Exploring The Predictive Capabilities of AlphaFold Using Adversarial Protein Sequences

Ismail R. Alkhouri, *Member, IEEE*, Sumit Jha, *Senior Member, IEEE*, Andre Beckus, *Member, IEEE*, George Atia, *Senior Member, IEEE*, Susmit Jha, *Senior Member, IEEE*, Rickard Ewetz, *Senior Member, IEEE*, and Alvaro Velasquez, *Member, IEEE*

***Abstract*—Protein folding neural networks (PFNNs) such as AlphaFold predict remarkably accurate structures of proteins compared to other approaches. However, the robustness of such networks has heretofore not been fully explored. This is particularly relevant given the broad social implications of such technologies and the fact that biologically small perturbations to non-critical residues of a protein sequence do not typically lead to drastic changes in the protein structure. Our study demonstrates that, similar to adversarial methods in machine learning, small changes to protein sequences can result in significant differences in the predicted protein structures using AlphaFold as determined by large distance measures. Despite this, our findings using multiple protein sequences suggest that AlphaFold is able to accurately predict the domain structure and folding regions of a protein. To gauge structural differences, we employ two alignment-based measures (root-mean-square deviation (RMSD) and the Global Distance Test (GDT) similarity), and one alignment-free measure, which is an effective Graph-based Structure Representation (GraSR) method. We prove that the problem of minimally perturbing protein sequences is NP-complete. Based on the well-established BLOSUM62 sequence alignment scoring matrix, we generate adversarial sequences. In our experimental evaluation, we consider 111 proteins (including 29 COVID-19 sequences) in the Universal Protein resource (UniProt), a central resource for protein data. Our findings suggest that, despite the high RMSD values returned by AlphaFold, it is capable of handling the BLOSUM adversarial sequences considered in our analysis, as evidenced by the preservation of the folded regions and the GraSR results.**

***Impact Statement*—The ability to obtain 3D structures of proteins is crucial for advancing our understanding of their functionalities, and Alphafold, a machine learning-based system, has demonstrated remarkable success in predicting these structures. However, the adoption of advanced machine learning models and artificial intelligence systems like protein folding neural networks (PFNNs) poses potential security and safety threats. Our investigation of the impact of adversarial protein sequences on the predictions made by PFNNs, including Alphafold, will inform the development of safer and more secure protein folding technologies, advancing our understanding of protein functionalities and contributing to the ongoing exploration of the potential of machine learning-based systems.**

***Index Terms*—Protein Folding Neural Networks, AlphaFold, BLOSSUM62 Distance, Adversarial Protein Sequences, Neural Networks Robustness**

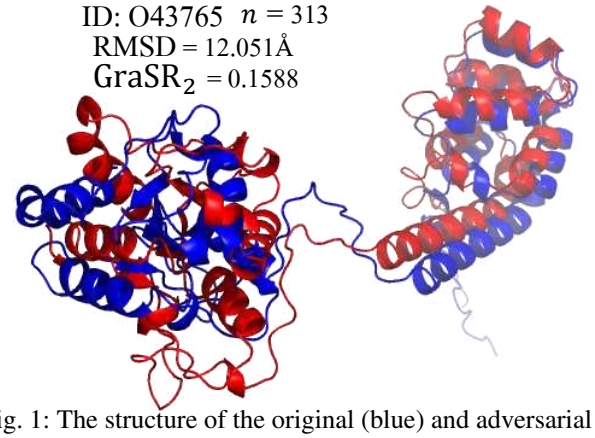

Fig. 1: The structure of the original (blue) and adversarial (red) sequences predicted using AlphaFold for the Small glutamine-rich tetratricopeptide repeat-containing protein alpha sequence. The length of the protein sequence is denoted by n. For structures, the Root Mean Square Deviation (RMSD) is given in Angstroms (equal to 10^{-10} meters and denoted by Å) after their alignment using PyMol [1], and GraSR_2 is the l_2 distance of the Graph Structure Representation (GraSR) structural vector descriptors.

I. Introduction

Proteins form the building blocks of life as they enable a variety of vital functions essential to life and reproduction. Naturally occurring proteins are bio-polymers typically composed of 20 amino acids and this primary sequence of amino acids is well known for many proteins, thanks to high-throughput sequencing techniques. However, in order to understand the functions of different protein molecules and complexes, it is essential to comprehend their three-dimensional (3D) structures. Until recently, one of the grand challenges in structural biology has been the accurate determination of the 3D structure of the protein from its primary sequence. Such accurate predictive protein folding promises to have a profound impact on the design of therapeutics for diseases and drug discovery [2].

Ismail R. Alkhouri is with the Computational Mathematics, Science, and Engineering Department at Michigan State University, and the Electrical Engineering and Computer Science Department at the University of Michigan Ann Arbor. The work was done while at the University of Central Florida (UCF). E-mails: alkhour3@msu.edu,ismailal@umich.edu.

Sumit Jha is with the Knights Foundation School of Computing and Information Sciences, Florida International University. Email: jha@cs.fiu.edu.

Andre Beckus is with the Information Directorate, Air Force Research Laboratory. Email: andre.beckus@us.af.mil.

George Atia is with the Electrical and Computer Engineering (ECE) Department, and the CS Department, UCF. Email: george.atia@ucf.edu.

Susmit Jha is with the CS Laboratory, SRI International. Email: susmit.jha@sri.com.

Rickard Ewetz is with the ECE Department, UCF. Email: rickard.ewetz@ucf.edu.

Alvaro Velasquez is with the CS Department, University of Colorado, Boulder. Email: alvaro.velasquez@colorado.edu.

AlphaFold [3] achieved unparalleled success in predicting protein structures using neural networks and remains first at the Critical Assessment of protein Structure Prediction (CASP14), which corresponds to year 2020, competition. While AlphaFold has been celebrated as a major advancement in structural biology [4], its ability to predict the structure of adversarially perturbed sequences has yet to be fully examined.

