Summer Review 5

FBGAN

Anvita Gupta, James Zou

Stanford University

Reviewed by : Arshdeep Sekhon

[1]Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

The Task: GANs for synthetic biology

- generate genes (protein sequences) that can encode proteins with specific properties
  - generate antimicrobial peptides: lower molecular weight peptides with less than 50 amino acids
  - optimize secondary structure for peptides– alpha helical peptides

Definitions

- secondary structure: the three dimensional form of local segments of proteins.
- important for protein functions
- two types: alpha[50 amino acids] and beta

The Dataset generation

- Uniprot database: Select 3655 proteins with length $5 - 50$ residues[1]
- Uniprot database: protein sequence records with functional information
- cluster by sequence similarity
- Select one from each cluster as a representative protein
- convert into cDNA sequences
- get a codon for each amino acid, start codon, stop codon[2]

[1] limit the length to 50 to avoid long-term dependencies + observation of secondary structure etc.

[2] The start codon marks the site at which translation into protein sequence begins, and the stop codon marks the site at which translation ends.

GAN

Loss function:

$$\min_G \max_D V(D,G) = E_{x \in P_{data}(x)}[log(D(x))] + E_{z \in P(z)}[log(1 - D(G(z)))] \quad (1)$$

WGAN more stable during training.

- 5 residual layers.
- 2 $1 - D$ convolutions of 5×1
- Use Gumbel Softmax instead of Softmax

