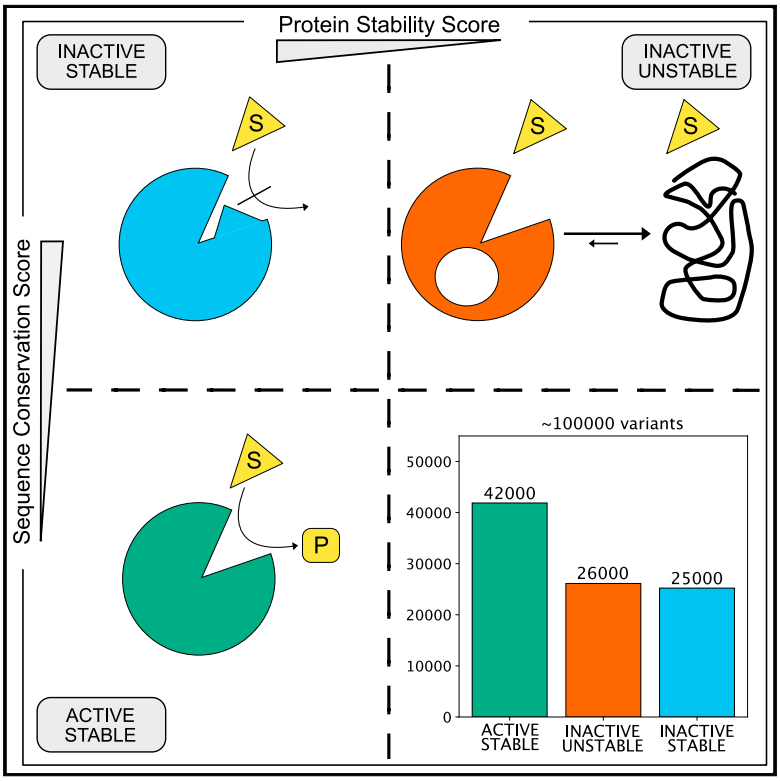
Cell Reports

Article

Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation

Graphical abstract



Highlights

- We analyze ca. 150,000 variant effects obtained by multiplexed assays
- We use protein stability and sequence conservation to predict variant effects
- About half of the loss-of-function variants lose function with loss of stability

Authors

Magnus Haraldson Høie, Matteo Cagiada, Anders Haagen Beck Frederiksen, Amelie Stein, Kresten Lindorff-Larsen

Correspondence

amelie.stein@bio.ku.dk (A.S.), lindorff@bio.ku.dk (K.L.-L.)

In brief

Missense variants in which a single amino acid is changed in a protein underlie a large number of genetic diseases. By combining experimental measurements of ca. 150,000 variant effects with analyses of protein stability and sequence conservation, Høie et al. train a machine learning model to predict variant effects.

Høie et al., 2022, Cell Reports *38*, 110207
January 11, 2022
https://doi.org/10.1016/j.celrep.2021.110207

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation

Magnus Haraldson Høie,[1] Matteo Cagiada,[1] Anders Haagen Beck Frederiksen,[1] Amelie Stein,[1,*] and Kresten Lindorff-Larsen[1,2,*]

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark

[2]Lead contact

*Correspondence: amelie.stein@bio.ku.dk (A.S.), lindorff@bio.ku.dk (K.L.-L.)

https://doi.org/10.1016/j.celrep.2021.110207

SUMMARY

Understanding and predicting the functional consequences of single amino acid changes is central in many areas of protein science. Here, we collect and analyze experimental measurements of effects of > 150,000 variants in 29 proteins. We use biophysical calculations to predict changes in stability for each variant and assess them in light of sequence conservation. We find that the sequence analyses give more accurate prediction of variant effects than predictions of stability and that about half of the variants that show loss of function do so due to stability effects. We construct a machine learning model to predict variant effects from protein structure and sequence alignments and show how the two sources of information support one another and enable mechanistic interpretations. Together, our results show how one can leverage large-scale experimental assessments of variant effects to gain deeper and general insights into the mechanisms that cause loss of function.

INTRODUCTION

The ability to predict and understand the effects of amino acid changes to protein structure, stability, and function plays central roles in a number of areas of protein science. For example, improving protein function or stability is key in many biotechnological applications, and the ability to understand and predict loss of function is of central importance in disease biology. Large-scale human genome sequencing efforts are revealing millions of missense variants that change the amino acid sequences of proteins, but we do not yet know the functional consequences for most of these variants.

Studies of the effects of single amino acid changes also present opportunities to test our understanding of the protein structure-function relationships. While a large fraction of single amino acid substitutions in a given protein are relatively well tolerated, there is a subset that has significant detrimental consequences (Schaafsma and Vihinen, 2017; Gray et al., 2017). Pinpointing which variants are in the detrimental group, and the biochemical and biophysical mechanisms underlying loss of fitness, is important, for example, for assessing pathogenicity of so-called variants of uncertain significance (Richards et al., 2015) and understanding the mechanistic origins of disease.

Recent technological advances have enabled high-throughput assays that can quantify changes in activity, stability, or other protein properties of interest for thousands of variants in a single experiment. Such multiplexed assays of variant effects (MAVEs; also called deep mutational scanning experiments) also provide large sets of data with systematic fitness profiles of variants (Fowler and Fields, 2014), often providing both mechanistic insight and systematic assessment of computational models for predicting variant effects. MAVEs are often performed using customized assays, which, however, also means that functional scores (sometimes called fitness scores) extracted from different MAVEs are often not directly comparable with one another without further normalization. The data generated by MAVEs have been shown to help predict the status of pathogenic and benign variants, while also serving as useful benchmarks for computational variant classification methods (Livesey and Marsh, 2020; Frazer et al., 2021). More generally, the data can also provide detailed insight into general aspects of protein structure and function (Gray et al., 2017; Ahler et al., 2019; Stein et al., 2019; Dunham and Beltrao, 2021; Chiasson et al., 2020; Starr et al., 2020; Cagiada et al., 2021; Amorosi et al., 2021).

A given amino acid change may affect multiple properties or functions of a protein, and by combining different assays it may be possible to disentangle which substitutions affect which of those properties and functions. As most proteins need to be folded to function, assays probing protein stability and cellular abundance have received special attention. Thus, a specific type of MAVE termed VAMP-seq (variant abundance by massively parallel sequencing) has been developed to probe cellular protein abundance (Matreyek et al., 2018) and has been shown to correlate with measurements of protein stability (Matreyek et al., 2018; Suiter et al., 2020). While the detailed relationship between protein stability and abundance is complicated

and not fully understood (Hingorani and Gierasch, 2014; Stein et al., 2019), we and others have shown that unstable proteins are often targeted for proteasomal degradation leading to lowered cellular abundance, and a large number of disease-causing variants are proteasomal targets and are found at low cellular levels (Meacham et al., 2001; Yaguchi et al., 2004; Olzmann et al., 2004; Ron and Horowitz, 2005; Yang et al., 2011, 2013; Arlow et al., 2013; Nielsen et al., 2017; Chen et al., 2017; Nielsen et al., 2017, 2020; Scheller et al., 2019; Abildgaard et al., 2019). By combining the results from a VAMP-seq experiment with a MAVE probing protein activity it is possible to distinguish between variants that cause loss of function due to lowered abundance and those that change the intrinsic activity of a protein (Jepsen et al., 2020; Chiasson et al., 2020; Cagiada et al., 2021; Amorosi et al., 2021). These experiments and analyses suggest that a relatively large fraction of variants that cause of loss of function are due to loss of stability and resulting degradation in the cell.

Although it is possible to perform multiplexed assays on multiple genes in a single experiment (Després et al., 2020; Jun et al., 2020; Hanna et al., 2021; Cuella-Martin et al., 2021), we are not yet able to probe all possible variants in all proteins by experiments. Thus, computational methods are important to predict and understand variant effects, and in some cases they may be even be more accurate than MAVEs for this purpose (Jepsen et al., 2020; Frazer et al., 2021). Most variant effect predictors are based on features extracted from evolutionary conservation of homologous proteins, biophysical calculations based on structure, and general knowledge of amino acid properties (Yue et al., 2005; Kumar et al., 2009; Adzhubei et al., 2010; Casadio et al., 2011; De Baets et al., 2012; Kircher et al., 2014; Choi and Chan, 2015; Ioannidis et al., 2016; Ancien et al., 2018; Wagih et al., 2018; Gerasimavicius et al., 2020; Livesey and Marsh, 2020). As some of these methods have been trained and tested on classification of clinical variants, it has been argued that comparison against data from MAVEs provides a useful and unbiased alternative to benchmark such methods (Livesey and Marsh, 2020). In such tests, it has been shown that various sequence-based approaches, including deep-learning methods, can achieve very high accuracy (Riesselman et al., 2018; Livesey and Marsh, 2020). Indeed, we and others have successfully applied sequence analysis and biophysical stability calculations for identification and analysis of pathogenic variants, although these methods were not trained on clinical variants (Pey et al., 2007; Yin et al., 2017; Nielsen et al., 2017; Gray et al., 2018; Scheller et al., 2019; Cline et al., 2019; Abildgaard et al., 2019; Jepsen et al., 2020; Frazer et al., 2021).

