



DEEP MUTATIONAL SCANNING OF THE GREEN FLUORESCENT PROTEIN

This project investigates the impact of singular and multiple mutations on the DMS score of green fluorescent protein (GFP). The analysis begins by examining the effects of mutation position and amino acid substitutions on the DMS scores of single mutants. Additionally, the study explores the physiological properties of neighboring amino acids, to identify causality for a lowered DMS score. Moreover, this project delves into the phenomenon of epistasis in GFP, investigating the interplay of multiple mutations and identifying buried amino acid residues that exhibit enhanced epistatic effects. Furthermore, the influence of the number of mutations on the DMS score is analyzed, leading to the development of weighted rankings to identify mutations with positive epistatic effects. Finally, PyRosetta is utilized to predict the difference in free energy (ΔG) between the unfolded and folded protein. The findings provide valuable insights into the factors influencing the DMS score in GFP mutants and have implications for protein engineering and design.

Roman Kurley, Lisa Duttenhöfer, Rebecca Ress
and Tianxin Angela Ma
Supervisor: Prof. Dominik Niopek, Jan
Mathony and Benedict Wolf

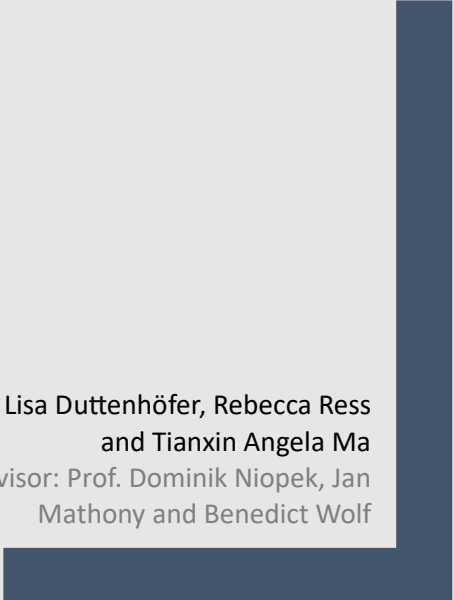


Table of Contents

Introduction	2
Methods.....	3
Single mutations	3
Epistasis.....	3
Mutants with sequential mutations	3
Structure Analysis	4
Ranking mutations according to their epistatic effect	4
Protein stability prediction with PyRosetta.....	5
Results.....	6
Single mutations	6
Epistasis.....	7
Mutants with sequential mutations	7
Structure Analysis	7
Ranking mutations with positive epistatic effects.....	8
Protein stability prediction with PyRosetta.....	8
Discussion.....	9
Single mutations	9
Epistasis.....	9
Mutants with sequential mutations	9
Structure Analysis	10
Ranking mutations with positive epistatic effects.....	10
Protein stability prediction with PyRosetta.....	10
Appendix.....	12
References.....	17

Introduction

The Green Fluorescent Protein (GFP) is a naturally occurring protein that exhibits a unique property: it can absorb blue light and emit green fluorescent light. GFP was first discovered in the jellyfish species *Aequorea victoria* and has since become a valuable tool in molecular biology and biomedical research (Fu *et al.*, 2015).

The structure of GFP consists of 238 amino acids arranged in a beta-barrel. Within the barrel, there is a chromophore, a small group of amino acids that are responsible for the fluorescent properties of GFP: serine, tyrosine, and glycine. Through a process called cyclization, these amino acids undergo a series of chemical reactions to produce the chromophore (Ong *et al.*, 2011).

It is not surprising, that GFP has a highly conserved 3D structure (Zimmer *et al.*, 2014) and can barely be improved when it comes to functionality and size decrease. However, having a strict stable structure is not always advantageous. In this project, the mutational effects leading to mutated GFP proteins were analyzed to first understand the physiological effects on the protein and later on determine how to improve GFP's tolerance towards mutations. This would help to understand how new modified versions of GFP can be created that are more resistant towards possibly deleterious mutations, whilst preserving normal levels of fluorescence activity.

Deep mutational scanning (DMS) is an experimental technique used to understand how genetic mutations affect the function of genes or proteins. It involves introducing mutations and measuring their impact on various traits or functions using high-throughput methods (Fowler and Fields, 2014). This approach is particularly relevant to our GFP project as it can help us explore how specific mutations in the GFP gene influence the fluorescence properties of the protein, providing insights into its structure and function.

The dataset used in this project was drawn from the paper "local fitness landscape of the green fluorescent protein" (Sarkisyan *et al.*, 2016a). Sarkisyan *et al.* used random mutagenesis to create different GFP mutants with missense mutations. These were then filtered, removing non-sense mutations and duplicates. The remaining mutants were sorted into groups depending on fluorescence activity and comparison with an RFP normalization score. In the end, a DMS score was calculated based on fluorescence intensity for every mutant.

In the first part of the project, the analysis focused primarily on single mutations. The main objective was to determine and quantify the relative impact of the mutation position versus the newly created amino acid on the resulting DMS score. Furthermore, the physiological impact of the change in amino acid to the fluorescence activity was determined by analyzing amino acid-specific parameters such as the volume of side chains (VSC), net charge index of side chains (NCISC), polarity (P1) and solvent-accessible surface area (SASA).

Moving the focus to mutants containing multiple mutations, the dataset was filtered in order to identify mutants that exhibit overlapping mutations. It was discovered that certain mutations have the ability to reverse the deleterious effects of other mutations. This observation is a consequence of epistasis, primarily observed in buried amino acid residues. It is suggested that amino acid residue interactions are the underlying cause of epistasis.

To find mutations having reasonably high positive epistatic effects, all mutations were ranked according to their ability to increase GFP's tolerance towards mutations. Ranking calculations were computed with weighted factors, and the impact of mutation count on protein fitness was included.

The stability of proteins is important as it can affect proper folding and functional performance. Protein stability can be characterized by thermodynamic stability, which refers to the resistance against denaturation, as well as kinetic stability, which measures the resistance of a protein against irreversible inactivation (Liu *et al.*, 2019). Since no experimentally determined parameters capturing stability were available in the dataset used, we utilized the software platform PyRosetta to predict a stability parameter, enabling us to investigate the relationship between function and stability in mutated GFP.

Methods

Single mutations

Roman Kurley

This project utilized Python programming for data analysis. The dataset consisted of 51,714 GFP mutants, of which only 1,083 had a single mutation, resulting in limited data for the analysis of single mutations. The project encompassed several stages of analysis to gain insights into the dataset of GFP mutants:

Data cleaning was performed initially, involving the removal of redundant columns that did not contribute crucial information. A new data frame was created, specifically isolating the single mutations for subsequent analysis.

An overview and data exploration phase followed, where the analysis was divided into two parts: positional impact and the impact of the new amino acid. Scatter plots were generated using Matplotlib to visualize the global distribution of DMS scores, providing a general understanding of the data. Statistical parameters such as mean and median were calculated to assess central tendencies. Heatmaps and scatter plots were utilized to identify relevant amino acids and positions, taking into account known GFP domains like the chromophore. Furthermore, boxplots and violin plots were employed to visualize the distribution width of DMS scores for each amino acid.

A series of statistical tests were conducted to delve deeper into the data. One-way analysis of variance (ANOVA) was utilized to identify significant differences between multiple groups, including new amino acids and positions, with their respective DMS scores. Eta-squared tests were employed to quantify the proportion of variance explained by each group. To control the family-wise error rate, a Bonferroni correction was applied. The normality assumption of the data was examined using Shapiro-Wilk tests and q-q plots. Additionally, Mann-Whitney U tests were performed to compare groups separately for new amino acids and positions. The Mann-Whitney U test, a nonparametric test, ranks the observations and assesses if the sum of ranks for one group significantly differs from the sum of ranks for the other group, thus determining if there is a significant difference between their medians. A Kruskal-Wallis test was conducted to evaluate overall differences between groups, while a Friedman test analyzed the dependency between samples.