The main contribution of this paper is to investigate the impact of adversarial protein sequences on AlphaFold's performance. First, we present the problem of adversarial attacks on Protein Folding Neural Network (PFNN) and prove it is NP-complete. To identify a space of similar protein sequences used in constructing adversarial perturbations, we use sequence alignment scores [5], such as those derived from Block Substitution Matrices (BLOSUM62). For the output structures, we leverage standard metrics commonly used in CASP, including the root-mean-square deviation (RMSD) and the Global Distance Test (GDT) similarity measure between the predicted structure and the structure of its adversarially perturbed sequence. Second, we generate examples where slight variations in protein sequences result in significantly different 3D protein structures, as measured by large distance metrics. To supplement our analysis, we utilize an alignment-free method, Graph-based protein Structure Representation learning (GraSR) [6], which indicates that AlphaFold preserves the underlying domain and folding structures of the protein despite the observed differences in structure when using distance metrics. However, we do not make any claims about AlphaFold's susceptibility to adversarial sequences, as further research is needed to draw definitive conclusions. Our study provides insights into the marked differences in 3D protein structures resulting from small sequence variations while preserving domain and folding structures. These insights can help guide further investigations in this area. See Figure 1 and its caption for an example.

Moreover, we conduct two experiments investigating the choice of the BLOSUM threshold and the use of the prediction, per-residue, confidence information obtained from AlphaFold. Our experiments show that different input protein sequences have very different adversarial robustness as determined by the RMSD (GDT-TS) in the protein structure predicted by AlphaFold. These values range from 1.011Å (0.43%) to 49.531Å (98.8%) when the BLOSUM62 distance between the original and adversarial sequences is bounded by a threshold of 20 units with a hamming distance of 5 residues only.

II. Summary and Related Work

Nearly four decades ago, it was observed that two protein structures with 50% sequence identity align with an RMSD of around 1 Å from each other [7]. Additionally, even proteins with 40% sequence identity and at least 35 aligned residues align within an RMSD of approximately 2.5 Å [8]. This raises the question: Should highly accurate PFNNs [9], [10] be able to predict similar structures when only a few residues in the input sequence are changed? The phenomenon of sequence-similar proteins producing similar structures have been observed in larger studies [11]. As with almost any rule in biology, a small number of counterexamples to the conventional wisdom of similar sequences leading to similar structures do exist, wherein even small perturbations can potentially alter the entire fold of a protein. However, such exceptions are not frequent and often lead to exciting investigations [12], [13].

Manipulating the multiple sequence alignment step of AlphaFold has been studied in [14] using in silico mutagenesis. However, there, the goal is not to study the robustness of the protein folding neural networks, but rather to enhance the prediction capability of AlphaFold in terms of the intrinsic conformational heterogeneity of proteins. The authors in [15], present a method that manipulates inputs to obtain diverse distinct structures that are absent from the AlphaFold training data. Using membrane proteins, the authors show that their method enhances the multiple sequence alignment step while generating more accurate structures.

In general, it has been demonstrated that AlphaFold predictions are not stable and should not be trusted with mutated sequences (not wild-type sequences) [16], [17]. However, evaluating AlphaFold capabilities in handling BLOSUM-based adversarial sequences, such as the ones in this paper, are yet to be explored.

The work in [18] is aimed at generating adversarial sequences in order to cause significant damage to the output predicted structure of RosettaFold [10], which, according to CASP, is the second best protein folding neural network. However, the authors only show results for a few proteins and do not consider all the standard metrics for measuring the output structures. In contrast, in this paper, we present results for more than 100 sequences, derive a complexity proof for the problem of finding adversarial protein sequences, and, based on the CASP competition, utilize all the standard metrics for measuring the output structures.

A. Robustness Metric using Adversarial Attacks

Given a protein sequence of n residues, denoted as $S = s_1 s_2 \dots s_n$, with an associated three-dimensional structure $\mathcal{A}(S) = (x_1, y_1, z_1), \dots, (x_n, y_n, z_n)$, we define a set of biologically similar sequences, denoted as $\mathcal{V}$, using the Block Substitution Matrices (BLOSUM) [5]. Then, we utilize adversarial attack techniques [19] on PFNNs within this space of similar sequences to identify a sequence $S_{\text{adv}} \in \mathcal{V}$ that produces a maximally different three-dimensional structure $\mathcal{A}(S_{\text{adv}})$. We then compute the RMSD, GDT, and GraSR between the structures for the original and adversarial inputs ($\mathcal{A}(S)$ and $\mathcal{A}(S_{\text{adv}})$).

B. BLOSUM Similarity Measures

Given two sequences of n residues $S = s_1 s_2 \dots s_n$ and $S' = s'_1 s'_2 \dots s'_n$, in which every residue s_i (or s'_i) is from the set $\mathcal{X} = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ of amino acids, a natural question is how to compute the sequence similarity D_{seq} between these proteins. A naive approach would be to count the number of residues that are

different, i.e., the Hamming distance. However, an analysis of naturally occurring proteins shows that not all changes in residues have the same impact on protein structures. Changes to one type of residue are more likely to cause structural variations than changes to another type.

Early work in bioinformatics focused on properties of amino acids and reliance on genetic codes. However, more modern methods have relied on the creation of amino acid scoring matrices that are derived from empirical observations of frequencies of amino acid replacements in homologous sequences [20], [21]. The original scoring matrix, called the PAM250 matrix, was based on empirical analysis of 1572 mutations observed in 71 families of closely-related proteins that are 85% or more identical after they have been aligned. The PAM1 model-based scoring matrix was obtained by normalizing the frequency of mutations to achieve a 99% identity between homologous proteins. These results were then extrapolated to create the PAM10, PAM30, PAM70 and PAM120 matrices with 90%, 75%, 55%, and 37% identity between homologous proteins.

Another interesting approach [5] to understanding protein similarity is the direct counting of replacement frequencies using the so-called Block Substitution Matrices (BLOSUM). Instead of relying solely on sequences of homologous proteins that are relatively harder to find, the BLOSUM approach focuses on identifying conserved blocks or conserved sub-sequences in a larger variety of proteins potentially unrelated by evolutionary pathways and counts the frequency of replacements within these conserved sub-sequences. BLOSUM62 (Figure 2), BLOSUM80 and BLOSUM90 denote block substitution matrices that are obtained from blocks or subsequences with at least 62%, 80%, and 90% similarity, respectively. The BLOSUM matrix $[B_{ij}]$ is a matrix of integers where each entry denotes the similarity between residue of type $b_i \in \mathcal{X}$ and type $b_j \in \mathcal{X}$. See Figure 2.

We identify the space of biologically similar sequences $\mathcal{V}$ for a given protein sequence S with respect to the BLOSUM distance. Based on the BLOSUM distance, we examine the predicted structures for similar sequences. In particular, we verify if the structural measures between the predicted structure $\mathcal{A}(S)$ and the structure of the adversarial sequence $\mathcal{A}(S_{\text{adv}})$ are large or small. We adopt a sequence similarity measure that counts replacement frequencies in conserved blocks across different proteins.

