EPiC Series in Computing

Volume 83, 2022, Pages 53–65

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

Analysis of AlphaFold2 for Modeling Structures of Wildtype and Variant Protein Sequences

Anowarul Kabir[1], Toki Tahmid Inan[1], and Amarda Shehu[1,2,3,4*]

[1] Dept. of Computer Science, George Mason University, Fairfax, VA 22030, USA
akabir4@gmu.edu, tinan@gmu.edu, amarda@gmu.edu
[2] Center for Advancing Human-Machine Partnerships, George Mason University
[3] Dept. of Bioengineering, George Mason University
[4] School of Systems Biology, George Mason University
[5] Mailing Address: MS 4A5, 4400 University Dr., Fairfax, VA 22030, USA

Abstract

ResNet and, more recently, AlphaFold2 have demonstrated that deep neural networks can now predict a tertiary structure of a given protein amino-acid sequence with high accuracy. This seminal development will allow molecular biology researchers to advance various studies linking sequence, structure, and function. Many studies will undoubtedly focus on the impact of sequence mutations on stability, fold, and function. In this paper, we evaluate the ability of AlphaFold2 to predict accurate tertiary structures of wildtype and mutated sequences of protein molecules. We do so on a benchmark dataset in mutation modeling studies. Our empirical evaluation utilizes global and local structure analyses and yields several interesting observations. It shows, for instance, that AlphaFold2 performs similarly on wildtype and variant sequences. The placement of the main chain of a protein molecule is highly accurate. However, while AlphaFold2 reports similar confidence in its predictions over wildtype and variant sequences, its performance on placements of the side chains suffers in comparison to main-chain predictions. The analysis overall supports the premise that AlphaFold2-predicted structures can be utilized in further downstream tasks, but that further refinement of these structures may be necessary.

1 Introduction

The road to AlphaFold2 was long and arduous. Many findings led to ResNet [13] and, more recently, to AlphaFold2 [4]. AlphaFold2 has been reported to predict a tertiary structure of a given protein amino-acid sequence with high accuracy [12]. This seminal development promises to allow molecular biology researchers to advance various structure-centric studies linking sequence, structure, and function, including studies focusing on the impact of sequence mutations on stability, fold, and function. One can now foresee utilizing AlphaFold2 to predict the tertiary structures of available wildtype and mutated protein sequences and then build on

*Corresponding author

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB2022 (EPiC Series in Computing, vol. 83), pp. 53–65

the predicted structures to support downstream tasks of predicting stability, function, ligand binding, and other molecular interactions.

This paper evaluates the performance of AlphaFold2 in predicting accurate tertiary structures of wildtype and mutated sequence variants of protein molecules. Proteins are selected from S^{sym} [10], a benchmark dataset used by researchers modeling the impact of mutations on stability. We utilize several global and local structure indicators and carry out detailed analyses to evaluate the performance and precision of AlphaFold2 in reproducing not only the tertiary structure of a wildtype sequence but also the tertiary structure of its possibly many variants. We expand the setting to an ensemble analysis; that is, we utilize AlphaFold2 to generate not just one, but several tertiary structures from a given amino-acid sequence.

The evaluation yields many interesting observations. For instance, utilizing various global structure comparison measures, it shows that AlphaFold2 performs similarly on wildtype and variant sequences with regards to the main chain. Specifically, the placement of the main chain of a protein molecule is highly accurate, with no differences observed between wildtypes and variants. AlphaFold2 also reports similar confidence in its predictions over both wildtype and variant sequences. Utilizing local structure assessment measures, such as secondary structure and side-chain placement at the mutation site, shows that accuracy suffers in comparison to main-chain predictions. We draw three major conclusions from the evaluation carried out in this paper. First, this study supports the overall premise that AlphaFold2-predicted structures can be utilized in further downstream tasks. However, downstream tasks where accurate placement of the side chain or accurate secondary structure at the mutation site are essential necessitate further refinement of AlphaFold2-predicted structures.

2 Methods

We first relate some details (and analysis) on the benchmark dataset we employ and then describe the various global and local structure assessment measures that we utilize to evaluate AlphaFold2-predicted structures.

2.1 Dataset

The S^{sym} [10] dataset is comprised of 684 mutations, half of which are direct and half are augmented following the hypothetical reverse mutation [6]. We consider only direct mutations, since both direct and reverse mutations refer to the same proteins. Direct mutations are further divided into two sets: stabilizing ($\Delta\Delta G \geq 0$) and destabilizing ($\Delta\Delta G < 0$). Fig. 1(a) illustrates the distribution with respect to $\Delta\Delta G$ shows that the dataset is biased towards stabilizing mutations. The dataset is comprised of 15 wildtype and their corresponding 342 variant proteins. Only 6 entries in the S^{sym} dataset (wildtype PDB IDs: 1qit, 2f0d, 1ihb; Variant PDB IDs: 1iob, 1mx2, 1lav) are found in the training dataset of AlphaFold2-advance ColabFold (the public version of AlphaFold2 we utilize in this paper). This indicates that there is very little, if any, data leakage. The lengths of protein sequences in the S^{sym} dataset are mostly in the range $[58, 162]$ amino acids; there are few variants for a protein 396 amino acids long. The length distribution is shown in Fig. 1(b). Fig. 1(c) shows that the dataset contains mutations at almost any position along the sequence (in the range of $[2, 200]$).

