Global analysis of multi-mutants to improve protein function

Johansson, Kristoffer E; Lindorff-Larsen, Kresten; Winther, Jakob R

Published in:
Journal of Molecular Biology

DOI:
10.1016/j.jmb.2023.168034

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Global Analysis of Multi-Mutants to Improve Protein Function

Kristoffer E. Johansson * Kresten Lindorff-Larsen * and Jakob R. Winther *

Linderstrom-Lang Centre for Protein Science, Section for Biomolecular Sciences, Department of Biology of (University of Copenhagen), Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark

Correspondence to Kristoffer E. Johansson, Kresten Lindorff-Larsen and Jakob R. Winther: kristoffer.johansson@bio.ku.dk (K.E. Johansson), lindorff@bio.ku.dk (K. Lindorff-Larsen), jrwinther@bio.ku.dk (J.R. Winther) @kristofferenoee (K.E. Johansson), @lindorfflarsen (K. Lindorff-Larsen) https://doi.org/10.1016/j.jmb.2023.168034

Abstract

The identification of amino acid substitutions that both enhance the stability and function of a protein is a key challenge in protein engineering. Technological advances have enabled assaying thousands of protein variants in a single high-throughput experiment, and more recent studies use such data in protein engineering. We present a Global Multi-Mutant Analysis (GMMA) that exploits the presence of multiply-substituted variants to identify individual amino acid substitutions that are beneficial for the stability and function across a large library of protein variants. We have applied GMMA to a previously published experiment reporting on >54,000 variants of green fluorescent protein (GFP), each with known fluorescence output, and each carrying 1–15 amino acid substitutions (Sarkisyan et al., 2016). The GMMA method achieves a good fit to this dataset while being analytically transparent. We show experimentally that the six top-ranking substitutions progressively enhance GFP. More broadly, using only a single experiment as input our analysis recovers nearly all the substitutions previously reported to be beneficial for GFP folding and function. In conclusion, we suggest that large libraries of multiply-substituted variants may provide a unique source of information for protein engineering.

Introduction

A major challenge in practical applications of proteins is the engineering of protein stability while at the same time maintaining the function of the protein. New developments in biotechnology are continuously applied to address this challenge with high-throughput methods currently in focus. Synthesis, screening and sequencing may today be performed for thousands of protein variants in parallel via Multiplexed Assay of Variant Effects (MAVE) also known as deep mutational scanning.1,2 Such experiments can identify loss-of-function variants with high accuracy, but are often unable to gauge more subtle effects. For example, stabilizing or mildly destabilizing substitutions are likely to have a minor, if any, detectable impact on the function of a stable protein. This is known as threshold robustness3 and cause such substitution effects more difficult to identify. Additive contributions to stability or catalysis are particularly attractive for protein engineering4,5 and the ability to extract such from MAVE experiments could impact on the field substantially. Stabilizing and folding-enhancing substitutions may also provide robustness to other mildly destabilizing substitutions and thus enhance evolvability towards novel functions.6,7

When a protein library contains a high fraction of variants with two or multiple amino acid
substitutions, statistical models have been used to investigate how the effect of individual substitutions combine. One important technical consideration is the mathematical framework to capture (non-)additivity of the variant effects. As a simple example, from standard thermodynamics \( \Delta G = -R \ln(K) \) one would expect that the stability effects of two independent substitutions are additive when quantified by free energy differences, whereas the corresponding equilibrium constants \( K \) would not be additive, but rather multiplicative. Thus, more general models typically consider additive single-substitution effects, known as a protein fitness potential \( ^{9} \) or latent variable \( ^{8,10} \) and transforms the sum of these to the assayed quantity via various non-linear functions to describe the data generated by a MAVE. \(^{3-11} \) One study further showed that a thermodynamic model could be used to quantify the effects of single-substitutions on protein binding and structural stability from experimental measurements of the effects of both single and double mutants on binding. \(^{11} \) This model was later shown to capture structural stability accurately. \(^{12} \) The measured function of a protein will often depend on the amount of folded protein and the fitness potential is thus related to free energies \(^{5,9,11} \) or statistical potentials. \(^{13} \)

Here, we demonstrate how a fitness potential may be used directly in protein engineering and present a method, Global Multi-Mutant Analysis (GMMA), that enables the identification of amino acid substitutions that have a general enhancing effect with little dependency on the sequence context, and thus substitutions with the potential to be additively combined. We highlight the benefit of informing single-substitution effects by variants with multiple amino acid substitutions, here referred to as multi-mutants. The central idea is to identify beneficial substitutions by their ability to compensate other substitutions that decrease function for example by destabilizing the protein. For example, while a stabilizing substitution may not have a measurable effect on function in the background of an already stable protein, it can be identified in the background of one or more destabilizing substitutions that decrease function for example by destabilizing the protein. This concept of compensating prior destabilization is well known in the field of protein engineering e.g. as the “partner potentiation”, \(^{14} \) “disrupt-and-restore” \(^{15} \) or similar principles formulated previously, \(^{16,17} \) but here applied to a large library of diverse multi-mutants that has the advantage of statistical averaging.

