weights) will modify the values to become values in accordance with the ϕ and ψ angle topologies observed in the training dataset. A simple straight chain of 150 valines is then computationally constructed and the ϕ and ψ angles are applied to each amino acid in order, resulting in the appearance of helical secondary structures. Any tailing loops at the COOH end will be cut out since it interferes with the next step. This structure is then relaxed using the FastRelax protocol which moves the ϕ and ψ angles randomly in its attempt to find the lowest scoring configuration compacting the structure in the process. A filter is applied to determine whether the final structure is ideal or not, its parameters are detailed in the Methods section. If the structure passes the filter the script exits, otherwise it repeats the whole process.

Results

The dataset was named PS_Helix_500 due to the fact that the features used were the ϕ and ψ angles, only strictly helical protein structures were used, and each structure was augmented 500 times.

Minimal hyperparameter search was needed and was performed manually. The neural network was trained on the dataset for 18,000 epochs (further training collapsed the network, i.e. all outputs were exactly the same structure) with a generally sloping down mean loss as shown in [Figure 3](#) indicating that the network got better at generating novel data compared to the original training data. The network was used to generate a sequence of ϕ and ψ angles for 25 structures. In other words, random numbers were generated then pushed through the network where they were modified (using the network's trained weights) to become the ϕ and ψ angles for 150 residues. This is the angle profile. Using PyRosetta, a simple straight 150-valine chain was constructed and used as a primary structure. The generated ϕ and ψ angles were applied to this primary structure (each amino acid's angles were changed according to the angle profile) resulting in a folded structure clearly showing helical secondary structures with loops between them. The last tailing loop at the COOH end was truncated, resulting in structures with variable sizes. The loop angles were within the area of the angles found in the dataset 1, but because they were generated independent of the thermodynamic stability of the structure, the structure did not come together into compact a topology. To push the structure into a thermodynamic energy minima it was relaxed using the PyRosetta FastRelax function which compacted the backbone topology while still having valines as a temporary placeholder sequence. This was repeated 25 times to result in the 25 structures in [Figure 4](#). For comparison five control structures were generated using the *de novo* design protocol by the previous paper⁵. [Figure 1B](#) shows the Ramachandran plot of the 25 generated structures, where red are the amino acids within helices having angles clustering around the same location as [Figure 1A](#) (in the fourth quadrant), as is desired for an α -helix, which has ideal angles around (-60° , -45°), our results had an angle range for the helices ($-127.4^\circ < \phi < -44.7^\circ$, $-71.3^\circ < \psi < 30.6^\circ$) not including the outliers, the five control structures show angles within the same region (purple for helices and black for loops). The structure generation setup was not perfect at achieving an ideal

Figure 3. The training loss. The mean loss of the whole network over epoch, for 18,000 epochs, showing a general downward trend. This indicated that in subsequent epochs the network got better at generating structures.

structure every time, so a filter was deployed to filter out sub-optimal structures by choosing a structure that had more residues within helices than within loops, was not smaller than 80 residues, and had more than 20% residues comprising its core. Due to the random nature of the structure generation the efficiency of the network is variable, while generating the 25 structure in [Figure 4](#) the fastest structure was generated after just 4 failed attempts, while the slowest structure was generated after 3834 failed attempts giving a success rate for the network between 25.0% at its best and 0.025% at its worst, and this took between ~1 minute and ~6 hours to generate a single structure with the desired characteristics utilising just 1 core on a 2011 MacBook Pro with 4 core Intel i7-2620M 2.70GHz CPU, 3GB RAM, and 120GB SSD. For comparison the protocol by 5 took ~60 minutes for each of the control structures on the same machine. The protocol is summarised in [Figure 2](#), and the results are compiled in [Figure 4](#) showing all 25 structures and the 5 controls.

These 25 structures had on average 84.7% of their amino acids comprising their helices, along with an average of 29.9% of their amino acids comprising their cores, [Table 2](#) shows the Rosetta packing statistic for all 25 structures, along with the five controls, the top7 protein, and the five structures from the⁵ paper all showing similar packing values.