In the neighbourhood analysis, a new dataset¹ containing various properties of each amino acid was obtained and merged with the existing data. This enabled the creation of neighbourhoods of length 7 around the mutation site, allowing for a contextual analysis of the amino acid in relation to its surrounding amino acids. Neighbourhoods without the mutation (original sequences) were also determined to compare the effects of the mutation on the neighbourhood properties, which could potentially impact the DMS score.

Finally, multiple sequence alignments were performed using Clustal Omega to analyze different GFP variants and colors of fluorescent proteins, which were taken from an extensive database on fluorescent proteins "FPbase"². The MUSCLE algorithm in the web version was employed to generate phylogenetic distance trees, providing visualizations of the evolutionary relationships between GFP variants. Due to time constraints, further exploration of the phylogenetic trees was not pursued.

Through these comprehensive analyses, the project aimed to uncover valuable insights into the GFP mutant dataset, shedding light on the positional and amino acid impacts on the DMS score, exploring neighborhood properties, and investigating the evolutionary relationships between GFP variants.

Epistasis

Mutants with sequential mutations

Tianxin Angela Ma

In order to analyze the dataset, the mutants with sequential mutations were filtered, and paths were constructed using them. However, due to the limitations of the dataset, the maximum path length that could be identified was 4. These paths included a mutant with mutation A (single mutation), a mutant with mutations A and B (double mutation), a mutant with mutations A, B, and C (triple mutation), and a mutant with mutations A, B, C, and D (quadruple mutation).

All structural information of GFP was obtained from (Chudakov *et al.*, 2010). Amino acid residues were categorized based on their orientation as either buried (in) or surface-exposed (out). Each amino acid position was assigned to one of these two orientations.

We focused on analyzing mutants with double mutations in this part. Based on the specific mutations present, these mutants were classified into three groups: "in-in" (both amino acid residues buried), "in-out" (one buried and one surface residue), and "out-out" (both residues surface-exposed). Additionally, two other mutants were considered for comparison, each containing one mutation from the double mutation mutants. If one or both of these single mutants were missing, the corresponding double mutation mutant was excluded from the analysis.

The mean of both DMS scores was calculated for every single mutation mutant pair. This mean score was considered as the expected DMS score if they are combined, neglecting epistatic effects. The mean score was then compared to the actual score obtained for the respective double mutation mutant. Cases with a higher actual score were classified as positive epistatic, indicating a positive joint effect, while cases with a lower score were considered negative epistatic.

Ranking mutations according to their epistatic effect

Lisa Duttenhöfer

Because the dataset in interest primarily contained mutants with multiple mutations, the exclusive effect of the number of mutations on the DMS score was analyzed prior to further analysis concerning epistatic relations. Building the mean DMS scores for each dataset segment with identical mutation counts and displaying them in a boxplot allowed for investigation of the effect the number of mutations had on the fitness of GFP (figure 12). By comparing the median and the quantiles of each mutation count group and performing a Mann-Whitney-U Test comparing the segments, the role of the mutation amount per mutant got more transparent for further examinations.

To obtain more specific knowledge about the existing mutations in the dataset, they were extracted and analyzed regarding their distinct role in the mutant's DMS scores. Calculating the variance of the DMS scores from all mutants containing a specific mutation with a specific mutation count and relating it to the respective number of available data in a scatterplot made it possible to compare mutations and their impact on the mutant in respect of the reliability of their data (figure 13).

In this context, the variance was used to quantify a mutation's impact on the mutant's DMS score. For example, given a group of mutants all containing a specific mutation A and a set number of mutations, the variance of the resulting DMS scores is small if A has notable impact on the DMS score, regardless of the existing mutations, and is big if the DMS scores vary despite the constant presence of A. It did not, however, show the nature (positive or negative epistatic) of the effect itself.

To characterize the mutations by their ability to stabilize negative mutations, rankings were established using several factors contributing to the mutant's DMS scores.

Next to the frequency a mutation appears within the used dataset, an approach to estimate the reliability of the calculation, a new factor was incorporated, characterizing the effect of the mutation more distinctly. The DMS score difference (DSD) is the calculated difference between the DMS score mean of all mutants containing and not containing the mutation in question, quantifying the impact of a mutation's presence in a mutant. To consider the influence of the number of mutations per mutant and the resulting decreasing DMS score, a weighted DSD was computed, considering the mean DMS score of the whole dataset in the mean calculation of the DSD. Therefore, it presents a measurement for the impact of the mutation on the DMS score while considering the distorting effect the variable number of mutations has on calculations regarding epistasis.

Because the variance calculation had to be as robust as possible, various methods were tried also considering the impact of the number of mutations. First, all mutation count groups were considered the same (var_1) (eq. 1a), later only mutants with 2 to 7 mutations were involved (var_2) (eq. 2b). var_3 weights the variances of the groups with the mean DMS scores, while var_4 considers the amount of data available of their respective group prior to computing the mean (eq. 3c+d). The different methods can be used to emphasize certain factors of the analyses.

A crucial factor of the fitness of a protein, being the stability of the same, led us to generate ΔG (free energy) predictions for each of the mutants using *PyRosetta*. To include the mutation's effect on free energy, the ΔG difference was built in the same manner as the DSD and included in a ranking calculation (eq. 4 e).

To emphasize the reliability of the data, a cutoff of ten was set for the minimal amount of available data for a mutation to be considered in the ranking. By combining the factors in various ways, a total of eleven rankings were computed, rating all 1810 mutations occurring in the dataset according to their impact on epistatic events.

Equation 1: methods used for variance calculation: var_1 and var_2 including all or part of the mutation count groups and computing the mean, var_3 and var_4 computing the weighted mean of all group's variances with either the mean DMS score (DS) or the AOD of the corresponding group.

$$\begin{aligned} (a) \quad var_1 &= \frac{\sum_{i=2}^{15} var_i}{n} \\ (b) \quad var_3 &= \frac{\sum_{i=2}^{15} var_i * meanDS_i}{n} \\ (c) \quad var_2 &= \frac{\sum_{i=2}^7 var_i}{n} \\ (d) \quad var_4 &= \frac{\sum_{i=2}^{15} var_i * AOD_i}{n} \end{aligned}$$

Equation 2: score calculations used for the rankings. DSD: DMS score difference, AOD: amount of data available in dataset; var in $score_3$ and $score_4$ can be calculated in various ways (Equation 15)

$$\begin{aligned} (a) \quad score_1 &= DSD * AOD \\ (b) \quad score_2 &= DSD_{weighted} * AOD \\ (c) \quad score_3 &= DSD * \frac{1}{var} * AOD \\ (d) \quad score_4 &= DSD_{weighted} * \frac{1}{var} * AOD \\ (e) \quad score_5 &= DSD_{weighted} * \frac{1}{var} * AOD * DGD \end{aligned}$$

