Application of ESMFold and AlphaFold for Prediction of Protein Stability Upon Mutation

Jack Ringer
Computer Science
University of New Mexico
jrin42@unm.edu

***Abstract*—Over the past few years there have been great developments in regards to the protein-folding problem, with DeepMind's AlphaFold [1] achieving a massive breakthrough in the challenging CASP14 competition [9]. Despite the breakthrough there remain challenges closely related to the protein folding problem that AlphaFold has not addressed, including the prediction of mutational effects [6]. Many believe that methods adapted from natural language processing, such as Meta AI's ESM model [10], can address these gaps [3]. This work investigates how predictions from ESMFold compare to AlphaFold in predicting protein stability change using the structure-based stability predictors DDGun [13] and ACDC-NN [2]. These methodologies are investigated using a dataset of 184 unique proteins and 3428 point-mutations. In addition, predicted structures of mutants from ESMFold are used to explore potential improvement of $\Delta\Delta G$ predictions. Our results indicate that computationally-generated wild-type structures can improve the prediction accuracy of DDGun and ACDC-NN over using purely sequence-based information. However, we find no significant difference in stability prediction when using either AlphaFold or ESMFold structures, and no improvement in prediction accuracy when incorporating mutant-structures from ESMFold as additional input. Our results highlight the need for further research on protein stability prediction methods and we offer several directions on future work. Code and additional information may be found at the following GitHub repository: https://github.com/Jack-42/BIOM505_Spring2023.**

I. Introduction

Single-point mutations occur when a single nucleotide base pair is swapped with another in the DNA or RNA sequence of an organism's genome. Although many point mutations are benign, they can cause functional changes in proteins which can lead to disease and other serious health issues [15].

The unfolding free energy difference ($\Delta\Delta G$, measured in kcal/mol) between a wildtype (most typical) and mutant protein ($\Delta\Delta G = \Delta G_{wildtype} - \Delta G_{mutant}$) is a metric to quantify how a mutation will impact protein stability [13]. An important property of $\Delta\Delta G$ is antisymmetry, for which the change in free energy in a direct variant (wild-type to mutant) is equal and opposite the change in free energy in the reverse variant ($\Delta\Delta G_{MW} = -\Delta\Delta G_{WM}$) [14]. Given that a protein's function depends on its stability, determining the effects of point mutations on protein stability can provide insight into the causes of disease and inform drug development [17].

[1]There are two versions of AlphaFold, AlphaFold 1 and AlphaFold 2. Within this work all mentions of "AlphaFold" refer to the more recent AlphaFold 2 model.

While in vivo and in vitro experimental techniques are the gold standard for evaluating the effects of protein variants, experimental validation is costly and the resources required to experimentally determine these effects for all known variants is prohibitively expensive [21]. Computational methods help us to overcome this limitation by predicting the impacts of different variants in silico.

The protein folding problem recently had a major breakthrough with AlphaFold, which achieved accuracy competitive with experimental methods in the CASP14 competition [9]. Despite AlphaFold's success in protein structure prediction, the model itself is not validated for mutation prediction and, in particular, is not able to tell whether a point mutation is destabilizing or not [6]. While AlphaFold may be modified to predict the effect of point mutations (e.g., see [7], [1]), some experts believe that predictive methods adapted from fields such as natural language processing may prove more useful for determining the effects of mutations [3]. Meta's ESM (Evolutionary Scale Modeling) approach, which takes advantage of transformer models, is particularly promising [10].

Concurrently, there have been developments in protein-stability prediction models. Although many computational methods exist, within this work the tools DDGun [13] and ACDC-NN [2] are selected to generate stability predictions using structures from AlphaFold and ESMFold. The use of these tools is motivated by the fact that both are structural-based, both were designed to obey the antisymmetry property, and both were found to achieve the best performance on a novel dataset in a recent survey of protein stability prediction methods [14].

II. Methods

A. Stability data collection and cleaning

The dataset of $\Delta\Delta G$ experiments was assembled by combining entries from FireProtDB [20] and ThermoMutDB [25]. These databases contain experimental $\Delta\Delta G$ values for thousands of mutations across hundreds of wild-type proteins. Each wild-type protein is uniquely identified by it's UniProt identification [5].

Prior to generating structures and making predictions from the proteins in these databases, it was necessary to remove irrelevant, inconsistent, and duplicated entries. Irrelevant entries included those that were not point mutations (i.e., multiple

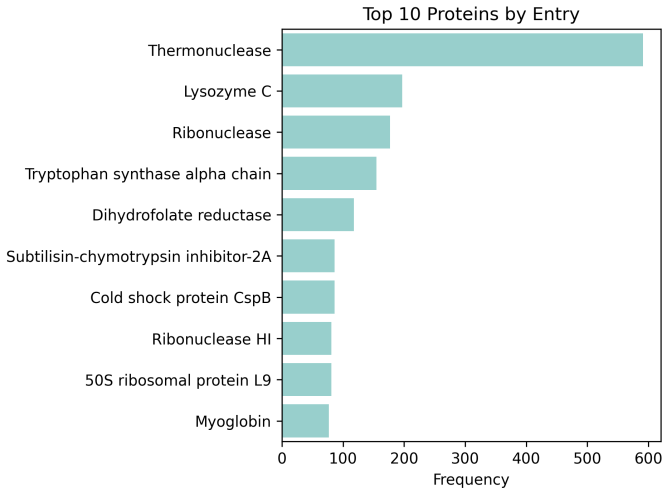

Fig. 1. Top 10 most common wild-type proteins represented in the final dataset.