Experiments and computational analyses such as those discussed above are now beginning to provide a consistent picture of the effects of variants on protein stability and function. Variants that cause substantial loss of stability are generally found at low protein levels in the cell, and thus often lead to a loss-of-function phenotype and disease. Hence, when a variant is predicted to be highly destabilizing, it is likely to be non-functional. The reverse, however, does not hold true. Variants that do not perturb protein stability may still cause loss of function via other mechanisms, such as perturbing active-site residues in an enzyme or key interaction sites for binding. Such effects can often be captured by evolutionary sequence analyses, which are potentially sensitive to all conserved molecular mechanisms that lead to loss of function.

We recently quantified how variants may affect protein activity and abundance and the relationship between the two (Jepsen et al., 2020; Cagiada et al., 2021). By analyzing two MAVEs, one probing activity and another probing abundance, for two proteins (PTEN and NUDT15) we found that about one-third of all possible variants cause loss of function, and about half of the loss-of-function variants also have low cellular abundance. Predictions of protein stability and analyses of sequence conservation demonstrated that many of these effects could be recapitulated by computational methods. Specifically, we found that analyses of sequence conservation captured general aspects leading to loss of function, whereas calculations of protein stability were more strongly correlated with experimental measures of protein abundance. We also found that variants that cause loss of function, but not lowered abundance, were enriched in the active sites of PTEN and NUDT15, whereas variants that cause loss of function via decreased protein abundance were more often found in the core of the protein structure (Cagiada et al., 2021).

Here, we aim to provide further insight into the relationship between variant effects on protein stability and function, and how computational predictions of changes in thermodynamic stability and analysis of sequence conservation may be used to predict variant effects. We have collected 39 datasets previously generated by MAVEs on 29 proteins, and analyze these using analyses of sequence conservation and predictions of protein stability. We train a machine learning model that uses these features as input to predict variant effects, and show that both sequence conservation and predictions of protein stability contribute to prediction accuracy. We use our model to provide a global view of the relationship between stability and function, and help pinpoint which variants lose function due to loss of stability. Together, our results show how loss of stability is an important contributor to loss of function, and point to future improvements for predictions of variant effects.

RESULTS

Analyzing variant effects from MAVEs by calculations of stability and conservation

We first aimed to quantify how well analyses of protein stability and sequence conservation are able to capture experimental measurements of variant effects. We thus collected 39 datasets generated by MAVEs on 29 proteins from the literature (Table S1). As we aimed to use the data in a globally trained machine learning model, we used rank normalization to bring the original variant scores reported by the individual studies onto a common scale (s_{exp}), where $s_{exp} \sim 1$ corresponds to wild-type-like activity in the experiment and $s_{exp} \sim 0$ corresponds to variants with low activity in the assay used in the MAVE.

For each of the 29 proteins we used Rosetta (Park et al., 2016; Leman et al., 2020; Frenz et al., 2020) to predict changes in thermodynamic stability ($\Delta\Delta G$) for each of the 19 possible variants at each of the positions resolved in the experimental structures. Here, $\Delta\Delta G = 0$ corresponds to the same stability as the wild-type protein and variants with $\Delta\Delta G > 0$ correspond to those that

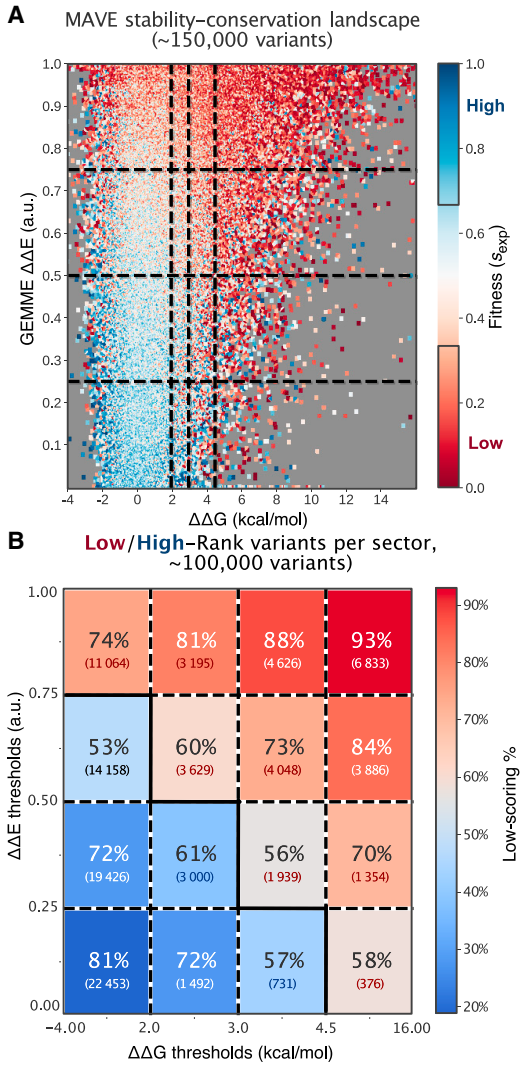


Figure 1. Stability and conservation score trends across variants

(A) Analysis of the variant fitness landscape across over 150,000 variants in 39 MAVE experiments, by GEMME ΔΔE and Rosetta ΔΔG score per variant, colored by normalized MAVE fitness score.

(B) Percentage of high-fitness (top 33% of individual MAVE experimental scores, blue) or low-fitness (bottom 33%, red) variants per sector in the fitness landscape, and number of variants per sector. The middle tertile (33rd–66th percentile) of variants are excluded here.

are predicted to destabilize the protein. We also built a multiple sequence alignment for each of the proteins and analyzed these using GEMME (Global Epistatic Model for predicting Mutational Effects) (Laine et al., 2019). Following rank normalization for each protein, the resulting GEMME ΔΔE scores quantify how likely each of the 19 substitutions are in terms of what has been observed in the evolutionary record, with ΔΔE ≈ 0 corresponding to very conservative substitutions and ΔΔE ≈ 1 corresponding to variants that are extremely rare or absent from the alignment and hence predicted to be disruptive. In total, we thus collected triplets of s_{exp}, ΔΔG and ΔΔE for 154,808 single amino acid variants covering 10,012 positions in the 29 proteins (Figure 1A).

We have previously shown that changes in protein stability (ΔΔG) and multiple-sequence-alignment-based conservation scores (ΔΔE) correlate with changes in cellular function or stability in selected proteins (Nielsen et al., 2017; Scheller et al., 2019; Abildgaard et al., 2019; Cagiada et al., 2021). This also holds for the 154,808 variants in the 29 proteins studied here (Figure 1A), so that variants with large ΔΔE or ΔΔG scores tend to show low fitness (low s_{exp}). In particular, variants for which both ΔΔG and ΔΔE values are high almost always show loss of function, while those with low scores for both usually show wild-type-like fitness. As expected and discussed above, variants that do not substantially perturb stability (low ΔΔG) can both have low and high values of s_{exp}, because substitutions may affect function through other mechanisms than loss of stability and abundance. From the stability-conservation landscape, however, it appears that such effects are captured by the ΔΔE scores so that, even for variants with low ΔΔG, conservative substitutions (low ΔΔE) tend to be associated with high fitness, and high ΔΔE with low fitness.

To quantify the power of ΔΔE and ΔΔG for classifying variants, we extracted the top and bottom third of variant scores (s_{exp}) as the subsets that are most clearly associated with wild-type-like and loss-of-fitness phenotypes, respectively. Next, we divided our MAVE stability-conservation landscape into a total of 16 sectors. The normalized ΔΔE scores have evenly distributed thresholds of 0.25, 0.50, and 0.75, while the ΔΔG thresholds are at 2.0, 3.0, and 4.5 kcal/mol (Figure 1). Inspection of the sectors confirms that extreme values of ΔΔE and ΔΔG can classify variants well into the high or low fitness categories, while moderate values have lower classification power (Figure 1B). For example, when ΔΔE<0.25 and ΔΔG<2.0 kcal/mol, 81% of the variants are in the high fitness category, and, at the opposite end, when ΔΔE>0.75 and ΔΔG>4.5 kcal/mol, 93% of the variants are in the low fitness category.

