Benchmarking the Robustness of Protein Folding Neural Networks: A COVID-19 Case Study Using AlphaFold

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 explored. This is particularly relevant given the broad social implications of such technologies and the fact that biologically small perturbations in the protein sequence do not generally lead to drastic changes in the protein structure. In this paper, we demonstrate that AlphaFold does not exhibit such robustness despite its high accuracy. This raises the challenge of detecting and quantifying the extent to which these predicted protein structures can be trusted. To measure the robustness of the predicted structures, we utilize (i) the root-mean-square deviation (RMSD) and (ii) the Global Distance Test (GDT) similarity measure between the predicted structure of the original sequence and the structure of its adversarially perturbed version. Based on the well-established BLOSUM62 sequence alignment scoring matrix, we generate adversarial protein sequences and show that the RMSD between the predicted protein structure and the structure of the original sequence are very large when the adversarial changes are bounded by (i) 20 units in the BLOSUM62 distance, and (ii) five residues (out of hundreds or thousands of residues) in the given protein sequence. We consider all 125 COVID-19 proteins in the Universal Protein resource (UniProt), a central resource for protein data managed by the European Bioinformatics Institute, Swiss Institute of Bioinformatics, and the US Protein Information Resource. These result in an overall GDT similarity test score average of around 31%, demonstrating a substantial drop in the performance of AlphaFold. This is the first paper demonstrating the susceptibility of AlphaFold to adversarial sequences and we present these results as the first set of benchmarks on the robustness of PFNNs.
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 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 [Chan et al. 2019]. AlphaFold [Jumper et al. 2021a] achieved unparalleled success in predicting protein structures using neural networks and remains first at the Critical Assessment of protein Structure Prediction (CASP2020) competition. While this has been touted as a breakthrough for structural biology [Bagdonas et al. 2021], the robustness its predictions has not yet been explored. The main contribution of this paper is to provide the first set of benchmarks demonstrating the susceptibility of AlphaFold to adversarial sequences by generating several examples where protein sequences, which vary in only five residues out of hundreds or thousands of residues, result in very different 3D protein structures. We use sequence alignment scores [Henikoff and Henikoff 1992] such as those derived from Block Substitution Matrices (BLOSUM62) to identify a space of similar protein sequences used in constructing adversarial perturbations. For the output structures, we leverage the standard metrics commonly used in CASP, namely (i) the root-mean-square deviation (RMSD) and (ii) the Global Distance Test (GDT) similarity measure between the predicted structure and the structure of its adversarially perturbed sequence. See Figure 1 for two examples.

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. Hence, our proposed approach is a first step in the direction of identifying protein sequences on which the predicted 3D structure cannot be trusted.
2 Summary and Related Work

PFNNs [Jumper et al., 2021b; Baek et al., 2021] should be expected to obey the natural observation that biologically small changes in the sequence of a protein usually do not lead to drastic changes in the protein structure. Almost four decades ago, it was noted that two structures with 50% sequence identity align within approximately 1 Å RMSD from each other [Chothia and Lesk, 1986]. Two proteins with even 40% sequence identity and at least 35 aligned residues align within approximately 2.5 Å [Sander and Schneider, 1991]. The phenomenon of sequence-similar proteins producing similar structures have also been observed in larger studies [Rost, 1999]. 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 [Cordes et al., 2000; Tuinstra et al., 2008].

2.1 Robustness Metric using Adversarial Attacks

The similar-sequence implies similar-structure paradigm dictates that PFNNs should make robust predictions. Given a protein sequence of $n$ residues $S = s_1 s_2 \ldots s_n$ with a three-dimensional structure $A(S) = (x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$, we define a notion of biologically similar sequences $\mathcal{V}$ using Block Substitution Matrices (BLOSUM) [Henikoff and Henikoff, 1992], and then employ formulations of adversarial attacks [Goodfellow et al., 2018] on PFNNs within this space of similar sequences to identify a sequence $S_{adv} \in \mathcal{V}$ that produces a maximally different three-dimensional structure $A(S_{adv})$. We then compute the RMSD and GDT between the structures for the original and adversarial inputs ($A(S)$ and $A(S_{adv})$), and use these metrics as the robustness measure. If the RMSD (GDT) is small (high), the response of the PFNN is deemed robust; a large (small) RMSD (GDT) indicates that the predicted structure is not robust.