GAN

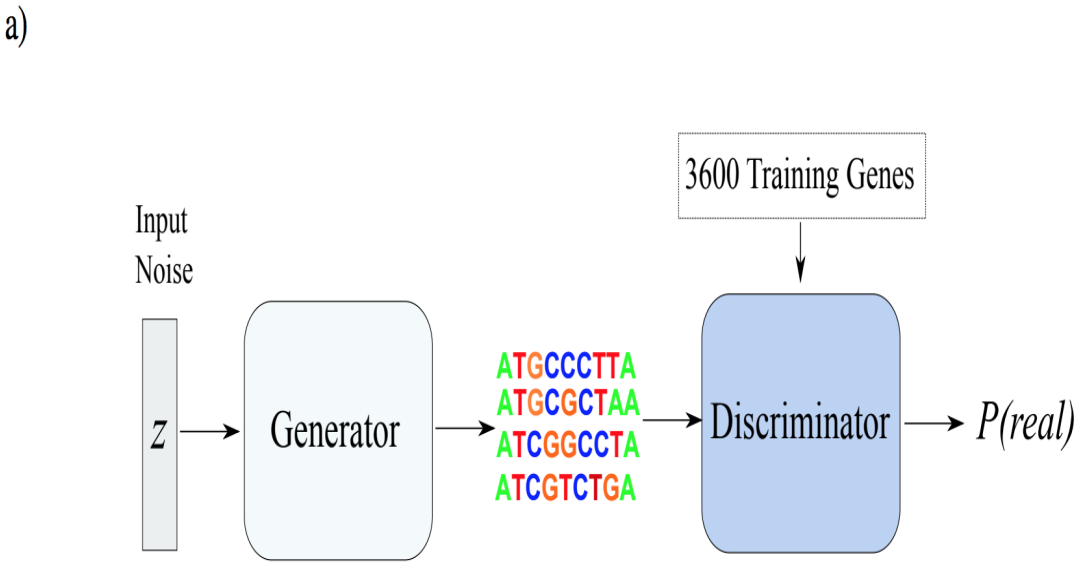


Figure: the GAN model

- Train GAN to produce valid sequences for a few epochs

The Analyzer

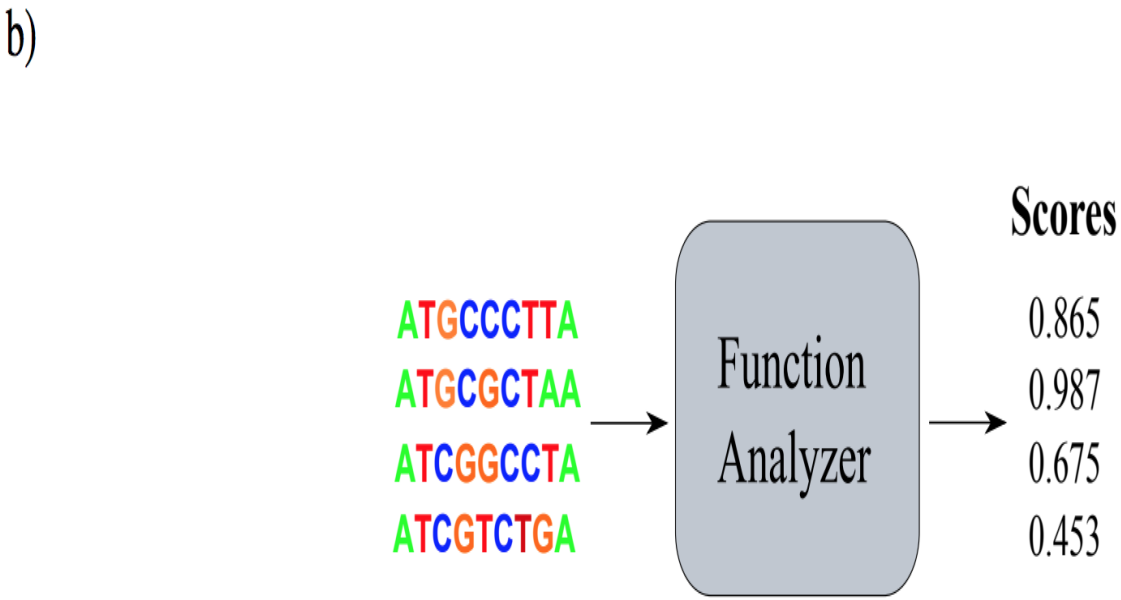


Figure: Function Analyzer

- Analyzer to select sequences that are desirable properties
- Pretrain Analyzer

The model

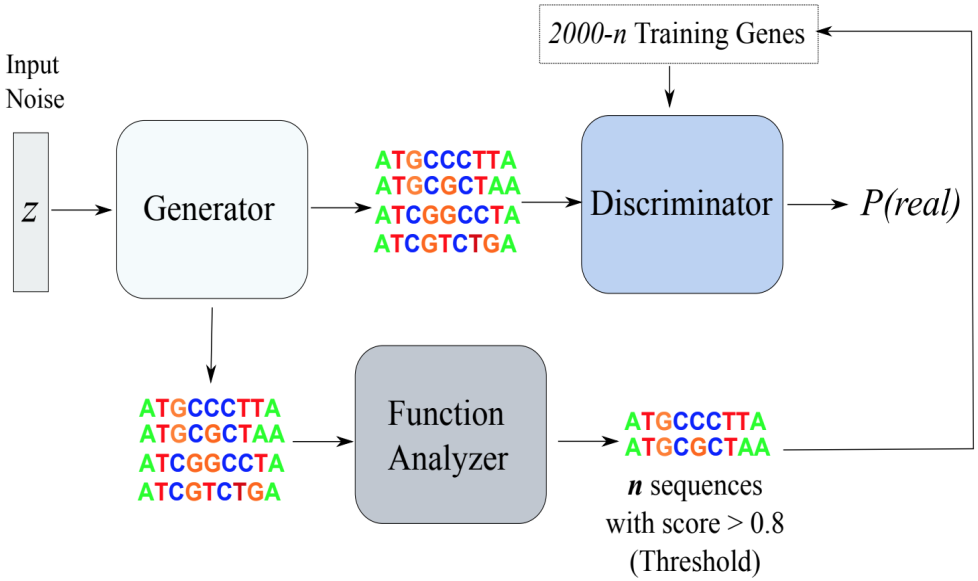


Figure: The model

Use feedback mechanism to select sequences that are desirable properties

Analyzer training

- 2600 experimentally verified antimicrobial properties from APD3 Database
- negative set from Uniprot
- Translate to cDNA and train Analyzer
- can be potentially non differentiable