Protein stability prediction with PyRosetta

Rebecca Ress

To predict the Gibbs free energy differences (ΔG) between the folded and unfolded states of GFP, we employed the PyRosetta-4 version. The crystal structure of the GFP protein (PDB ID: 2wur) was obtained from the Protein Data Bank³. The structure was parsed using the PDBParser module from the Biopython library and then converted into a pose object using the PyRosetta library. To predict the ΔG values, we followed the "Cartesian $\Delta\Delta G$ " protocol initially described by Park, Bradley et al. (2016) and adapted it for PyRosetta. The "Cartesian $\Delta\Delta G$ " protocol is part of the Rosetta software and combines relaxation techniques and energy minimization to improve the prediction of Gibbs free energy (ΔG) by considering both conformational and intermolecular interactions. We used the FastRelax algorithm implemented in PyRosetta, which combines molecular dynamics and energy minimization to enhance the stability and optimize the energetic state of the protein. The "ref2015" scoring function was used. To investigate the effects of amino acid mutations on protein stability, we introduced the mutations into the protein's pose. The code employs the MinMover algorithm to perform energy minimization and optimize the protein's energy. To specify which residues should be movable during this process, a MoveMap object is created and configured. By setting the backbone (bb) and side chain (chi) atoms of all residues as movable, the algorithm can adjust their positions to find a more favorable energy state for the protein. The $\Delta\Delta G$ values were computed as the difference between the ΔG of the wildtype and mutated pose objects. To evaluate the correlation between $\Delta\Delta G$ values and DMS scores, we employed the Spearman rank correlation, which is robust against non-linear relationships between $\Delta\Delta G$ values and DMS scores. For additional analysis, we assigned specific secondary structural elements (helix, loop, or beta-strand) to positions in the protein's amino acid sequence. This information was obtained from the UniProt website⁴. To identify positions with significantly different distributions of $\Delta\Delta G$ values compared to neighbouring positions, we performed a Wilcoxon rank-sum test.

Results

Single mutations

Roman Kurley

The primary objective of this part of the project was to comprehensively investigate the influence of mutation position and newly formed amino acids on the DMS score. Additionally, a detailed analysis of amino acid properties within small neighborhoods was conducted to shed light on potential factors contributing to lowered DMS scores. To begin, a heatmap (figure 2) was generated to visualize the limited availability of data. Notably, less than a quarter of the mathematically possible single mutation combinations were present for analysis. This observation highlighted the challenges posed by the vast number of mutation possibilities and the selective nature of the dataset. Moving forward, a scatterplot (figure 3) was employed to gain broader insights into the distribution patterns across amino acids. The scatterplot revealed distinct clusters above and below specific DMS score thresholds, indicating the presence of significant hotspots in the data. Remarkably, arginine (R) and proline (P) exhibited the highest mutation frequency within the dataset, prompting further focused investigation on these specific amino acids. In order to quantify the impact of each amino acid on the DMS score, both the mean and median values were calculated (figure 4). The results demonstrated that proline displayed a significantly lower average and median DMS score compared to other amino acids. This finding suggests that proline substitutions may have a greater influence on reducing the overall DMS score. Conversely, arginine showed a slightly lower average DMS score but a broader distribution pattern, indicating a wide range of DMS scores associated with this amino acid. This was visualized in a boxplot (figure 5). To provide a more comprehensive understanding of the relationship between position and new amino acids, various statistical tests were performed. The results obtained from the eta-squared test, in conjunction with the ANOVA analysis, revealed that approximately 58% of the variance in DMS score could be attributed to the position of the mutation, while only 7% of the variance could be explained by the specific new amino acid (table 1). Furthermore, an assessment of the data's adherence to the assumption of normal distribution was conducted using the Shapiro-Wilks test and a q-q plot (figure 6). Both tests indicated significant deviations from the expected normal distribution, suggesting the presence of non-normality in the dataset. Given these insights, a Mann-Whitney U-test was conducted to compare the two groups position and new amino acid. The resulting scatterplot (figure 7) displayed a broad range of p-value distributions without a clear discernible pattern. However, more data points with significant differences ($p < 0.05$) between position and new amino acid were visible. This lack of distinct patterns indicated the absence of a straightforward relationship between the two groups, but the higher amount of significantly different data points hinted at the possibility of a positional impact. Additionally, a Kruskal-Wallis test (table 2) was performed to evaluate the overall differences among the impact of positions and new amino acids. The test results indicated a significant difference within the position group, suggesting that specific positions may have a more pronounced effect on the DMS score. However, no significant differences were observed when comparing positions with new amino acids. To confirm the dependency between new amino acids and position, a Friedman test (table 2) was conducted, which is specifically designed for dependent samples. The results of the Friedman test were consistent with those of the Kruskal-Wallis test, reinforcing the notion of interdependence between new amino acids and position. The analysis of the neighborhood properties involved the creation of a dedicated dataframe (table 3). This allowed for a detailed examination of the differences between mutated and unmutated sequences in terms of various neighborhood characteristics. Specifically, mutations exhibiting remarkably low DMS scores were subjected to thorough investigation. Moreover, multiple sequence alignments (MSA) were performed on different GFP variants and fluorescent proteins. However, due to the availability of comprehensive data on FPbase, the focus shifted to phylogenetic tree analysis for the selected variants. Although phylogenetic trees were obtained for the chosen variants, the wealth of additional information and precise hereditary relationships provided by FPbase rendered further exploration unnecessary. By conducting an extensive analysis encompassing mutation positions, new amino acids, neighborhood properties, and phylogenetic relationships, this project sought to elucidate key factors influencing the DMS score and uncover potential mechanisms underlying the observed variations in protein functionality.

Epistasis

Mutants with sequential mutations

Tianxin Angela Ma

Paths with a maximum length of 4 were identified (figure 1). Mutations, in general, are considered deleterious for the functionality of GFP. The decrease of the DMS score by adding up mutations is therefore expected and shown in (figure 12). A continuous decrease in the DMS score can also be observed for mutants with sequential mutations with an increasing number of mutations, as illustrated in figure 8. However, in some cases, the DMS scores deviated from this trend.

For instance, in figure 9, the score initially decreased with the addition of mutations F130L and E142V. However, the score increased again upon introducing the fourth mutation, K214E. This suggests that specific mutations can reverse the deleterious effects of other mutations, indicating the presence of epistasis.

Epistasis refers to the interaction between genes, which can have varying meanings depending on the context. Initially, it was described as a masking effect, where a variant at one locus prevents the manifestation of another variant at a different locus. Later, it was understood as a biological interaction between or within proteins, as in our case. (Cordell, 2002)

Mathematically, epistasis can be defined as a deviation from additivity. However, the choice of scale becomes crucial in this case. Since there is insufficient information regarding the additive effects of mutations on the logarithmic brightness of fluorescence, and our goal was to increase GFP's tolerance towards mutations rather than enhance fluorescence brightness. We considered an increased DMS score of a mutant compared to the mean score of all the mutants it comprises as positive epistasis, and conversely, the opposite case as negative epistasis.

For example, if mutant A with mutation A had a score of 3.5, mutant B with mutation B had a score of 3.7, and mutant C, containing mutations A and B, had a score of 3.7, we would classify this as positive epistasis since the mean score of A and B (3.6) is lower than the score of C (3.7).

In further analysis, we concentrated on positive epistasis since we aimed to increase GFP's tolerance toward mutations.

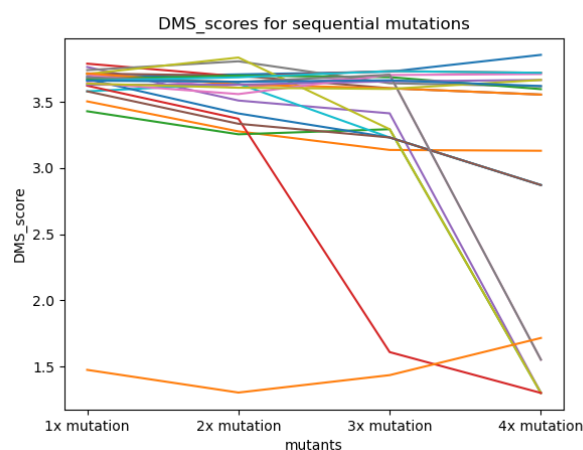


Figure 1: DMS scores for all 22 paths containing sequential mutations. x-axis: number of mutations each mutant has. y-axis: DMS score. Mutants from the same path are connected with lines. Each color stands for one path.