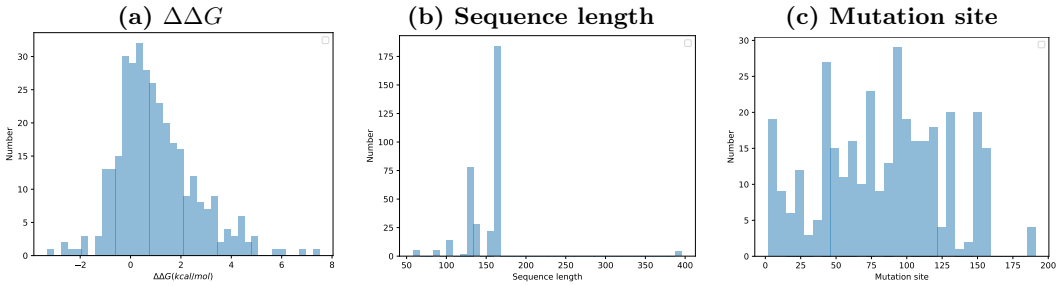

Figure 1: Data distribution corresponding to (a) the theromodynamic stability change ($\Delta\Delta G$) upon mutations, (b) sequence length, and (c) mutation position.

2.2 AlphaFold2 Protocol

We run the AlphaFold2-advance implementation from ColabFold [8]; we refer to this protocol as AlphaFold2 for convenience. For each protein sequence, we select a fixed set of parameters. We use the faster *mmseq2* multiple sequence alignment (MSA) method over jackhmmer, as recommended by ColabFold. The *max-recycles* is set to 3; this is an important parameter for long sequences, or when there are multiple homooligomers. All our variant proteins are single chains (so we set *homooligomer* to 1); however, since some wildtype proteins contained multiple chains, *homooligomer* is set to ≤ 2. We do not use any templates for modeling protein structures. Important parameters in our setup are *is-training* and *num-samples*, which enable us to generate multiple predictions for a single sequence and so evaluate the possible diversity of the solution space. The *is-training* parameter leverages the stochasticity of AlphaFold2, and *num-samples* indicates the number of random seeds to apply. With these parameters, we generated 45 structures for each protein sequence. The top five PLDDT-ranked structures are further relaxed using the Amber-relax protocol; as we will describe later, PLDDT is a measure of prediction confidence. A summary of all parameters and their values is presented in Table 1.

Table 1: Parameter values for utilizing AlphaFold2 to generate 45 structures for a sequence.

Parameter	**Value**	**Parameter**	**Value**
homooligomer	1 (variant) or $\leq$2 (wildtype, per chain)	msa_method	mmseq2
add-custom-msa	false	pair-mode	unpaired
rank-by	PLDDT	num-models	5
use-ptm	true	num-ensemble	1
max-recycles	3	tol	0
is-training	true	num-samples	8
subsample-msa	true	num-relax	Top5

2.3 Metrics to Evaluate Global Structure Quality

We first use global structure quality metrics to compare the quality of a predicted structure to the ground-truth structure (that can be found in the Protein Data Bank (PDB) [1]). For instance, an AlphaFold2-generated structure for the wildtype sequence is compared to the experimentally-available structure for the wildtype in the PDB. Similarly, an AlphaFold2-generated structure for a variant sequence is compared to the experimentally-available structure

for that variant in the PDB.

The evaluation of global structure quality focuses on the main-chain carbon (CA) atoms, as in Critical Assessment of protein Structure Prediction (CASP). We employ five popular metrics that allow global comparisons of two tertiary structures: Root-Mean-Square-Deviation (RMSD) [7], Template Modeling Score (TM-Score) [15], Global Distance Test-Total Score (GDT-TS) [14], its more "high-accuracy" variant GDT-HA, and MaxSub score [9, 11]. RMSD is a dissimilarity metric, where lower values correspond to better proximity. While informative, higher values are harder to interpret properly, as the metric depends on the length of the chain (number of amino acids). In general, sub-angstrom values (≤ 1Å) are related with exceptional model accuracy.

TM-Score and GDT-TS are similarity metrics, where higher values mean better proximity; the latter two provide a score in $[0, 1]$. TM-score weights smaller distances more than larger distances and so makes the overall score more sensitive to global similarity than to local structural variations. In general, a TM-score no lower than 0.5 indicates high model accuracy.