We have applied our analysis to previously published experimental data that reports the fluorescence of >54,000 variants of green fluorescent protein (GFP) each containing 1–15 amino acid substitutions with a total of 1,879 unique single-amino-acid substitutions observed in the experiment. \(^{5} \) GMMA was inspired by observations first made in the original analysis of this data that showed that the overall fitness landscape to a large extent can be described using effects on protein stability. \(^{9} \) Briefly described, the GFP variants were generated using error-prone PCR (epPCR) and expressed in fusion with a red-fluorescent protein to correct for variations in expression levels. \(^{9} \) Then, fluorescence activated cell sorting (FACS) was used to divide the cells into eight fractions based on their level of green fluorescence, and each fraction was sequenced to link an intensity of fluorescence with a genotype. \(^{9} \) We have applied GMMA to the results of this experiment to estimate accurately the effects of single-substitutions. Among these, we identify substitutions that are beneficial for both function and stability and thus directly applicable to engineering. As validations of the approach, we show that the model accurately predicts experimental variant effects that were not used in training, and show that the top six substitutions may progressively be combined in new GFP variants that achieve a more than 6-fold enhancement of cellular fluorescence compared to the starting point when expressed in \( \textit{E. coli} \) cells. All tested GFP variants designed by GMMA were more fluorescent than the starting sequence which is a notable improvement of a similar study using a more complicated model. \(^{18} \) Finally, we show that the method identifies a large number of substitutions that have previously been described to be beneficial for improving GFP.

Results and discussion

Concept

We first illustrate briefly the qualitative concept that inspired the development of GMMA (Figure 1). The original report of the GFP data describes the ability of the functional protein to tolerate many random mutations before losing function (Figure 1(e)). This analysis shows that the “wild type” (or reference sequence) on average can tolerate approximately four amino acid substitutions before the average brightness of variants drops to 50%. Inspired by this, we next considered the subset of variants that carries a specific substitution, e.g. V163A, and find that in this background the protein can tolerate on average an additional five substitutions before the average brightness drops to 50% (Figure 1(f)). Thus, from this simple analysis it may be concluded that V163A possesses some enhancing property across all 677 sequence contexts in the subset of GFP variants that carries V163A.

Fitness potentials may quantify this concept which may be described by an analogy with traditional measurements of protein stability using chemical denaturant. \(^{19} \) In this analogy, the protein is inactivated, rather than unfolded, by introducing substitutions rather than chemical denaturant. Thus, we describe the inactivation as a two-state process where the system, a “variable protein”, is
in equilibrium between an active and inactive state and where a substitution, \( s \), affects a fitness potential, \( F \), in the same way as heat or chemical denaturant is typically considered to affect the stability of a protein (Figure 1(a)):

\[
F_v = F_{wt} + \sum_{s} \Delta F_s
\]  

(1)

Here, \( F_{wt} \) is the fitness potential of the “wild-type” or reference sequence and \( \Delta F_s \) is the effect of the substitution, \( s \), on the fitness potential of a
variant, \( F_v \). At fitness potential zero, the active and inactive states are equally populated, and the assay readout should be half of the maximal. A predicted assay readout is calculated from the fitness potential in analogy with the fraction of folded protein via Boltzmann probabilities (see Methods). This formulation is similar to previous global models \(^{3,8–11,20,21}\) but here applied with the particular aim of identifying generally enhancing substitutions. Throughout this report, we will return to the analogy between the fitness potential and conventional structural stability and highlight some important similarities and differences between these.

The GMMA thus comprises a set of equations, each describing the activity of a variant, and carrying a number of parameters, \( \Delta F_{\text{in}} \) together with the global wild-type fitness potential, \( F_{\text{WT}} \) (and possible baseline parameters). GMMA also requires a mathematical description that relates the fitness potential of a variant to the measured quantity (here the brightness of the fluorescence in a FACS experiment); in this work we combine a biophysical model with a phenomenological description (see Methods). To analyze the resulting system of equations, it is important to have more data (variants) than parameters (substitutions), which is possible with a multi-mutant library where a set of unique substitutions may be mixed in many different ways to make a larger set of variants. It is also important that all equations are coupled, and this may be tested by analyzing an undirected bipartite graph in which the protein variants constitute one layer of nodes (Figure 1(b), circles) and the unique substitutions the other layer (Figure 1(b), squares). This substitution-variant-graph formalism may be used to study many aspects of the multi-mutant library. For example, GMMA identifies enhancing substitutions by their ability to compensate negative effects, and it is important that the variant library has a distribution of substitutions-per-variant that covers the transition region where the system loses activity, for the same reason that chemical denaturation experiments should have several measurements in the denaturation region (Figure 1(c)).

The inactive state results from a number of detrimental amino acid substitutions, irrespective of mechanism, and is thus somewhat broadly defined. Some detrimental substitutions cannot be compensated by other enhancing substitutions, e.g. removal of a crucial functional side chain, and we term these irreversible fatal substitutions (IFS). IFS may be related to the assayed function of the protein and cellular stability, or for GFP, related to the chromophore maturation reaction. In principle, IFS have an infinite impact on the fitness potential but in practice they will be estimated to an arbitrary high value. Thus, in contrast to conventional structural stability measurements, a distinct advantage of GMMA is that it only identifies substitutions that support the assayed function.

In the interpretation of the GFP data it is important to realize that the deactivation-via-substitutions is qualitatively different from conventional in vitro unfolding experiments where unfolding is induced after the irreversible chromophore maturation. Indeed, one of the best-known enhanced variants of GFP is named superfolder GFP (sfGFP) and this name refers to the ability to fold and mature before creating non-fluorescent inclusion bodies.\(^{22}\) For GMMA, the considered equilibrium between the active and inactive states, may likewise lie before the maturation reaction with only the active state being able to mature. Furthermore, the physical situation may closer resemble a steady state where active protein is removed from the equilibrium by irreversible maturation, possibly in a chaperone dependent fashion,\(^{23}\) and inactive protein is removed, e.g. by protease degradation or aggregation. In this scenario, the kinetics of maturation and other cellular processes could also influence the outcome of GMMA, highlighting how GMMA probes the biophysical quantities that are important in a cellular context.