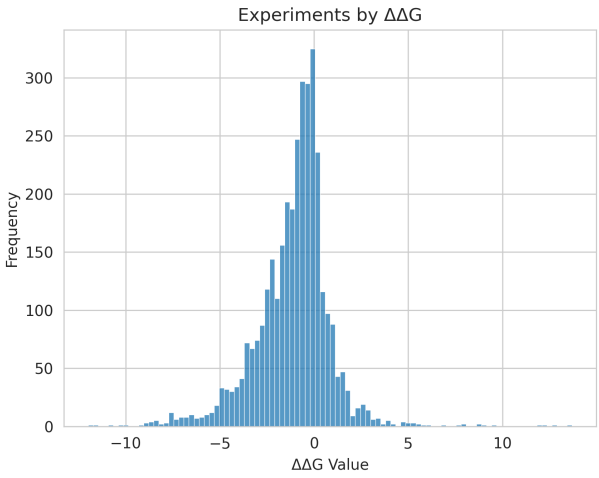

Fig. 2. Histogram of experimental $\Delta\Delta G$ values in the final dataset. Positive $\Delta\Delta G$ is considered to be stabilizing whereas negative $\Delta\Delta G$ is destabilizing.

mutations). Entries which were missing values in critical fields (chain, mutation, UniProt ID, and $\Delta\Delta G$) along with those with inconsistent mutation codes were removed [2]. All $\Delta\Delta G$ values in FireProt were multiplied by -1 to ensure consistency, as $\Delta\Delta G > 0$ is considered stabilizing by Thermomut, DDGun, and ACDC-NN whereas $\Delta\Delta G < 0$ is considered stabilizing by FireProt [20].

Duplicates were also accounted for in the data cleaning process. An entry is considered to be a duplicate of another when both entries have identical UniProt IDs and identical mutation information (position and mutated residue). Importantly, for each set of these duplicates *only the first entry is kept*. Note that the average (or median, weighted average, etc) of $\Delta\Delta G$ values for duplicates is not used. This is because different $\Delta\Delta G$ values for the same mutation could be found for a variety of reasons (such as differences in environmental conditions or random errors) which could not be reasonably accounted for using the data from ThermoMut and FireProt. After removing duplicate entries from each individual database, entries from ThermoMut that are duplicates of FireProt are removed (this choice was arbitrary, we also could have filtered FireProt entries that were duplicated from ThermoMut). The two (now unique) databases were then combined to create a final dataset.

Due to the computational expense involved in both structure prediction and profile generation, proteins with 400 or less residues were kept. Some unanticipated errors with generated profiles of some wild-type proteins also caused the dataset to be reduced (see section II-C).

The final dataset consisted of 3428 mutations and 184 unique wild-type proteins. It is worth noting that not all wild-type proteins are equally represented within the database, as there are a select few proteins which account for hundreds of entries whereas many wild-types only have one or two mutation entries (see Figure 1). Figure 2 provides a histogram of experimental $\Delta\Delta G$ values in the dataset. Additional visualizations are provided in the Appendix (Section VI).

[2]Examples of inconsistent entries include those where the mutation position was greater than the protein's sequence length, the indicated wild-type residue was not present at the indicated position in the wild-type sequence, and codes which provided more than one mutant amino acid. It is worth noting that many of these problematic entries came from ThermoMutDB. Other works have noted some of the problems with ThermoMut, for example see [8].

B. Structure Prediction

Both AlphaFold and ESMFold are used within this work to generate structure-based predictions. AlphaFold and ESMFold are computational methods which predict a protein's 3D structure given only its sequence. Both methods are based upon deep learning approaches and were trained on hundreds of thousands of experimental structures. The paragraphs below provide a more in-depth overview of these two methods and how they differ.

AlphaFold views a folded protein as a spatial graph, where nodes correspond to residues and edges connect residues close in proximity [22]. The model consists of a system of sub-networks which come together to form an end-to-end model. In generating a prediction, the input protein sequence is first used to generate a multiple sequence alignments (MSAs) using external databases and tools. These MSAs compare sequences of similar proteins and identify regions of similarity. In conjunction with this step, pair-representations are constructed from the input protein sequence. These pair-representations capture information relating each residue in the protein sequence [9]. The constructed MSA and pair-representation are then used as input to two attention-based modules. These modules iteratively pass information between one another in order to refine the relationships between amino acid residues within the protein (shown as the green array in Figure 3) along with the relationships between each amino acid and the related proteins found by sequence alignment (red array in Figure 3). The refined representations are then fed to a structure module which calculates the predicted 3D structure [9].

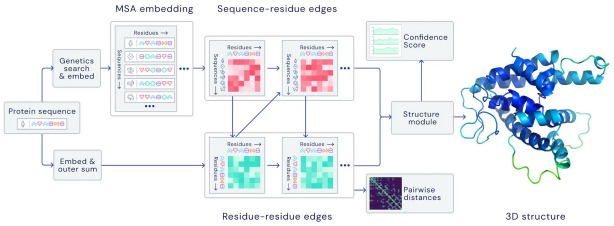

Fig. 3. Overview of AlphaFold. The two attention-based modules can be seen in the middle of the figure. Taken from [22].

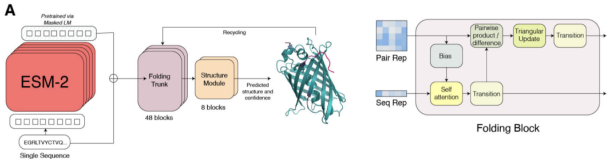

Fig. 4. Overview of ESMFold. Taken from [10].

ESMFold consists of an ESM-2 language model followed by a folding trunk and structure module (shown in Figure 4). The ESM-2 language model is trained by treating the structure-prediction problem as a masked language modeling objective [10]. The masked language modeling objective is a common pre-training technique for large language models where the model must predict masked words within a sentence given the context of the rest of the sentence (e.g., predicting "dog" in the sentence "I took my ' ' to the park today"). In the context of proteins, ESM-2 is trained to predict the identity of randomly masked amino acids given the rest of the sequence as context [10]. The pre-trained ESM-2 model is then used within ESMFold to extract representations of an input protein sequence. These representations are fed to a folding trunk and structure module, which use these representations to construct a predicted protein structure. Similar to how AlphaFold uses an iterative process to refine its own internal representations, ESMFold uses a "recycling" process before providing a final output. Perhaps the most noteworthy difference with ESMFold is that the model does not rely on generating MSAs in order to make a prediction, allowing for significant improvements in inference time over AlphaFold [10].