More generally, the two-dimensional fitness landscapes illustrate the partial interdependency of ΔΔE and ΔΔG. Most notably, it is clear that evolution selects strongly against destabilizing variants so there are almost no cases with high values of ΔΔG and low values of ΔΔE (only 4% of the variants have ΔΔE<0.25 and ΔΔG>3.0 kcal/mol). Also, as discussed above, low values of ΔΔG may be associated both with high and low values of s_{exp}, whereas variants with larger values of ΔΔG tend to have low s_{exp}. Thus, focusing on the stable variants (ΔΔG<2.0 kcal/mol) and only the top/bottom third of the fitness scores, we find that 38% of these variants have ΔΔE>0.5 (Figure 1A) and are thus more likely to be non-functional. Focusing on the variants that the GEMME analysis suggest are incompatible with what has been observed through evolution (ΔΔE>0.5), and thus are more likely to be non-functional, we find that 47% are predicted to be unstable (ΔΔG>2.0 kcal/mol; Figure 1B). Thus, in line with our previous analysis (Cagiada et al., 2021), these results suggest that approximately half of the non-conservative substitutions are selected against due to stability effects.

We have previously found that low-ΔΔG (structurally stable) but high-ΔΔE (evolutionarily constrained) variants are often found in clusters near active sites and other functionally important regions in the proteins PTEN and NUDT15 (Cagiada et al., 2021). We here examined two additional proteins (CBS and HMGCR) to understand the distribution and location of

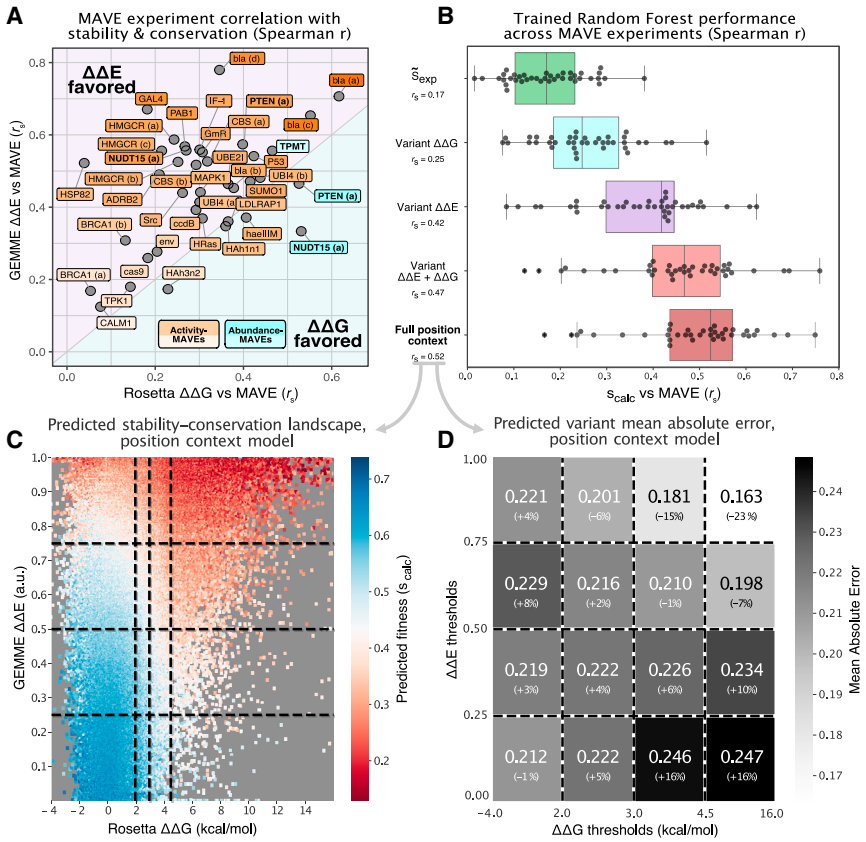


Figure 2. Correlations and predictions of variant effects

(A) Spearman correlation coefficients between GEMME ΔΔE or Rosetta ΔΔG, and the experimental variants effects (s_{exp}) for each of the 39 MAVEs.

(B) Correlation between predicted and experimental variant effects from a baseline ($\tilde{s}_{exp}$) and four different machine learning models. Apart from the first row, each of the rows correspond to one of four different random forest models that used different features as input (indicated by the labels and described in detail in the main text and STAR Methods). For each class of model, each point corresponds to one of the 39 sets of data, and the correlation coefficients were calculated in a leave-one-protein-out cross-validation across 29 proteins and 39 MAVE datasets. The boxplots show the first, second, and third quartiles and extend to 1.5 times the difference between the first and third quartiles (capped at the minimum and maximum values, respectively).

(C) Predicted fitness landscape in leave-one-protein-out cross-validation of the full position-context random forest model.

(D) MAE and relative difference from the mean MAE from the position-context random forest model.

low-ΔΔG, high-ΔΔE variants. We find that stable but evolutionarily constrained variants in these two proteins tend to locate close to known active-site residues, co-factors, or at protein-protein interaction sites (Figure S1). Further, known active-site residues had low values of ΔΔG and high values of ΔΔE. Together with previous analyses (Cheng et al., 2005; Cagiada et al., 2021), these results suggest that many low-ΔΔG, high-ΔΔE variants are found at functionally important sites that are constrained by evolution (high ΔΔE), but not due to intrinsic structural stability (low ΔΔG).

Training and benchmarking random forest models

Having established that there is an overall relationship between calculated values of ΔΔG and ΔΔE and experimental variant effects (s_{exp}) and that both independently contribute valuable information (Figure 1), we decided to train a machine learning model to predict variant effects from ΔΔG and ΔΔE. Before doing so, we analyzed how well the data generated by the individual MAVEs correlate with the calculated values of ΔΔG and ΔΔE (Figure 2A).

We find a broad range of Spearman correlation coefficients (r_s) between the MAVE scores (s_{exp}) and our calculated ΔΔG and ΔΔE scores, in line with previous observations of considerable variation in correlation with variant effect predictors (Livesey and Marsh, 2020). Correlations range from low ($r_s \sim 0.1$ on BRCA1 and Calmodulin-1 [CALM1] for both ΔΔE and ΔΔG) to relatively high ($r_s \sim 0.6 - 0.8$) when predicting multiple independent MAVEs on β-lactamase (bla) (Figure 2A). Overall and as expected, experimental variant effects that correlate well with ΔΔG also correlate well with ΔΔE. In line with the results shown above (Figure 1) and previous analyses (Nielsen et al., 2017; Abildgaard et al., 2019; Jepsen et al., 2020; Livesey and Marsh, 2020; Gerasimavicius et al., 2020; Cagiada et al., 2021), we find that variant effects tend to be more strongly correlated with ΔΔE than ΔΔG. Two notable outliers to this observation are the abundance-based MAVEs (VAMP-seq) for PTEN (Matreyek et al., 2018) and NUDT15 (Suiter et al., 2020), which are more strongly correlated with our analysis of protein stability than conservation, labeled as PTEN (a) and NUDT15 (a) in Figure 2A (Cagiada et al., 2021). Interestingly, the VAMP-seq data for the protein TPMT (Matreyek et al., 2018) is slightly more strongly correlated with ΔΔE than ΔΔG. PTEN and NUDT15 variants have also been assayed for their respective biochemical functions (Mighell et al., 2018; Suiter et al., 2020), and the resulting s_{exp} scores correlate better with ΔΔE and less well with ΔΔG than the corresponding protein abundance scores. The MAVEs of respective biochemical function are labeled as PTEN (b) and NUDT15 (b) in Figure 2A (Cagiada et al., 2021). While the reasons for the low correlation between ΔΔE and ΔΔG with several of the experiments remain unclear; possible explanations include inaccuracies in stability calculations, poor sequence alignments, or experimental assays that probe properties not directly related to stability or an evolutionary-conserved function of the protein. Indeed, similar variation in correlation between the outcome of MAVEs and ΔΔE-like and ΔΔG scores have previously been observed (Hopf et al., 2017; Livesey and Marsh, 2020; Dunham and Beltrao, 2021).

Next we constructed a simple baseline model that captures effects of substituting each of the 20 protein-coding amino acids for another by averaging over the normalized scores from all variants in our dataset ($\tilde{s}_{exp}$; Figure S2). Similar to previous observations (Gray et al., 2017), this substitution matrix captures well-known biochemical patterns. In particular, we observe that substitutions of hydrophobic residues—often found in the protein core—with charged or polar residues on average leads to a large loss of fitness, while changes from a polar to a hydrophobic residue on average does not cause a substantial loss of fitness. In general, we note that the $\tilde{s}_{exp}$ matrix is distinctly asymmetric. In addition to the effects of polar and apolar residues, we see that substitutions to proline are substantially more detrimental to function than substitutions from proline, and substitutions from cysteine affect fitness more than substitutions to cysteine. Such asymmetry is not present in, for example, the BLOSUM62 substitution matrix (Hess et al., 2016) (Figure S3), which captures average evolutionary effects without taking directionality into account. Thus, the $\tilde{s}_{exp}$ and BLOSUM62 matrices are only modestly correlated (Figure S4), and we prefer the $\tilde{s}_{exp}$ matrix as our null model. Nevertheless, although $\tilde{s}_{exp}$ captures the typical single amino acid effects observed in a MAVE, it lacks information about structural and sequence context, and thus is overall a poor predictor of experimental variant effects (mean r_s = 0.17; Figure 2B, green).