GDT-TS scores, typically reported in %, indicate an average over numbers of "spatially-similar" amino acids at various proximity thresholds. Specifically, the GDT score is calculated as the largest set of corresponding CA atoms within a defined distance cutoff, after iterative superimposition of two structures under comparison. As in CASP, the GDT total score (GDT-TS) score we report is the average result of cutoffs at 1, 2, 4, and 8Å. GDT-TS values 0.6 and higher indicate models of good quality, with values of 0.8 indicating exceptional accuracy. GDT-HA is a more accurate version of GDT-TS and is computed over smaller distance cutoffs (half the size of the ones used in GDT-TS). In this way, GDT-HA more heavily penalizes larger deviations between structures.

The MaxSub score is also a similarity metric and varies in $[0, 1]$. The MaxCluster search algorithm identifies the maximal subset (MaxSub) of corresponding amino acids (CA atoms) that can be superimposed within a given distance threshold d. This process is iterated four times using a distance threshold of $d/4$, $d/2$, $3d/4$, and d, until no more amino acids can be added to the final set of M amino acids. The MaxSub score is then calculated as $1/N \sum_1^M [1/1 + (d_i^2/d^2)]$, where N is the total number of amino acids, d is the distance threshold, d_i is the distance between corresponding amino acids, and M is the number of amino acids identified by the algorithm.

Each of these metrics give us a summary value with which to characterize an AlphaFold2-generated structure. As described earlier, we utilize AlphaFold2 to generate a set of tertiary structures for a given sequence (wildtype or variant). So, we obtain distributions of these values. The evaluation in Section 3 summarizes the distribution of AlphaFold2-generated structures for the wildtype (by comparing them to the wildtype ground-truth) and then separately the distribution of AlphaFold2-generated structures for a variant (by comparing them to the ground-truth for the variant).

2.4 Metrics to Evaluate Prediction Confidence

For each amino acid, AlphaFold2 reports a confidence measure, referred to as PLDDT. This is useful information, as it allows us to ask whether AlphaFold2 is more confident on predicted amino acids for wildtype structures over variant structures. We first narrow this question to a specific amino acid, where there is a point mutation and summarize each AlphaFold2-generated structure with the PLDDT score for a target amino acid. So, let us suppose for instance, that various mutations exist for a wildtype sequence at position i. Section 3 relates the distribution of the PLDDT values of amino acid at position i over AlphaFold2-generated structures for the

wildtype sequence, and compares this distribution with the PLDDT values of amino acid at position i over AlphaFold2-generated structures for the various variant sequences (aggregating structures predicted for various sequences). We refer to this analysis as a 0-neighborhood confidence score, as we can expand this analysis to a k-neighborhood confidence score; in the latter, we average the PLDDT scores over amino acids $i-k, \ldots, i, \ldots i+k$.

2.5 Methodology to Evaluate Local Structure Quality

We investigate the accuracy of AlphaFold2 on local structure. Specifically, we focus on an amino acid at position i for which there are one or more mutations. Consider an AlphaFold2-generated structure for a wildtype sequence. First, instead of using RMSD to compare it to the PDB-available structure, we carry out the following protocol: we align only the backbone region that anchors the side chain of amino acid i; that is, the N, CA, and C atoms. Once these three atoms are optimally superimposed, we then report the RMSD over the side-chain atoms in an AlphaFold2-generated structures to the corresponding ones in the PDB-available structure. This process can be repeated over all AlphaFold2-generated structures for the wildtype sequence to obtain a distribution of such values. This analysis can also be carried out over AlphaFold2-generated structures for a variant sequence, using the PDB-available structure for variant under investigation as reference. Second, we assess the secondary structure at position i, as predicted via DSSP [3, 5]. We compare the DSSP-calculated secondary structure for the amino acid of interest in the AlphaFold2-predicted structure versus the ground-truth structure. We provide statistics for each of the 7 common secondary structure (α-helix, β-sheet, bend, strand, turn, 3-10 helix, and none/coil).

3 Results

The evaluation detailed below consists of three sets of analyses, analysis of global structure, analysis of prediction confidence, and analysis of local structure.

3.1 Global Structure Quality Analysis

Fig. 2 summarizes the global structure quality assessment along RMSD, TM-Score, GDT-TS, GDT-TA, and MaxSub-Score. Wildtype predictions are assessed separately from variant ones. As described in Section 2, since several structures are computed for a sequence, their distances from the ground truth are summarized in terms of the average; this is used as proxy of global structure quality. Fig. 2 shows the prediction quality for each sequence, using green for wildtype sequences and red and orange for variants; red is used to denote the destabilizing variants, and orange for the stabilizing variants. The bars show the standard deviation. The dataset is ordered by protein sequence length in ascending order.

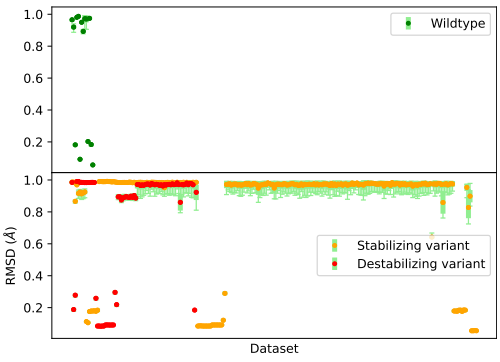

Figure 2: Global structure dissimilarity assessment: RMSD (Å).