2.2 BLOSUM Similarity Measures

Given two sequences of $n$ residues $S = s_1 s_2 \ldots s_n$ and $S' = s'_1 s'_2 \ldots s'_n$, 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 of residue.

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 [Dayhoff et al., 1978; Jones et al., 1992]. 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 [Henikoff and Henikoff, 1992] 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, 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 $b_j \in \mathcal{X}$.
We identify the space of biologically similar sequences $V$ for a given protein sequence $S$ with respect to the BLOSUM distance. We expect the predicted structures for the similar sequences to be similar. If there is a large RMSD (or small GDT) between the predicted structure $\mathcal{A}(S)$ and the structure $\mathcal{A}(S_{adv})$ of the adversarial sequence $S_{adv} \in V$, it would reflect a lack of robustness in the prediction of the network. We adopt a sequence similarity measure that counts replacement frequencies in conserved blocks across different proteins.

3 Approach

Our approach to evaluating the robustness of PFNNs is based on two main ideas: (i) the existence of adversarial examples in PFNNs that produce adversarial structures possibly very different from the original structure, and (ii) the use of BLOSUM matrices for identifying a neighborhood of a given sequence that are biologically similar and hence expected to have similar 3D structures. We utilize the RMSD and GDT 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, AlphaFold is the model of focus.

3.1 Sequence Similarity Measures

Given two sequences $S = s_1s_2 \ldots s_n$ and $S' = s'_1s'_2 \ldots s'_m$, the BLOSUM distance between the two sequences is given by Equation (1) below. For an illustrative example of $D_{seq}$, see Figure 2.

$$D_{seq}(S, S') = \sum_{i \in [n]} (B_{s_i, s'_i} - B_{s_i, s'_j})$$

(1)

Figure 2: The BLOSUM62 matrix $B$ (top). The original sequence $S$ is followed by 10 sequences generated by changing 4, 5, 6, and 7 residues (bottom). The sequences are samples from the space in (2) with different values of $L$ and $H$. The distance $D_{seq}$ is calculated using (1).
3.2 Output Structural Measure

Given a sequence of \( n \) residues \( S = s_1 s_2 \ldots s_n \), its three dimensional structure \( \mathcal{A}(S) \) is an ordered \( n \)-tuple of three-dimensional co-ordinates \((x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)\). 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) [Zemla 2003].

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, Schrödinger and DeLano to align \( \mathcal{A}(S') \) with regard to the target structure \( \mathcal{A}(S) \), which denotes the predicted structure of the original sequence \( S \). Let the aligned structure be denoted by \( \mathcal{A}(S)' = (\hat{x}'_1, \hat{y}'_1, \hat{z}'_1), \ldots, (\hat{x}'_n, \hat{y}'_n, \hat{z}'_n) \). Then, the RMSD, measured in Å, is obtained as

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

where

\[
d(\mathcal{A}(S)_i, \mathcal{A}(S')_i) = (x_i - \hat{x}'_i)^2 + (y_i - \hat{y}'_i)^2 + (z_i - \hat{z}'_i)^2
\]

represents the 3D carbon-alpha coordinates of the \( i \)-th residue. We use the carbon-alpha coordinates as they are the standard approach in CASP [Zemla 2003].

Another standard metric for protein structure is the GDT similarity measure which is introduced by Zemla [2003] and commonly used in the CASP competition along with the RMSD. In some cases, the latter is known to be sensitive to outliers [Zemla 2003]. The GDT score returns a value in \([0, 1]\) where 1 means identical structures, and is computed with respect to four thresholds, \( \delta_j \), as

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

where the thresholds \( \delta_1, \delta_2, \delta_3, \) and \( \delta_4 \) 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 \( 1(\cdot) \) is the indicator function that returns 1 only if the event in its argument is satisfied. In \([5]\), each \( j \in [4] \) reflects the number of residues in the structures where the distance is below \( \delta_j \).