The LSTM network was not the only network that was tested, an MLP-GAN²⁵ was also tested on the same dataset, hyperparameter search was performed through a grid search then a random search to finally reach the following hyperparameters

for the generator network: 4 layers with the following nodes 256, 512, 1024, and batch normalization with 0.15 momentum between the layers, and an output node with the shape of the dataset (150, 2). This network employed the ReLU activation for the hidden layers and the TanH activation function for the output layer with the binary crossentropy loss and the Adam optimiser with a 0.0004 learning rate. The discriminator network had the following hyperparameters: 3 layers with the following nodes 512, 256, and 1 node for the output, and 25% dropout rate between the layers. This network employed the ReLU activation for the hidden layers and the sigmoid activation function for the output layer with the binary crossentropy loss and the Adam optimiser with a 0.002 learning rate. The architecture was trained for 100,000 epochs with a mini batch size of 512 and the results were not as good as the LSTM [Figure 5](#); generating structures that did have helices and loops but much more sparse and kinked. Thus the LSTM is the preferred architecture for this particular dataset.

Discussion

In this paper we outlined how a neural network architecture can design a compact helical protein backbone. We attempted to test our network to generate structures that included sheets, but that failed, mainly due to the wide variation in the loop angles that did not compact a structure to bring the strands together, which sheets rely on more to develop compared to helices.

We demonstrated that the ϕ and ψ angles were adequate features to design a protein backbone topology only without a

Figure 4. The designed structures. This figure shows all 25 structures that were generated using the neural network, displayed using PyMOL²⁶. It can be seen that all structures have compact helical structures of variable topologies. The 5 control structures at the bottom were generated using the protocol from the⁵ paper showing better helices but similar compactness.

Table 2. Structure packing scores: This table summarises the packing score of each structure, calculated as the average from 30 measurements using the PyRosetta output `packstat` function. The controls were the structures generated using the modified protocol from⁵. For additional comparison the top7 (PDB ID: 1QYS)⁴ protein packing stat is measured along with the five structures from the protocol paper by⁵.

Structure	Packing Score	Structure	Packing Score	Structure	Packing Score	Structure	Packing Score
1	0.610	11	0.612	21	0.803	top7	0.479
2	0.657	12	0.673	22	0.758	2KL8	0.447
3	0.647	13	0.520	23	0.489	2LN3	0.630
4	0.632	14	0.713	24	0.693	2LTA	0.627
5	0.548	15	0.698	25	0.705	2LV8	0.700
6	0.663	16	0.660	Control 1	0.564	2LVB	0.635
7	0.655	17	0.756	Control 2	0.534		
8	0.731	18	0.788	Control 3	0.568		
9	0.638	19	0.720	Control 4	0.555		
10	0.734	20	0.649	Control 5	0.501		

sequence. Although we understand the distribution of angles in the Ramachandran plot (the distribution of helix and sheet angles) the neural network's power comes in the form of combining several dimensional spaces to take a better decision. Given its observation of natural proteins, it can calculate the ideal combination of angles that results in deciding the number and lengths of helices, and the loops between them, that will still result in a compactly folded protein backbone.

The reason we concentrated our efforts on generating a backbone only is because once a backbone is developed it can be sequence-designed using other protocols, such as RosettaDesign^{4,14} or the protocol from¹³.

Though our network had a wide variation of success rates, from high to low, that was due to the random nature of the setup, which was our target to begin with (to randomly generate backbone topologies rather than directly design a specific pre-determined topology). Generating multiple structures and auto filtering the suboptimal ones provided an adequate setup, this achieved our goal of *de novo* helical protein backbone design within a reasonable time (1–6 hours) on readily available machines.

As a control we used a slightly modified *de novo* design protocol from⁵, which also performs backbone design (resulting in an all-valine backbone topology) followed by sequence design of that backbone. It has numerous advantages such as generating better helices, but the user must still pre-determine

the topology to be generated (decide the number of helices, their lengths and locations, and the lengths of the loops between them), while this neural network automatically takes that decision and randomly generates different backbones, which can be very useful for database generation (see below).

The neural network is available at this [GitHub](#) repository which includes an extensive README file, a video that explains how to use the script, as well as the dataset used in this paper and the weights files generated from training the neural network.

For future work we are currently working on an improved model that uses further dataset dimensions and features that will allow the design of sheets.

There are many benefits to generating numerous random protein structures computationally. One benefit can be in computational vaccine development, where a large diversity of protein structures as scaffolds are required for a successful graft of a desired motif^{27–29}. Using this setup, combined with sequence design, a database of protein structures can be generated that provides more variety of scaffolds than what is available in the protein databank, especially that this neural network is designed to produce compact backbones between 80 and 150 amino acids, which is within the range of effective computational forward folding simulations such as AbinitioRelax structure prediction³⁰.