In addition to pure structure prediction, both models also provide an estimate of their prediction confidence using predicted local distance difference test (pLDDT) scores. The pLDDT score provides an estimate of how well a prediction would agree with an experimental structure based on the local distance difference test [9], [11]. Scores range from 0 to 100, with values closer to 100 indicating high degrees of confidence whereas those closer to 0 indicate low confidence. In general, a pLDDT score greater than 90 is considered a highly confident prediction, between 70 and 90 is confident, and below 70 is low confidence [23].

Within the context of this work, both ESMFold and AlphaFold are used to compute the predicted 3D structures of wild-type proteins. Given the expense of running AlphaFold, we opted to use publicly available predictions from the AlphaFold Protein Structure Database [23]. For ESMFold the publicly available code-base is used to generate predictions using the default parameters [18]. In regards to predictions of mutant-structures, only ESMFold was used as the AlphaFold Database does not contain structures for mutant proteins and (again) AlphaFold was found to be too computationally expensive to make predictions using available computing resources. GPU Computing resources at at the University of New Mexico's Center for Advanced Research Computing (CARC) were used to generate predictions for ESMFold, using a single node equipped with an Nvidia Tesla K40M.

C. Stability Prediction

Within this work we apply the tools ACDC-NN and DDGun to estimate $\Delta\Delta G$ values for single-point mutations. ACDC-NN is a data-driven approach using neural network architectures whereas DDGun is an untrained method that uses evolutionary and statistical potentials. Both of these methods are capable of taking in the protein's wild-type structure, protein profile (see paragraph below), and the mutation information (wild-type residue, mutant residue, position) to generate a predicted $\Delta\Delta G$ value for a given mutation.

In addition to structural information, both ACDCN-NN and DDGun rely on evolutionary information derived from alignment files generated by HH-Blits [16]. HH-Blits is part of the open-source software HH-Suite and generates alignment files by aligning hidden Markov model (HMM) profiles during an iterative search process to find analogous homologs [19]. These HMM profiles are provided by the Uniclust database, which generates profiles by clustering the UniProt database into clusters of globally alignable sequences [12]. Given the computational expense of running alignment searches for hundreds of protein sequences, we use computational resources CARC to perform sequence alignments using HH-Blits.

After generating the alignment file, the contained alignments are converted into a protein profile. For a sequence of length N, the alignment file is used to generate an Nx20 profile matrix P, where $P(i, j)$ corresponds to the frequency of residue j at sequence position i [2], [13]. P is then used as the source of evolutionary information for both DDGun and ACDC-NN.

As well as using wild-type structures and profiles as input, ACDC-NN offers the capability to also include mutant-structures and profiles as additional input information. For this work we use this capability to investigate if mutational structures generated by ESMFold can improve prediction accuracy in comparison to only using wild-type structures. DDGun and ACDC-NN can also generate predictions using only sequence information (i.e., the wild-type protein sequence, mutation information, and profile as input), and we use these as base-cases to test if computationally-generated structures can potentially improve prediction accuracy in instances where only sequence information is available (e.g., proteins which do not have experimentally-determined structures, or cases where experimental structures do not have complete coverage).

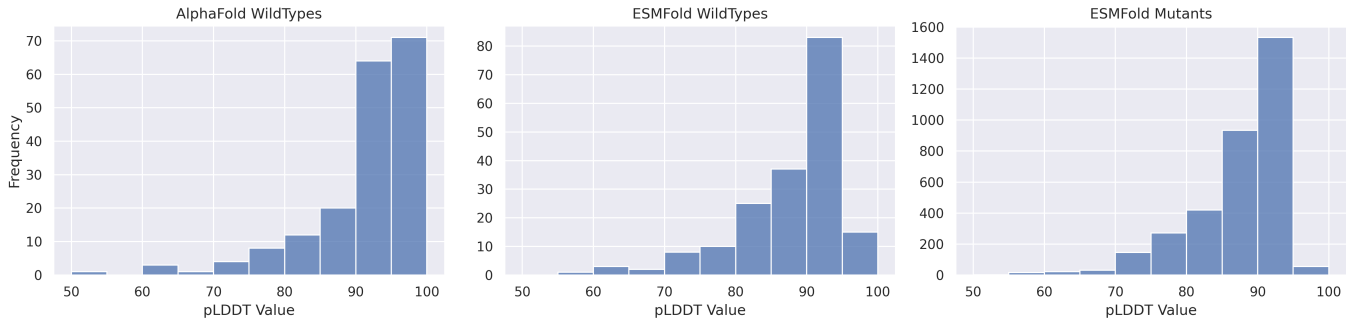

Fig. 5. Histogram of average pLLDT values for AlphaFold on wildtypes (left) and ESMFold on wildtypes (middle) and mutants (right). Average pLLDT values are calculated on a per-protein basis.

III. Results

A. Structure Prediction

As shown by Figure 5 and Table I, the average pLDDT scores for both AlphaFold and ESMFold were both above 0.87 and scores fell into the "confident" or "highly confident" range in a majority of cases. AlphaFold had higher confidence over ESMFold in predicting wild-type structures, particularly due to the fact that a far larger number of AlphaFold's predictions had pLDDT scores > 95 in comparison to ESMFold. In comparison to predictions for wild-types, ESMFold gave lower pLDDT scores for predictions on mutant structures (although the median score still falls in the "confident" range for the mutant case).