III. Approach

Our approach to evaluating the robustness of PFNNs as a machine learning model to minimally perturbed inputs is based on two main ideas: (i) the existence of adversarial examples in PFNNs that produce adversarial structures that are possibly at a large distance from the original structure, and (ii) the use of BLOSUM for identifying a neighborhood of a given sequence comprising biologically similar sequences, and hence expected to yield similar 3D structures. We utilize the RMSD, GDT, and GraSR between the structure of an original protein sequence and the structure of the adversarial sequence as a measure of robustness of a protein folding network on the given input. In this work, we focus on the state-of-the-art AlphaFold model, the winner of the 1st place in CASP2020.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Fig. 2: The BLOSUM62 matrix.

A. Sequence Similarity Measures

Given two sequences $S = s_1 s_2 \dots s_n$ and $S' = s'_1 s'_2 \dots s'_n$, the BLOSUM distance between the two sequences is given by Equation (1) below.

$$D_{\text{seq}}(S, S') = \sum_{i \in [n]} \left(B_{s_i s_i} - B_{s_i s'_i} \right) . \tag{1}$$

For an illustrative example of D_{seq}, see Figure 3.

B. Output Structural Measure

Given a sequence of n residues $S = s_1 s_2 \dots s_n$, its three dimensional structure $\mathcal{A}(S)$ is an ordered n-tuple of three-dimensional co-ordinates $(x_1, y_1, z_1), \dots (x_n, y_n, z_n)$. Our goal is to utilize a structural distance measure that captures the variations in the two structures $\mathcal{A}(S)$ and $\mathcal{A}(S')$ and is invariant to rigid-body motion. Therefore, in this work, we use standard structural distances, namely the RMSD, measured in Å, and the GDT with its two variants: (i) the Total Score (TS) and (ii) the High Accuracy (HA) [22].

Given the output structure of the adversarial sequence $\mathcal{A}(S')$, an alignment algorithm is employed before computing the RMSD and GDT measures between the two structures of interest. We use the alignment procedure implemented in PyMOL [1] to align $\mathcal{A}(S')$ with regard to the target structure $\mathcal{A}(S)$. Let the aligned structure be denoted by $\hat{\mathcal{A}}(S') = (\hat{x}'_1, \hat{y}'_1, \hat{z}'_1), \dots, (\hat{x}'_n, \hat{y}'_n, \hat{z}'_n)$. Then, the RMSD, measured in Å, is obtained as

$$\text{RMSD}(\mathcal{A}(S), \hat{\mathcal{A}}(S')) = \sqrt{\frac{1}{n} \sum_{i \in [n]} d(\mathcal{A}(S)_i, \hat{\mathcal{A}}(S')_i)} , \tag{2}$$

where $d(\mathcal{A}(S)_i, \hat{\mathcal{A}}(S')_i) = (x_i - \hat{x}'_i)^2 + (y_i - \hat{y}'_i)^2 + (z_i - \hat{z}'_i)^2$ and $\mathcal{A}(S)_i$ represents the 3D carbon-alpha coordinates of the i^{th} residue. Using the carbon-alpha coordinates is the standard approach in CASP [22].

Another standard metric for gauging the similarity of protein structures is the GDT similarity measure, introduced by [22]

Original and Adversarial Sequences

		D_{seq}	D_{ham}
S	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG		
S'_1	DVPSMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	32	4
S'_2	FGCYMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	20	4
S'_3	MDLFMRFFTLGSITAQPIRVPNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	12	4
S'_4	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLWAYKLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	42	4
S'_5	MDLFMRFFVIAAVTAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	17	5
S'_6	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASDIERRLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	49	5
S'_7	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATRVLTMKKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	18	6
S'_8	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQAVEFQLVLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	57	6
S'_9	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASLPFGWLVIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLACMYISMYSHLLLVAAG	23	7
S'_{10}	MDLFMRFFTLGSITAQPVKIDNASPASTVHATATIPLQASDIGIINGIGVAFLAVFQSATKIIALNKRWQLALYKGFQFICNLLLLFVTIYSHLLLVAAG	65	7

Fig. 3: The original sequence S is followed by 10 sequences generated by changing 4, 5, 6, and 7 residues. While the BLOSUM distance may not specifically focus on protein functionality, when comparing two protein sequences, the BLOSUM distance offers a metric to compute a biologically relevant distance in contrast to the Hamming distance.

and commonly used in the CASP competition along with the RMSD. In some cases, the latter is known to be sensitive to outliers [22]. The GDT score returns a value in $[0, 1]$ where 1 indicates identical structures, and is computed with respect to four thresholds, δ_j, as

$$\text{GDT}(\mathcal{A}(S), \hat{\mathcal{A}}(S')) = \frac{1}{4n} \sum_{j\in[4]} \sum_{i\in[n]} \mathbf{1}\big(d(\mathcal{A}(S)_i, \hat{\mathcal{A}}(S')_i) < \delta_j\big) \,, \tag{3}$$

where the thresholds δ_1, δ_2, δ_3, and δ_4 for TS (HA) are given by $1(0.5)$, $2(1)$, $4(2)$, and $8(4)$ for j equals to 1, 2, 3, and 4, respectively, and $\mathbf{1}(\cdot)$ is the indicator function. In (3), each $j \in [4]$ reflects the number of residues in the structures for which the distance is less than δ_j.

In addition to distance metrics, we use an alignment-free method called Graph-based protein Structure Representation learning (GraSR) [6]. GraSR is a learning-based approach that aims at circumventing the challenges associated with sequence segmentation and feature engineering, which is achieved by leveraging deep neural networks to automatically learn structural representations. GraSR combines graph NNs (GNNs) and long short-term memory (LSTM) units, where protein structures are initially represented as graphs using intra-residue distances. Subsequently, an encoder is optimized through a contrastive learning framework. The concept behind this is that GNNs have a greater capacity to learn both global and local geometric features of residues. See Figure 1 in [6] for an example. In short, GraSR is a trained NN that produces a 400-dimensional vector representation (descriptor) of a protein structure $g(\mathcal{A}(S)) \in \mathbb{R}^{400}$, incorporating domain rotations. This means that the features from GraSR are obtained from a forward pass of a pre-trained GNN.

To measure the distance between the original and adversarial structural representations, we use the l_p-norm with $p \in \{2, \infty\}$, defining GraSR_p as:

$$\text{GraSR}_p(\mathcal{A}(S), \mathcal{A}(S')) = \|g(\mathcal{A}(S)) - g(\mathcal{A}(S'))\|_p \,, \tag{4}$$

where $g(\mathcal{A}(S))$ is the GraSR descriptor vector for structure $\mathcal{A}(S)$, and $\|\cdot\|_p$ is the l_p-norm.