Overall, RMSD values show very low dissimilarity between ground-truth and AlphaFold2 predicted structures for all proteins regardless of proteins types and sequence length. Even the quality value spread is very low. Therefore, all 45 predictions for each protein converge to the same ground-truth protein structure. However, the values show a two-tailed distribution where RMSD values are either close to 0Åor 1Å, with very few predictions in between. TM-Score, GDT-TS, GDT-HA, and MaxSub-Score additionally illustrate that AlphaFold2 predictions have high similarity compared to ground-truth structure. The values are close to 1 in most cases; GDT-HA shows more spread in structure quality. Overall, these metrics agree that AlphaFold2 predictions, whether on wildtype or variant sequences, are very close to the ground truth (the experimentally-available structure).

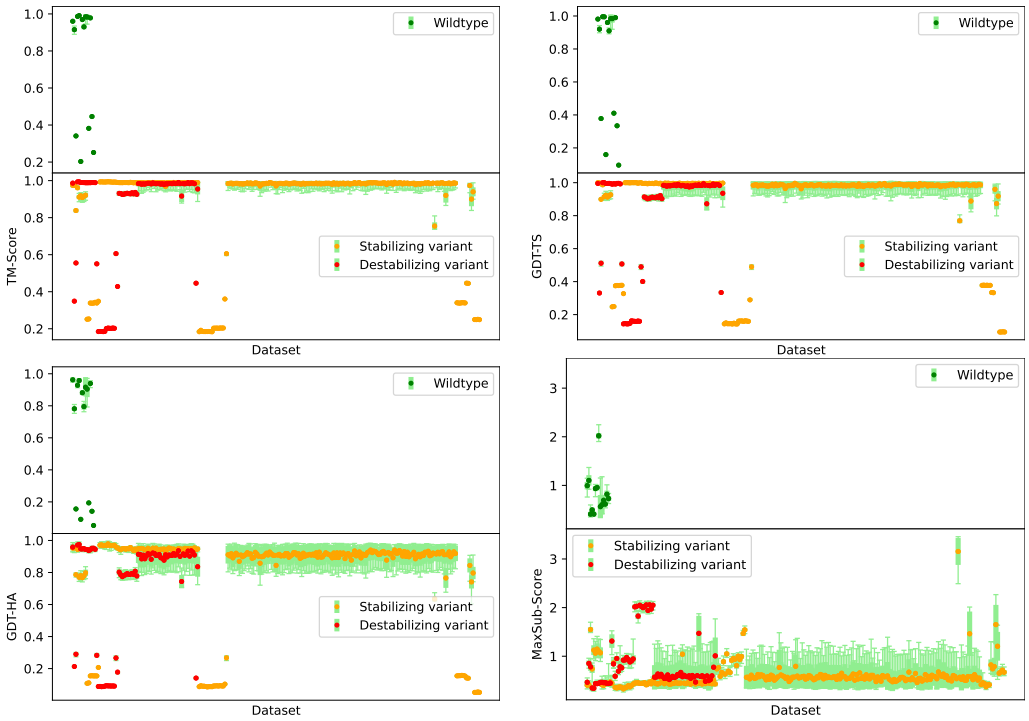

Figure 2: Global structure similarity assessment: TM-Score, GDT-TS, GDT-HA and MaxSub-Score.

3.2 Aggregate Prediction Confidence Analysis

AlphaFold2 provides PLDDT and PTM-Score for each predicted structure given a protein sequence and ranks them by PLDDT score. We utilize these scores as described in Section 2. Fig. 3 shows the confidence distribution for wildtype and variant proteins (separately). Again, green is used for wildtype sequences, red for destabilizing variants, and orange for stabilizing variants. Fig. 3 shows that there are no differences between wildtypes and variants (and no differences between stabilizing and destabilizing variants); AlphaFold2 reports similarly-high confidence for both groups. In particular, AlphaFold2 reaches 90% PLDDT score in most cases; no structure is predicted with less than 87% confidence.

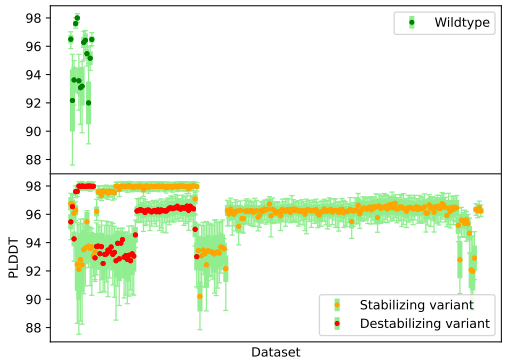

Figure 3: Global prediction confidence analysis. The whole-structure PLDTT confidence score obtained via AlphFold2 is averaged over computed structures; bars show standard deviation and range.