Model	Case	Mean pLDDT	Median pLDDT
AlphaFold	Wild	91.13	93.51
ESMFold	Wild	87.78	90.51
ESMFold	Mutant	87.33	88.50

TABLE I
OVERVIEW OF PLDDT VALUES FOR ALPHAFOLD AND ESMFOLD

Figure 7 shows four sample wild-type predictions from AlphaFold (top) and ESMFold (bottom), along with the corresponding Uniprot IDs. Images were generated using Py3DMol [24].

B. Stability Prediction

An overview of results is shown in Table II. Here the Pearson correlation coefficient (r), mean-absolute error (MAE), and root-mean square error (RMSE) are used to measure the quality of predicted $\Delta\Delta G$ values against ground-truth (i.e., experimental) values [3].

As can be seen from Table II, there was little difference in prediction quality when using AlphaFold structures compared to ESMFold. In comparison to using only sequence-based information, the use of computational structures improved prediction quality for both DDGun and ACDC-NN (although the effect for ACDCN-NN was small).

[3]Note that confidence intervals are not included for these metrics because (due to unavoidable limitations) our dataset excludes much of the structure space of all .pdbs and thus it is difficult to view our dataset as a "sample" of some larger population

Method	Structures From	r	MAE	RMSE
DDGun-Seq	-	0.4084	1.3178	1.9211
DDGun	AlphaFold	0.4617	1.2489	1.8631
DDGun	ESMFold	0.4611	1.2495	1.8636
ACDC-NN-Seq	-	0.4593	1.2406	1.8547
ACDC-NN	AlphaFold	0.4754	1.2299	1.8452
ACDC-NN	ESMFold	0.4765	1.2307	1.8442
ACDCN-NN-Mut	ESMFold	0.4721	1.2457	1.8606

TABLE II
ASSESSMENT OF EACH METHODOLOGY ON OUR DATASET. "-SEQ" IS USED TO INDICATE METHODS WHICH ONLY USED SEQUENCE-BASED INFORMATION, AND "-MUT" WHERE STRUCTURES OF MUTANT PROTEINS WERE USED AS ADDITIONAL INFORMATION. BOTH THE RMSE AND MAE ARE EXPRESSED IN (KCAL/MOL).

Figure 6 provides $\Delta\Delta G$ parity plots (i.e., experimental vs predicted values) using the methodologies investigated in this work. In general $\Delta\Delta G$ predictions were compressed around 0, with all methodologies struggling to make predictions for larger experimental values of $|\Delta\Delta G|$.

C. pLDDTs and Prediction Quality

Here we present our findings on the relationship between the average pLDDT values of structures and the $\Delta\Delta G$ prediction error for both DDGun and ACDC-NN. Given that pLDDT values are an estimate of structure quality, one may expect that higher pLDDT values positively correlate with the accuracy of $\Delta\Delta G$ predictions.

Interestingly, such a relationship was not seen. Table III provides the correlation between average pLDDT score for each structure and the RMSE for corresponding $\Delta\Delta G$ predictions. One would expect there to be a negative correlation between error (RMSE) and pLDDT value, but in fact no significant relationship was found between the two values as all p-values are above 0.05.

Method	Structures From	r	p-value
DDGun	AlphaFold	0.0757	0.3071
DDGun	ESMFold	0.0944	0.2023
ACDC-NN	AlphaFold	0.0920	0.2140
ACDC-NN	ESMFold	0.1070	0.1482
ACDCN-NN-Mut	ESMFold	0.0278	0.1041

TABLE III
CORRELATION BETWEEN PLDDT VALUES AND CORRESPONDING RMSE OF $\Delta\Delta G$ PREDICTIONS.

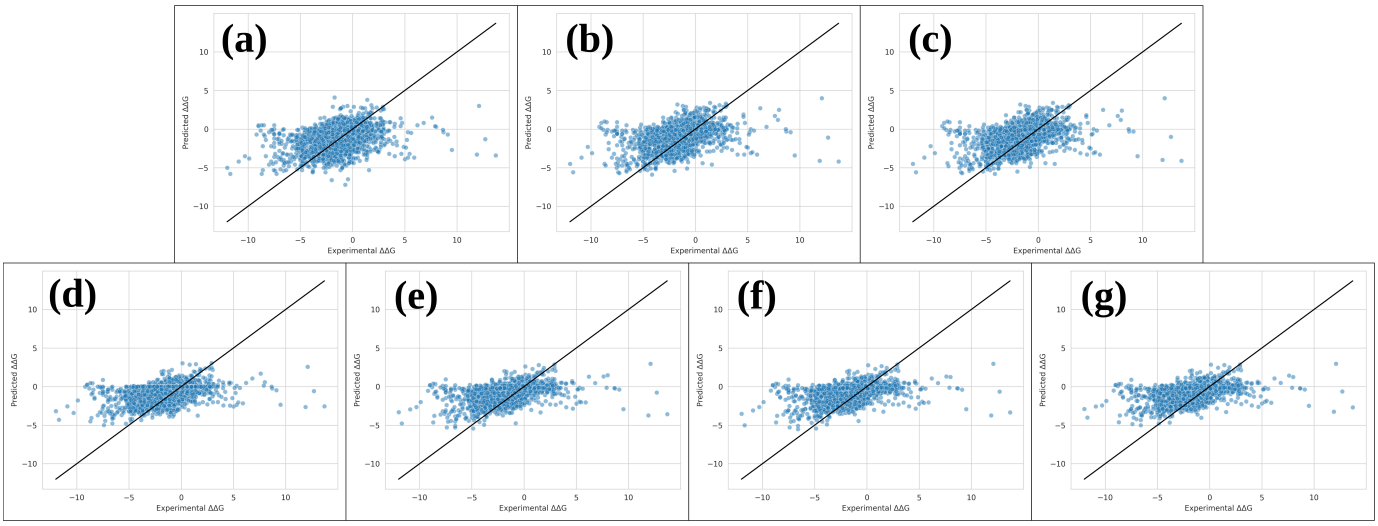

Fig. 6. Parity plots showing experimental $\Delta\Delta G$ values (x-axis) vs. predicted $\Delta\Delta G$ values (y-axis) for each methodology. Perfect predictions would follow the black line. The labels for each plot correspond to methodology as follows: **(a)** DDGun-Seq **(b)** DDGun using AlphaFold wild-type structures **(c)** DDGun using ESMFold wild-type structures **(d)** ACDC-NN-Seq **(e)** ACDC-NN using AlphaFold wild-type structures **(f)** ACDCN-NN using ESMFold wild-type structures **(g)** ACDCN-NN using ESMFold wild-type and mutant structures