Figure 5. Results from an MLP-GAN. This figure shows 25 structures generated from an MLP-GAN, it can be inferred that these structures are not as good as the ones generated from an LSTM since they are much more fragmented; with smaller kinked helices and numerous loops between them.

Data availability

Source data

The entire Protein Data Bank was downloaded using this command:

```
$ rsync -rlpt -v -z --delete --port=33444 rsync.wwpdb.org
:: ftp/data/structures/divided/pdb/.dataset
```

After which it was processed and filter as described in the *Methods* section.

Software availability

Source code available from: <https://sarissabban.github.io/RamaNet/>.

Archived Source code at time of publication: [10.5281/zenodo.3755343](https://zenodo.org/record/3755343)

License: MIT License.

Author contributions

Corresponding author: Sari Sabban sari.sabban@gmail.com
 Author roles Sari Sabban: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation. Author roles Mikhail Markovsky: Investigation, Methodology, Software, Writing – Review & Editing.

Acknowledgements

The corresponding author would like to thank the High Performance Computing Center at King Abdulaziz University for making available the Aziz high performance computer where the corresponding author was able to augment the dataset and perform the Abinitio folding simulations.

References

- Huang PS, Boyken SE, Baker D: **The coming of age of *de novo* protein design.** *Nature*. 2016; **537**(7620): 320–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dougherty MJ, Arnold FH: **Directed evolution: new parts and optimized function.** *Curr Opin Biotechnol*. 2009; **20**(4): 486–91.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Huang PS, Ban YEA, Richter F, et al.: **Rosettaremodel: a generalized framework for flexible backbone protein design.** *PLoS One*. 2011; **6**(8): e24109.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kuhlman B, Dantas G, Ireton GC, et al.: **Design of a novel globular protein fold with atomic-level accuracy.** *Science*. 2003; **302**(5649): 1364–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Koga N, Tatsumi-Koga R, Liu G, et al.: **Principles for designing ideal protein structures.** *Nature*. 2012; **491**(7423): 222–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grigoryan G, Degrado WF: **Probing designability via a generalized model of helical bundle geometry.** *J Mol Biol*. 2011; **405**(4): 1079–100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Harbury PB, Plecs JJ, Tidor B, et al.: **High-resolution protein design with backbone freedom.** *Science*. 1998; **282**(5393): 1462–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Huang PS, Oberdorfer G, Xu C, et al.: **High thermodynamic stability of parametrically designed helical bundles.** *Science*. 2014; **346**(6208): 481–485.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Joh NH, Wang T, Bhate MP, et al.: **De novo design of a transmembrane α -helix bundle.** *Science*. 2014; **346**(6216): 1520–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Regan L, DeGrado WF: **Characterization of a helical protein designed from first principles.** *Science*. 1988; **241**(4868): 976–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Thomson AR, Wood CW, Burton AJ, et al.: **Computational design of water-soluble α -helical barrels.** *Science*. 2014; **346**(6208): 485–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Alford RF, Leaver-Fay A, Jeliazkov JR, et al.: **The Rosetta all-atom energy function for macromolecular modeling and design.** *J Chem Theory Comput*. 2017; **13**(6): 3031–3048.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang J, Cao H, Zhang JZH, et al.: **Computational protein design with deep learning neural networks.** *Sci Rep*. 2018; **8**(1): 6349.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Murphy GS, Mills JL, Miley MJ, et al.: **Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core.** *Structure*. 2012; **20**(6): 1086–96.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Senior AW, Evans R, Jumper J, et al.: **Improved protein structure prediction using potentials from deep learning.** *Nature*. 2020; **577**(7792): 706–710.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Radford A, Metz L, Chintala S: **Unsupervised representation learning with deep convolutional generative adversarial networks.** *arXiv*. 2015.
[Reference Source](#)
- Bishop CM: **Mixture density networks.** Technical Report, Aston University, Birmingham, (Unpublished). 1994.
[Reference Source](#)
- Alzantot M, Chakraborty S, Srivastava M: **Sensegen: A deep learning architecture for synthetic sensor data generation.** In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, IEEE. 2017; 188–193.
[Publisher Full Text](#)
- Tyka MD, Keedy DA, André I, et al.: **Alternate states of proteins revealed by detailed energy landscape mapping.** *J Mol Biol*. 2011; **405**(2): 607–18.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kingma DP, Ba J: **A method for stochastic optimization.** *arXiv preprint*, 434(arXiv: 1412.6980). 2017.
[Reference Source](#)
- Chaudhury S, Lyskov S, Gray JJ: **Pyrosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta.** *Bioinformatics*. 2010; **26**(5): 689–91.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cock PJ, Antao T, Chang JT, et al.: **Biopython: freely available python tools for computational molecular biology and bioinformatics.** *Bioinformatics*. 2009; **25**(11): 1422–3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Joosten RP, te Beek TAH, Krieger E, et al.: **A series of PDB-related databases for everyday needs.** *Nucleic Acids Res*. 2011; **39**(Database issue): D411–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Touw WG, Baakman C, Black J, et al.: **A series of PDB-related databanks for everyday needs.** *Nucleic Acids Res*. 2015; **43**(Database issue): D364–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goodfellow IJ, Pouget-Abadie J, Mirza M, et al.: **Generative Adversarial Networks.** *arXiv 1406.2661*. 2014.
[Reference Source](#)
- The PyMOL Molecular Graphics System. Version 1.8. 2015.
[Reference Source](#)
- Correia BE, Bates JT, Loomis RJ, et al.: **Proof of principle for epitope-focused vaccine design.** *Nature*. 2014; **507**(7491): 201–206.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Azoitei ML, Ban YEA, Julien JP, et al.: **Computational design of high-affinity**