3.3 Local Prediction Confidence Analysis

To detect any differences in PLDTT confidence scores due to the mutation, we repeat the confidence analysis over wildtypes and mutated variants separately, taking into account 0-, 1-, 2-, and 3-neighborhoods around the mutation site (as described in Section 2. Fig. 4 summarizes the results. AlphaFold2 is very confident at the mutation site and over a neighborhood around the mutation site. The mutation site prediction confidence is more than 90% in most cases, with only in a few cases below 80% but not less than 50%. No particular differences are observed over wildtype (green), stabilizing variants (orange), and destabilizing variants (red).

Suppose that for each wildtype there are n_{wt} variants. We now compute the mutation site PLDDT confidence for each variant and group them by their corresponding wildtype protein. Fig. 5 relates this assessment via boxplots and shows that even for different neighborhood windows AlphaFold2 is pretty confident at all mutation sites.

3.4 Local Structure Quality Analysis

3.4.1 Side-chain Placement Accuracy

We also evaluate the dissimilarity of side chains at the mutation site (between computed and experimentally-available structure) using RMSD; we focus on the top-ranked AlphaFold2 prediction in this case (among the 45 computed structures for a sequence). Again, green is used for the wildtypes, red for destabilizing variants, and orange for stabilizing variants. Results are related in the top panel of Fig. 6. When focusing only on the side chain at the mutation site (0-neighborhood), RMSD values vary in the $[0, 2.0]$Årange for all sequences; no particular differences are observe among the groups (wildtype, destabilizing variants, and stabilizing variants). When the analysis is expanded to larger neighborhoods around the mutation site (averaging the RMSD over the side chains), the RMSD values obtained become more compact and vary in $[0, 0.8]$Å (data not shown). The bottom panel of Fig. 6 juxtaposes the PLDDT values at the mutation site (0-neighborhood) with the side-chain RMSDs (calculated as described above). The analysis for the wildtype sequences is shown on the left, and the analysis for the variants is shown on the right. The calculated Pearson and Spearman correlations [2] are similarly low, suggesting that there is no correlation between the prediction confidence and the side-chain RMSD at the mutation site.

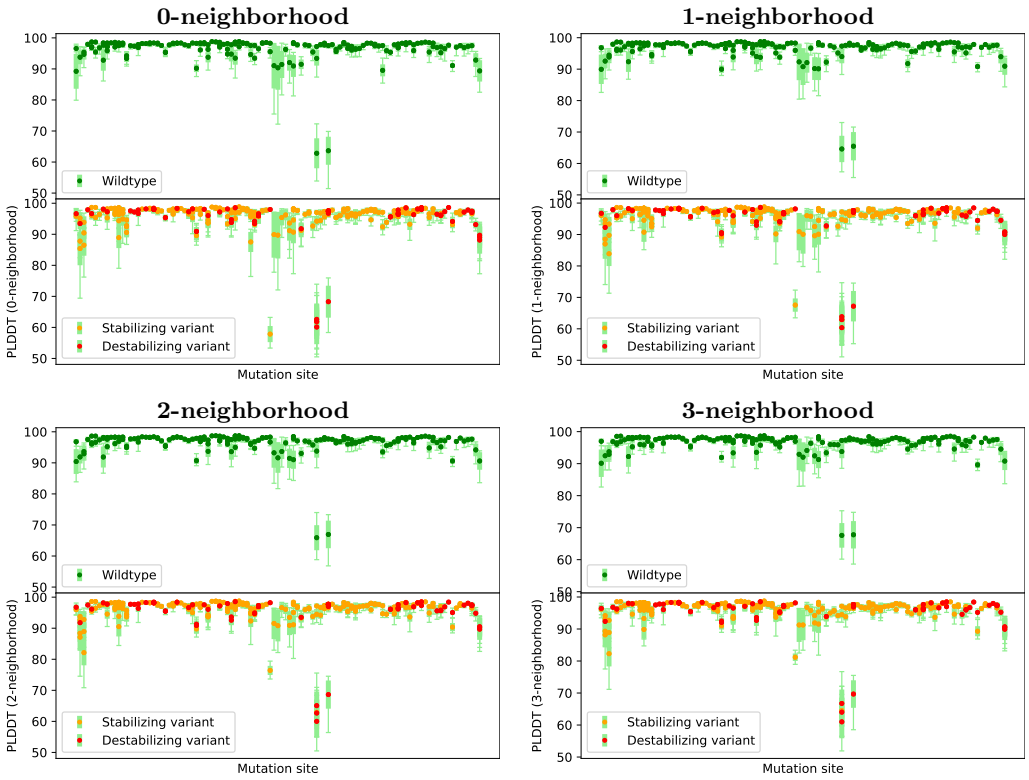

Figure 4: Local prediction confidence analysis at and around the mutation sites (over increasing neighborhoods).

3.4.2 Secondary Structure Accuracy

Fig. 7 organizes the data based on the 7 possible secondary structures at a mutation site in the ground-truth structure. If all predicted structures reproduce the secondary structure at the mutation site as in the ground-truth structure, then the score is 45 (for 45 Alphafold2-predicted structures). Deviations from this score indicate wrong secondary structure among the predicted structures. Fig. 7 plots the average, range, and standard deviation and shows that the average ranges from the high 20s to 45 in most cases (with the 3-10 helix having the lowest value).