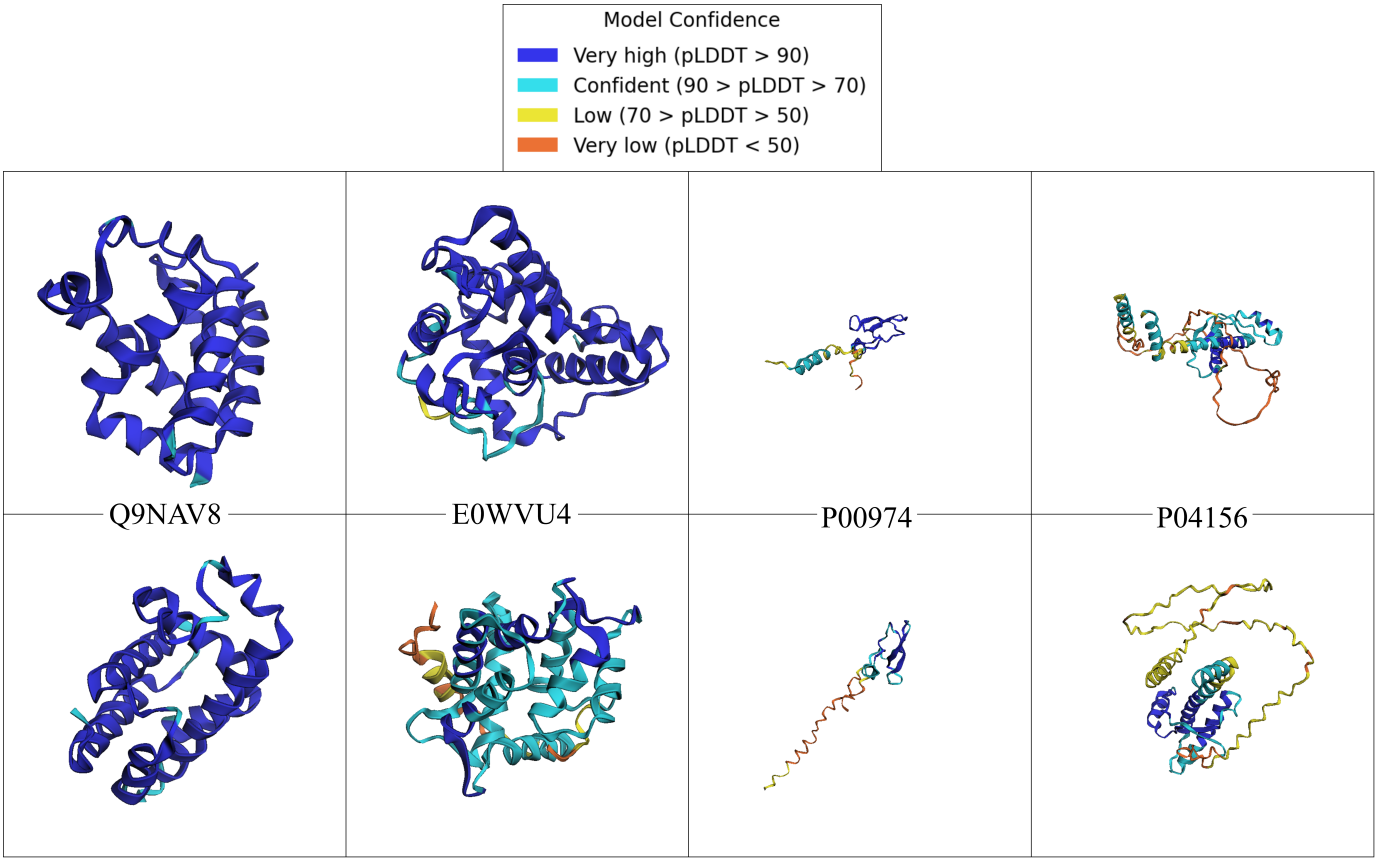

Fig. 7. Example wild-type predictions from AlphaFold (top row) and ESMFold (bottom row). Two confident predictions (left) and two less-confident predictions (right) were intentionally highlighted. The UniProt ID of the corresponding sequence is indicated for each prediction pair.

IV. Discussion

Overall, in this work we investigated how AlphaFold and ESMFold could be applied to predict protein stability from a dataset of 184 proteins and 3428 point-mutations using DDGun and ACDC-NN. Our results indicate that structures from both AlphaFold and ESMFold can be applied towards predicting $\Delta\Delta G$ values using ACDC-NN and DDGun. No significant difference was found between using wild-type structures from AlphaFold or ESMFold. Incorporating computational structures improved predictions over using purely sequence-based information (although for ACDCN-NN the improvement was small). Adding predicted mutant structures from ESMFold as additional information for ACDC-NN was found to not be beneficial.