- epitope scaffolds by backbone grafting of a linear epitope. *J Mol Biol.* 2012; **415**(1): 175–192.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Azoitei ML, Correia BE, Ban YE, *et al.*: **Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold.** *Science.* 2011; **334**(6054): 373–376.
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Rohl CA, Strauss CEM, Misura KMS, *et al.*: **Protein structure prediction using Rosetta.** *Methods Enzymol.* 2004; **383**: 66–93.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 24 September 2020

<https://doi.org/10.5256/f1000research.29106.r71608>

© 2020 Ludwiczak J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jan Ludwiczak

Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

Stanislaw Dunin-Horkawicz

Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

In their paper, Sabban and Markovsky propose a novel computational method for the creation of the backbones of the alpha-helical protein structures. The authors designed and trained a generative neural network using dihedral angles of the natural protein structures with the aim of creating protein backbones with presumably novel topologies. As the network efficiency is not ideal the authors use additional post-processing steps to assure that the obtained backbones are helical and well-packed.

This is an interesting concept that falls into the wide range of applications of machine learning techniques to study the sequence-structure relationships in proteins. At the same time, in my opinion, some aspects of the work and the underlying method should be clarified and improved before the manuscript will be suitable for indexing.

My biggest concern is that the authors do not compare their approach to the current state-of-the-art of applying generative ML models in the protein design/backbone generation problems. The authors did not refer to recent relevant papers, e.g. Anand and Huang (2018¹), Anand *et al.* (2019²) and Eguchi *et al.* (2020³). It would be important to see the description of the method in a broader context, i.e. what was already done and how the authors' approach and contribution compares with this, and what the advantages and weaknesses of their method are.

Another aspect connected with the first point is the method benchmarking. In the current version of the manuscript the authors present, in the results section, 25 backbone architectures generated with the proposed algorithm and compare them with the 5 structures generated with the protocol adapted from the Koga *et al.* (2012) paper and 5 experimental structures from PDB. The comparison is made in terms of the packing scores, distributions of the dihedral angles, and visual

inspection. I think that the paper would benefit a lot by including a more quantitative benchmark and comparison with other methods (e.g. those mentioned above), given that their source codes are available.

For a more quantitative benchmark the authors could reach for other computational validation methods - for example, design the sequences compatible with their calculated backbones and check with the aid of in silico folding (e.g. Rosetta) whether these will fold to the stable structures (a similar approach was undertaken in Anand *et al.*, 2019²). A comparison of the designed sequence characteristics and foldability with other methods would be useful here to judge the method performance. Furthermore, the authors could also assess the novelty of the backbones generated with their method. Are the generated backbones representing novel architectures not seen in the experimental structures? Demonstrating that the method can in fact generate novel architectures, not previously seen in the PDB (and which, after sequence design, are able to fold in silico to the predicted topology) would certainly make the method more convincing.

The aforementioned concerns deal mostly with the biological/protein design aspects of the work. However, I also agree with several concerns regarding the machine learning part raised by Reviewer #1, in particular, the justification of the selected architecture and loss functions used (this part was largely corrected by the authors in the amended version of the manuscript – ver. #2).