Structure Analysis

Tianxin Angela Ma

Out of the total number of 12,777 double mutation mutants, 11,954 of them met the requirement of having existing single mutation mutants that contributed to the composition of the double mutation mutant. The deviation between the actual score and the mean score for positive and negative epistasis is shown in figure 10 and 11.

The analysis revealed greater epistatic effects in mutants where both mutations occurred at positions with buried amino acid residues. This observation held true for both positive and negative epistasis, although the effect was more substantial in the case of negative epistasis. (figure 10 and 11).

Considering that the essential chromophore for fluorescence is located at the center of the protein (Chudakov *et al.*, 2010), it is reasonable to expect that buried amino acid residues would interact more intensely with the chromophore. This explains the stronger epistatic effects observed in mutants with buried amino acid residues compared to those with surface residues or a combination of buried and surface residues. The stronger negative epistasis compared to positive epistasis can be attributed to the fact that mutations, in general, are considered deleterious.

Ranking mutations with positive epistatic effects

Lisa Duttenhöfer

To evaluate the extent of influence of the mutation count on the DMS score, a boxplot, as shown in figure 12 was generated, displaying a gradient decrease of the fitness with increasing mutation count. Considering the results of the boxplot and a *Mann-Whitney-U-test* (p-value = 0.01), we set seven as the mutation count from which the number of mutations in the protein has more weight than relations of epistatic nature, as already implied by Sarkisyan *et al.* (2016b).

Figure 13 shows the scatterplots from the variance analyses per mutation and mutation count and displays a clear decrease of available data with an increasing mutation count. Looking at example mutations from our investigation marked in the plot, low variances and a high amount of used values were distinctive. Mutations declared highly stabilizing by Johansson *et al.* (2023), showed low variances but less available data, which indicates a loss of information caused by an incomplete mutational dataset.

The generated rankings shown in table 4 show varying results depending on the calculation method used. Mutations such as V163A, K214E, G232R, S175G, I171V, and K113R are in the Top 20 in almost every ranking, marking them as the most positive influencing in our dataset. For K214E we can confirm a positive epistatic effect from the path analysis (figure 9). The weighted rankings show differing results due to additional emphases set through the mean DMS score or the available data. Looking at the factors making up the rankings directly, the critical factor in this analysis is the missing data for the majority of values, with the cutoff for how often a mutation appears in the dataset, altering the results significantly (table 5). However, setting the cutoff at an appearance number of ten, the rankings gave an overview of the mutations with the most positive effect in the dataset setting various focal points. Rankings involving the *PyRosetta* generated values for ΔG significantly differ from the other rankings, although V163A appears as the Top 5 mutation again. Compared to the results from Johansson *et al.* (2023) there are few commonalities, suggesting further investigation into the role of ΔG in the DMS score is necessary.

Protein stability prediction with PyRosetta

Rebecca Röss

We used the crystal structure of GFP for the prediction, which consists of 236 amino acids. Therefore, 1079 single mutations were used rather than 1083. The predicted ΔG value for the wildtype protein was determined to be -953.838 kcal/mol. All single mutations exhibited remarkably similar ΔG values, ranging from -783.393 kcal/mol to -783.379 kcal/mol. However, our investigation found no significant correlation between the $\Delta\Delta G$ values and the DMS scores. The Spearman correlation coefficient was calculated as 0.17, with a p-value of 5.3×10^{-8} , indicating a lack of strong association. This observation is visually represented in figure 14. Furthermore, we observed patterns in the distribution of mutations across the protein sequence. Mutations with lower $\Delta\Delta G$ values were predominantly localized in the initial segment of the protein sequence. About 50% of the Mutations were primarily observed within amino acid positions 60 to 180. Conversely, mutations with higher $\Delta\Delta G$ values were predominantly observed in the latter half of the amino acid sequence. Moreover, our analysis revealed distinct distributions of ΔG values at specific positions within the protein sequence.

Discussion

Single mutations

Roman Kurley

The discussion section aims to interpret and discuss the findings of this part of the project, which focused on investigating the influence of mutation position, new amino acids, neighborhood properties, and phylogenetic relationships on the DMS score of GFP mutants. Our analysis revealed several key findings that contribute to our understanding of protein functionality and provide valuable insights into the GFP mutant dataset. The heatmap (figure 1) analysis highlighted the limited availability of data, emphasizing the need for caution in generalizing the results. Despite this limitation, our study was able to identify significant trends and patterns within the available dataset. The scatterplot (figure 2) analysis showed distinct clusters above and below specific DMS score thresholds, indicating the presence of significant hotspots in the data. Specifically, arginine and proline exhibited the highest mutation frequency, prompting further investigation. The boxplot (figure 4) analysis further supported the impact of these amino acids, showing lower DMS scores for proline compared to other amino acids. Statistical tests revealed that the position of the mutation explained a substantial portion of the variance in DMS scores, whereas the specific new amino acid accounted for a smaller proportion. However, the presence of significant deviations from the assumption of normality in the dataset adds some uncertainty to these results. The Mann-Whitney-U-test (figure 6) did not reveal a clear pattern in the relationship between position and new amino acids but a higher amount of data points with significant differences between the impact of position and amino acid resulted, hinting at a significant difference between the two groups that has to be analyzed further. The Kruskal-Wallis test (table 2) supported these findings, showing no significant difference between the impact of positions and new amino acids. The dependency between new amino acids and position was further confirmed by the Friedman test (table 2). The analysis of neighborhood properties (table 3) provided insights into the differences between mutated and unmutated sequences, highlighting potential factors contributing to lowered DMS scores. However, further 3D structure analysis is required to confirm the effects of neighborhood properties on the interactions of different amino acid residues and their impact on the DMS score. By comprehensively investigating mutation positions, new amino acids and neighborhood properties, this project contributes to our understanding of the factors influencing the DMS score and provides insights into the variations in protein functionality. The limitations of this study include the limited availability of data, reliance on computational analysis, and the absence of experimental validation. Future research should focus on expanding the dataset, incorporating experimental approaches, and conducting structural analyses to gain further insights into the mechanisms underlying the observed variations in the DMS score.

Epistasis

Mutants with sequential mutations

Tianxin Angela Ma

The occurrence of unexpected score increments suggested the presence of epistatic effects in GFP. To obtain further valuable data for analysis, a combination of directed and random mutagenesis can be employed. Specific mutations that appear to have positive epistatic effects on other mutations can be inserted using directed mutagenesis. This process can be repeated with multiple mutations of interest. Random mutagenesis can be performed following the directed mutagenesis to introduce additional mutations. The epistatic effects of these mutations can then be analyzed. This approach offers an advantage over the data set we used (Sarkisyan *et al.*, 2016b) as it provides more specific data tailored to the investigation of epistatic effects. While the data set by Sarkisyan *et al.* was valuable for obtaining a general overview of the protein's properties, the acquired information with further experiments, as described before enables the collection of more targeted data without generating excessive, unnecessary data in the process.

Due to the limited dataset, less than 2% of all possible double mutation mutants were included in the analysis. To provide more evidence for stronger epistatic effects of buried amino acid residues, a larger amount of data would be necessary. One potential approach to gathering this additional data is directed mutagenesis to generate every possible double mutation mutant. Pursuing this method could achieve a more thorough understanding of epistatic interactions based on the orientation of each amino acid residue.

Ranking mutations with positive epistatic effects

Lisa Duttenhöfer

Throughout the analyses of the variances and the effect of mutations on the DMS score of the protein, it was possible to gain knowledge about specific mutations that showed a notably positive impact on the fluorescence of GFP. The role of the number of mutations per mutant, without considering epistatic correlations, is significant, as shown in figure 12.