**Model estimation**

The global fit of the effects of the individual substitutions is complex for at least two reasons. First, the global parameters including baselines (Figure 1(a), black line) are estimated simultaneously with the fitness potential of all individual variants (Figure 1(a), abscissa values), which may result in a rough optimization surface featuring many local optima. Second, many substitutions are observed in only few variants and may be poorly determined with greater uncertainties that are combined with otherwise well-determined substitution effects. To address these complexities, we use a three-step estimation procedure (Figure 1(d)). The first two steps focus on achieving initial estimates of global parameters (Figure 1(e)) and individual effects (Figure 1(f)) using a mean-field approach. The final global optimization in step three enhance consistency by allowing the fit to converge to the nearest optimum (Figure 1(g)).

Previous analyses showed that the fluorescence of a variant, to first order, simply depends on the number of substitutions.\(^9\) Thus, in the first step, an initial global wild-type potential, \( F_{\text{WT}} \), is estimated together with the average effect of a substitution, \( \langle \Delta F \rangle \) (Figure 1(e)). This average effect replaces the sum in Eq. (1):

\[
F_v = F_{\text{WT}} + N_v \langle \Delta F \rangle
\]  

where, \( N_v \) is the number of substitutions in variant \( v \). While this assumption is rather crude, it is only used in the estimates of initial values in step two, and it becomes robust with an increasing number of variants of each \( N \)-mutant (i.e. a variant with \( N \)
amino acid substitutions). The average effect, \( \langle \Delta F \rangle \), may be biased by IFS (which are in principle infinite) and these are therefore not used in the fit of \( F_{wt} \) and \( \langle \Delta F \rangle \) in steps one and two (see Methods).

In the second step, effects of individual substitutions are estimated by an approach similar to that used in the first step (Figure 1(f)). For each substitution, \( s \), \( \Delta F_s \) is estimated from the subset of variants that contains that substitution. The effect is included in Eq. (2) by replacing one term of the average effect with the specific effect of \( s \) to be estimated with the remainder of the terms described by an average effect:

\[
F_v \approx F_{wt} + \Delta F_s + (N_v - 1) \langle \Delta F \rangle
\]  

A robust fit to Eq. (3) requires a sufficient number of \( N \)-mutants (to estimate the average brightness) for several different values of \( N \), but in contrast to above we only consider the subset of variants that contains the substitution \( s \). To illustrate the required number of variants per substitution, Figure 1(f) shows three examples of fits to Eq. (3) where it is clear that the 14 observed variants containing L178V result in average values of the \( N \)-mutant brightness that do not fit the model well. In contrast, the enhancing V163A and detrimental C48S substitutions, both fit the model well with 677 and 194 observed variants, respectively. These considerations are central for the final uncertainty estimation used to identify accurately estimated substitution effects.

When all initial effects have been estimated by this mean-field approach, we perform a global optimization in step three using damped least-squares optimization from which uncertainties are calculated (see Methods).

The three-step estimation strategy ensures that effects that are poorly estimated due to few observations, typically < 10–20 (Figure 1(f) and supplementary Figure S1), do not affect the initial estimation for substitutions with good data. A 5-fold cross validation shows that the model is able to predict the brightness of GFP variants not used in training accurately with an RMSD on brightness of 0.25 (approx. 10% of the dynamic range from 1.4 to 3.8) and a Spearman’s correlation coefficient \( r_s = 0.87 \) (Figure 2). Notably, the most substituted variants are predicted well with 91% of active variants (brightness >2.7) correctly predicted as active for variants with six to eight substitutions and 73% for variants with nine or more substitutions. Thus, a strength of GMMA is that all the data are analyzed globally, thus increasing the robustness of inferring the effects of substitutions that are individually less well determined.

The GMMA formalism, including the substitution-variant graph, enables a description of library features that are useful for learning substitution effects. The GFP library has on average 33 variants per substitution and the cross validation (Figure 2) and subsampling experiments (supplementary Figure S1) suggests that a library approximately half the size would work as well. The distribution of variants in the GFP library

---

**Figure 2.** Five-fold cross validation of GFP multi-mutants where each variant was randomly assigned to one of five groups. (a) Root mean square deviation (RMSD) and mean absolute error (MAE) of predicted brightness for the full model fitted to all data (black) and the mean of the training and test data of the five cross validations (red and blue respectively). The error bars show the standard deviation from the five models each trained on ≈80% of data and the dashed line indicates the experimental standard deviation of the wild-type brightness. (b) Predicted vs. experimental brightness of all variants, where the prediction of each variant is from the cross-validation when used as test data. The variants are colored according to the number of substitutions to show that even the mostly mutated variants are well predicted.
resembles a randomly connected graph, consistent with a library generated by epPCR (supplementary Figure S2). In order for GMMA to achieve accurate substitution effects from a single experiment, it is critical to have a variant library of sufficient size and with appropriate connectivity; the substitution-variant graph provides a convenient formalism to study this.

With the analogy of the fitness potential with unfolding free energy discussed above, it is interesting to consider to what extent the wild-type fitness potential, $F_{wt}$, relates to the thermodynamic stability as measured by chemical unfolding. We estimate the potential of the reference sequence (avGFP + F64L) to be $F_{wt} = -2.8$ using room temperature and a gas constant corresponding to units of kcal/mol. This is substantially smaller in magnitude than values $< -10$ kcal/mol reported from unfolding experiments of GFP (avGFP + Q80R:F99S:M153T:V163A); however, unambiguously determining the stability of GFP is challenged by at least one folding intermediate of stability $-3.7$ kcal/mol and very slow unfolding kinetics. The comparison is further complicated by in vitro unfolding experiments being qualitatively different from the equilibrium probed by GMMA. As discussed above, GMMA may probe the stability of the pre-mature form of GFP and its ability to covalently and irreversibly mature the fluorophore in the context of a cell, whereas unfolding experiments deactivates the mature protein in vitro. Since maturation is spontaneous, the pre-mature state is likely less stable compared to the mature protein and thus, GMMA could probe a less stable structure compared to unfolding experiments of the mature protein.