Analyzer for Antimicrobial Peptide Coding genes

- Classifier
- Input: Gene Sequences Output: 1/0 codes for AMP or not
- Positive Set of 2600 AMPs from APD3 Database
- Negative set of 2600 random peptides from Uniprot
- translated to cDNA

Analyzer: Architecture

- RNN architecture: 2 GRU layers of 128 size
- Last time step to dense layer
- sigmoid activation function: whether gene belongs to positive class or not

Analyzer: Check secondary structure

- Wrapper around PSIPRED secondary structure predictor
- PSIPRED: predict secondary structure of each aminoacid
- Wrapper: gene sequence to protein sequence to PSIPRED
- predicts structure of amino acids inside the protein: total number of alpha helix tagged residues: choose above a certain cutoff
- If gene to protein not possible: output 0

Results: Generate protein coding genes

- Train GAN to produce ≤ 156 nucleotides
- Correct gene if start codon, some codons, stop codins
- Before training, 3.125% sequences follow the correct gene structure
- After training, 77.08% sequences follow correct structure

Results: Generate protein coding genes

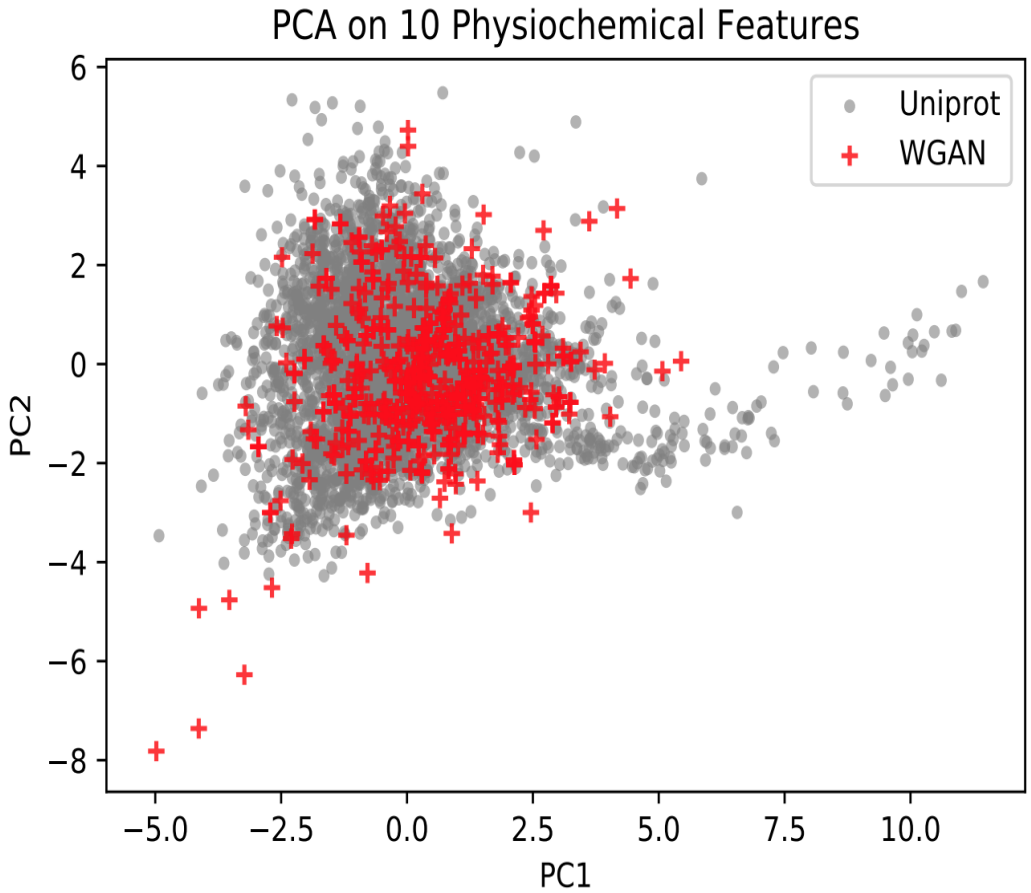


Figure: A set of 500 valid genes were sampled from the trained WGAN, and 10 physiochemical features were calculated for the proteins encoded by the synthetic genes. The same 10 features were also calculated for the cDNA sequences from Uniprot proteins. PCA was performed on the features of the natural cDNA