Given that AlphaFold is considered to be more accurate than ESMFold [4], it is surprising that there was not a larger difference between the two methods when applied towards predicting protein stability. Considering this fact as well as the lack of correlation between prediction error and pLDDT value, one may conclude the particular errors from computational structures did not have a significant impact on final $\Delta\Delta G$ prediction. There are several potential explanations for why this may be the case. One explanation is that DDGun only extracts from structures differences in interaction energy (measured by the Bastolla statistical potential) and the relative solvent accessibility of residues, and so it may be possible that errors from AlphaFold and ESMFold did not significantly impact the calculation of these measures [13]. For ACDC-NN it is more difficult to say, as this method uses neural networks to extract structure features and these networks are notoriously considered "black-boxes" (i.e., it is is not clear how they reach a particular decision or what exact features they extract). However, it's worth noting that ACDC-NN was pre-trained on predictions from DDGun, and so it seems likely that the model would replicate some of the biases of DDGun [2]. An additional hypothesis is that sequence-based information, in particular the evolutionary profiles used by both DDGun and ACDC-NN, provide much of the valuable information for estimating $\Delta\Delta G$ values and thus deflate the importance of structure accuracy for $\Delta\Delta G$ prediction. After all, while structural information did improve $\Delta\Delta G$ predictions over purely sequence-based methods, the improvement was not necessarily massive (especially for ACDC-NN). However, all of these hypotheses remain untested, and additional work would need to be done to accept or reject each of them.

The lack of improvement in $\Delta\Delta G$ prediction accuracy when incorporating mutant structures was somewhat surprising as well. It is possible that mutant-structures from ESMFold were not particularly accurate and thus hindered the accuracy of ACDC-NN. There is some evidence of this from our results, as both the MAE and RMSE slightly increase over the wild-type ESMFold case, when at worst incorporating mutant structures should provide the same level of performance. However, this hypothesis is at odds with our finding that there was no significant relationship between pLDDT value and prediction accuracy when using mutant-type structures (see Table III). Additional work would be necessary to explore this area.

Many of our results replicate several findings from a 2022 survey of protein stability prediction methods [14]. In particular, we replicate the findings that computational structures modestly improve prediction performance over sequence-based methods (although the report used RosettaFold rather than AlphaFold or ESMFold) and that computational predictions of $\Delta\Delta G$ cluster around 0 (i.e., computational methods over-predict little change in stability). Our metrics do not exactly mirror those found in [14], but this is not surprising given that a different dataset was used. As well, the authors in [14] do not find a significant difference between using computational and experimental structures, which is related to our finding that the lower accuracy of ESMFold structures (in comparison to AlphaFold) did not have a significant impact on final $\Delta\Delta G$ prediction accuracy.

Our results (as well as those from the literature) highlight that significant developments need to be made in protein stability prediction before computational methods can be reliably deployed in practical settings. Given the challenges associated with predicting protein stability and the limited availability of experimental data, future work may include improving theoretical understanding of the relationship between protein stability and external factors and/or constructing model architectures that are able to learn more from relatively smaller datasets. Concurring with [14], it seems one of the biggest steps towards improving existing frameworks will be properly modifying the distribution of $\Delta\Delta G$ predictions to prevent compression around 0.

V. Acknowledgement

Much thanks to my advisor Vincent Metzger for his guidance and input over the course of this project. We'd also like to acknowledge the UNM Center for Advanced Research Computing (CARC) for allowing us to use computational resources for several aspects within this project.

References

[1] Mehmet Akdel et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29:1056 – 1067, 2021.

[2] S Benevenuta, C Pancotti, P Fariselli, G Birolo, and T Sanavia. An antisymmetric neural network to predict free energy changes in protein variants. *Journal of Physics D: Applied Physics*, 54(24):245403, 2021.

[3] Ewen Callaway. After alphafold: protein-folding contest seeks next big breakthrough. *Nature*, 613(7942):13–14, Dec 2022.

[4] Ewen Callaway. Alphafold's new rival? meta ai predicts shape of 600 million proteins. *Nature*, 611(7935):211–212, Nov 2022.

[5] The Uniprot Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 46:2699 – 2699, 2018.

[6] DeepMind. Frequently asked questions. https://alphafold.ebi.ac.uk/faq.

[7] Matteo P. Ferla, Alistair T. Pagnamenta, Leonidas Koukouflis, Jenny C. Taylor, and Brian D. Marsden. Venus: Elucidating the impact of amino acid variants on protein function beyond structure destabilisation. *Journal of Molecular Biology*, page 167567, Mar 2022.

[8] Shahid Iqbal, Fuyi Li, Tatsuya Akutsu, David Benjamin Ascher, Geoffrey I. Webb, and Jiangning Song. Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. *Briefings in bioinformatics*, 2021.

[9] John M. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[11] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, Aug 2013.

[12] Milot Mirdita, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45(D1):D170–D176, Nov 2016.

[13] Ludovica Montanucci, Emidio Capriotti, Yotam Frank, Nir Ben-Tal, and Piero Fariselli. Ddgun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics*, 20, 2019.

[14] Corrado Pancotti, Silvia Benevenuta, Giovanni Birolo, Virginia Alberini, Valeria Repetto, Tiziana Sanavia, Emidio Capriotti, and Piero Fariselli. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics*, 23(2), Jan 2022.

[15] Medline Plus. How can gene mutations affect health and development?: Medlineplus genetics. https://medlineplus.gov/genetics/understanding/mutationsanddisorders/mutationscausedisease/, Mar 2021.

[16] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature Methods*, 9(2):173–175, Dec 2011.

[17] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio. Inps-md: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics*, 32(16):2542–2544, Apr 2016.

[18] Tom Sercu et al. Evolutionary scale modeling. https://github.com/facebookresearch/esm, Oct 2022.

[19] Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1), Sep 2019.

[20] Jan Stourac et al. Fireprotdb: database of manually curated protein stability data. *Nucleic Acids Research*, 49:D319 – D324, 2020.

[21] Alexey Strokach, Tian Yu Lu, and Philip M. Kim. Elaspic2 (el2): Combining contextualized language models and graph neural networks to predict effects of mutations. *Journal of Molecular Biology*, 433(11):166810, May 2021.