3.5 Visualization of Selected Structures

We now visualize some of the structures and the side chains at the mutation sites. We select three sets of structures (top-ranked, predicted by AlphaFold2 for wildtype and variant sequences) that span the spectrum of side-chain RMSD values (measured as described above) and categorized in Best RMSD, Medium RMSD, and Worst RMSD. Fig. 8 shows the top-ranked structure in red, superimposed over the ground truth in green (main-chain atoms superimposed). Side chains at the mutation site are shown in stick-and-balls representation. The backbone is drawn in transparent, with secondary structures visible. Fig. 9 then focuses on the (top-ranked) computed versus the experimentally-available side chain at the mutation site. To prepare these figures, the N, CA, and C atoms are the mutation site in the computed structure are opti-

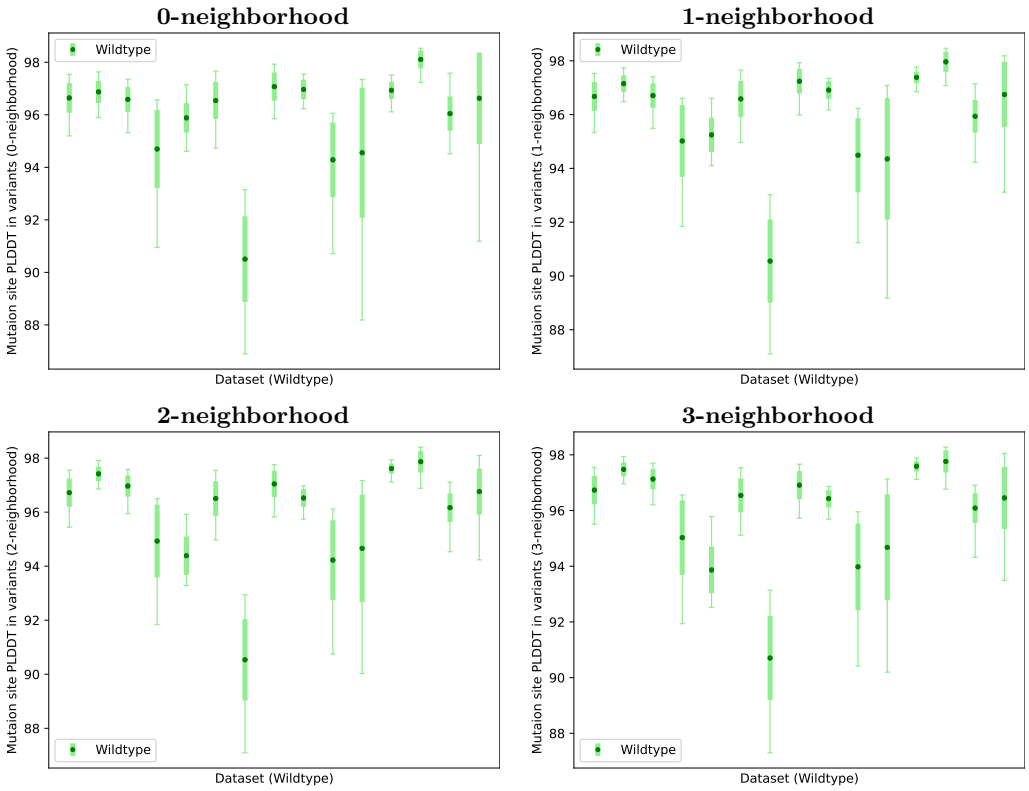

Figure 5: Boxplots relate the grouped local prediction confidence distribution at mutation sites for variants.

mally superimposed over the corresponding atoms in the experimentally-available structure. The obtained transformation is then applied to the heavy side-chain atoms at the mutation site in the computed structure, and the resulting side chain is shown in red, together with the experimentally-available side chain, which is shown in green.

4 Conclusion

This is an exciting time for structural biology and beyond. The ability to predict a high-quality tertiary structure on demand for a given protein sequence opens up many avenues of molecular biology research. One we focus on here is the ability of AlphaFold2 to support studies aiming to predict the impact of mutations on stability, function, interactions, and more. Towards this goal, this paper analyzes the quality of tertiary structures generated via AlphaFold2 for wildtype and mutated variants sampled from a benchmark dataset. Various global and local structure analyses yield many useful observations and support the premise that AlphaFold2-predicted structures can assist downstream tasks. However, when accurate secondary structure and placement of the mutated side chain is essential, further refinement of AlphaFold2-predicted structures is necessary to improve precision.