C. Adversarial Attacks on PFNNs

Small carefully crafted changes in a few pixels of input images cause well-trained neural networks with otherwise high accuracy to consistently produce incorrect responses in domains such as computer vision [23], [24], [25], [26]. Given a neural network $\mathcal{A}$ mapping a sequence S of residues to a three-dimensional geometry $\mathcal{A}(S)$ describing the structure of the protein, we seek to obtain a sequence S' such that the sequence similarity measure $D_{\text{seq}}(S, S')$ between S and S' is small and some structural distance measure $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S'))$ is maximized. This can be achieved by solving the following optimization problem

$$\max_{S'} D_{\text{str}}\left(\mathcal{A}(S), \mathcal{A}(S')\right) \text{ s.t. } D_{\text{seq}}(S, S') \leq L \,. \tag{5}$$

In our experiments, we set $L = 20$ and D_{str} as the RMSD measure. Given the discrete nature of the input sequences, well-known methods for generating adversarial examples (e.g. gradient-based methods) fail to produce valid and accurate results. As such, we propose a solution based on a brute-force exploration in the space of biologically similar sequences that, given a sequence of interest S with n residues, can be defined as

$$\mathcal{V}_{L,H}(S) = \{S' \in \mathcal{X}^n \mid D_{\text{seq}}(S, S') \leq L \text{ and } D_{\text{ham}}(S, S') \leq H\} \,, \tag{6}$$

where $\mathcal{X}^n$ is the set of all possible sequences over $\mathcal{X}$ of length n, D_{ham} is the hamming distance, and H is a pre-defined threshold. For long sequences, the search space can be extensively large. Therefore, we select random samples from $\mathcal{V}_{L,H}(S)$ and choose the sequence that returns the maximum value based on the RMSD measure. Our approach to generating adversarial sequences falls under the class of black-box attacks. This means that we only have access to the output of the network [27].

It is worth noting that the inference time of complex protein folding systems, which apply multiple processing and alignment steps prior to the use of any neural network, such as AlphaFold is extremely high compared to NN-based image classifiers. The forward pass of such systems involves a large number of computations. This fact, along with the discrete nature of the input space, are the bottleneck of developing more complex black-box attacks [28], which in general require

a high number of queries. Given the high computational cost of generating the structure of a single sequence in AlphaFold and the discrete nature of the input sequence, we believe that these factors contribute to the absence of other baseline attack methods for comparison in our experimental results.

IV. Complexity

In this section, we formalize the problem of generating an adversarial attack for PFNNs and establish its complexity.

Definition 1 (PFNN Adversarial Attack (PAA) Problem)**.** *Given a learning model $\mathcal{A}(.;\theta) : \mathcal{X}^n \to (\mathbb{R}\times\mathbb{R}\times\mathbb{R})^n$ mapping residues to 3-dimensional coordinates and parameterized by θ, a sequence $S \in \mathcal{X}^n$, and a sequence alignment scoring matrix B, find an input sequence $S' \in \mathcal{X}^n$ such that $D_{\text{seq}}(S, S') \leq L$ and $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) \geq U$, where the bounds L and U and distance functions d and D are given.*

We prove that the PAA problem is **NP-complete**. This establishes that, in general, there is no polynomial-time solution to the PAA problem unless **P = NP**. Due to this complexity and for ease of presentation, we adopt simple perturbation attacks for our experiments in the next section. We begin by defining the **NP-complete** problem to be reduced to an instance of the PAA problem.

Definition 2 (CLIQUE Problem)**.** *Given an undirected graph $G = (V, E)$ and an integer k, find a fully connected sub-graph induced by $V' \subseteq V$ such that $|V'| = k$.*

Theorem 1. *The PFNN Adversarial Attack (PAA) problem in Definition 1 is **NP-complete**.*

Proof: It is easy to verify that the PAA problem is in **NP** since, given a solution sequence S', one can check whether the constraints $D_{\text{seq}}(S, S') \leq L$ and $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) \geq U$ are satisfied in polynomial time. It remains to be shown whether the PAA problem is **NP-hard**. We establish this result via a reduction from the CLIQUE problem in Definition 2. Given a CLIQUE instance $\langle G = (V, E), k\rangle$ with $|V| = n$ and $|E| = m$, we construct its corresponding PAA instance $\langle \mathcal{A}(.\,;\theta), S, B, L, U\rangle$ as follows. Without loss of generality, let us consider a restricted version of the PAA problem where there are only two residue types $\{N, K\}$ with the corresponding BLOSUM62 sub-matrix $B' = 6 \cdot I$, where I denotes the identity matrix. Following the one-hot representation of residues adopted in [9], any input tensor over $\{N, K\}$ is represented as a one-hot encoding $S^{\text{in}} \in (\mathbb{B} \times \mathbb{B})^n$ to be used as an input tensor to $\mathcal{A}$, where $s^{\text{in}}_{i0} = 1$ ($s^{\text{in}}_{i1} = 1$) denotes that residue s^{in}_i is of type N (K). Let $S = (N, N, \ldots, N)$ denote the all-N sequence. We set $L = 6k$ and $U = \frac{k(k-1)}{2}\sqrt{\frac{3}{n}}$. The connectivity structure of $\mathcal{A}$ is derived from the edges E in the CLIQUE instance as follows. The first column of the input tensor corresponding to s^{in}_{i0} for all $i \leq n$ is disconnected from the network and the second column corresponding to s^{in}_{i1} is connected to $\mathcal{A}$ such that, for each edge $(v_i, v_j) \in E$, we have a connection from s^{in}_{i1} and s^{in}_{j1} to each of the three outputs in the first three-dimensional coordinate of $\mathcal{A}(S^{\text{in}})_1$. All connections have a weight of unity and this defines the parameters θ of the model $\mathcal{A}$. Therefore, without loss of generality, we are only considering the first of the n output three-dimensional coordinates $\mathcal{A}(S^{\text{in}})_1$. In particular, these values keep track of the number of edges induced by the vertices in G corresponding to the non-zero entries in $s^{\text{in}}_{11}, \ldots, s^{\text{in}}_{1n}$. We now prove that there is a clique of size k in G if and only if there is a feasible solution $S^{\text{in}} = S'$ to the reduced PAA instance.