Making it more specific, figure 13 shows the effect of all available mutations and makes it possible to subsume the mutations in terms of the reliability and the variance of the resulting DMS scores. However, the data being distorted because of the considerably small amount of data available compared to a complete mutational landscape made it hard to make a quantitative statement on the effect of specific mutations, especially for higher mutation counts.

The rankings, made to predict the ability of a specific mutation to improve a mutant's robustness towards other mutations, consist of several statistical factors of the dataset, making the result's interpretation dependent on which aspect is in focus. While the calculated DSD emphasizes the size of the mutation's effect on the DMS score, the variance, depending on its calculation method, also gives additional information about the composition of the data segment. The ranking results in table 4 show several mutations consistently at Top ranking scores, implying them to have a highly positive effect on the fitness of GFP. The ΔG values generated by *PyRosetta* and the resulting rankings couldn't be validated by any literature. With Johansson *et al.* (2023) presenting different results and stating the role of ΔG in stability as too small to be considered without further processing, additional calculations are needed to continue the analyses in that aspect. This way, it is possible to set a focus and make predictions about which mutations have a stabilizing effect on a mutant, all within the limitations of the available data.

Protein stability prediction with PyRosetta

Rebecca Röss

This part focused primarily on the thermodynamic stability of proteins, as limited research and prediction models are available for kinetic stability (Sanchez-Ruiz, 2010). We chose to use the parameter of Gibbs free energy differences (ΔG) due to the abundance of prediction models and studies related to it (Potapov *et al.*, 2009). Regarding the predicted ΔG values, our main emphasis was on single mutations. This was influenced by the fact that prediction models, such as the Rosetta software suite, are biased towards specific mutations, particularly destabilizing ones (Fang, 2023), as they were validated using datasets with an overrepresentation of such mutations (Frenz *et al.*, 2020). Therefore, the accuracy of predictions for multiple mutations may be compromised. In our analysis, we observed a negative ΔG value for the wildtype protein, indicating a stable folded protein structure consistent with the expected behavior of a well-functioning GFP. Interestingly, all single mutations exhibited remarkably similar ΔG values, falling within the range of -783.3936 kcal/mol to -783.3790 kcal/mol. While these ΔG values are significantly less negative than that of the wildtype, they demonstrate a consistent impact of amino acid substitutions on protein stability within a similar range. It is worth noting that prediction models tend to overpredict destabilizing mutations due to inherent biases. However, when assessing the relationship between the predicted $\Delta\Delta G$ values and the experimental DMS scores, our analysis revealed no significant correlation. Although several interesting observations were made, such as the concentration of mutations with lower $\Delta\Delta G$ values in the first half of the protein sequence and the prevalence of mutations with higher $\Delta\Delta G$ values in the rear region of the sequence, it is important to acknowledge the limitations of our dataset. Our analysis was based

on a relatively small dataset, which may limit the statistical power of our findings. Additionally, prediction models don't always yield perfect results and may deviate from actual measurements. To enhance the validity of our predictions, experimental measurement of ΔG values for individual mutated proteins and refinement of the PyRosetta software's predictions would be beneficial.

Appendix

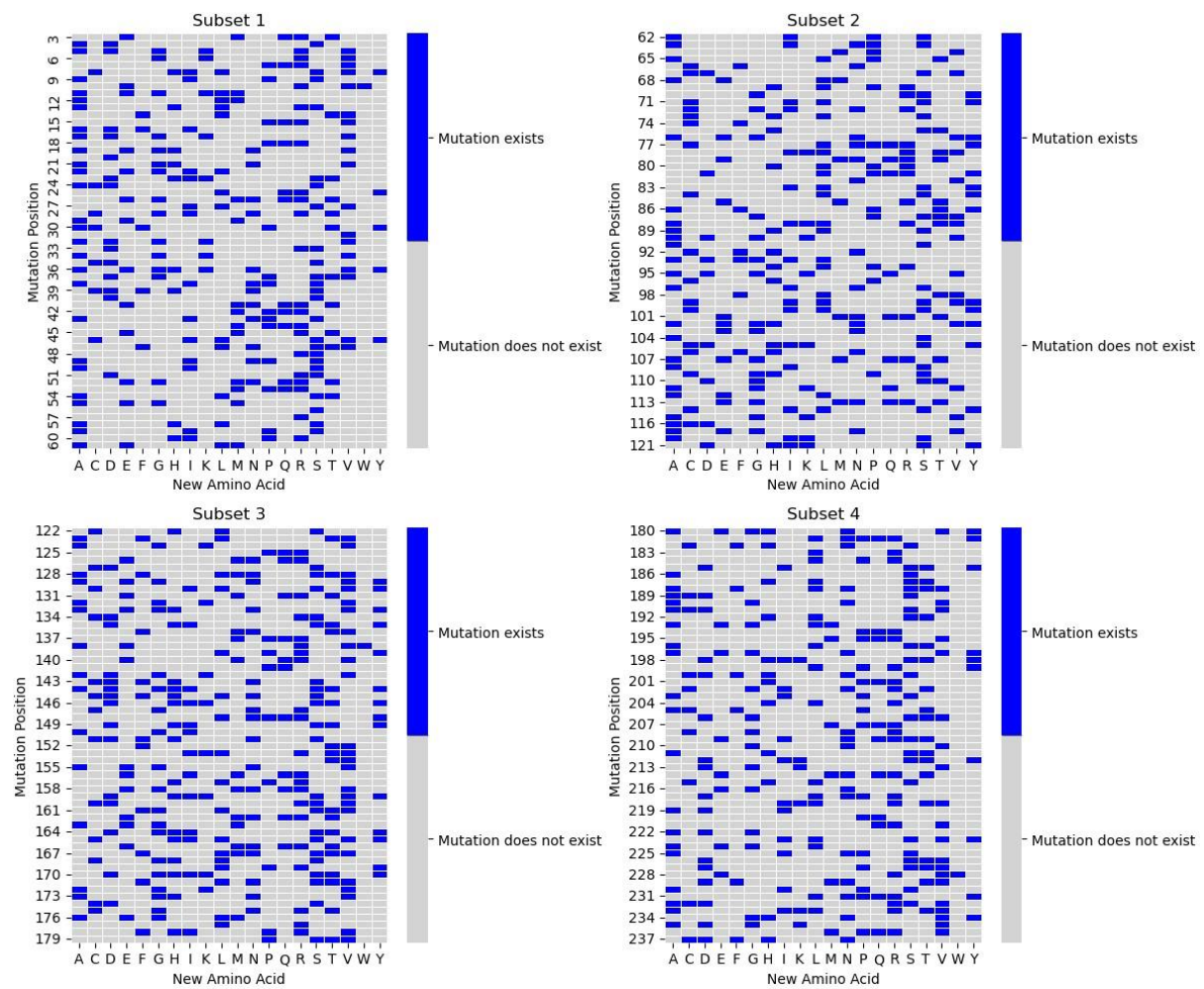


Figure 2: All single mutation combinations available in our dataset for each position (y-axes) and each new amino acid (x-axes) depicted in a subset of 4 heatmaps.