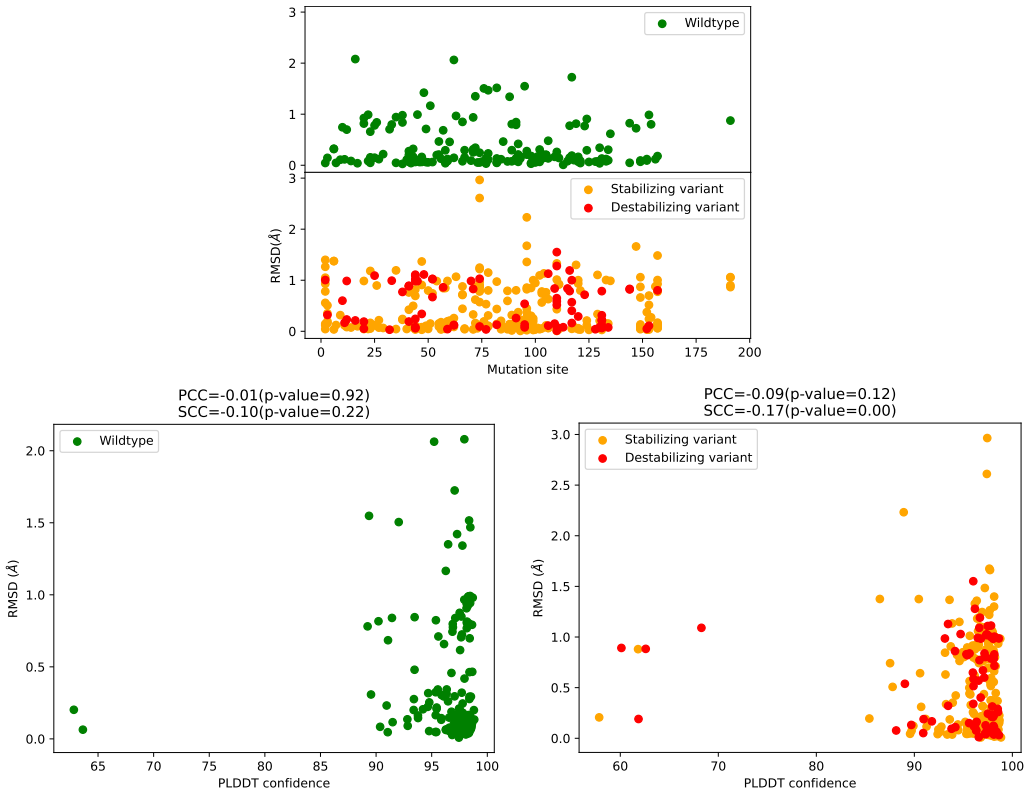

Figure 6: Top panel: Side chain analysis at mutation site. Bottom panel: Correlation analysis between PLDDT confidence and side-chain RMSD ($\AA$) at mutation site. No correlation is observed.

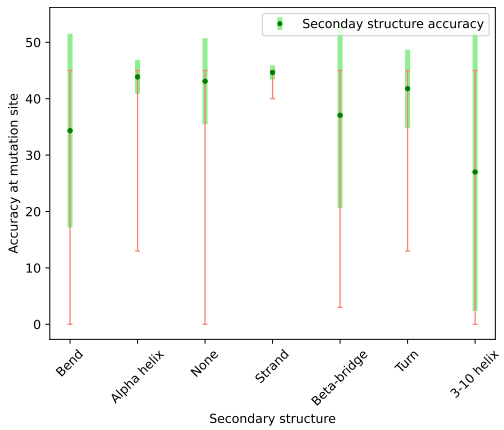

Figure 7: Secondary structure analysis at mutation site.