($\implies$) Assume there is a clique of size k in G. We can derive a feasible solution S' to the reduced PAA instance as follows. For every vertex $v_i \in V$ (not) in the clique, let ($s'_{i0} = 1$) $s'_{i1} = 1$. Since S is the all-N sequence, its corresponding one-hot encoding consists of $s_{i0} = 1$ for all $1 \leq i \leq n$. Thus, the corresponding BLOSUM62 distance is

$$D_{\text{seq}}(S, S') = \sum_{1\leq i\leq n} (6 - 6 \cdot \mathbf{1}(s_i \neq s'_i)) = 6k\,. \tag{7}$$

This satisfies the sequence alignment constraint defined by $D_{\text{seq}}(S, S') \leq L = 6k$. Furthermore, the solution S' induces outputs of $x'_1 = y'_1 = z'_1 = k(k-1)/2$, leading to an RMSD of U. Without loss of generality, we omit the alignment step in computing the RMSD and therefore assume that $\mathcal{A}(S') = \hat{\mathcal{A}}(S')$. The corresponding RMSD distance $D_{\text{str}}(\mathcal{A}(S), \hat{\mathcal{A}}(S'))$ in output predictions is presented below. Recall that $x_1 = y_1 = z_1 = 0$ for the the all-N sequence S because its corresponding column in the one-hot encoding is disconnected from the network.

$$\begin{aligned} D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) &= \sqrt{\frac{1}{n}\sum_{i\in[n]} d(\mathcal{A}(S)_i, \hat{\mathcal{A}}(S')_i)} \\ &= \sqrt{\frac{1}{n}\left[3\left(0 - \frac{k(k-1)}{2}\right)^2\right]} = \frac{k(k-1)}{2}\sqrt{\frac{3}{n}}\,. \end{aligned} \tag{8}$$

Thus, the constraint $D_{\text{str}}(S, S') \geq U = \frac{k(k-1)}{2}\sqrt{\frac{3}{n}}$ is satisfied.

($\impliedby$) We prove the contrapositive. That is, if there is no clique of size k in G, then the reduced PAA instance is infeasible. We proceed by showing that there must be exactly k non-zero entries in the column vector $\{s'_{i1} | i \leq n\}$ in order to satisfy constraints $D_{\text{seq}}(S, S') \leq L = 6k$ and $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) \geq U$ and that, if there is no clique of size k, then there is no choice of k non-zero entries in $\{s'_{i1} | i \leq n\}$ that will satisfy these constraints. Let k' denote the number of non-zero entries in $\{s'_{i1} | i \leq n\}$. To satisfy $D_{\text{seq}}(S, S') \leq L = 6k$, it follows that $k' \leq k$. If $k' < k$, then the maximum value of $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S'))$ is $\frac{k'(k'-1)}{2}\sqrt{\frac{3}{n}} < \frac{k(k-1)}{2}\sqrt{\frac{3}{n}}$ and denotes to the case where the k' non-zero entries correspond to a clique of size k' in G. The strict inequality is due to the monotonically increasing nature of this equation. Therefore, it must be that $k = k'$ and we have outputs $x'_1 = y'_1 = z'_1 = k(k-1)/2$ as before. Suppose that the k' non-zero entries in $\{s'_{i1} | i \leq n\}$ do not correspond to a clique in G. Then the values x'_1, y'_1, and z'_1 output by $\mathcal{A}$ and corresponding to the number of edges induced by the chosen non-zero entries would be strictly less than $k(k-1)/2$. Therefore, we would have $D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) < U$. This proves that the reduced PAA is infeasible. ■

TABLE I: RMSD results for the three considered categories in the experiment conducted in Section V.A.

Seq. ID	n	Category	RMSD	μ_{all}	μ_{diff}	μ'_{all}	μ'_{diff}
Q01629	132	MIN.	6.02	64.63	32.44	60.99	36.99
Q01629	132	AVG.	19.92	64.63	64.75	63.77	69.57
Q01629	132	MAX.	19.906	64.63	66.99	90.21	90.19
Q5BJD5	291	MIN.	14.023	82.23	38.86	81.22	37.79
Q5BJD5	291	AVG.	14.232	82.23	82.24	81.17	77.23
Q5BJD5	291	MAX.	13.567	82.23	98.17	82.42	98.1
P59595	422	MIN.	24.74	68.25	29.13	67.57	31.5
P59595	422	AVG.	28.164	68.25	69.44	68.69	69.04
P59595	422	MAX.	24.62	68.25	96.14	67.51	96.44
P59633	154	MIN.	21.67	44.82	27.14	44.15	38.75
P59633	154	AVG.	21.52	44.82	45.1	43.8	42.26
P59633	154	MAX.	23.13	44.82	61.26	46.13	54.84
P0DTC9	419	MIN.	25.593	68.39	28.46	67.9	28.83
P0DTC9	419	AVG.	21.767	68.39	68.37	68.5	70.83
P0DTC9	419	MAX.	23.685	68.39	97.1	68.64	96.94

TABLE II: RMSD results when $L \in \{20, 30, 40\}$ for the experiment in Section V.B.

Seq. ID	n	L	RMSD	μ_{all}	μ_{diff}	μ'_{all}	μ'_{diff}
Q14653	427	20	18.87	79.76	92.92	79.46	86.29
Q14653	427	30	22.42	79.76	93.15	77.45	64.12
Q14653	427	40	28.28	79.76	90.49	79.42	69.026
Q5BJD5	291	20	14.311	82.23	89.77	80.6	80.64
Q5BJD5	291	30	15.708	82.23	59.26	83.13	43.53
Q5BJD5	291	40	17.132	82.23	62.02	83.21	62.83
P59595	422	20	24.321	68.25	91.44	67.05	89.51
P59595	422	30	30.139	68.25	93.142	67.44	89.29
P59595	422	40	30.675	68.25	46.87	66.4	29.33
P0DTC9	419	20	26.51	68.39	80.32	68.09	80.316
P0DTC9	419	30	26.27	68.39	68.05	68.61	65.18
P0DTC9	419	40	31.33	68.39	40.52	67.76	35.56
P07711	333	20	7.09	93.68	92.4	93.2	81.12
P07711	333	30	8.52	93.68	95.91	92.95	92.69
P07711	333	40	9.246	93.68	95.91	92.85	95.76
Q9Y397	364	20	11.184	84.24	97.35	83.85	95.81
Q9Y397	364	30	11.828	84.24	95.91	83.51	85.416
Q9Y397	364	40	14.222	84.24	95.91	83.71	89.79

V. Experimental Results

For our experimental setup, we use the default settings of the latest version of AlphaFold[1]. This includes the initial multi-sequence alignment (MSA) step, the five-model ensembles predictions, recycling, output confidence ranking, and amber relaxation. For further details about each step, we refer the reader to [3] and its supplementary information. We include results from using the high-accuracy full database configuration of the initial AlphaFold MSA step along with the less accurate (and faster) reduced database option. In order to compute the RMSD and GDT, we need to employ an alignment algorithm. In this paper, we use the built-in alignment PyMOL procedure [1]. The parameters of PyMOL alignment are selected using the default settings, which include an outlier rejection cutoff of 2, a maximum number of outlier rejection cycles of 5, and the use of the structural superposition step. We note that these outliers only impact the calculations of the RMSD.