[22] The AlphaFold Team. Alphafold: a solution to a 50-year-old grand challenge in biology. https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology, Nov 2020.

[23] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021.

[24] Aaron Virshup. avirshup/py3dmol. https://github.com/avirshup/py3dmol, Apr 2016.

[25] Joicymara S. Xavier et al. Thermomutdb: a thermodynamic database for missense mutations. *Nucleic Acids Research*, 49:D475 – D479, 2020.

VI. Appendix

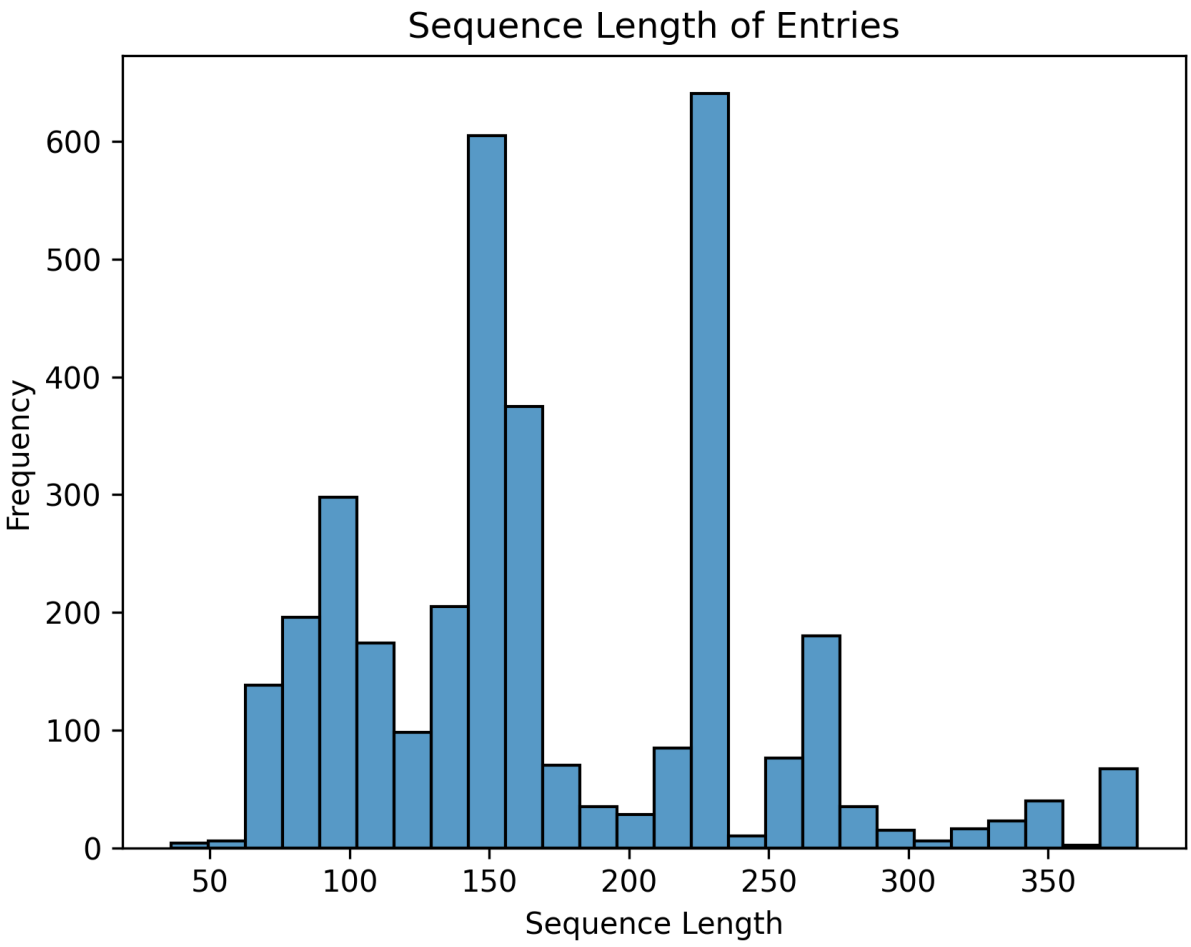

Fig. 8. Histogram of protein sequence-lengths in our dataset.