3.3 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 [Croce et al. 2020; Andriushchenko et al. 2020; Bai et al. 2020; Croce and Hein 2021]. 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

\[
\max_{S'} D_{\text{str}}(\mathcal{A}(S), \mathcal{A}(S')) \text{ subject to } D_{\text{seq}}(S, S') \leq L.
\]

In our experiments, we set \( L = 20 \) and \( D_{\text{str}} \) is 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

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

where \( \mathcal{X}^n \) is the set of all possible sequences over \( \mathcal{X} \) of length \( n \), \( D_{\text{ham}} \) is the hamming distance, and \( H \) is a predefined threshold. For long sequences, the search space can be extensively large. Therefore,
we select random samples from $V_{L,H}(S)$ 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 [Papernot et al., 2017]. We leave the derivation of more sophisticated attacks (e.g. whitebox attacks) and the use of generated adversarial examples for adversarial training as avenues for future work that can leverage the benchmarks proposed herein.
Table 1: RMSD, GDT-TS, and GDT-HA with $L = 20$ and $H = 5$. The complete list is placed in the supplementary material. The run-time results are measured in days. The average columns correspond to 20 adversarial samples per each protein ID.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Similarity (%)</th>
<th>RMSD</th>
<th>GDT-TS (%)</th>
<th>GDT-HA (%)</th>
<th>run-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q01015</td>
<td>99.87</td>
<td>1.15</td>
<td>99.79</td>
<td>99.79</td>
<td>0.110</td>
</tr>
<tr>
<td>Q9N2D1</td>
<td>99.97</td>
<td>1.12</td>
<td>99.77</td>
<td>99.77</td>
<td>0.111</td>
</tr>
<tr>
<td>Q01513</td>
<td>99.83</td>
<td>1.17</td>
<td>99.73</td>
<td>99.73</td>
<td>0.112</td>
</tr>
<tr>
<td>Q9N3F2</td>
<td>99.94</td>
<td>1.18</td>
<td>99.74</td>
<td>99.74</td>
<td>0.113</td>
</tr>
<tr>
<td>Q01011</td>
<td>99.88</td>
<td>1.21</td>
<td>99.71</td>
<td>99.71</td>
<td>0.114</td>
</tr>
<tr>
<td>Q9N3G1</td>
<td>99.95</td>
<td>1.22</td>
<td>99.72</td>
<td>99.72</td>
<td>0.115</td>
</tr>
<tr>
<td>Q01012</td>
<td>99.89</td>
<td>1.23</td>
<td>99.73</td>
<td>99.73</td>
<td>0.116</td>
</tr>
<tr>
<td>Q9N3H1</td>
<td>99.96</td>
<td>1.24</td>
<td>99.74</td>
<td>99.74</td>
<td>0.117</td>
</tr>
<tr>
<td>Q01013</td>
<td>99.90</td>
<td>1.26</td>
<td>99.76</td>
<td>99.76</td>
<td>0.118</td>
</tr>
<tr>
<td>Q9N3I1</td>
<td>99.91</td>
<td>1.27</td>
<td>99.77</td>
<td>99.77</td>
<td>0.119</td>
</tr>
<tr>
<td>Q01014</td>
<td>99.92</td>
<td>1.28</td>
<td>99.78</td>
<td>99.78</td>
<td>0.120</td>
</tr>
<tr>
<td>Q9N3J1</td>
<td>99.93</td>
<td>1.29</td>
<td>99.79</td>
<td>99.79</td>
<td>0.121</td>
</tr>
<tr>
<td>Q01015</td>
<td>99.94</td>
<td>1.30</td>
<td>99.80</td>
<td>99.80</td>
<td>0.122</td>
</tr>
<tr>
<td>Q9N3K1</td>
<td>99.95</td>
<td>1.31</td>
<td>99.81</td>
<td>99.81</td>
<td>0.123</td>
</tr>
<tr>
<td>Q01016</td>
<td>99.96</td>
<td>1.32</td>
<td>99.82</td>
<td>99.82</td>
<td>0.124</td>
</tr>
<tr>
<td>Q9N3L1</td>
<td>99.97</td>
<td>1.33</td>
<td>99.83</td>
<td>99.83</td>
<td>0.125</td>
</tr>
<tr>
<td>Q01017</td>
<td>99.98</td>
<td>1.34</td>
<td>99.84</td>
<td>99.84</td>
<td>0.126</td>
</tr>
<tr>
<td>Q9N3M1</td>
<td>99.99</td>
<td>1.35</td>
<td>99.85</td>
<td>99.85</td>
<td>0.127</td>
</tr>
</tbody>
</table>

Note: The columns represent the RMSD, GDT-TS, and GDT-HA values, the average run-time, and the number of runs. The protein IDs are listed in alphabetical order. The similarity values are calculated using the sequence identity matrix.
4 Benchmarking Results

For our experimental setup, we use the default configuration of AlphaFold. This includes the initial multi-sequence alignment step, the five-model ensembles predictions, recycling, output confidence ranking, and amber relaxation. For further details about each step, we refer the reader to Jumper et al. [2021a] and its supplementary information. 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 without outlier rejections. These outliers only impact the calculations of RMSD.