Our adversarial sequences are generated by randomly sampling 20 sequences from the set $\mathcal{V}_{L,H}$ in (6) with $H = 5$ and $L = 20$. Then, we pick the sequence that returns the maximum value in RMSD structural distance. We use an AMD EPYC 7702 64-Core Processor with 1 TiB of RAM and NVIDIA A100 GPU. We generate adversarial sequences against the considered protein sequences from the UniProt database considered by AlphaFold in [29]. The original fasta (file extension for protein sequences) sequence files are available online[2]. Additionally, we generate adversarial sequences against most of the the UniProt (Universal Protein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases [30]). We have made our code available online[3].

A. Confidence Experiment

Given a sequence S, per residue, AlphaFold generates an estimate of its prediction confidence in the form of a value in $[0, 100]$. This value is called the predicted Local Distance Test (pLDDT) and represents the predicted value on the lDDT-Cα metric [31].

In this subsection, we answer the following question. Do the substituted residues based on their low (or high) confidence scores impact the resulting RMSD between the original and adversarial structure prediction? Phrased differently, in terms of the RMSD, we illustrate the impact of using the prediction confidence scores of every residue of the predicted structure of the original sequence in determining the location of the residues to be altered in the adversarial sequence generation method presented in the previous section. As such, five, not cherry picked, randomly selected sequences are used. Then, the locations of the 5 residues to be altered are taken based on three categories as follows. Residues are selected with confidence values near the (*i*) minimum confidence score (MIN. category), (*ii*) the average score (AVG. category), and (*iii*) the maximum confidence score (MAX. category). Results are presented in Table I. Our analysis shows that, in general, the RMSD of the output structure is not affected by the choice of substituting residues with low or high confidence scores. Therefore, in our method, the positions of the substituted residues are chosen without considering the confidence scores.

B. BLOSUM Threshold Experiment

In this subsection, we want to investigate how a change in the bound on biological similarity, w.r.t. the BLOSUM distance, changes the adversarial sequence. In other words, we show the impact of using different BLOSUM thresholds in set $\mathcal{V}_{L,H}$. As such, we randomly select 6 sequences and generate adversarial sequences by configuring the BLOSUM threshold, L, to be 20, 30, and 40 (we use strict equalities to ensure the exact BLOSUM distance) and set $H = 5$. For each case, we obtain the RMSD after alignment as reported in the fourth column of Table II. Furthermore, we present the average confidence percentage level of the prediction of the original (adversarial) sequence as reported by AlphaFold and denoted by μ_{all} (μ'_{all}). Additionally, in the 6th and 8th columns, we report the average confidence values for the residues that are different between the original and adversarial sequences. These are denoted by μ_{diff} and μ'_{diff}, respectively.

[1] https://github.com/deepmind/alphafold

[2] https://ftp.uniprot.org/pub/databases/uniprot/pre_release/covid-19.fasta

[3] https://github.com/ialkhouri/PFNN_Attacks

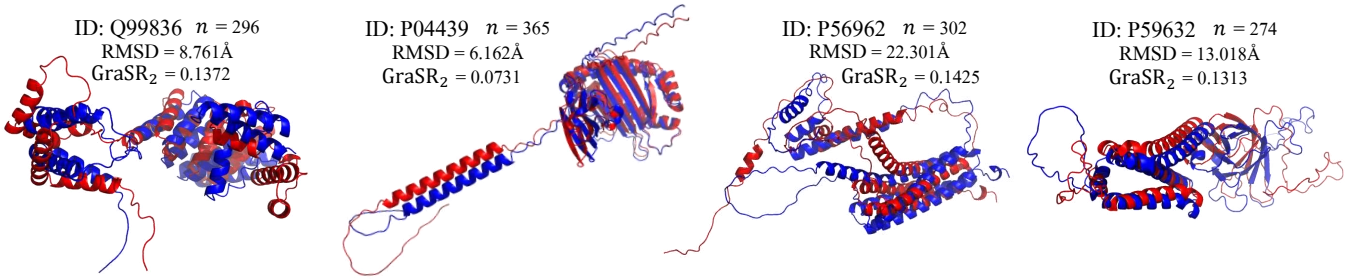

Fig. 4: The structures of the original (blue) and adversarial (red) sequences from AlphaFold. The 3D plots, aligned using PyMol [1], are for proteins O43765 (first), P04439 (second), P56962 (third), and P59632 (fourth). For structure differences, the RMSD values are reported. The structures of the complete list of sequences are attached in the provided anonymous link.

TABLE III: RMSD, GDT-TS, and GDT-HA results using the full database AlphaFold configuration with $L = 20$ and $H = 5$. The average columns correspond to 20 adversarial samples for each protein ID. The complete table is placed in the supplementary material.

Seq. ID	n	Similarity (%)	RMSD	Avg. RMSD	GDT-TS (%)	Avg. GDT-TS (%)	GDT-HA (%)	Avg. GDT-HA (%)	run-time (days)
O43765	313	98.4026	14.438	9.1741± 3.7576	13.9776	35.4832± 18.3219	2.8754	17.6358 ± 14.6829	1.6068
P56962	302	98.3444	22.301	15.8695 ± 3.6513	12.3344	18.6921 ± 4.0056	3.4768	5.803 ± 1.9135	0.5959
P04439	365	98.6301	6.162	3.7942 ± 1.1511	47.7397	68.2705 ± 11.0327	25.0	45.774 ± 11.4072	0.6429
Q99836	296	98.3108	8.761	5.2907 ± 2.3055	24.1554	46.6723 ± 22.0819	7.6858	26.2584 ± 20.8061	0.6246
P59632	274	98.1752	13.018	8.4704 ± 2.4464	24.8175	41.0401 ± 14.257	9.0328	21.6834 ± 12.183	0.5214

TABLE IV: GraSR results using the full database AlphaFold configuration with $L = 20$ and $H = 5$. RMSD results are added for comparison. The complete list is given in the supplementary material.