With the experimental and computational variant data we proceeded to train a set of machine learning models to predict s_{exp} from our set of available ΔΔG, ΔΔE, and $\tilde{s}_{exp}$ scores. We chose to use random forest models because of their robustness to outliers and noise, minimal need to adjust model hyperparameters, as well as the possibility to easily extract information about the extent to which the different features are used in the model decision process (Breiman, 2001; Bernard et al., 2009). We term the values predicted by the model s_{calc}.

We used a leave-one-protein-out procedure for training, selecting one protein for validation, and training on the normalized MAVE, ΔΔE, and/or ΔΔG data from all other proteins in our overall set. Thus, when more than one set of experiments had been performed on a single protein, we excluded the extra datasets during training. We assess the model by calculating the correlation between s_{calc} predicted from the resulting random forest with the s_{exp} values from all MAVEs on the protein that was left out in training, looping over all proteins one at a time.

First, we trained three random forest models using as inputs only the variant ΔΔG, only the variant ΔΔE, or both. In line with the variation in the correlation to the input data, we observe a range of correlation coefficients from these models, with the combined ΔΔG and ΔΔE model correlating with the normalized MAVE scores with a median r_s = 0.47 (Figure 2B, orange). The ΔΔG-only model (r_s = 0.25, Figure 2B, cyan) only performs slightly better than $\tilde{s}_{exp}$ (green), and the ΔΔE-only model (r_s = 0.42, Figure 2B, purple) is again closer to capturing the experimental outcomes. We note here that several of the correlation coefficients observed in these analyses (Figure 2B) are smaller than those obtained when correlating directly with the s_{exp} values (Figure 2A). While the direct correlations essentially represent 39 individual, "protein-specific" models, we aimed to build a single model generalizing across different proteins and residue contexts, capturing the relationship between, e.g., ΔΔG and s_{exp} on a single scale.

As variant effects depend both on the specific substitutions but also the context within the protein, we trained a more complex "position-context" model that also takes into account the scores for other substitutions at the given protein position. Specifically, in addition to the ΔΔE and ΔΔG value for the variant to be predicted, we input the entire set of 20 ΔΔG and 20 ΔΔE values at a position, representing the stability and conservation scores for all single amino-acid variants at the given position, as well as the mean score for these. We further add three features from the baseline model, $\tilde{s}_{WT\rightarrow Mut}$ (the average score when mutating from the wild type to the specific variant [mutant] residue), $\tilde{s}_{WT\rightarrow Any}$ (the average score for substitutions from the specific wild-type amino acid to any of the other 19), and $\tilde{s}_{Any\rightarrow Mut}$ (the average score when mutating each of the 19 other amino acids to the specific variant amino acid) (see more detailed discussion in STAR Methods). The resulting position-context random forest model yields an improved performance (median r_s = 0.52, Figure 2B, brown). The model also successfully recapitulates the trends in the stability-conservation landscape (Figures 2C and S7). Looking more closely at how the model performs in different sectors of this landscape, we find that predictions generally have a similar mean absolute error (MAE) in most sectors (Figures 2D and S6) with MAE ≈0.22, with somewhat more accurate predictions for highly destabilizing (ΔΔG>3.0) and non-conservative (ΔΔE>0.50) substitutions (MAE 0.16–0.21).

Role of stability and conservation in the predictions

With the final model in hand and having shown that it recapitulates the experimentally observed stability-conservation landscape relatively well (Figures 1A, 2C and S5), we proceeded to analyze the properties of the model. Looking at the predictions of the >150,000 variant effects from the leave-one-protein-out model, we find that the position-context model predicts fitness outcomes with a Spearman correlation coefficient of 0.50 (Figure 3A). The most dense region has a sigmoidal shape, with predictions of the experimentally derived s_{exp} range 0.5–1.0 seeming largely indistinguishable by the model. Indeed, the model is particularly good at separating the approximately one-third of the variants that cause substantial loss of function (36% have s_{exp}<0.3) from those with activities closer to wild type. We note here that the noisy nature of data generated by many MAVEs, as well the rank-normalization that we used for each protein, likely mean that many variants with fitness scores comparable with wild type end up being spread out with s_{exp} in the range ≈0.3–1.0, in line with our previous observations of about one-third of variants causing loss of function (Cagiada et al., 2021). Thus, we suggest that the model is better suited at capturing the bimodal distribution seen in many MAVEs (Gray et al., 2017). While calculations of ΔΔG and ΔΔE may capture more subtle effects on function in individual proteins (Nielsen et al., 2021), such effects appear difficult to extract in our global analysis. In the absence of the model's ability to distinguish between such effects, the predictions plateau at a relatively narrow distribution of values that minimizes the mean square error between s_{exp} and s_{calc} for these variants.

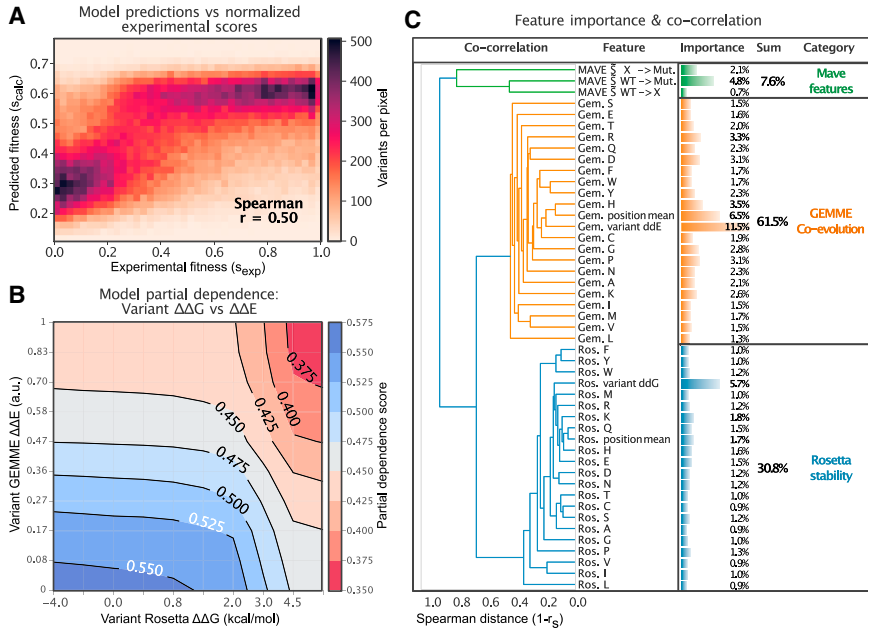


Figure 3. Interpretation of the position-context random forest model

(A) Density plot of predicted variant scores versus normalized experimental fitness scores in leave-one-protein-out cross-validation.

(B) Partial dependence plot for the position-context model trained on all variants. Most of the variation in fitness scores is explained along the GEMME ΔΔE axis, with Rosetta ΔΔG only affecting variant scores at values greater than ≈2. Colors indicate the predicted partial score for the given simulated variant ΔΔG and ΔΔE score sampled across the entire dataset, but does not take into account effects from the remaining features. Note the non-linearity of the scale of the axes.

(C) Position-context model feature importance and dendrogram.

To gain further insight into the relative importance of the stability and conservation parameters for predicting the variant scores, we calculated the partial dependencies in the position-context random forest model. These partial dependencies quantify how the model's predictions of s_{calc} depend on the chosen features, here the ΔΔG and ΔΔE for the specific variant, marginalizing over the remaining features in the model (Molnar, 2019). For variants that at most cause a modest change in stability (ΔΔG<2.0 kcal/mol), the predictions are most strongly influenced by ΔΔE. This observation is in line with the notion that, for stable variants, ΔΔE is a relatively good predictor of s_{exp} (Figure 1A). In contrast, for more destabilizing variants (ΔΔG>3.0 kcal/mol) the partial dependence varies with both ΔΔG and ΔΔE, in line with the finding that both of these are useful quantities to help predict variant effects in this part of the stability-conservation landscape (Figures 1 and 2).

We then proceeded to examine the impact of all features in the random forest model. We stress that this so-called feature importance only quantifies how much each feature is used overall in the prediction of s_{calc}, but does not directly describe how and when these features are used (Strobl et al., 2007). We calculated the correlation between each pair of features and used these to cluster and build a dendrogram of the features (Figures 3C and S7). The features fall into three overall categories corresponding to ΔΔG, ΔΔE, and $\tilde{s}_{exp}$.