Feedback analyzer: Results

- training accuracy $= 0.9447$ and validation accuracy $= 0.8613$
- test accuracy $= 0.842$

Results: Feedback-Loop to Optimize Antimicrobial Properties

- After GAN and AMP analyzer are trained, link with feedback loop
- analyzer selects sequences with $P(AMP) > 0.8$ and feed into discriminator as real sequences
- Replace n oldest with selected n newest

Criteria to evaluate selected genes

- does the Analyzer predict more sequences antimicrobial over time?
- are the generated genes similar to real AMP wrt properties and sequences of proteins?

- After 60 epochs, nearly all predicted as anitmicrobial
- 93.3% of the generated sequences after closed loop training have correct gene structure

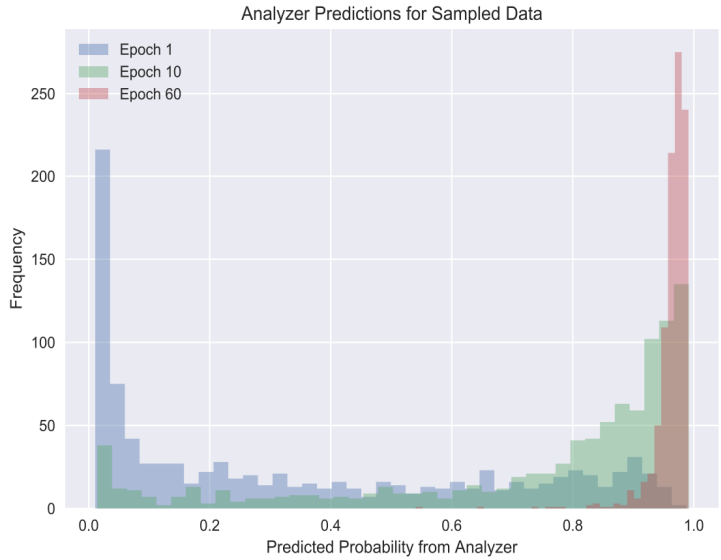
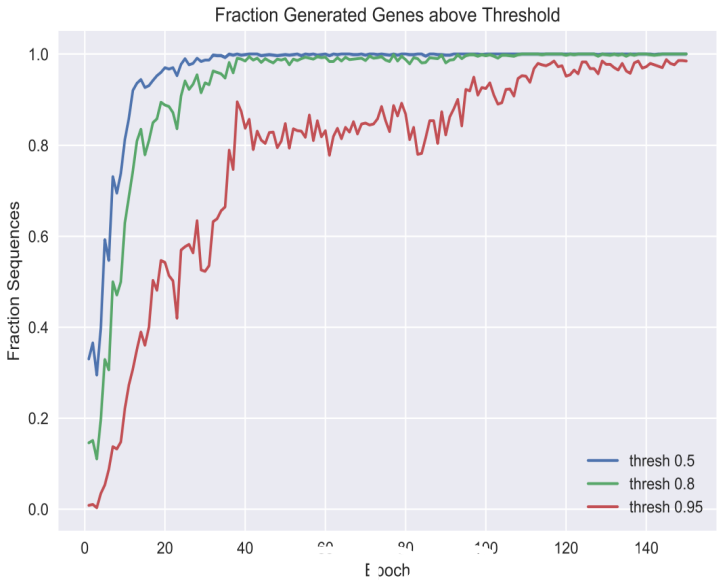

Analyzer Predictions for Sampled Data
Epoch 1
Epoch 10
Epoch 60
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability from Analyzer

Figure:


Fraction Generated Genes above Threshold
Fraction Sequences
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
120
140
thresh 0.5
thresh 0.8
thresh 0.95