Seq. ID	n	RMSD	GraSR$_2$	GraSR$_\infty$
O43765	313	14.438	0.1588	0.0229
P56962	302	22.301	0.1425	0.0251
P04439	365	6.162	0.0731	0.0111
Q99836	296	8.761	0.1372	0.0204
P59632	274	13.018	0.1313	0.0184

We observe that, in general, when the BLOSUM threshold distance increases, the RMSD also increases. This means that biologically increased distance in the input space, in general, causes higher changes in the output predictions of AlphaFold. In terms of the confidence scores, we observe that the change in the overall average confidence between the original and perturbed sequence is not significant. However, in almost all the considered cases, we notice that the prediction confidence of the altered residues has reduced for the adversarial sequence when compared to the ones reported for the original sequence.

C. UniProt Case Studies

We apply our adversarial approach to 111 publicly available protein sequences as of the time of this writing per the UniProt database (including 29 COVID-19 sequences) using AlphaFold full database configuration. Additionally, in the supplementary material, we provide complete results using the reduced AlphaFold configuration. The BLOSUM62 distance between the original and adversarial sequences is at most 20, thus they are biologically close to each other w.r.t the employed distance measure [7], [8]. Given the long list of the considered sequences, we describe only the following. SGTA_HUMAN Small glutamine-rich tetratricopeptide repeat-containing protein alpha (O43765), HLAA_HUMAN HLA class I histocompatibility antigen, A alpha chain (P04439), STX17_HUMAN Syntaxin-17 (P56962), AP3A_SARS ORF3a (P59632), and MYD88_HUMAN Myeloid differentiation primary response protein MyD88 (Q99836). The cases covered include homo sapiens and severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) organisms which provide a wide variety of proteins. The considered sequences vary in length as they range from $n = 22$ to $n = 2511$.

Figures 1 and 4 show the aligned predicted structures of the proteins described earlier where the original sequence is given in blue and the adversarial sequence is given in red. Our results indicate that even small changes to the input sequence can lead to significant differences in the predicted output structures, regardless of the structure of the original sequence. These perturbed inputs can be considered as adversarial examples for a machine learning model, in the sense that they are similar to the original input but result in significant changes in the output. However, from a biological perspective, we note that AlphaFold is able to effectively preserve the domain structure of the protein as evidenced by the visual similarity of the folded regions in the structures of the original and perturbed sequences, and reported by the GraSR metric that, unlike RMSD and GDT, accounts for domain rotations. It is important to note that the information obtained from these structures and their robustness can also depend on the specific application they are used for. In some cases, scientists only focus on the folding shape of the 3D structure of the complete or a portion of the sequence of interest [16].

The resulting structural distances (similarities) measured in Å (percentage) are given in terms of the RMSD (GDT-TS) in the fourth (sixth) column of Table III for the full database configuration. Furthermore, we report the results

TABLE V: Prediction confidence results using the full database AlphaFold configuration with $L = 20$.

Seq. ID	n	RMSD	μ_{all}	σ_{all}	μ_{diff}	σ_{diff}	μ'_{all}	σ'_{all}	μ'_{diff}	σ'_{diff}
O43765	313	14.438	80.221	19.634	94.786	1.027	80.554	19.423	93.71	1.392
P56962	302	22.301	69.172	23.753	96.516	0.409	69.342	23.759	96.54	0.463
P04439	365	6.162	86.845	18.995	44.23	2.704	86.921	19.068	44.678	3.968
Q99836	296	8.761	81.213	13.817	78.914	7.454	80.918	13.971	72.198	6.253
P59632	274	13.018	58.367	18.783	66.136	2.029	57.364	18.794	60.926	4.362

TABLE VI: Overall Prediction and attack results for the reduced and full database configurations of AlphaFold.

Configuration.	Avg. n	Std. n	Avg. μ_{all}	Std. μ_{all}	Avg. RMSD	Std. RMSD	Avg. GDT	Std. GDT	Avg. run-time	Std. run-time
reduced database	480.53	416.66	78.22	10.96	15.31	11.24	34.08	28.39	0.68	0.59
full database	410.73	336.63	78.25	10.23	14.78	11.18	34.95	28.16	0.86	0.63

using GDT-HA in the eighth column. The high similarity between the original and adversarial sequences is observed from the third column. The similarity percentage is calculated as $100(n - D_{\text{ham}}(S, S'))/n$, where $D_{\text{ham}}(S, S') \leq H = 5$. The complete results of all the considered proteins, including reduced AlphaFold configuration, are provided in the supplementary material. In addition, the results of GraSR_2 and GraSR_∞ are reported in Table IV.

As observed from the RMSD and GDT results in Table III, a small change in the input sequence corresponding to the substitution of only five residues cause AlphaFold to predict structures that are highly divergent from the predicted structure of the original sequence. We remark that further analysis becomes protein-specific and contingent upon the specific application at hand. Some researchers may prioritize examining the folding characteristics of the 3D structure for the entire protein or a specific region of interest. The implications and significance of such alterations can vary based on the context and objectives of the study. The last column in Table III reports the total execution time (in days) of running the 20 adversarial sequences that were randomly selected from the set $\mathcal{V}_{L,H}$, which is shown to scale with the sequence length. We only select 20 samples given the long time incurred by AlphaFold to predict the output structure.

Additionally, in Table V, we report the average (deviation) prediction confidence results as for all the residues (designated with subscript 'all') and for the 5 altered residues (subscript 'diff'). The standard deviation is denoted σ. We observe that, independent of the average prediction confidence, the RMSD between the original and adversarial predicted structures is always high. This is noted for both the full and reduced database configurations of AlphaFold. Moreover, we observe that AlphaFold predicts the adversarial structure with similar confidence values to the original sequence (e.g., see the 4th and 8th columns in both tables). The same observation holds for the entire sequence and for the altered residues (columns 6 and 10). Although we observed significant changes in the RMSD and GDT values, the similarity in the confidence values and the preservation of most of the folding regions with high confidence suggest that AlphaFold is performing as expected in general. This is also observed from the results of Table IV.

In the last two tables of the supplementary material, we present a breakdown of the GDT scores between the structures of the original and perturbed sequences based on the prediction confidence scores of the original sequence, using the regions (1 to 4) defined by AlphaFold. As can be seen, the GDT scores are generally low across all regions.