We here remind the reader that the model uses stability and conservation effects of all possible substitutions at a specific position even when predicting the effect of a specific variant. Thus, for example, when predicting the effect of a specific isoleucine to alanine substitution, the model will use as features the effect of changing that isoleucine residue to all other 19 protein-coding amino acids. As expected, however, the feature-importance calculations show that the model has its largest contribution from the specific substitution, with other substitutions playing a smaller role (Figure 3C). In addition to using information about the specific substitution, the average ΔΔE score also has a large feature importance, which we take to mean that the model uses this to determine whether a specific position is overall restrictive in the types of substitutions that have been seen during evolution. In addition to these effects, substitutions to positively charged amino acids (e.g., ΔΔE to arginine and histidine and ΔΔG to lysine) have greater than average contributions, and we speculate that these values provide additional structural context on whether, e.g., a position is buried or not (as substitutions to large positively charged residues are generally disfavored at buried positions). We note that effects of substitutions to histidine and asparagine have previously been shown to correlate most strongly with other substitutions at the same position (Gray et al., 2017).

Adding up the individual contributions of the features from the three classes ($\tilde{s}_{exp}$, ΔΔE, and ΔΔG), we find that the ΔΔE features have roughly twice the importance as ΔΔG features, with the total importance being substantially greater than from $\tilde{s}_{exp}$ (Figure 3C). Looking at the values for the specific variant predicted, the values are 11.8%, 5.7%, and 4.8% for ΔΔE, ΔΔG, and $\tilde{s}_{exp}$, respectively, although we note caveats in interpreting these numbers very precisely (Strobl et al., 2007). Thus, in line with analyzing simpler models using only a subset of the features (Figure 2B), this analysis shows that all three classes of features contribute to the performance of the final model, with ΔΔE carrying the greatest weight.

A global map of variant effects

The results above further support previous observations that analyses of sequence conservation are overall more informative than stability calculations to predict variant affects as probed by most MAVE experiments (Jepsen et al., 2020; Livesey and Marsh, 2020; Gerasimavicius et al., 2020; Cagiada et al., 2021). Nevertheless, they also show that stability calculations can help improve overall prediction accuracy and provide mechanistic insight into the relative roles of stability and conservation in explaining variant effects. Specifically, as argued previously

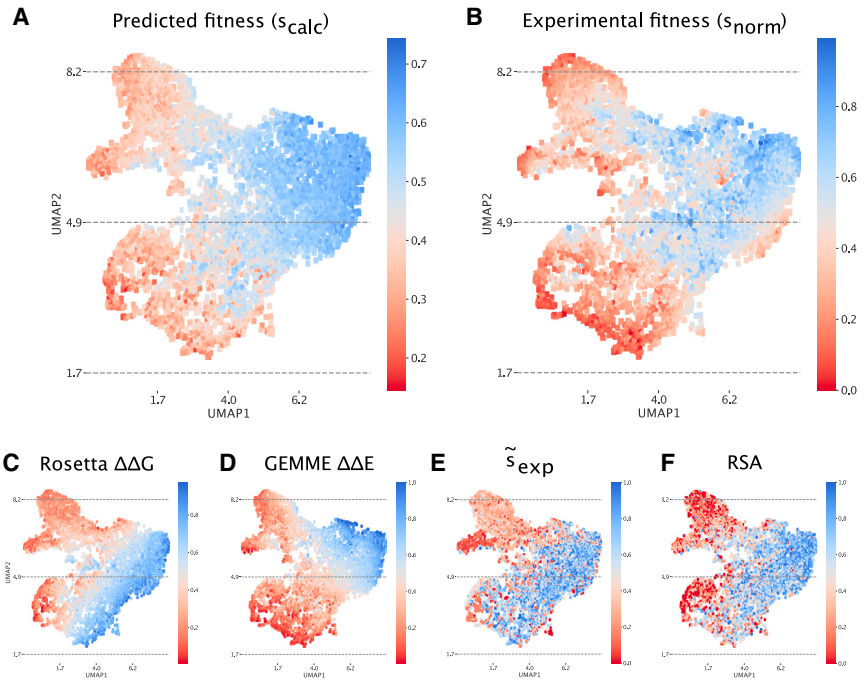


Figure 4. UMAP projection of over 6,500 positions with data for at least 15 variants in the MAVEs

The maps are colored by different properties and are locally averaged using a convolutional kernel (see STAR Methods). All scores were normalized between 0 and 1 to ease comparisons.

(A) Position-averaged predicted variant scores from the position-context random forest model.

(B) Position-averaged experimental fitness scores.

(C) Position-averaged ΔΔG values.

(D) Position-averaged ΔΔE values.

(E) $\tilde{s}_{exp}$.

(F) Relative solvent accessibility from protein structure.

(Stein et al., 2019; Jepsen et al., 2020; Cagiada et al., 2021), we can use ΔΔE calculations as a proxy for capturing a broad range of effects on biological function, and ΔΔG as the subset that involves specifically stability and abundance.

We thus used the data that we collected to create global maps of variant effects, and analyzed these in terms of fitness, stability, and conservation. Inspired by previous work (Dunham and Beltrao, 2021), we visualized the 6,500 amino acid positions in our dataset with more than 15 values of s_{exp} and corresponding values of ΔΔG and ΔΔE (Figure 4). Specifically, we used Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction (McInnes et al., 2018) to represent all positions in a two-dimensional map where positions with similar profiles of s_{exp}, ΔΔG, and ΔΔE are located close to one another (see STAR Methods). Finally, we color coded this map using the predicted and calculated position-averaged fitness scores, the position-averaged values of ΔΔG and ΔΔE, values for $\tilde{s}_{exp}$, and the relative solvent accessibility (RSA) of the position (Figure 4).

As expected from the fact that the model was trained to predict the experimental data, the map colored by s_{calc} (Figure 4A) and s_{exp} (Figure 4B) closely resemble one another, although with a few differences highlighting the imperfection of the model. In particular, the maps reveal two regions (top left and bottom left of the maps) that are enriched in low-fitness variants. The map of stability effects (Figure 4C) shows a gradual change moving from the top left to the bottom right, and illustrates that, in particular, the top left part of the map corresponds to positions where variants on average are destabilizing. In contrast, the map colored by evolutionary scores (Figure 4D) reveals a gradient from the bottom left to the top right, and identifies many of the same regions as the experimental map in terms of low scores. Comparing the maps colored by the ΔΔG and ΔΔE scores reveals two regions (top left and part of the bottom left) with amino acid positions where it appears that variants cause loss of function due to loss of stability. Comparison with the map indicating solvent accessibility (Figure 4F) shows clearly that many of these positions are buried inside the protein structure. The same comparisons also show that a group of positions near the bottom are enriched in positions where variants lose function in ways that are not due to stability but that can nonetheless be discovered using analyses of the evolutionary record. Finally, looking at the map colored by the average $\tilde{s}_{exp}$ scores (Figure 4E) shows similarities to both the stability map (Figure 4C) and solvent accessibility map (Figure 4F), in line with the fact that these are buried positions and that $\tilde{s}_{exp}$ captures basic physico-chemical aspects of amino acid chemistry and protein structure (Figure S2).

DISCUSSION

We have analyzed the relationship between experimental measurements of variant effects on protein function and computational analysis of protein stability and conservation. We collected over 150,000 measurements from multiplexed assays of variant effect in 29 proteins and compared them with predictions of changes in protein stability (ΔΔG) and evolutionary conservation (ΔΔE). Our goal was 2-fold. First, we aimed to examine how well these computed scores could predict variant effects, and, second, we wanted to shed further light on how often changes in protein stability may perturb protein function.

In general, and in line with previous observations, we find that our analysis of conservation of sequences through evolution (ΔΔE) is more strongly correlated with the experimental measurements than predictions of changes in thermodynamic stability (ΔΔG) (Figure 2A). We note here that, although each of the 39 experiments provides a systematic and comprehensive analysis of variant effects, they were conducted using very different assays for selection/screening and in different organisms. Together with uncertainties in the experimental and predicted values, this likely explains the substantial variation that we

observe in agreement between experimental scores and computed values. Thus, we refrain in general from analyzing individual proteins and variants and focus on the global agreement. As also previously observed (Cagiada et al., 2021), the VAMP-seq experiments that specifically probe protein abundance tend to be more strongly correlated with stability effects than measurements probing protein activity. Finally, we note that, while discrepancies between experimental measurements of variant effects and computational predictions can point to shortcomings in the prediction models, these may also reveal aspects of the protein's function that the assay was not sensitive to.

Our global stability-conservation landscape (Figure 1A) reveals the interdependency of stability and conservation. As evolution tends to disfavor unstable proteins (Bloom et al., 2006; Echave and Wilke, 2017), we find that variants with high ΔΔG tend to have large values of ΔΔE and generally low fitness (low s_{exp}). The reverse relationship, however, is not true because being stable is a necessary, but not sufficient, criterion for being functional. Examining the stable variants (low ΔΔG) we find that our analysis of the evolutionary record (ΔΔE) can be used to predict with reasonable accuracy whether a variant retains function or not (Figure 1B). In our recent analysis of data probing activity and abundance in NUDT15 and PTEN, we showed that about half of the variants that lose function do so because they become unstable and are found at low abundance. In line with this, we find that roughly half of the variants that have high ΔΔE scores also have high values of ΔΔG, and hypothesize that many of the remaining variants (low ΔΔG, high ΔΔE) lose function because of substitutions in functionally important residues (Cheng et al., 2005; Echave, 2019; Cagiada et al., 2021).