Results: Similarity of generated vs exprimental sequences

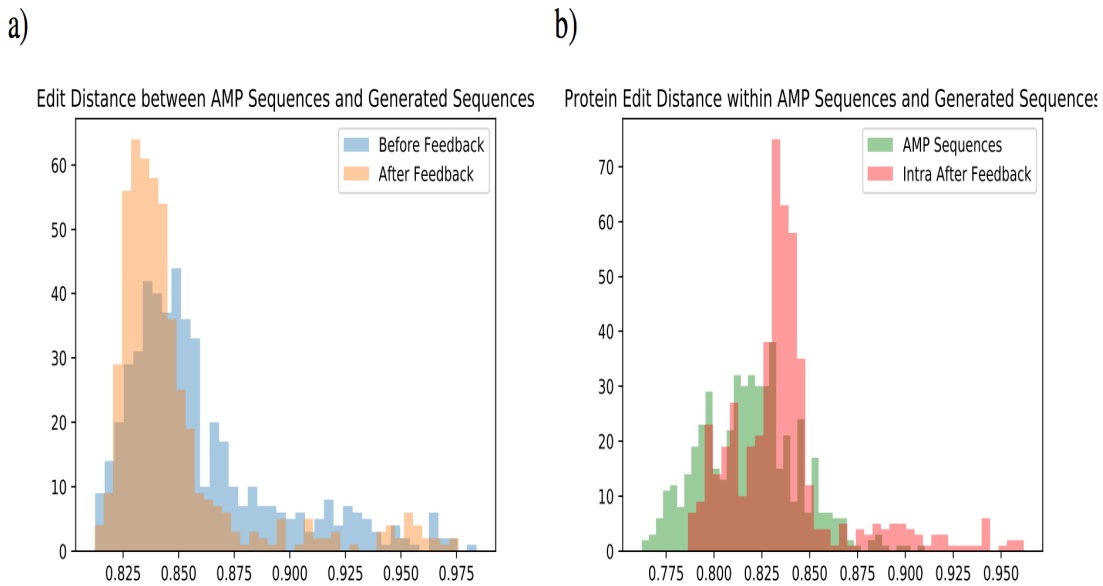


Figure: Edit distance results

- a larger proportion of sequences with a lower edit distance from the AMP sequences
- the sequences after feedback have a higher edit distance within themselves than the antimicrobial sequences do with each other

Results: physiochemical properties

- the proteins encoded by the closed-loop sequences shift to be closer to the positive antimicrobial peptides in five out of ten physiochemical properties such as Length, Hydrophobicity, and Aromaticity, and remains as similar as the sequences without feedback for properties such as Charge and Aliphatic index.
- This is true even though the analyzer operated directly on the gene sequence rather than these physiochemical properties
- the feedback mechanism did not directly optimize the physiochemical properties that show a shift.

	Positive AMP	Before Feedback	After Feedback
Length	32.37 ± 17.983	21.419 ± 13.190	36.992 ± 16.978
Molar Weight	3514.0068 ± 1980.59	2419.032 ± 1479.013	4023.584 ± 1848.048
Charge	3.8575 ± 2.979	2.356 ± 2.447	2.708 ± 2.249
Charge Density	0.00123 ± 0.00084	0.00127 ± 0.00138	0.00091 ± 0.00096
pI	10.2697 ± 2.046	10.143 ± 2.444	9.474 ± 1.844
Instability Index	27.174 ± 26.717	37.791 ± 35.697	53.145 ± 29.495
Aromaticity	0.0822 ± 0.0602	0.0642 ± 0.0695	0.0775 ± 0.066
Aliphatic Index	91.859 ± 47.236	84.397 ± 45.681	84.889 ± 34.837
Boman Index	0.770 ± 1.500	1.801 ± 1.721	0.888 ± 1.155
Hydrophobicity Ratio	0.435 ± 0.128	0.390 ± 0.144	0.441 ± 0.109

Results: Secondary Structure with Black-Box PSIPRED Analyzer

- Use Secondary Structure Anayzer
- secondary structure more attractive to optimize for since it arises in short peptides of length less than 50
- gene sequences with more than 5 alpha-helical residues were input back into the discriminator as real data.
- After 43 epochs of feedback, the helix length in the generated sequences was significantly higher than the helix length without feedback and the helix length of the original Uniprot proteins,