Our adversarial sequences are generated by randomly sampling 20 sequences from the set $V_{L,H}$ in (5) 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 COVID-19 protein sequences from the UniProt database considered by AlphaFold in Jumper et al. [2020]. The original fasta (file extension for protein sequences) sequence files are available online. Additionally, we generate adversarial sequences against all the UniProt (Universal Protein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases) and UniProt. Our code along with the adversarial sequences are also made available.

4.1 COVID-19 Case Studies

We apply our adversarial approach to all 130 publicly available COVID-19 protein sequences as of the time of this writing per the UniProt database. The BLOSUM62 distance between the original and adversarial sequences is at most 20, thus they are biologically close to each other. 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), PHB2_HUMAN Prohibitin-2 (Q99623), and MYD88_HUMAN Myeloid differentiation primary response protein MyD88 (Q99836). The cases covered include homo sapiens and severe acute

---

1 We use the latest version of AlphaFold (V2) available at https://github.com/deepmind/alphafold
3 https://github.com/ialkhouri/PFNN_Attacks
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 3 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. We observe that, independent of the predicted structure of the original sequence, a small change in the input sequence results in significant changes in the output structures. 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 1. Furthermore, we report the results using GDT-HA in the eighth column. The high similarities between the original and adversarial sequences are observed in the third column of the table. 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 proteins are provided in the supplementary material.

As observed from the RMSD and GDT results in Table 1, small changes in the input sequence corresponding to only five residues cause AlphaFold to predict structures that are highly divergent from the predicted structure of the original sequence. The last column in Table 1 reports the total execution time (in days) of running the 20 adversarial sequences that we randomly select from 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.

In Figure 4, we plot the RMSD, GDT-TS, and GDT-HA between the predicted structures of 50 randomly selected COVID19 proteins from the considered set and their adversarially generated versions. We only select 50 to better visualize the graph. We observe that the GDT and RMSD values are almost always consistent for the original and perturbed sequences. This means high (low) RMSD values correspond to low (high) GDT scores. The major observation is drawn from the reported average values of 14.2Å, 31.7%, and 17.3% for the RMSD, GDT-TS, and GDT-HA measures, respectively. In this paper, we use these values to represent the overall robustness of AlphaFold in handling adversarial sequences. It is important to note that, in CASP14 (year 2020), AlphaFold achieved a median GDT-TS score of 92.4%, and 88% of their predictions fall under RMSD = 4Å. These results are computed when comparing the predicted and ground truth structures. The CASP14 AlphaFold results shed light on the significance of the values reported in Table 1 and Figure 4 as they show how small changes in the input sequences can damage the predictions. This leads to the main message of this paper that even under a basic approach to generate perturbations of the input protein sequence, AlphaFold is, in general, not robust.

5 Conclusions & Future Work

The groundbreaking progress made in recent years on the prediction of protein folding structures promises to enable profound advances in the understanding of diseases, the mapping of the human proteome, and the design of drugs and therapeutics. However, until these predictions are shown to be robust, we argue that the grand challenge of predictive protein folding persists. In this paper, we have presented the first work in this direction by demonstrating that Protein Folding Neural Networks (PFNNs) are often susceptible to adversarial attacks in the form of minor perturbations to the input protein sequence. These perturbations can induce great changes in the predicted protein structure and the resulting lack of robustness precludes the adoption of such PFNNs in safety-critical applications. We have employed standard protein structural distance and similarity to measure the robustness of AlphaFold. While the perturbation methods employed in this paper were basic for the purposes of illustrating the lack of robustness of PFNNs, the benchmarks established herein can be readily used as a baseline for future work on adversarial attacks to and the robustness of PFNNs. We will explore the derivation of more sophisticated adversarial attacks for future work and plan to leverage these contributions to the design of adversarial training schemes that leverage the perturbed sequences and their outputs in order to train more robust PFNNs.

https://predictioncenter.org/casp14/index.cgi
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]