We also tested the robustness of the conclusions and values obtained by GMMA using synthetic data. Specifically, we generated multi-mutant data resembling those obtained for GFP and asked whether GMMA could recover the variant effects that were used to generate the data (see Methods). This analysis showed that it is possible to recover quantitatively the variant effects when the global model that relates the fitness potential to the experimental readout is accurately known. Importantly, however, the analysis also revealed that GMMA can robustly recover the ranking of the variant effects and distinguish those that are beneficial from those that are not, even when the global model is imperfect (supplementary Figure S3). In that case, however, the scale of the estimated fitness potential might not be accurate.

**GFP substitution effects.** Using GMMA we obtain accurate estimates for 1,107 substitution effects (59% of the 1,879 present in the library) with 80% found to decrease and 8% enhance the fitness potential, and the remainder to have an insignificant or neutral effect (Figure 3). In the following we highlight some points that can be learned about GFP from these.

The majority of enhancing substitutions are found at solvent exposed positions (63/83; 76%; Figure 3) together with almost all substitutions with insignificant effect, i.e. $\Delta F_s$ close to zero (126/140; 90%). Both percentages are enrichments compared to the overall fraction of exposed substitutions (712/1,107; 64%). Nearly all substitutions at buried positions have detrimental effects (361/395; 91%) with the notable exception
of the top-ranking substitution, V163A, which enhances the potential 38% more than the second best substitution, K166Q, and has low uncertainty due to a large amount of observed variant contexts in the data (supplementary Figure S4). Another indication that V163A is highly beneficial is that another substitution at this position, V163G, is found to be the 12th most enhancing substitution.

A group of positions with enrichment of enhancing substitutions have Phe, Ile and Leu as wild-type residues at solvent exposed positions for which 11/64 (17%) of substitutions are enhancing compared to 63/712 (9%) for exposed positions overall. With the exception of two Ile to Phe substitutions, the enhancing amino acids probed at these positions tend to slightly decrease hydrophobicity which is in apparent contrast to other engineering studies that find increased surface hydrophobicity to be stabilizing but fits well with the observation that such stabilizations tend to have poor additive contributions.

Surface positions with wild-type Glu have a significantly higher fraction of enhancing substitutions, 12/71 (17%), as compared to Asp with only 5/90 (6%) substitutions being enhancing. This may suggest a qualitative difference in the role of these two negatively charged amino acids (that otherwise have the same transitions in the codon table) in the context of a beta-barrel. Furthermore, four of the 15 most enhancing substitutions remove a Glu sidechain from the surface (Figure 4).

All 42 substitutions at all ten Pro residues in GFP are found to be detrimental, which is not surprising considering the structural role of Pro. Other positions that only show highly detrimental substitutions include the chromophore positions Y66 and G67 and the maturation-related sites R96 and E222.

The GMMA has some similarities to a previously described neural network that was used to fit the experimental data. Indeed the ΔFₜ values we infer from our GMMA correlate strongly with the neural networks’ hidden values for the 952 substitutions we could obtain from the authors of (supplementary Figure S5). Those values were shown to correlate to stability calculations by Rosetta, and by analogy we examined whether the variant effects obtained from GMMA could be rationalized by structural stability of the mature state of GFP. We thus used Rosetta to predict the change in thermodynamic stability for all individual substitutions, and overall find a relatively good correlation between GMMA and Rosetta (rₛ = 0.6; supplementary Figure S6(a))—a value that is higher than previously reported correlations between Rosetta stability calculations and the results of multiplexed assays of variant abundance—reinforcing the previous observation that structural stability is a substantial component of the fitness potential. However, when we focus on those substitutions identified by GMMA as most beneficial we find that most are not identified by the structural stability calculations since none of the top 15 substitutions from these two methods overlap (supplementary Figure S6(a)). We hypothesized that the global analysis of all variants in GMMA improves our ability to estimate the effects of individual substitutions compared to the measurements of the variants carrying just those individual substitutions—in particular for those substitutions that are most beneficial to function. This hypothesis is supported by our finding that the Rosetta calculations correlate more strongly with the results of GMMA than of the measured brightness of singly-substituted variants (rₛ = 0.6 vs rₛ = 0.4, respectively; supplementary Figure S6).

**Engineering GFP based on GMMA.** Having used GMMA to identify individual substitutions that can suppress the detrimental effect of other substitutions, we examined whether these substitutions alone could be combined to improve GFP. Thus, we progressively inserted the 15 most enhancing substitutions into GFP in sets of three (Figure 4(b)) and measured fluorescence levels during expression in *E. coli* at 41 °C. It is important to note that we use the measurement of fluorescence directly in cells because this closely resembles the conditions of the assay used to generate the data. The elevated temperature was applied to enhance the possible associated differences in thermal stability. As references we compare to the reference GFP sequence used in the screen (WT) and sfGFP (see Methods).