Results: Secondary Structure with Black-Box PSIPRED Analyzer

- helix length was greater with feedback than without feedback
-

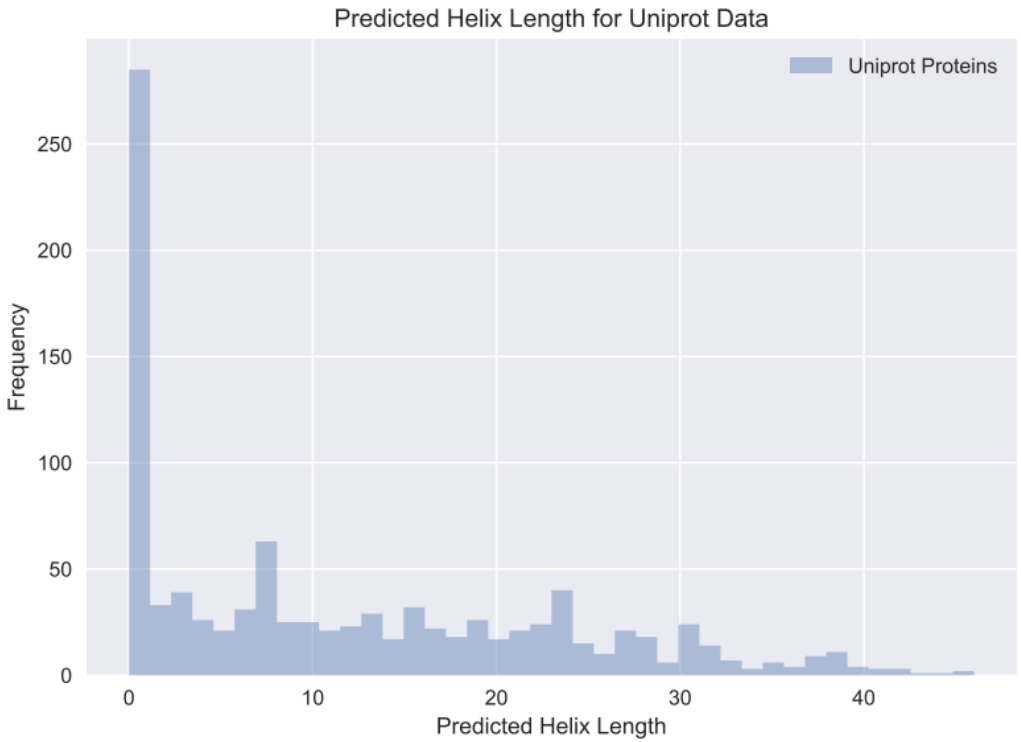


Figure: Before

Results: Secondary Structure with Black-Box PSIPRED Analyzer

- helix length was greater with feedback than without feedback
-

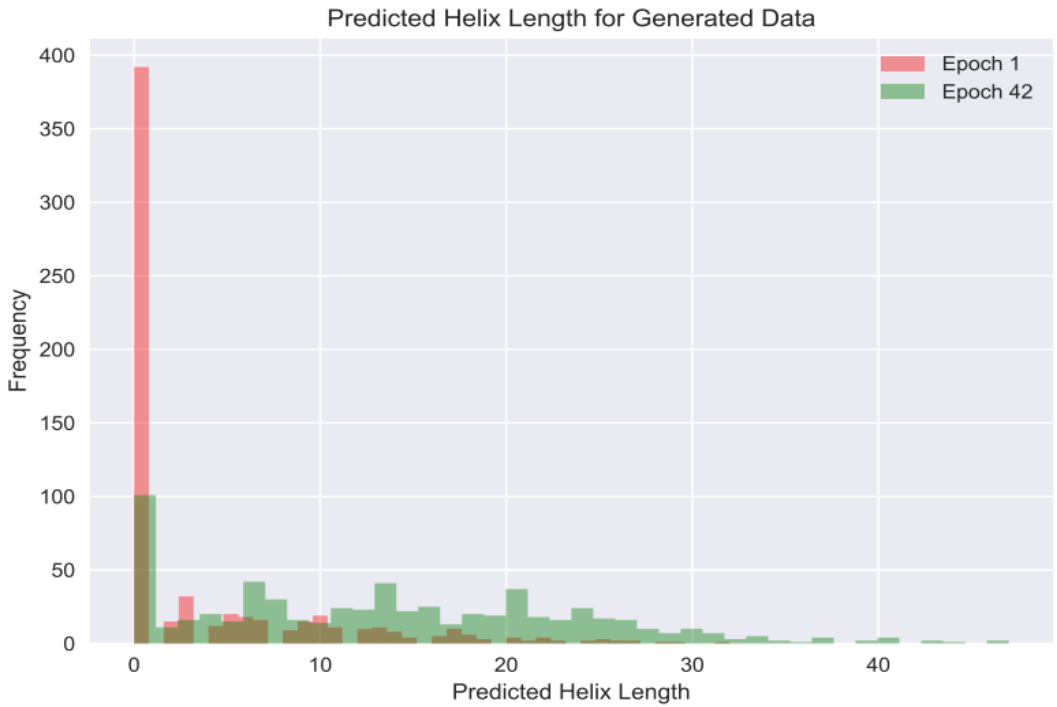


Figure: After FBGAN

Results: Secondary Structure

- these 3D peptide structures were produced by ab initio folding from our generated gene sequences, using knowledge-based force field template-free folding from the QUARK server