Based on the partial correlation between both ΔΔE and ΔΔG with s_{exp}, we built a prediction model with these parameters as input. The final position-context model reaches an accuracy that surpasses that of models that solely use ΔΔE or ΔΔG (Figure 2B). This suggests that, although ΔΔE calculations to some extent capture stability effects, they can still be improved by explicitly including calculated values of ΔΔG. Indeed, the model is most accurate for variants that both ΔΔE and ΔΔG suggest would be non-functional (Figure 2D).

The random forest model recapitulates key aspects of the global stability-conservation landscape (Figure S7), including the interdependency of ΔΔG and ΔΔE. A more detailed analysis of the random forest model shows that ΔΔE is indeed the more informative quantity when ΔΔG is small, whereas both ΔΔG and ΔΔE contribute to accuracy when ΔΔG is large (Figure 3B). This is also reflected in the feature-importance analysis, which shows that overall the various ΔΔE terms contribute roughly twice the importance of the ΔΔG terms among the decision trees (Figure 3C). Similar effects are also seen in our global maps of variant effects (Figure 4), which reveal an almost orthogonal gradient of scores for ΔΔG and ΔΔE, and indicate how the combination of these two maps contribute to the final predicted scores. We also show that including contextual information about effects of other substitutions at the same position further increases correlation with experimental scores. We speculate that this context captures additional information about the structural, biochemical, and functional requirements at this position. These results are in line with the recent observation that the positional profiles from MAVEs contain a rich set of information beyond that of individual substitutions (Dunham and Beltrao, 2021).

There are several possibilities for future improvements of the prediction methods used here. First, a number of methods have recently been developed to analyze sequence information (Riesselman et al., 2018; Alley et al., 2019; Frazer et al., 2021; Hsu et al., 2021) and could be tested instead of the GEMME method that we used. Second, while the Rosetta method that we used to predict ΔΔG is among the most accurate methods for stability prediction, it is computationally expensive and requires access to protein structures. Thus, future work is needed on assessing the utility of predicted structures (both from template-free and template-based methods), as well as testing and developing more computationally efficient methods for predicting stability effects.

Another area for further development is to analyze and improve predictions in areas of intermediate values of ΔΔG and ΔΔE, where errors in the predicted values would have a greater impact. While our results overall conform to the expected relationship between ΔΔE and ΔΔG, more subtle effects can modulate this relationship. For example, it has been observed that active-site residues may actually be destabilizing (Shoichet et al., 1995), and more detailed stability calculations may thus aid in detecting such effects.

The types of analyses presented here may also be used to improve or interpret experiments. For example, when constructing an experimental assay, it might be useful to compare experimental variant effects with calculations of ΔΔE to examine whether the assay is sensitive to evolutionarily conserved functions. Also, it has recently been demonstrated that biophysical ambiguities prevent accurate predictions of how substitutions combine to affect phenotype (Li and Lehner, 2020), and we suggest that biophysical models and predictions of variant effects may help alleviate some of these ambiguities. Finally, a random forest model combining sequence and structural features has been used to predict pathogenicity of human missense variants (Ponzoni et al., 2020).

In summary, our large-scale analysis contributes to the mounting evidence for the important role for loss of stability in loss of function. In addition to the improvements observed when combining conservation analysis with structure-based stability predictions, our analyses also help pinpoint those variants that lose function due to stability. Such information helps provide mechanistic insight into specific variants and proteins but may also be a starting point for developing therapies. In particular, variants that lose function due to loss of abundance, but whose intrinsic function is not otherwise perturbed, would be ideal targets for approaches that aim to restabilize or otherwise restore protein levels (Balch et al., 2008; Kampmeyer et al., 2017; Stein et al., 2019; Henning et al., 2021).

Limitations of the study

There are some limitations in our study. First, there may be biases and limitations arising from the specific choice of experimental data that we analyzed. We focused only on soluble

proteins, and our selection criteria further included availability of the experimental measurements and an experimentally determined structure of the protein. As proteins with known structures tend to have deeper multiple sequence alignments (Orlando et al., 2016), the latter requirement might lead to subtle biases in our analyses and represent more accurately calculated conservation scores. Second, the experimental data come from a range of assay types and organisms across the domains of life, and future analyses will be needed to examine how well our methods extrapolate to other proteins and assays. Third, our analysis of whether low-ΔΔG (structurally stable) but high-ΔΔE (evolutionarily constrained) variants are enriched in known active sites is currently based on a small number of proteins, and it remains unclear to what extent these observations hold more generally. Fourth, the method we use to predict protein stability is imperfect and has been benchmarked on a biased dataset (Frenz et al., 2020), leaving it unclear exactly how well it performs for the wider range of mutations analyzed here.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - ΔΔG calculations using Rosetta
  - Evolutionary conservation analysis using GEMME
  - Collecting and normalizing data from MAVEs
  - Baseline substitution model
  - Features and random forest models
  - Model analysis
  - Relative solvent accessibility
  - UMAP projection of MAVE positions
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.110207.

ACKNOWLEDGMENTS

Our research is supported by the Protein Interactions and Stability in Medicine and Genomics (PRISM) center funded by the Novo Nordisk Foundation (NNF18OC0033950, to A.S. and K.L.-L.) and a grant from the Lundbeck Foundation (R272-2017-4528, to A.S.)

AUTHOR CONTRIBUTIONS

Conceptualization, M.H.H, A.S., and K.L.-L.; methodology, M.H.H., M.C., A.H.B.F., A.S., and K.L.-L.; formal analysis, M.H.H., M.C., A.H.B.F., and A.S.; data curation, M.H.H., A.S., and K.L.-L.; writing – original draft, M.H.H., A.S., and K.L.-L.; visualization, M.H.H. and M.C.; supervision, A.S. and K.L.-L.; project administration, A.S. and K.L.-L.; funding acquisition, A.S. and K.L.-L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 19, 2021
Revised: October 1, 2021
Accepted: December 13, 2021
Published: January 11, 2022

SUPPORTING CITATIONS

The following references appear in the supplemental information: Adkar et al., 2012; Bandaru et al., 2017; Brenan et al., 2016; Dandage et al., 2018; Deng et al., 2012; Doud and Bloom, 2016; Findlay et al., 2018; Firnberg et al., 2014; Giacomelli et al., 2018; Haddox et al., 2016; Jacquier et al., 2013; Jiang, 2019; Jones et al., 2020; Kelsic et al., 2016; Kitzman et al., 2015; Lee et al., 2018; Melamed et al., 2013; Mishra et al., 2016; Ribeiro et al., 2018; Rockah-Shmuel et al., 2015; Spencer and Zhang, 2017; Starita et al., 2015; Stiffler et al., 2015; Styczynski et al., 2008; Sun et al., 2020; Weile et al., 2017.

REFERENCES

Abildgaard, A.B., Stein, A., Nielsen, S.V., Schultz-Knudsen, K., Papaleo, E., Shrikhande, A., Hoffmann, E.R., Bernstein, I., Gerdes, A.-M., Takahashi, M., et al. (2019). Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. eLife *8*, e49138.

Adkar, B., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M., Gokhale, R., and Varadarajan, R. (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. Structure *20*, 371–381.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Ahler, E., Register, A.C., Chakraborty, S., Fang, L., Dieter, E.M., Sitko, K.A., Vidadala, R.S.R., Trevillian, B.M., Golkowski, M., Gelman, H., et al. (2019). A combined approach reveals a regulatory mechanism coupling SRC's kinase activity, localization, and Phosphotransferase-Independent functions. Mol. Cell *74*, 393–408.e20.

Alley, E.C., Khimulya, G., Biswas, S., Alquraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods *16*, 1315–1322.

Amorosi, C.J., Chiasson, M.A., McDonald, M.G., Wong, L.H., Sitko, K.A., Boyle, G., Kowalski, J.P., Rettie, A.E., Fowler, D.M., and Dunham, M.J. (2021). Massively parallel characterization of cyp2c9 variant enzyme activity and abundance. Am. J. Hum. Genet. *108*, 1735–1751.

Ancien, F., Pucci, F., Godfroid, M., and Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Scientific Rep. *8*, 1–11.

Arlow, T., Scott, K., Wagenseller, A., and Gammie, A. (2013). Proteasome inhibition rescues clinically significant unstable variants of the mismatch repair protein msh2. Proc. Natl. Acad. Sci. *110*, 246–251.