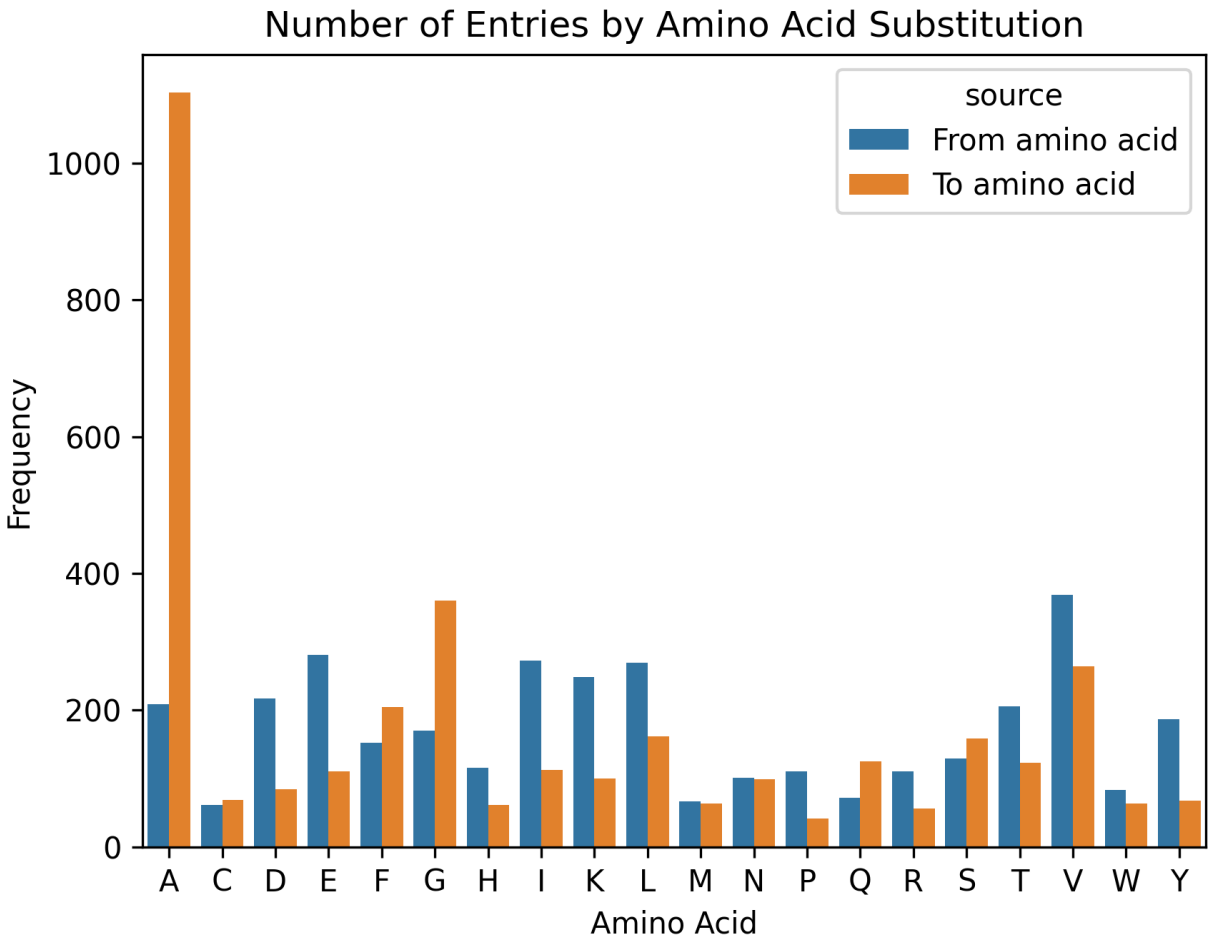

Fig. 9. Number of substitutions (both to and from) for each amino acid in our dataset.

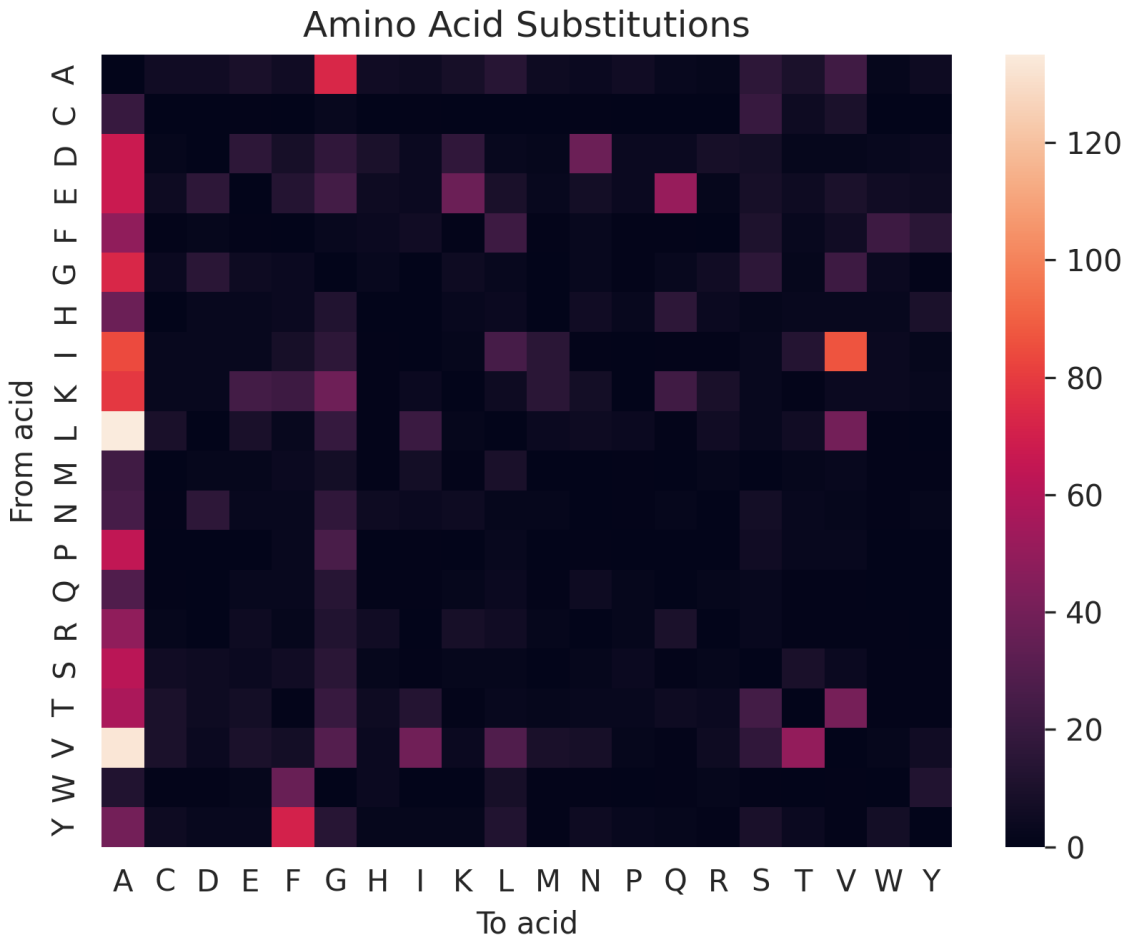

Fig. 10. Heatmap of amino acid substitutions in our dataset.