We find that all variants show fluorescence levels that are higher than the reference sequence (Figure 5 and supplementary Figure S7) with the Top6 variant giving as high an in-cell signal as sfGFP. Notably, Top6 and sfGFP only have two substitutions in common, V163A and A206V, (Figure 4(b) and supplementary Table S1) whereas sfGFP contains 10 substitutions relative to WT. The enhanced brightness of the cells expressing these variants may reflect a multitude of mechanisms including maturation kinetics, chaperone dependence, expression levels,
quantum yield, spectral properties, etc. Without further investigation of this, we highlight that GMMA aim to enhance the function that is assayed in the experiment and at the conditions applied. It is not sensitive to a particular mechanism except that substitutions should be able to rescue variants that are inactivated by other mutations. We expect the increased fluorescence mainly to reflect a difference in protein levels (quantum yield of eGFP is already 0.6) and thus to be a proxy for in vivo maturation and stability. While the Top12 and Top15 variants have lower fluorescence than Top6 and 9 they are still at least as bright as the reference wild type.

These experiments demonstrate two important points of GMMA. First, that at least the six top-ranking substitutions appear to have a substantial additive positive effect, with Top6 giving significantly more signal than Top 3. One might speculate that the leveling off of Top9 and the decreased signal of Top12 and Top15 relative to Top6 may be caused by context-specific effects, due e.g. to the structural proximity between I171V and E172A or T203I, S205T and A206V (Figure 4). The latter combination, however, appears to be close to optimal in other high-throughput experiments. Second, none of the 15

![Figure 4. Positions (a) of the top 15 enhancing substitutions (b) shown on the structure of GFP (PDB entry 1EMM). Position 163 has two substitutions within top 15 and rank 9 is not resolved in this structure. Substitutions known from sfGFP are marked in bold in (b).](image)

![Figure 5. Experimental test of multiply-substituted GFP variants composed of top ranking substitutions identified by GMMA. Error bars show the standard deviation between four replicates.](image)
most highly ranking substitutions are highly deleterious (all variants are fully functional). This observation supports the robustness of the outcome of GMMA since these enhancing substitutions have all been observed in a diverse set of multi-mutants in which they have been judged to play an enhancing role.

In addition to the calculated uncertainties, GMMA offers transparent analysis of the final fits of individual substitutions (supplementary Figure S4). These reveal that effects of the top 5 ranking substitutions are well-determined by the data, whereas rank 6 (T43N) is only observed in nine variants and thus we have less confidence in this. While V163A and A206V are both present in sfGFP, the high fluorescence of Top6 suggests that K166Q, E172A and V68M may be promising substitutions for GFP engineering. Continued experimental investigation of these specific effects is interesting, but beyond the scope of the present study.

Known GFP substitutions. To examine whether the substitutions we identify using GMMA more generally have enhancing and background-insensitive effects, we compare these to substitutions that are known to be beneficial for GFP in multiply substituted contexts found in the literature22,23,32–35 (see Methods). Known substitutions are in general found to be enhancing or with insignificant effect (Figure 6(a)) and GMMA outperforms both Rosetta stability calculations and experimental measurements of single-substitution effects in the identification of known substitutions (Figure 6(b)). Known substitutions that are estimated by GMMA to have a detrimental effect (supplementary Table S1) may indeed also be accurate because they were originally selected for other purposes. Perhaps most notable are the chromophore substitution S65T, selected for spectral properties, and C48S, introduced to avoid cysteine oxidation in extracellular environments. Likewise, the substitutions Q80R and H231L, here estimated to be slightly detrimental, are historical substitutions caused by early PCR errors which are still present in some synthetic genes of GFP variants.30

We furthermore consider 136 substitutions found in GFP variants deposited to the PDB. As expected, these are found to be enhancing or with insignificant effect. Of the 136, only four substitutions were found to have a slightly detrimental effect (ΔF < |Fwt|/2) which confirms that GMMA robustly avoids deleterious substitutions also on a larger set of substitutions. In total, we find that 19 of the top 30 enhancing substitutions obtained from GMMA are known and that none of the known substitutions are estimated to inactivate GFP (ΔF > |Fwt|).

The effects estimated by GMMA correlate strongly with the observed brightness of singly-substituted variants (Figure 6(a); r = -0.77). This, however, is mainly caused by the fact that the averaged FACS readout and GMMA analysis agree on low brightness variants. In contrast, when the goal is the more challenging one of...
identifying enhancing variants, we see a much weaker correlation between the observed brightness and GMMA ($r_s = -0.26$ for variants identified by GMMA as significantly enhancing; see Methods) because individual enhancing variants have only modest effects on brightness when introduced in an already stable background (Figure 6(a), insert).

**Conclusions**

We have here introduced a concept to quantify enhancing effects of individual substitutions through a global analysis of high-throughput assay measurements of multi-mutants. Our work builds on previous experiments probing the fitness landscape of GFP and analyses that show that most apparent epistatic effects can be explained by combining (weakly) destabilizing substitutions. We extend these analyses by focusing on the ability to improve function by finding “suppressors” that diminish the effects of destabilizing substitutions. We thus demonstrate that the effects of single amino acid substitutions may be better determined by analysis of multi-mutants and have presented a Global Multi-Mutant Analysis (GMMA) that implements this.

Our fitness potential is shown to correlate with structural stability (supplementary Figure S6), with the important addition, e.g. compared to in vitro unfolding experiments, that substitutions that decrease function appear as destabilizing. Our model is based on thermodynamics, Eq. (4), and the fitness potential may be interpreted as a free energy of (un)folding under the given assumptions; most notably the absence of couplings between substitutions and the global map between the fitness potential to the assay readout. Recent large-scale stability measurements confirm that even among residues that are close in space, nearly all stability effects are close to additive. Our computational experiments using synthetic data furthermore indicate that the ranking of substitutions and zero-point (fraction of enhancing substitutions) are robust towards inaccuracies in the global model and thus, that GMMA can be interpreted as a free energy with a system dependent scale.