Balch, W.E., Morimoto, R.I., Dillin, A., and Kelly, J.W. (2008). Adapting proteostasis for disease intervention. Science *319*, 916–919.

Bandaru, P., Shah, N.H., Bhattacharyya, M., Barton, J.P., Kondo, Y., Cofsky, J.C., Gee, C.L., Chakraborty, A.K., Kortemme, T., Ranganathan, R., et al. (2017). Deconstruction of the Ras switching cycle through saturation mutagenesis. eLife *6*, e27810. https://doi.org/10.7554/eLife.27810.

Bernard, S., Heutte, L., and Adam, S. (2009). Influence of hyperparameters on random forest accuracy. In International Workshop on Multiple Classifier Systems (Springer), pp. 171–180.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. *103*, 5869–5874.

Breiman, L. (2001). Random forests. Machine Learn. *45*, 5–32.

Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N., Zhu, C., Bagul, M., Goetz, E., et al. (2016). Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants. Cell Rep. *17*, 1171–1183.

Cagiada, M., Johansson, K.E., Valanciute, A., Nielsen, S.V., Hartmann-Petersen, R., Yang, J.J., Fowler, D.M., Stein, A., and Lindorff-Larsen, K. (2021). Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. Mol. Biol. Evol.

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. Hum. Mutat. *32*, 1161–1170.

Chen, L., Brewer, M.D., Guo, L., Wang, R., Jiang, P., and Yang, X. (2017). Enhanced degradation of misfolded proteins promotes tumorigenesis. Cell Rep. *18*, 3143–3154.

Cheng, G., Qian, B., Samudrala, R., and Baker, D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res. *33*, 5861–5867.

Chiasson, M.A., Rollins, N.J., Stephany, J.J., Sitko, K.A., Matreyek, K.A., Verby, M., Sun, S., Roth, F.P., DeSloover, D., Marks, D.S., et al. (2020). Multiplexed measurement of variant abundance and activity reveals CKOR topology, active site and human variant impact. eLife *9*, e58026.

Choi, Y., and Chan, A.P. (2015). Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics *31*, 2745–2747.

Cline, M.S., Babbi, G., Bonache, S., Cao, Y., Casadio, R., Cruz, X., Díez, O., Gutiérrez-Enríquez, S., Katsonis, P., Lai, C., et al. (2019). Assessment of blind predictions of the clinical significance of brca1 and brca2 variants. Hum. Mutat. *40*, 1546–1556.

Cuella-Martin, R., Hayward, S.B., Fan, X., Chen, X., Huang, J.-W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C., et al. (2021). Functional interrogation of DNA damage response variants with base editing screens. Cell *184*, 1081–1097.

Dandage, R., Pandey, R., Jayaraj, G., Rai, M., Berger, D., and Chakraborty, K. (2018). Differential strengths of molecular determinants guide environment specific mutational fates. PLoS Genet. *14*, e1007419.

De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. *40*, D935–D939.

Deng, Z., Huang, W., Bakkalbasi, E., Brown, N.G., Adamski, C.J., Rice, K., Muzny, D., Gibbs, R.A., and Palzkill, T. (2012). Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. J. Mol. Biol. *424*, 150–167.

Després, P.C., Dubé, A.K., Seki, M., Yachie, N., and Landry, C.R. (2020). Perturbing proteomes at single residue resolution using base editing. Nat. Commun. *11*, 1–13.

Doud, M.B., and Bloom, J.D. (2016). Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. Viruses *8*, 155.

Dunham, A.S., and Beltrao, P. (2021). Exploring amino acid functions in a deep mutational landscape. Mol. Syst. Biol. *17*, e10305. https://doi.org/10.15252/msb.202110305.

Echave, J. (2019). Beyond stability constraints: a biophysical model of enzyme evolution with selection on stability and activity. Mol. Biol. Evol. *36*, 613–620.

Echave, J., and Wilke, C.O. (2017). Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. Annu. Rev. Biophys. *46*, 85–103.

Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., and Rubin, A.F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. *20*, 1–11.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2014). Comprehensive, high-resolution map of a genes fitness landscape. Mol. Biol. Evol. *31*, 1581–1592.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Frazer, J., Notin, P., Dias, M., Gomez, A., Brock, K., Gal, Y., and Marks, D. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95.

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. Front. Bioeng. Biotechnol. *8*, 558247.

Gerasimavicius, L., Liu, X., and Marsh, J.A. (2020). Identification of pathogenic missense mutations using protein stability predictors. Scientific Rep. *10*, 15387.

Giacomelli, A.O., Yang, X., Lintner, R.E., McFarland, J.M., Duby, M., Kim, J., Howard, T.P., Takeda, D.Y., Ly, S.H., Kim, E., et al. (2018). Mutational processes shape the landscape of tp53 mutations in human cancer. Nat. News *50*, 1381–1387.

Gray, V.E., Hause, R.J., and Fowler, D.M. (2017). Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. Genetics *207*, 53–61.

Gray, V.E., Hause, R.J., Luebeck, J., Shendure, J., and Fowler, D.M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell Syst. *6*, 116–124.

Haddox, H.K., Dingens, A.S., and Bloom, J.D. (2016). Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. PLoS Pathog. *12*, e1006114.

Hanna, R.E., Hegde, M., Fagre, C.R., DeWeirdt, P.C., Sangree, A.K., Szegletes, Z., Griffith, A., Feeley, M.N., Sanson, K.R., Baidi, Y., et al. (2021). Massively parallel assessment of human variants with base editor screens. Cell *184*, 1064–1080.

Henning, N.J., Boike, L., Spradlin, J.N., Ward, C.C., Belcher, B., Brittain, S.M., Hesse, M., Dovala, D., McGregor, L.M., McKenna, J.M., et al. (2021). Deubiquitinase-targeting chimeras for targeted protein stabilization. bioRxiv. https://doi.org/10.1101/2021.04.30.441959.

Hess, M., Keul, F., Goesele, M., and Hamacher, K. (2016). Addressing inaccuracies in BLOSUM computation improves homology search performance. BMC Bioinformatics. *17*, 189.

Hingorani, K.S., and Gierasch, L.M. (2014). Comparing protein folding in vitro and in vivo: foldability meets the fitness challenge. Curr. Opin. Struct. Biol. *24*, 81–90.

Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–135.

Hsu, C., Nisono, H., Fannjiang, C., and Listgarten, J. (2021). Combining evolutionary and assay-labelled data for protein fitness prediction. bioRxiv. https://doi.org/10.1101/2021.03.28.437402.

Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). Revel: an ensemble method for predicting the pathogenicity of rare missense variants. Am. J. Hum. Genet. *99*, 877–885.

Jacquier, H., Birgy, A., Nagard, H.L., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., et al. (2013). Capturing the mutational landscape of the beta-lactamase tem-1. PNAS *110*, 13067–13072.

Jepsen, M.M., Fowler, D.M., Hartmann-Petersen, R., Stein, A., and Lindorff-Larsen, K. (2020). Classifying disease-associated variants using measures of protein activity and stability. In Protein Homeostasis Diseases (Elsevier), pp. 91–107.

Jiang, R.J. (2019). Exhaustive mapping of missense variation in coronary heart disease-related genes. TSpace. http://hdl.handle.net/1807/98076.

Jiangchun, L. (2018). Python partial dependence plot toolbox. https://github.com/SauceCat/PDPbox.

Jones, E.M., Lubock, N.B., Venkatakrishnan, A., Wang, J., Tseng, A.M., Paggi, J.M., Latorraca, N.R., Cancilla, D., Satyadi, M., Davis, J.E., et al. (2020). Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. eLife *9*, e54895.

Jun, S., Lim, H., Chun, H., Lee, J.H., and Bang, D. (2020). Single-cell analysis of a mutant library generated using CRISPR-guided deaminase in human melanoma cells. Commun. Biol. *3*, 1–12.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kampmeyer, C., Nielsen, S.V., Clausen, L., Stein, A., Gerdes, A.-M., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2017). Blocking protein quality control to counter hereditary cancers. Genes Chromosom. Cancer *56*, 823–831.

Kelsic, E.D., Chung, H., Cohen, N., Park, J., Wang, H.H., and Kishony, R. (2016). RNA structural determinants of optimal codons revealed by MAGE-seq. Cell Syst. *3*, 563–571.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kitzman, J.O., Starita, L.M., Lo, R.S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. Nat. Methods *12*, 203–206.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073.

Laine, E., Karami, Y., and Carbone, A. (2019). GEMME: a simple and fast global epistatic model predicting mutational effects. Mol. Biol. Evol. *36*, 2604–2619.

Lee, J.M., Huddleston, J., Doud, M.B., Hooper, K.A., Wu, N.C., Bedford, T., and Bloom, J.D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. PNAS *115*, E8276–E8285.

Leman, J.K., Weitzner, B.D., Lewis, S.M., Adolf-Bryfogle, J., Alam, N., Alford, R.F., Aprahamian, M., Baker, D., Barlow, K.A., Barth, P., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat. Methods *17*, 665–680.