For the considered dataset, the values presented in Table VI gauge the overall robustness of AlphaFold as an ML model to adversarial sequences. As indicated in the documentation of AlphaFold, for better accuracy, the full database configuration incurs a longer execution time compared to the reduced database configuration. The reported average values of the RMSD and GDT-TS measures are 14.78Å and 37.95%, respectively. In CASP14 (year 2020), AlphaFold achieved a median GDT-TS score of 92.4%, and 88% of their predictions fall under RMSD = 4Å [4]. These results are obtained by comparing the predicted structures and the groundtruth. The CASP14 AlphaFold results underscore the significance of the values reported in Tables III and VI, as they show how small changes in the input sequences could damage the predictions (See columns 6 to 9 in Table VI). The main conclusion from our study is that even though AlphaFold may return high RMSD values when viewed from a machine learning perspective, the preservation of the folded regions and the GraSR results of the biological structure with similar confidence scores, indicate that AlphaFold is generally able to handle the BLOSUM adversarial sequences that were considered in our analysis.

VI. Conclusion

Recent advancements in the prediction of protein folding structures hold great potential for understanding diseases, mapping the human proteome, and designing drugs and therapeutics. However, this paper argues that further examination of AlphaFold's performance against BLOSUM-based adversarial sequences is necessary. We present the first work in this area by showing that Protein Folding Neural Networks (PFNNs) are vulnerable to adversarial attacks through minor perturbations of the input protein sequence. Despite these perturbations causing significant changes in predicted protein structure as determined by large distance measures, AlphaFold predictions have, in general, preserved the biological structures. This is surprising, given that previous studies have suggested that Alphafold may not be well-suited for studying mutations. We employed standard protein structural distance and similarity measures to evaluate AlphaFold's outputs, and suggest that the results can be used as a baseline for future research on the robustness of PFNNs against adversarial attacks.

For future work, it is worth exploring novel and faster adversarial attacks for this discrete domain of protein sequences.

[4] https://predictioncenter.org/casp14/index.cgi

Recent work on adversarial attacks on large language models (LLMs) may hold some insights into how this can be achieved. Indeed, the attack surface of LLMs is also governed by discrete tokens that have been successfully attacked to bypass the safety guardrails of tools such as Chat-GPT [32].

ACKNOWLEDGEMENTS

We would like to thank Dr. Herve Roy (Biomedical Sciences at UCF) for insightful discussions. This work was supported in part by NSF Award CCF-2106339 and NSF CAREER Award CCF-1552497.

REFERENCES

[1] L. Schrödinger and W. DeLano, "Pymol."
[2] H. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, "Advancing drug discovery via artificial intelligence," *Trends in pharmacological sciences*, vol. 40, no. 8, pp. 592–604, 2019.
[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[4] H. Bagdonas, C. A. Fogarty, E. Fadda, and J. Agirre, "The case for post-predictional modifications in the alphafold protein structure database," *Nature Structural & Molecular Biology*, vol. 28, no. 11, pp. 869–870, 2021.
[5] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.
[6] C. Xia, S.-H. Feng, Y. Xia, X. Pan, and H.-B. Shen, "Fast protein structure comparison through effective representation learning with contrastive graph neural networks," *PLoS computational biology*, vol. 18, no. 3, p. e1009986, 2022.
[7] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins.," *The EMBO journal*, vol. 5, no. 4, pp. 823–826, 1986.
[8] C. Sander and R. Schneider, "Database of homology-derived protein structures and the structural meaning of sequence alignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 9, no. 1, pp. 56–68, 1991.
[9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[10] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021.
[11] B. Rost, "Twilight zone of protein sequence alignments," *Protein engineering*, vol. 12, no. 2, pp. 85–94, 1999.
[12] M. H. J. Cordes, R. E. Burton, N. P. Walsh, C. J. McKnight, and R. T. Sauer, "An evolutionary bridge to a new protein fold," *Nature Structural Biology*, vol. 7, no. 12, pp. 1129–1132, 2000.
[13] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman, "Interconversion between two unrelated protein folds in the lymphotactin native state," *Proceedings of the National Academy of Sciences*, vol. 105, no. 13, pp. 5057–5062, 2008.
[14] R. A. Stein and H. S. Mchaourab, "Modeling alternate conformations with alphafold2 via modification of the multiple sequence alignment," *bioRxiv*, 2021.
[15] D. Del Alamo, D. Sala, H. S. Mchaourab, and J. Meiler, "Sampling alternative conformational states of transporters and receptors with alphafold2," *Elife*, vol. 11, p. e75751, 2022.
[16] E. Callaway, "What's next for the ai protein-folding revolution," *Nature*, vol. 604, pp. 234–238, 2022.
[17] G. R. Buel and K. J. Walters, "Can alphafold2 predict the impact of missense mutations on structure?," *Nature Structural & Molecular Biology*, vol. 29, no. 1, pp. 1–2, 2022.
[18] S. K. Jha, A. Ramanathan, R. Ewetz, A. Velasquez, and S. Jha, "Protein folding neural networks are not robust," *arXiv preprint arXiv:2109.04460*, 2021.
[19] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 2018.
[20] M. Dayhoff, R. Schwartz, and B. Orcutt, "22 a model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, pp. 345–352, 1978.
[21] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Bioinformatics*, vol. 8, no. 3, pp. 275–282, 1992.
[22] A. Zemla, "Lga: a method for finding 3d similarities in protein structures," *Nucleic acids research*, vol. 31, no. 13, pp. 3370–3374, 2003.
[23] F. Croce, M. Andriushchenko, V. Sehwag, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," *arXiv preprint arXiv:2010.09670*, 2020.
[24] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European Conference on Computer Vision*, pp. 484–501, Springer, 2020.
[25] Y. Bai, Y. Zeng, Y. Jiang, Y. Wang, S.-T. Xia, and W. Guo, "Improving query efficiency of black-box adversarial attack," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 101–116, Springer, 2020.
[26] F. Croce and M. Hein, "Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers," *arXiv preprint arXiv:2103.01208*, 2021.
[27] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ACCS'17*, 2017.
[28] K. Mahmood, R. Mahmood, E. Rathbun, and M. Van Dijk, "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks," *IEEE Access*, 2021.
[29] J. Jumper, K. Tunyasuvunakool, P. Kohli, D. Hassabis, and A. Team, "Computational predictions of protein structures associated with covid-19," *DeepMind website*, 2020.
[30] "Uniprot: the universal protein knowledgebase in 2021," *Nucleic acids research*, vol. 49, no. D1, pp. D480–D489, 2021.
[31] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lddt: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, 2013.
[32] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.