A cross-validation analysis demonstrates that GMMA can be used to predict experimental results on a complex fitness landscape involving nine or more substitutions in many variants which is also confirmed experimentally by fluorescence measurements of multiply-substituted GFP variant that contains the top ranking GMMA substitutions. We show that the six best substitutions may be combined to design a GFP variant which, under the assay conditions, is highly fluorescent and identify at least three enhancing substitutions, K166Q, V68M and E172A, that are known but seem underestimated in the literature of enhanced GFP variants. Finally, we show that GMMA is able to identify a large number of the substitutions that are known to be beneficial to GFP in multiple-substituted contexts. A recent application of GMMA to the protein Thioredoxin using an in-cell folding sensor (performed after this work was completed) showed that the identified substitutions could be used to increase the stability of a very stable protein well beyond the dynamical range of the cellular assay. Thus, together these results suggest that GMMA will have a broad application to enhance protein stability and function.

**Data and code availability**

The code and dataset generated and analyzed during the current study are available in the GitHub repository, https://github.com/KULL-Centre/_2022_Johansson_GMMA_GFP.

**Methods**

We explore a model based on thermodynamic principles and assume (i) that the fitness potential can be described mathematically via an equilibrium between the active state, $A$, and the inactive state, $D$, and (ii) that the observed brightness, $B$, of a variant, $v$, is proportional to the fraction of the protein found in the active state:

$$B_v \propto \frac{[A]}{[A] + [D]} = \frac{1}{1 + \exp[F_v/RT]}$$

(4)

Here $[A]$ and $[D]$ are the concentrations of active and inactive protein respectively, $R$ is the gas constant and $T$ the temperature. We use a value for $RT$ that corresponds to a fitness potential in kcal/mol at room temperature. In an in vitro experiment in dilute solution, the fluorescence should be linearly proportional to the fraction of active protein. In the context of a cell and a FACS experiment this relation may, however, be more complex; we thus choose to follow the original report of the GFP data, and use the log-fluorescence as the brightness $B$, assumed to be proportional to the fraction of active protein. If the proportionality of the brightness and fraction of active protein is an accurate description, the fitness potential may be interpreted as a free energy that describes the equilibrium $A = D$. In a cellular context the relationship between thermodynamic stability and protein abundance may be more complex for example due to the presence of chaperones and proteases, and the fitted values of $F_{wt}$ and $\Delta F_s$ should therefore not be strictly interpreted as thermodynamic stability parameters.

We note that our approach is in principle not limited to additive effects and that couplings may be included in Eq. (1), provided that the data warrants estimation of these. Possibly due to the nature of epPCR, we were not able to find a double mutant cycle with sufficient data to make a case but cou-
plings could be estimated from variant libraries designed to study particular interactions. Also, we note that the model fits well with linear terms only, probably because many stability effects are highly additive.36 We carried out all data analyses using the R project for statistical computing with packages minpack.lm and igraph.

**Initial estimation**

The initial estimation of stability effects considers the average effect of substitutions, \( \langle \Delta F \rangle \), which is sensitive to highly destabilizing substitutions. Thus, we first detect irreversible fatal substitutions (IFS) that are inactive in all contexts and have an unlikely pattern of activity among multi-mutants, e.g. many inactive double mutants in the case of GFP (see supplementary Appendix 1). We stress that the IFS are detected directly from an analysis of the data and is not based on any pre-defined notions of variants or regions that are particularly relevant for function. Variants containing nonsense mutations are also generally expected to be IFS. We exclude 62 nonsense IFS and 51 missense IFS from the initial estimation together with 2,310 + 4,851 variants that contain these nonsense and missense IFS substitutions, respectively.

The wild-type fitness potential, \( F_{wt} \), is first estimated together with the average effect of all substitutions, \( \langle \Delta F \rangle \), using Eq. (5) which combines Eqs. (2) and (4). These are fitted to the average brightness of the N-mutants, \( B_N \), i.e. the average brightness of double-mutants, triple-mutants, etc.

\[
B_N = \frac{\alpha_a + \alpha_0 \exp(-F_{wt} + N\langle \Delta F \rangle)/RT}{1 + \exp(-F_{wt} + N\langle \Delta F \rangle)/RT}
\]  

(5)

Here, a constant baseline, \( \alpha_0 \), for the brightness of the inactive variants is fitted whereas a constant baseline for active variants, \( \alpha_a \), is not fitted independently but calculated from \( F_{wt} \) and the brightness of the wild-type sequence, \( B_{wt} \), during fitting:

\[
\alpha_a = B_{wt} + (B_{wt} - \alpha_0) \exp(-F_{wt}/RT)
\]  

(6)

This makes \( \alpha_a \) less sensitive to noise and outliers in the variant readout by relying on the data point \((F_{wt}, B_{wt})\) which is experimentally well-determined with 3,645 barcodes in the high-throughput assay, i.e. individual observations of wild-type nucleotide sequence (2,444) or synonymous sequences.7 A standard error of each parameter is calculated as the square root of the diagonal of the inverse Hessian matrix. The initial wild-type potential is estimated to \( F_{wt} = -1.8 \pm 0.2 \), the average effect of substitutions \( \langle \Delta F \rangle = 0.45 \pm 0.04 \), and \( \alpha_0 = 1.43 \pm 0.03 \) (Figure 1(e)). This results in \( \alpha_a = 3.8 \).