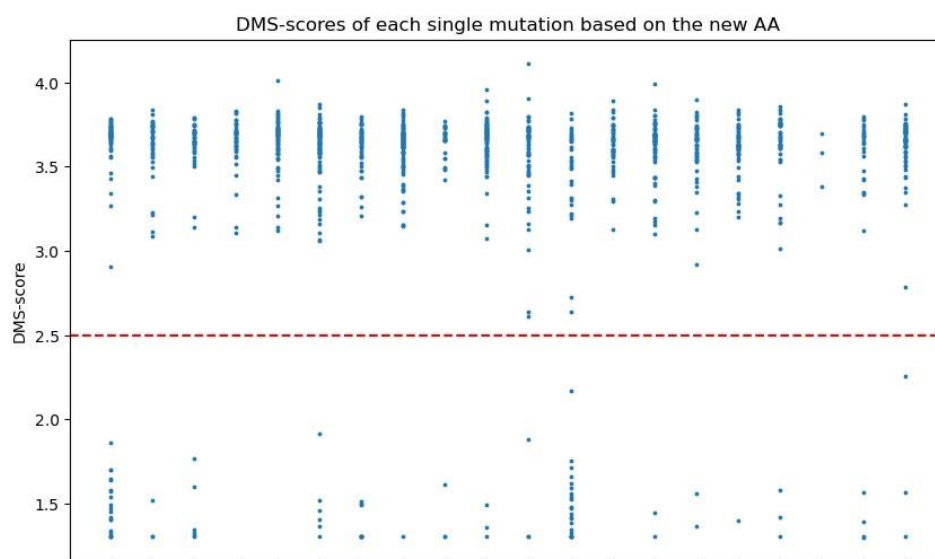


Figure 3: DMS scores of each single mutation in the data set plotted in a scatter plot based on the new amino acid (x-axis) and the respective DMS score (y-axis).

new_AA	Mean DMS by Amino Acid	Median DMS by Amino Acid
A	3.568873	3.635771
C	3.484644	3.659148
D	3.243770	3.618095
E	3.344491	3.647804
F	3.290031	3.638083
G	3.471302	3.641087
H	3.521980	3.655410
I	3.561838	3.627550
K	3.430530	3.661849
L	3.543811	3.654187
M	3.638662	3.670929
N	3.539638	3.658432
P	2.641112	3.214552
Q	3.452677	3.648993
R	3.101493	3.676710
S	3.465196	3.639505
T	3.635548	3.681731
V	3.557836	3.667263
W	3.552605	3.583799
Y	3.625013	3.649549

Figure 4: Mean and median DMS score calculated for each amino acid. Proline (P) and Arginine (R) are highlighted. The mean and median of proline is lowered, the mean of arginine is lowered, but the median is comparable to other amino acids.

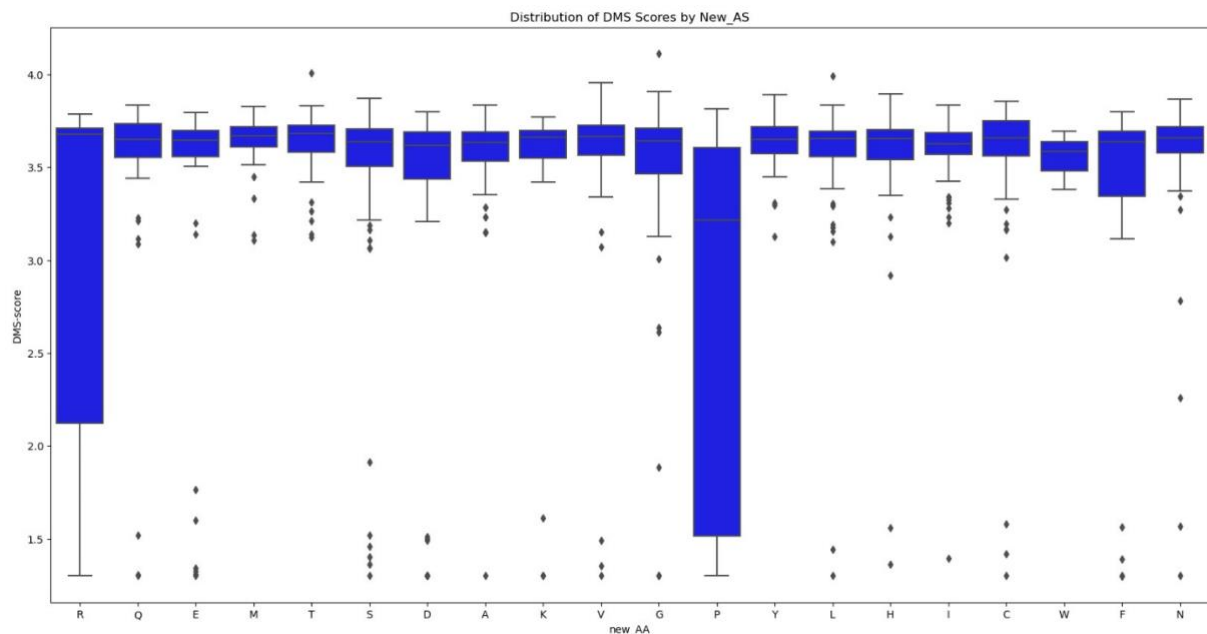


Figure 5: Distribution of data points for each amino acid (x-axis) and the respective DMS score (y-axis) in a boxplot. The black horizontal line marks the median.

Table 1: Results of ANOVA using grouped positions and grouped new amino acids. The results of the eta-squared test are shown below: 58% of variance is explained by position, 7% is explained by new amino acid. The rest of the variance is in the residual.

ANOVA_pos_new_AA:					
	df	sum_sq	mean_sq	F	PR(>F)
Position	232.0	243.851576	1.051084	6.003194	1.301208e-82
new_AA	19.0	29.717435	1.564076	8.933106	1.507908e-23
Residual	832.0	145.672832	0.175088	NaN	NaN
0.5816489450614979					
0.07088375323670729					

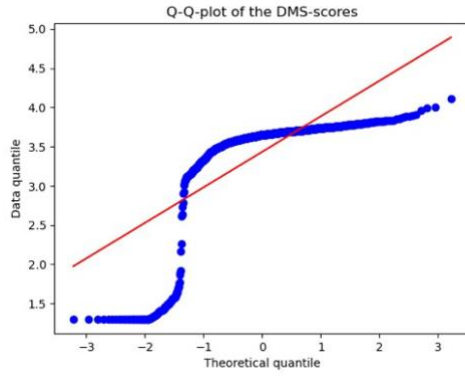


Figure 6: Q-Q-plot of the DMS scores. The red line represents the ideal normal distribution comparing the theoretical quantiles (x-axis) and the actual data quantiles (y-axis). The blue dotted line represents the distribution in this data set.

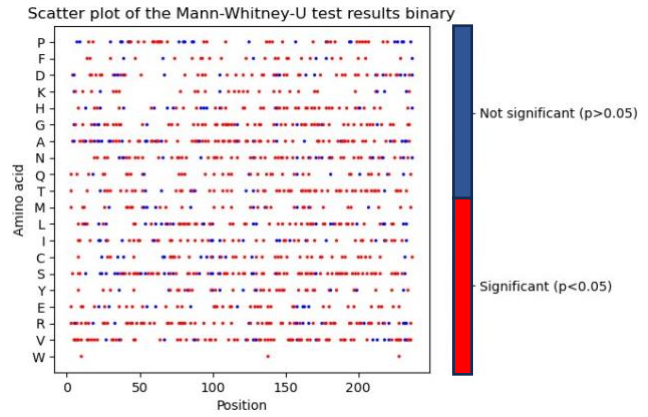


Figure 7: Scatter plot showing the results of the Mann-Whitney-U test for each amino acid (y-axis) and each position (x-axis) of the available single mutations. The data points are either significant ($p < 0.05$) in red or not significant ($p > 0.05$) in blue.

Table 2: Results of the Kruskal-Wallis tests and Friedman test. There is a significant difference between the positions only, whereas there is no significant difference between the position and amino acids. The Friedman test results show the dependency of the groups.

Kruskal-Wallis-test_pos
Test statistic: 577.8469817125756
P-value: 1.8399990857612868e-31

There is a significant difference between the positions.

Kruskal-Wallis-test_pos_new_AA
Test statistic: 1083.0
P-value: 0.49428529234990104

There is no significant difference between the positions and amino acids.