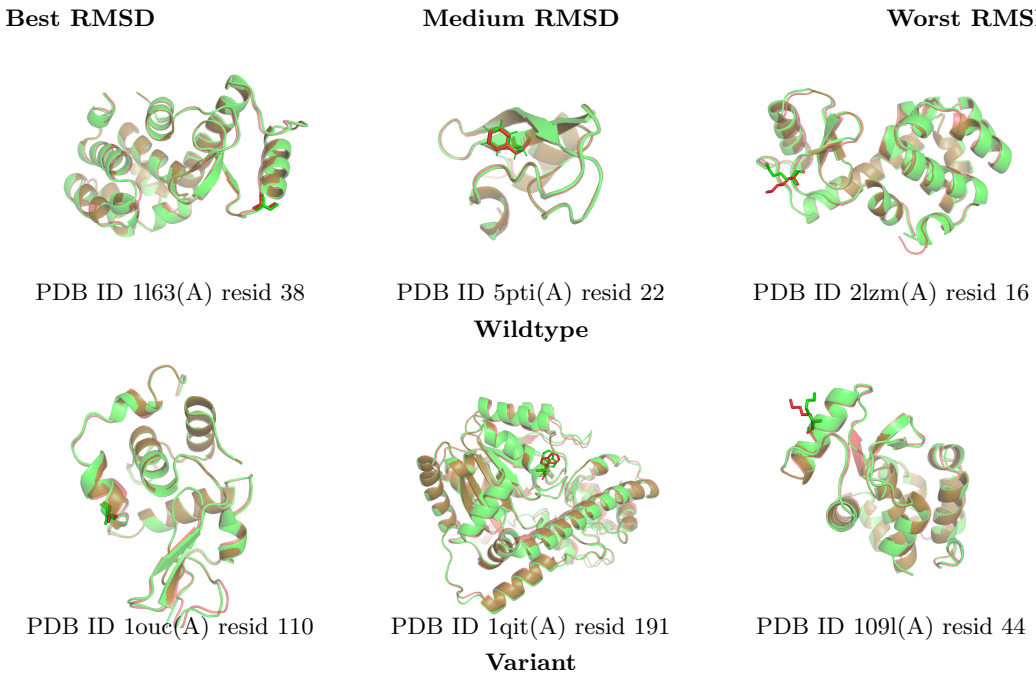

Figure 8: Local side chain analysis at mutation site. Green and red structure shows the ground-truth and top-ranked predicted structure, respectively. The stick and dot representation is used to highlight the side chain at the mutation site.

Acknowledgment

This work is supported in part by NSF Grant No. 1763233. Computations were run in part on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu). This material is additionally based upon work supported by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank Dr. Debswapna Bhattacharya for familiarizing us with AlphaFold2-advance.

References

[1] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[2] J. C. F. de Winter, S. D. P. Gosling, and J. Jeff. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273–290, 2016.

[3] R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend. A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39(Database):D411–D419, November 2010.

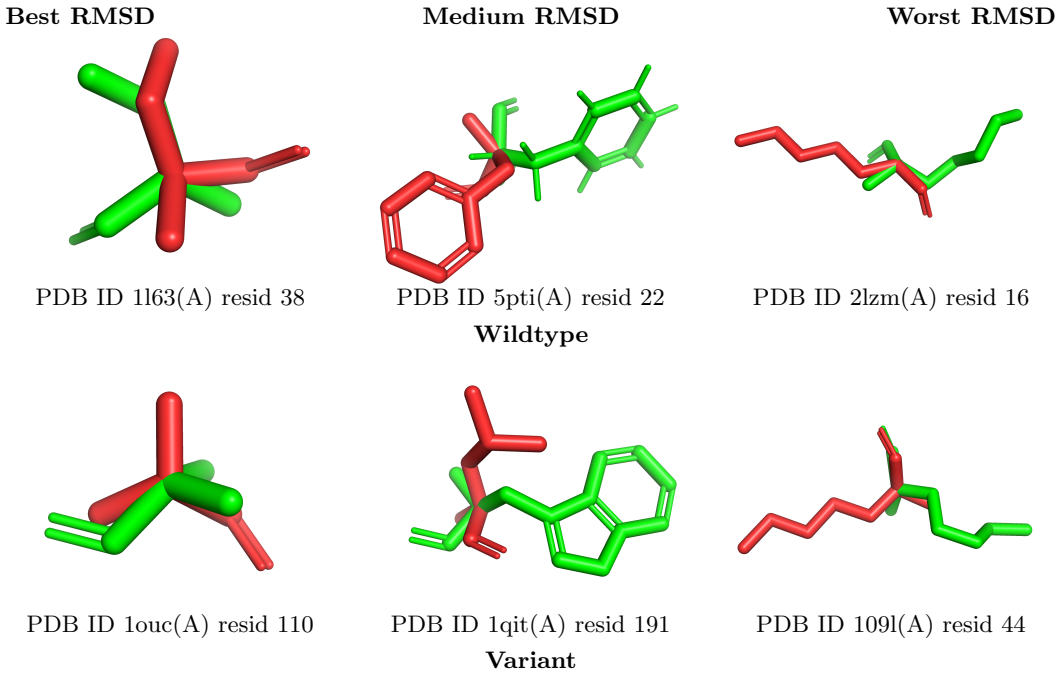

Figure 9: Local side chain analysis at mutation site. Green and red structure shows the ground-truth and top-ranked predicted side chains at the mutation site, respectively. The stick and dot representation is used.

[4] J. Jumper, R. Evans, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

[5] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[6] Yunqi Li and Jianwen Fang. PROTS-RF: A robust model for predicting mutation-induced protein stability changes. *PLoS ONE*, 7(10):e47247, October 2012.

[7] A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.*, 26(6):656–657, 1972.

[8] Milot Mirdita, Sergey Ovchinnikov, and Martin Steinegger. Colabfold - making protein folding accessible to all. *bioRxiv*, 2021.

[9] Kliment Olechnovič, Bohdan Monastyrskyy, Andriy Kryshtafovych, et al. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics*, 35(6):937–944, 2019.

[10] Fabrizio Pucci, Katrien V Bernaerts, Jean Marc Kwasigroch, and Marianne Rooman. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, 34(21):3659–3665, April 2018.

[11] Naomi Siew, Arne Elofsson, Leszek Rychlewski, and Daniel Fischer. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 2000.

[12] K. Tunyasuvunakool, J. Adler, Z. Wu, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596:590—596, 2021.

[13] J. Xu, M. McPartlon, and J. Lin. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Mach Intel*, 3:601–609, 2020.

[14] Adam Zemla. Lga: a method for finding 3d similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003.

[15] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.