In step 2, we use the values of \( F_{wt} \), \( \langle \Delta F \rangle \) and the baseline parameters from step 1 to estimate initial values of the individual substitutions, \( \Delta F_s \), from the subset of variants that contains the substitution, \( s \), using Eqs. (3) and (5) (Figure 1(f)). The use of a fixed value of \( \langle \Delta F \rangle \) makes the initial estimates robust and self-consistent for the global fit, and the approach is further supported by the observation that variant effects are mostly independent of the details of the background.41 Since we always only change the background with a single substitution, \( \langle \Delta F \rangle \) is not expected to vary much and in practice less than the uncertainty that results from small sets of variants. We require that all subsets have a diversity of at least 3 different multi-mutants that spans the transition region and gives a fit with standard error < 1.0 and an absolute deviation < 5.0 log fluorescence units. With this, we find that 56% of substitutions have sufficient data for this initial estimation. We then estimate the remaining initial \( \Delta F_s \), values from these well-determined effects.

**Graph analysis**

We pre-process the data for the global multi-mutant analysis and build a bipartite graph by assigning protein variants to one layer of nodes and another layer of individual substitutions. All variant nodes are linked to the substitution nodes that the variant is composed of. We check that all nodes are connected in the graph. If a subset of variants is composed of a subset of substitutions that does not occur in the rest of the variants, this graph becomes disconnected from the rest, and GMMA may be carried out on the subset alone (e.g. a single-mutant library is fully disconnected, GMMA cannot be applied to this). Single mutants do not inform the global fit more than any other variants and indeed 130 stability effects are estimated without the single mutant being observed. Substitution nodes with a single link represents substitutions that contributes one parameter and only one data point to the global analysis, referred to as hanging nodes. These do not inform the global optimization and, thus, the effect of these are calculated after the global optimization. In summary, the graph is cleaned for 7 disconnected node pairs, and 257 hanging substitution nodes together with 255 dependent variants nodes. Finally, the graph is examined to ensure that no pair of substitutions only occurs together as these would make them impossible to distinguish and should be reparametrized as a single effect. The graph analysis is independent of the initial estimates and considers all data, including IFS and non-sense substitutions.

**Global estimation**

With input from both the initial estimation and the graph analysis, we perform the global optimization. The global fit was not stable without some type of regularization (presumably because of parameter correlations) and thus, we apply a trust region as
implemented in the Levenberg-Marquardt least-squares algorithm with effects limited to the range –5 to 10. This is chosen over an elastic net type of regularization because we are interested in the fitted parameter values and these are not expected to follow a single normal distribution (e.g. centered at zero) and thus, a more complex prior would be needed, e.g. for ridge regularization. We optimize all effects $\{\Delta F_s\}$ and the wild-type fitness potential $F_{wt}$ to the nearest optimum (Figure 1(g)). The baselines are fixed at the values determined in the initial fit (1.4 and 3.8). With analytically calculated gradients, the global optimization of 1,616 parameters from 53,763 data points took 4–5 hours on a normal laptop.

The fit has a reduced chi-square $\chi^2 = 4.4$ which indicates that some parts of the data do not fit the model; however, we note that the cross-validation analysis shows that the data is robustly fitted and with at most a small level of overfitting (Figure 2). One contribution to an elevated $\chi^2$ could be the use of the distribution of the observed brightness of the wild-type sequence as a proxy for the uncertainties of all variants, which may indeed have higher uncertainties in the assay. Our fit uncertainties of all variants, which may indeed have higher uncertainties in the assay. Our fit

Uncertainties

In the global analysis, we estimate the uncertainties from the covariances in the inverted Hessian matrix. We calculate two error measures to judge the uncertainty and reliability of the effects. The first, $\delta_s$, is calculated using the log-fluorescence measurement uncertainty, reported to be $\delta_{B_{wt}} = 0.11$ for the wild-type sequence, $^9$ propagated via the covariance matrix diagonal and multiplied by 3 to get the 99.7% percentile (and to somewhat compensate for the expectation that the variants have higher uncertainty than the wild type):

$$\delta_s = 3 \frac{\left|\Delta F_s\right|}{\delta_{B_{wt}}}$$

(8)

A resampling experiment suggests that this measure captures the estimation accuracy well (supplementary Figure S1).

The second uncertainty, $\delta_{\Delta F_s}$, is used to filter out unreliable stability effects and is calculated from:

$$\delta_{\Delta F_s}^2 = \frac{\left(\frac{\partial \Delta F_s}{\partial B}ight)^2 \text{RSS}_s}{\text{DOF}_s V_s}$$

(9)

Again, the derivative is from the diagonal of the covariance matrix, $\text{RSS}_s$ is the residual sum-of-squares of the variants used to estimate substitution $s$, $\text{DOF}_s$ is the number of degrees-of-freedom and $V_s$ is the number of variants used to estimate substitution $s$. The last factor gives an error-of-the-mean type of uncertainty that compensates for the case where a lucky fit of few variants is penalized. The number of degrees of freedom for a substitution assumes that a uniform fraction of parameters is estimated together with $\Delta F_s$ from the $V_s$ data points

$$\text{DOF}_s = V_s (1 - S/V) - 2$$

where $S$ and $V$ are the total number of substitutions and variants respectively. We set a relatively conservative threshold and mark 772 (61%) effects with $\delta_{\Delta F_s} > 0.05$ as unreliable (a value that can be compared to $\langle \Delta F \rangle \approx 0.5$). This low threshold has been set by manual inspection of plots that show the fit of each substitution to its respective subset of variants (similar to Figure 1 (f)). We use a constant threshold here to facilitate a clear discussion. However, in a specific application, substitutions could be judged individually based on both uncertainty measures and such plots, since many of the 772 poorly estimated substitutions are still informative (as in supplementary Figure S4). Notably, 222 (12%) of all substitutions are exclusively or predominantly observed in inactive variants and are therefore only represented in the flat region of the model. Thus, all of these may reliably be identified as destabilizing or even well determined on a range, even though the reported point estimate itself is highly uncertain. Of the remaining 550 (29%) substitutions with uncertain effects, the majority (398, 21%) are caused by poor statistics with five or fewer observed variants. The conservative threshold does exclude some substitutions with more than 10 observations (61 or 3%), that are potentially stabilizing, e.g. L221V, Q80K or E6A, and may be interesting depending on the application.