Friedman-test_pos_new_AA
Test statistic: 1083.0
P-value: 0.49428529234990104

There is no significant difference between the positions and amino acids.

Table 3: Exemplary results of K113R with the neighbourhood properties of mutated and unmutated neighbourhood.

	Mutation	Neighbourhood	Molecular Weight	Residue Weight	pKa1
972	K113R	AEVRFEG	914.97	788.87	15.38
973	K113R-unmut	AEVKFEG	886.96	760.86	15.39
	pKb2	p14	H	VSC	P1
972	66.42	40.61	-0.64	443.5	63.3
973	66.33	39.59	0.39	438.5	64.1
					P2
					SASA
					NCISC
972					0.337986
973					0.312107

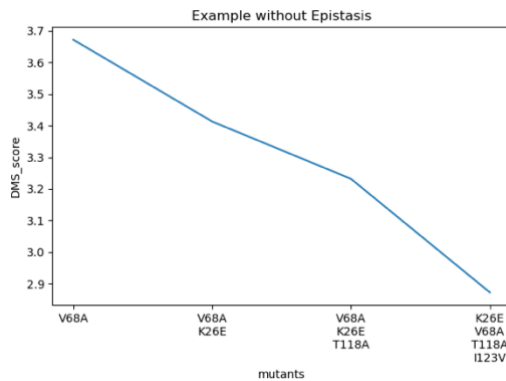


Figure 8: DMS scores for one example path containing sequential mutations shown in a line plot. x-axis: mutations within each mutant. y-axis: DMS score. In this case, (positive) epistasis is not present.

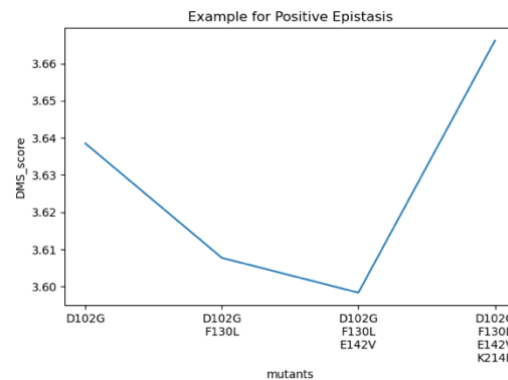


Figure 9: DMS scores for one example path containing sequential mutations shown in a line plot. x-axis: mutations within each mutant. y-axis: DMS score. Each color stands for one path. In this case, positive epistasis is present.

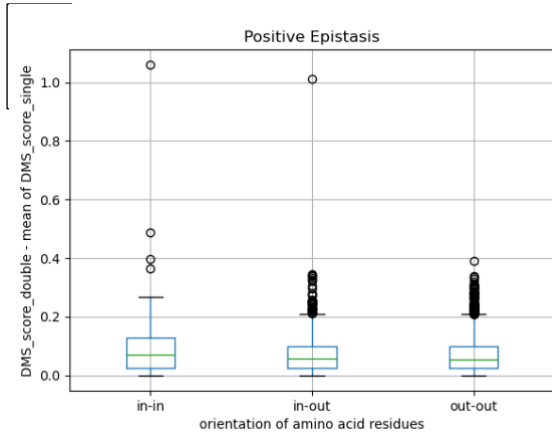


Figure 10: Distribution of deviation between measured DMS score of double mutation mutants and calculated DMS score (mean DMS score of single mutation mutants the double mutation mutant consists of) for positive epistatic mutants. Grouped by the orientation of both amino acid residues. in-in: both residues are buried. in-out: one buried and one surface residue. out-out: both are surface residues.

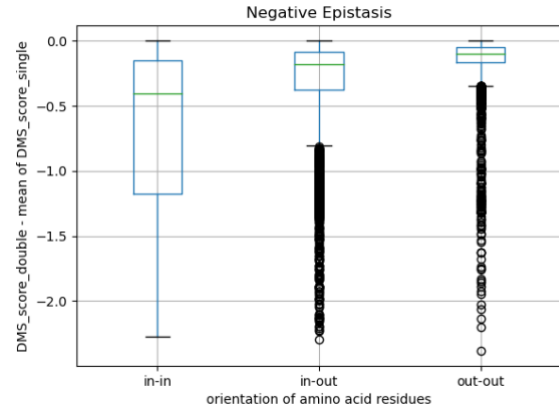


Figure 11: Distribution of deviation between measured DMS score of double mutation mutants and calculated DMS score (mean DMS score of single mutation mutants the double mutation mutant consists of) for negative epistatic mutants. Grouped by the orientation of both amino acid residues. in-in: both residues are buried. in-out: one buried and one surface residue. out-out: both are surface residues.

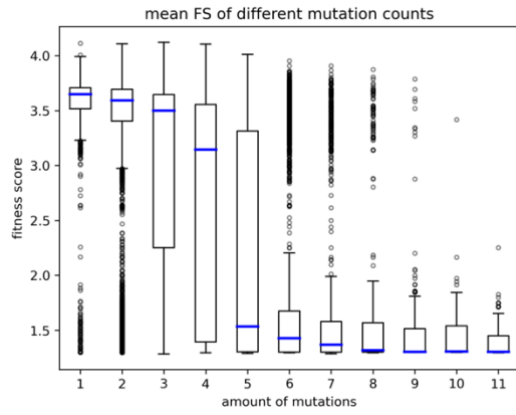


Figure 12: boxplot showing the impact of the number of mutations on the fitness score. In blue: median

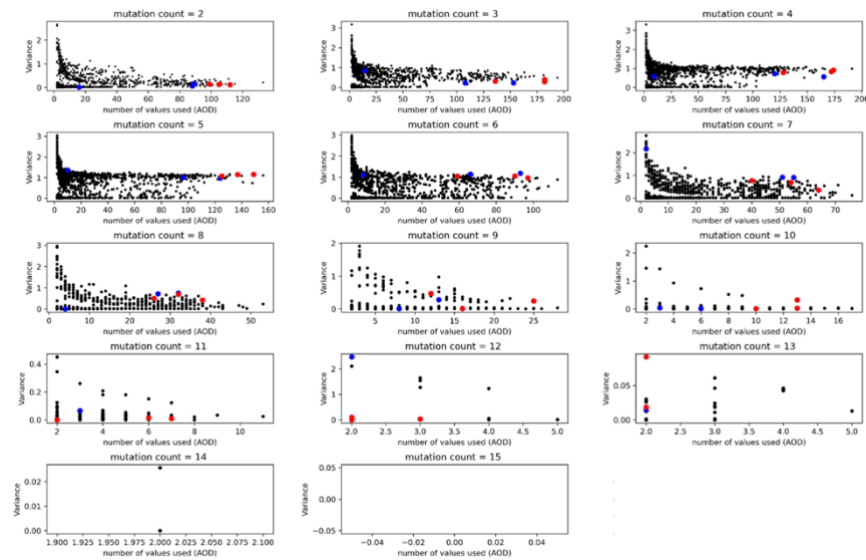


Figure 13: scatterplots per mutation count. Each dot presents one mutation. Marked in red are the mutations K214E and K113R identified as positive in pedigree analyses. Marked in blue are the mutations V163A, K166Q and I171V, which were established by Johansson et al. (2023) as highly stabilizing in epistatic relations in GFP. In total there are 1810 mutations.

Table 4 Ranking results using the equations in (equation 2) and a cut-off (mutation count) at a minimum amount of ten mutants containing the mutation. *a* being computed by a combined rank of the variance (var_2) and the amount of data for this mutation *b* based on $score_1$ *c* based on $score_2$ *d, f, h, j* based on $score_3$ with different methods of calculation for the variance ($var_1, var_2, var_3, var_4$) *e, g, i* based on the weighted $score_4$ and the variances 1 to 3 and *k* based on the weighted $score_5$. For the tables with the data making up the ranking scores see the html files in the github repository.