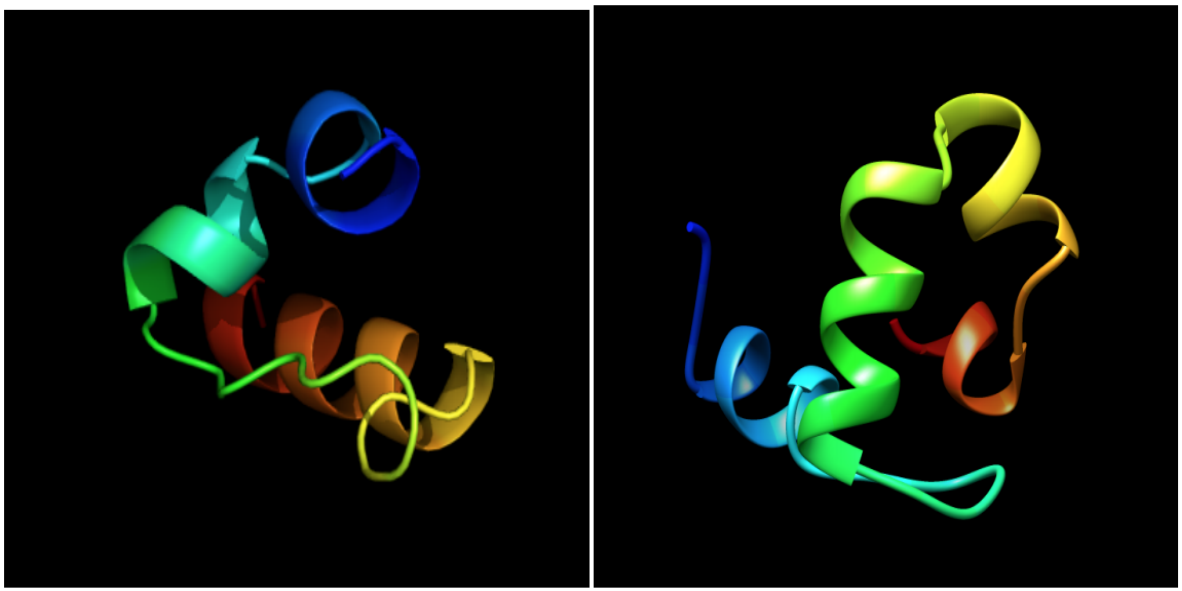


Figure 5: Example peptides from the synthetic genes output by our WGAN model with feedback from the PSIPRED analyzer. Both proteins show a clear helix structure. The peptide on the left was predicted to have 10 residues arranged in helices, while the peptide on the right was predicted to have 22 resides in helices; accordingly, the peptide on the right appears to have more residues arranged in helices.

Figure: 3d secondary structure