Synthetic data

Each of the 1615 substitutions considered in the global analysis (i.e. those not disconnected or hanging) are assigned their estimated effects if these pass the uncertainty threshold, and otherwise a random value drawn from a normal distribution resembling the accurate effects. Synthetic brightness values for 53,762 variants are calculated via Eq. (1), the value of $F_{wt}$ given above and one of three global models relating brightness to fitness potential: 1) the sigmoid function as used in the global fit with the parameters given above (Figure S3(A)), 2) a log-linear function $B_v = \log(-10 \times F_v + 20)$ for $F_v < 0$ and $B_v = \log(5)$ otherwise (Figure S3(B)). 3) a discontinuous, close to binary, log-linear function $B_v = \log(-20 \times F_v + 30)$ for $F_v < 0$ and $B_v = \log(5)$ otherwise (Figure S3(C)). The synthetic brightness is added a Cauchy distributed noise of scale 0.05 (compared to baselines 1.4 and 3.8) and explicit misclassification (true active variants with the brightness of an inactive or vice versa) of 158
random variants (similarly to estimated classification rate in the GFP data). Since the Cauchy distribution is not unlikely to generate samples of very high absolute values, brightness values above 4.5 are returned to the model value and values below 1.28 are set to 1.28 (minimum brightness observed in GFP data). GMMA of the three synthetic data sets is carried out as described above except that the initial mean-field estimate is omitted and random initial values are used instead.

**Classification**

For the sake of discussion and early IFS identification or inactive. Variants with brightness below 2.7, half-way between maximum and minimum observed brightness in the original data, are assigned as inactive and the rest as active.

We mark substitutions with low uncertainty in the fitness potential from GMMA (high/low uncertainty classification described above with the uncertainty calculation) as significantly enhancing if the effect plus uncertainty is less than zero, significantly detrimental if the effect minus uncertainty is greater than zero and insignificant otherwise. Substitutions with high uncertainty are marked as detrimental if the effect is greater than the wild-type potential and marked as unknown otherwise.

Solvent exposure categories are exposed or buried according to DSSP. 42

**Rosetta stability effects**

Rosetta scores were calculated according to protocols published in 43,44 using precompiled Rosetta version 3.13 of June 2021 with the REF15 score function (ref2015_cart) and the chromophore parameters distributed with Rosetta (CRO.params). Calculations were based on PDB entry 1EMM that covers position Leu-5 to Ile-227 both included. This was prepared by 20 independent and unrestrained minimizations using the command:

```
```

The lowest energy structure was used to evaluate all single variants using the command:

```
./cartesian_ddg -score:weights ref2015_cart -ddg:iterations 3-force_iterations false -ddg:score_cutoff 1.0 -bbnrs 1 -fa_max_dis 9.0 -extra_res_fa CRO.params
```

We used no special handling of substitutions involving proline and the substitution effect was calculated as the average of three repeats of the calculation. Results in Rosetta energy units (REU) were divided by 2.94 which have previously showed to correlate well with stability effects in kcal/mol. 43 Nonsense and chromophore mutations could not be evaluated and calculations at position Arg-80 were ignored because of the mismatch to our reference Gln-80.

**Known substitutions**

We focus on published variants that are relevant for our reference sequence (avGFP + F64L) and consider the substitutions that constitute sfGFP, 32 T-Sapphire GFP (tsGFP), 33 a split GFP (splitGFP), 34 a computationally-optimized GFP known as des1123 and a tryptophan chromophore variant called nowGFP 35 (supplementary Table S1). While many more variants exist in the literature, these five variants represent close to all substitutions. 32 We expect substitutions from these to be stabilizing or have an insignificant effect in GMMA. We note that some substitutions from nowGFP have been reported to support the tryptophan chromophore. 45,46

Additionally, we also consider 136 substitutions found in 147 structures of GFP, selected in the PDB to have >90% identity to our wild-type sequence. Since these variants have all been expressed and crystalized, we expect that these substitutions do not destabilize GFP substantially, i.e. lower expectations compared to substitutions from the enhanced variants mentioned above.

**Experimental testing**

The chromophore substitution S65T is known to change the spectral properties of GFP and since this is contained in sfGFP, we add this substitution to all variants (including the WT reference), in order to compare fluorescence levels measured using the same excitation wavelength of 488 nm.

Synthetic genes encoding the multiply mutated GFP variants were custom synthesized by Twist Bioscience in a pET29b expression vector and expressed using BL21(DE3) as host. The DNA sequence was codon optimized for E. coli for the reference sequence, and only varied at the positions of substitutions, where most common codons were used. Precultures were grown overnight and normalized for cell density to approximately OD600 of 3.0, respectively. Expression was carried out in 96-well plates which were inoculated with 5 µL in 45 µL LB medium containing kanamycin for plasmid selection. After incubation at 37 °C for two hours, IPTG was added to 1 mM and the temperature was shifted to 41 °C and fluorescence measured using a TECAN Infinite 200 Pro microplate reader over the next 12 hours. All measurements were done with two independent precultures, each with two technical replicates. The numbers reported in Figure 5 are averaged over the two last hours of the profiles in supplementary Figure S7.
References