Minor issue – It would be good to have the software requirements (names and versions of the packages) also in the machine-readable format in the repository (e.g. requirements.txt format) for the ease of software installation (now on the website - <https://sarisabban.github.io/RamaNet/> - only names of the packages are visible).

References

1. Anand N, Huang PS: Generative Modeling for Protein Structures. *32nd Conference on Neural Information Processing Systems (NeurIPS, 2018)*. 2018. [Reference Source](#)
2. Anand N, Eguchi R, Huang PS: Fully differentiable full-atom protein back-bone generation. *7th International Conference on Learning Representations (ICLR, 2019)*. 2019. [Reference Source](#)
3. Eguchi R, Anand N, Choe C, Huang P: IG-VAE: Generative Modeling of Immunoglobulin Proteins by Direct 3D Coordinate Generation. *bioRxiv*. 2020. [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Structural bioinformatics, machine learning

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 11 Nov 2020

Sari Sabban

Corrections:

1. As per reviewer's recommendation we added the linux_requirements.txt and python_requirements.txt files to the repository.
2. We added a paragraph in the introduction section about the related works the reviewer recommended.
3. We measured the RMSD of each of the generated structures compared to the structures in the dataset and added the data to the results section. This is to see whether or not the network generates structures within the dataset (overfitting). New figures added as well.
4. As the reviewer suggested, we added results showing 5 additional structures that were sequence designed and their folding simulated. These results were previously posted in the first iteration of the BioRxiv manuscript but we removed them for the following reason: we were unsure whether the sequence design / folding simulation was an adequate evaluation criteria for our particular results, reason is because the AbinitioRelax folding simulation does not always successfully fold natural structures (even natural crystal structures). Take the example between 1YN3 and 4WYH (where we tried to use as controls) here Abinitio successfully folds 1YN3 but not 4WYH, thus 1YN3 can be sequence designed and folded but 4WYH cannot be sequence designed and folded even though it is a natural helical protein structure (results can be provided). Given this, when we sequence designed and folded our structures we could not determine whether they were not successfully folding because their backbone are not logical, or because Abinitio fails at folding those particular backbones, or because good fragments could not be generated for those particular structures, or because the sequence design step failed, especially that the network is producing random structures. We did however get successful secondary structure predictions from PSIPRED after sequence design, further explanation is in the results section of the new manuscript version. Nonetheless we re-added just the Abinitio folding simulation results and we hope this will give further and more accurate insight into our setup.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 02 July 2020

<https://doi.org/10.5256/f1000research.25288.r64120>

© 2020 Valdenegro-Toro M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Matias Valdenegro-Toro

¹ Robotics Innovation Center, German Research Center for Artificial Intelligence, Bremen, Germany

² Robotics Innovation Center, German Research Center for Artificial Intelligence, Bremen, Germany

This paper is about a generative model for protein folding.

The main idea is to use a LSTM Generative Adversarial Network, trained on a dataset of real protein backbone angles, in order to generate novel protein designs. This makes sense, as GANs are regularly used as generative models to learn the probability distribution from samples, from where new samples can be generated. These methods are generally applied to images, but it is possible to apply them to other modalities such as sequences, text, etc. I find that this is a novel application that is a bit outside of the common uses of GANs.

Overall the paper looks good in terms of writing and technical content. I am not an expert in proteins, so I will comment on the Machine Learning parts, which are my expertise.

The authors propose the use of an LSTM-GAN, but they do not justify why they use it, in particular the LSTM architecture of a GAN, which is not that common. GANs are mostly used as generative models for images, and there are GAN versions for sequential data. My question is then, what is the motivation to use an LSTM specifically? From what I understand, the dimensionality of the problem is fixed, and protein samples do not have a varying number of folds or angles, in which case a sequential model such as an RNN or LSTM does not seem to be necessary. A simple multi-layer perceptron (MLP) could work equally well, it could be interesting if the authors tried this or performed an architecture search to justify their choice.

The neural network architectures for the discriminator and generator look fine, it would be nice if the authors mention how they reached these architecture, and if any grid/random search was performed.

The choice of activation function seems odd, since the sigmoid activation used in hidden layers can lead to vanishing gradient problems, while the current standard are ReLU activations and its many variations. The choice of the MDN also looks odd, since a GAN generator network generally outputs values in the output range directly, while the MDN outputs a probability distribution from where a sample needs to be drawn. I believe both choices have to be justified appropriately.