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>j</i>	<i>k</i>
1 T225A	1 V163A	1 V163A	1 V163A	1 E17D	1 V163A	1 E17D	1 G232R	1 E17D	1 G232R	1 D216V
2 N185S	2 I171V	2 I171V	2 G232R	2 G232R	2 G232R	2 G232R	2 V163A	2 G232R	2 V163A	2 G33S
3 N149S	3 S175G	3 F99L	3 I167V	3 V163A	3 I167V	3 V163A	3 I171V	3 V163A	3 I171V	3 T186A
4 M88L	4 K113R	4 K214E	4 S175G	4 K214E	4 S175G	4 K214E	4 S175G	4 I171V	4 S175G	4 S65P
5 K209R	5 I167V	5 K158R	5 K113R	5 K158R	5 K113R	5 S175G	5 I167V	5 K214E	5 I167V	5 V163A
6 N159D	6 D117G	6 N144D	6 K156R	6 S175G	6 K156R	6 K158R	6 K113R	6 F99L	6 K113R	6 D234G
7 S147G	7 K156R	7 S175G	7 N144D	7 K113R	7 I171V	7 K113R	7 K214E	7 S175G	7 K156R	7 T108P
8 F114L	8 N144D	8 T43A	8 K158R	8 N144D	8 K214E	8 N144D	8 K156R	8 K158R	8 K214E	8 G67S
9 N159S	9 K214E	9 K113R	9 K214E	9 N121S	9 K158R	9 I167V	9 T38A	9 T97A	9 D117G	9 A110D
10 E132G	10 D129G	10 K79R	10 I128T	10 I167V	10 N144D	10 F99L	10 D129G	10 K158R	10 T38A	10 S30P
11 K107E	11 K158R	11 F99S	11 D117G	11 K79R	11 D117G	11 N121S	11 N144D	11 K113R	11 K158R	11 S202G
12 S202G	12 N105S	12 N121S	12 D197G	12 N198D	12 I128T	12 T43A	12 D117G	12 T43A	12 I123V	12 T108S
13 T186A	13 I167T	13 K156E	13 I171V	13 F99L	13 H25R	13 I171V	13 N105S	13 N144D	13 D129G	13 F46L
14 V193A	14 I123V	14 D129G	14 N198D	14 T43A	14 I123V	14 N198D	14 I167T	14 F223S	14 N144D	14 Q69R
15 I188V	15 T38A	15 T97A	15 H25R	15 K156R	15 D197G	15 K79R	15 K158R	15 K79R	15 I167T	15 D197G
16 F46Y	16 I128T	16 F223S	16 K79R	16 K156E	16 D129G	16 K156R	16 I123V	16 I167V	16 N105S	16 L18Q
17 T118A	17 H25R	17 H25R	17 N170D	17 E172G	17 T38A	17 T97A	17 S72G	17 M153V	17 I128T	17 G104D
18 N146S	18 K79R	18 I171T	18 E172G	18 H25R	18 N198D	18 H25R	18 K79R	18 N121S	18 S72G	18 Y151H
19 K45R	19 K166R	19 I123V	19 T38A	19 K214R	19 M233V	19 K156E	19 I128T	19 D129G	19 M233V	19 L125Q
20 F165L	20 S72G	20 I167V	20 I123V	20 D133G	20 K79R	20 F99S	20 H25R	20 F99S	20 H25R	20 G31D

Table 5: Comparison of the ranking results of ranking *h* using the cut-off and not using it.

With the cut-off significantly altering the results.

<i>a</i>	<i>b</i>
1 E5R	G232R
2 N164C	V163A
3 I128M	I171V
4 E142Q	S175G
5 G232R	I167V
6 I229M	K113R
7 V163A	K156R
8 K162V	K214E
9 N164G	D117G
10 F165T	T38A
11 G4V	K158R
12 I171V	I123V
13 S175G	D129G
14 I167V	N144D
15 K113R	I167T
16 I167V	N105S
17 K113R	I128T
18 G174R	S72G
19 K214E	M233V
20 K156R	H25R

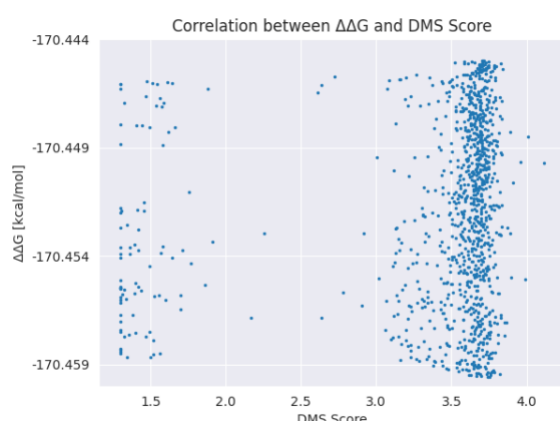


Figure 14: Scatter plot showing the correlation between the $\Delta\Delta G$ values and the DMS scores. Based on the plot, no significant correlation can be observed between the $\Delta\Delta G$ values and the DMS scores

References

- Chudakov, D. M., Matz, M. V., Lukyanov, S., and Lukyanov, K. A. (2010). Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol Rev* 90, 1103-1163.
- Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11, 2463-2468.
- Chudakov, D. M., Matz, M. V., Lukyanov, S., and Lukyanov, K. A. (2010). Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol Rev* 90, 1103-1163.
- Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11, 2463-2468.
- Fang, J. (2023). The role of data imbalance bias in the prediction of protein stability change upon mutation. *PloS one* 18, e0283727.
- Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods* 11, 801-807.
- Frenz, B., Lewis, S. M., King, I., DiMaio, F., Park, H., and Song, Y. (2020). Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Frontiers in bioengineering and biotechnology*, 1175.
- Fu, J. L., Kanno, T., Liang, S. C., Matzke, A. J., and Matzke, M. (2015). GFP Loss-of-Function Mutations in *Arabidopsis thaliana*. *G3 (Bethesda)* 5, 1849-1855.
- Johansson, K. E., Lindorff-Larsen, K., and Winther, J. R. (2023). Global Analysis of Multi-Mutants to Improve Protein Function. *Journal of Molecular Biology* 435, 168034.
- Liu, Q., Xun, G., and Feng, Y. (2019). The state-of-the-art strategies of protein engineering for enzyme stabilization. *Biotechnology advances* 37, 530-537.
- Ong, W. J., Alvarez, S., Leroux, I. E., Shahid, R. S., Samma, A. A., Peshkepija, P., Morgan, A. L., Mulcahy, S., and Zimmer, M. (2011). Function and structure of GFP-like proteins in the protein data bank. *Mol Biosyst* 7, 984-992.
- Potapov, V., Cohen, M., and Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein engineering, design & selection* 22, 553-560.
- Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., and Soylemez, O. (2016a). Local fitness landscape of the green fluorescent protein. *Nature* 533, 397-401.
- Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., *et al.* (2016b). Local fitness landscape of the green fluorescent protein. *Nature* 533, 397-401.
- Zimmer, M. H., Li, B., Shahid, R. S., Peshkepija, P., and Zimmer, M. (2014). Structural Consequences of Chromophore Formation and Exploration of Conserved Lid Residues amongst Naturally Occurring Fluorescent Proteins. *Chem Phys* 429, 5-11.

1: <https://www.kaggle.com/datasets/aleiopaullier/aminoacids-physical-and-chemical-properties?resource=download> [17.07.2023, 15:15]

2: <https://www.fpbases.org/> [17.07.2023, 15:15]

3: <https://www.rcsb.org/structure/2wur> [16.07.2023, 20:46]

4: <https://www.uniprot.org/uniprotkb/P42212/entry#structure> [16.07.2023, 20:42]