The authors use the Adam optimizer, which is standard and a good choice for a starting problem. There is a sentence in the paper which I find problematic: "The Adam optimiser had an MDN activation through time loss function defined to increase the likelihood of generating the next time step value." The problem here is that conceptually, an optimizer computes a gradient descent step, an optimizer does not have an activation or loss associated with it, so I am not sure what the authors mean here. The authors mention a L2 loss (mean squared difference), but this is not the loss that is used to train a GAN, even for continuous values.

The authors mention that for training a GAN, the binary cross-entropy loss is used, but this is similar and not exactly the loss that is used to train GANs. The correct loss actually consists of two loss functions, one for the generator, and another for the discriminator, and both losses use the output of the D and G networks together. A full description of the losses and its variations can be found in Section 3.2 of [1], and I believe the authors should describe the actual loss they used very carefully, including the interaction of the D and G networks into the loss, as it is the core of the method used to generate protein angles in their paper.

The description of the algorithm used to train a GAN looks okay, except for steps 3 and 6 since as I mentioned before, the loss used for training a GAN used both the D and G networks, so this should be explicitly mentioned. Step 4 and 5 are not strictly necessary since a GAN can train the D and G networks on the same batch.

I briefly analyzed the code linked in the paper, and I am not convinced that the authors are using a Generative Adversarial Network (GAN). The network implemented in the code looks more like an auto-regressive model, and I do not see adversarial training in it. Additional evidence is that in Figure 3 only one loss function is being shown, while as I mentioned before, a GAN uses two loss functions and generally both are plotted in order to visualize training. I believe this is a very important point in order to make the article scientifically sound, but not the only one.

I think the authors could also discuss what other learning-based methods they tried. Other methods such as Variational Autoencoders (VAEs) come to mind, which are also able to work as generative models.

I believe that many references are missing from this paper, in particular to the ML community. I have already mentioned the GAN tutorial [1], and additionally I would suggest the Adam optimizer paper [2], and the original Mixture Density Networks paper [3]. Please always make sure that you are citing the techniques you are using from another field.

[1]: <https://arxiv.org/abs/1701.00160>

[2]: <https://arxiv.org/abs/1412.6980>

[3]: <http://publications.aston.ac.uk/id/eprint/373/>

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 01 Sep 2020

Sari Sabban

After close inspection of the SenseGen neural network code (which was adopted from the referenced paper) it is clear that it is an LSTM network and not a GAN, thus the paper was changed to reflect this detail. The point of confusion was that the SenseGen neural network authors used a generator and a discriminator networks, but at the very end of their article they mentioned that they were not connected thus not trained adversarially.

To fix this, Fig 2 was changed along with all sections to describe an LSTM rather than a GAN architecture. The output results remain the same since the network was neither changed nor re-trained. The paper title was also updated. This should address the comments about the GAN architecture.

We did train an MLP-GAN (results which were not as good as the LSTM), as per the recommendation these results were added as a comparison at the end of the results section (fig 5).

Competing Interests: No competing interests

Comments on this article

Version 2

Author Response 03 Oct 2020

Sari Sabban

Corrections:

1. As per reviewer's recommendation we added the linux_requirements.txt and python_requirements.txt files to the repository.
2. We added a paragraph in the introduction section about the related works the reviewer recommended.
3. We measured the RMSD of each of the generated structures compared to the structures in the dataset and added the data to the results section. This is to see whether or not the network generates structures within the dataset (overfitting). New figures added as well.
4. As the reviewer suggested, we added results showing 5 additional structures that were sequence designed and their folding simulated. These results were previously posted in the first iteration of the BioRxiv manuscript but we removed them for the following reason: we were unsure whether the sequence design / folding simulation was an adequate evaluation criteria for our particular results, reason is because the AbinitioRelax folding simulation does not always successfully fold natural structures (even natural crystal structures). Take the example between 1YN3 and 4WYH (where we tried to use as controls) here Abinitio successfully folds 1YN3 but not 4WYH, thus 1YN3 can be sequence designed and folded but 4WYH cannot be sequence designed and folded even though it is a natural helical protein structure (results can be provided). Given this, when we sequence designed and folded our structures we could not determine whether they were not successfully folding because their backbone are not logical, or because Abinitio fails at folding those particular backbones, or because good fragments could not be generated for those particular structures, or because the sequence design step failed, especially that the network is producing random structures. We did however get successful secondary structure predictions from PSIPRED after sequence design, further explanation is in the results section of the new manuscript version. Nonetheless we re-added just the Abinitio folding simulation results and we hope this will give further and more accurate insight into our setup.

Competing Interests: None

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research