Li, X., and Lehner, B. (2020). Biophysical ambiguities prevent accurate genetic prediction. Nat. Commun. *11*, 1–11.

Livesey, B.J., and Marsh, J.A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol. Syst. Biol. *16*, e9380.

Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. *50*, 874–882.

McEntyre J, O.J. (2002). The NCBI Handbook, The BLAST Sequence Analysis Tool (National Center for Biotechnology Information (US)).

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, preprint arXiv:1802.03426.

Meacham, G.C., Patterson, C., Zhang, W., Younger, J.M., and Cyr, D.M. (2001). The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. Nat. Cel. Biol. *l*, 100–105.

Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. RNA *19*, 1537–1551.

Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. Am. J. Hum. Genet. *102*, 943–955.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. *45*, D170–D176.

Mishra, P., Flynn, J., Starr, T., and Bolon, D. (2016). Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. Cell Rep. *15*, 588–598.

Molnar, C. (2019). Interpretable machine learning, *Ch. 8.1 - partial dependence plot (PDP)*, ISBN: 9780244768522. https://christophm.github.io/interpretable-ml-book/.

Nielsen, S.V., Hartmann-Petersen, R., Stein, A., and Lindorff-Larsen, K. (2021). Multiplexed assays reveal effects of missense variants in MSH2 and cancer predisposition. PLoS Genet. *17*, e1009496.

Nielsen, S.V., Schenstrøm, S.M., Christensen, C.E., Stein, A., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2020). Protein destabilization and degradation as a mechanism for hereditary disease. In Protein Homeostasis Diseases, Angel L. Pey, ed. (Elsevier), pp. 111–125.

Nielsen, S.V., Stein, A., Dinitzen, A.B., Papaleo, E., Tatham, M.H., Poulsen, E.G., Kassem, M.M., Rasmussen, L.J., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2017). Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genet. *13*, e1006739.

Olzmann, J.A., Brown, K., Wilkinson, K.D., Rees, H.D., Huai, Q., Ke, H., Levey, A.I., Li, L., and Chin, L.-S. (2004). Familial Parkinson's disease-associated l166p mutation disrupts DJ-1 protein folding and function. J. Biol. Chem. *279*, 8506–8515.

Orlando, G., Raimondi, D., and Vranken, W. (2016). Observation selection bias in contact prediction and its implications for structural bioinformatics. Scientific Rep. *6*, 1–8.

Park, H., Bradley, P., Greisen, P., Jr., Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D., and DiMaio, F. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J. Chem. Theor. Comput. *12*, 6201–6212.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Machine Learn. Res. *12*, 2825–2830.

Pey, A.L., Stricher, F., Serrano, L., and Martinez, A. (2007). Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. Am. J. Hum. Genet. *81*, 1006–1024.

Ponzoni, L., Peñaherrera, D.A., Oltvai, Z.N., and Bahar, I. (2020). Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. *36*, 3084–3092.

Ribeiro, A.J.M., Holliday, G.L., Furnham, N., Tyzack, J.D., Ferris, K., and Thornton, J.M. (2018). Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res. *46*, D618–D623.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association For Molecular Pathology. Genet. Med. *17*, 405–423.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822.

Rockah-Shmuel, L., Toth-Petroczy, A., and Tawfik, D.S. (2015). Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. PLoS Comput. Biol. *11*, e1004421.

Ron, I., and Horowitz, M. (2005). ER retention and degradation as the molecular basis underlying Gaucher disease heterogeneity. Hum. Mol. Genet. *14*, 2387–2398.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlić, A., Quesada, M., Quinn, G.B., Westbrook, J.D., et al. (2010). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. *39*, D392–D401.

Schaafsma, G.C., and Vihinen, M. (2017). Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. Hum. Mutat. *38*, 839–848.

Scheller, R., Stein, A., Nielsen, S.V., Marin, F.I., Gerdes, A.-M., Di Marco, M., Papaleo, E., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2019). Toward mechanistic models for genotype-phenotype correlations in phenylketonuria using protein stability calculations. Hum. Mutat. *40*, 444–457.

Shoichet, B.K., Baase, W.A., Kuroki, R., and Matthews, B.W. (1995). A relationship between protein stability and protein function. Proc. Natl. Acad. Sci. *92*, 452–456.

Spencer, J.M., and Zhang, X. (2017). Deep mutational scanning of *S. pyogenes* cas9 reveals important functional domains. Scientific Rep. *7*, 1–14.

Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M., Parvin, J.D., Shendure, J., Fields, S., et al. (2015). Massively parallel functional analysis of BRCA1 ring domain variants. Genetics *200*, 413–422.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell *182*, 1295–1310.

Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and mechanistic models for disease-causing protein variants. Trends Biochem. Sci. *44*, 575–588.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 1–15.

Stiffler, M., Hekstra, D., and Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 β -lactamase. Cell *160*, 882–892.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics *8*, 1–21.

Styczynski, M.P., Jensen, K.L., Rigoutsos, I., and Stephanopoulos, G. (2008). BLOSUM62 miscalculations improve search performance. Nat. Biotechnol. *26*, 274–275.

Suiter, C.C., Moriyama, T., Matreyek, K.A., Yang, W., Scaletti, E.R., Nishii, R., Yang, W., Hoshitsuki, K., Singh, M., Trehan, A., et al. (2020). Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. Proc. Natl. Acad. Sci. U. S. A. *117*, 5394–5401.

Sun, S., Weile, J., Verby, M., Wu, Y., Wang, Y., Cote, A.G., Fotiadou, I., Kitaygorodsky, J., Vidal, M., Rine, J., et al. (2020). A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. Genome Med. *12*, 13.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Wagih, O., Galardini, M., Busby, B.P., Memon, D., Typas, A., and Beltrao, P. (2018). A resource of variant effect predictions of single nucleotide variants in model organisms. Mol. Syst. Biol. *14*, e8430.

Weile, J., Sun, S., Cote, A.G., Knapp, J., Verby, M., Mellor, J.C., Wu, Y., Pons, C., Wong, C., Lieshout, N.v., et al. (2017). A framework for exhaustively mapping functional missense variants. Mol. Syst. Biol. *13*, 957.

Yaguchi, H., Ohkura, N., Takahashi, M., Nagamura, Y., Kitabayashi, I., and Tsukada, T. (2004). Menin missense mutants associated with multiple endocrine neoplasia type 1 are rapidly degraded via the ubiquitin-proteasome pathway. Mol. Cell. Biol. *24*, 6569–6580.

Yang, C., Asthagiri, A.R., Iyer, R.R., Lu, J., Xu, D.S., Ksendzovsky, A., Brady, R.O., Zhuang, Z., and Lonser, R.R. (2011). Missense mutations in the NF2 gene result in the quantitative loss of merlin protein and minimally affect protein intrinsic function. Proc. Natl. Acad. Sci. *108*, 4980–4985.

Yang, C., Huntoon, K., Ksendzovsky, A., Zhuang, Z., and Lonser, R.R. (2013). Proteostasis modulators prolong missense VHL protein activity and halt tumor progression. Cell Rep. *3*, 52–59.

Yin, Y., Kundu, K., Pal, L.R., and Moult, J. (2017). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (human N-acetyl-glucosaminidase) and UBE2I (human SUMO-ligase) challenges. Hum. Mutat. *38*, 1109–1122.

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J. Mol. Biol. *353*, 459–473.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
MAVE experimental datasets	References Table S1, this paper	https://zenodo.org/record/5647208
Software and algorithms		
Code for analysis and training models	This paper	https://zenodo.org/record/5647208
Scikit-learn 0.24.2	Scikit-learn: Machine Learning in Python	https://scikit-learn.org/
Scipy 1.6.3	SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python	https://scipy.org/
DSSP 3.0.0	Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features	https://anaconda.org/salilab/dssp
Biopython 1.79	Biopython: freely available Python tools for computational molecular biology and bioinformatics	https://biopython.org/
GEMME	GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects	http://www.lcqb.upmc.fr/GEMME/Home.html
Rosetta, release July 2, 2020	The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design	https://www.rosettacommons.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the Lead Contact, Kresten Lindorff-Larsen (lindorff@bio.ku.dk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyses existing, publicly available data. A list of the datasets is available in Table S1. Data used in our analyses are available via https://doi.org/10.5281/zenodo.5647207 and https://github.com/KULL-Centre/papers/tree/main/2021/ML-variants-Hoie-et-al.
- Scripts to repeat our analyses are available via https://doi.org/10.5281/zenodo.5647207 and https://github.com/KULL-Centre/papers/tree/main/2021/ML-variants-Hoie-et-al.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

ΔΔG calculations using Rosetta

We searched the Protein Data Bank (Rose et al., 2010) using the BLASTp webserver (McEntyre J, 2002) with default settings using as query the sequences for which the MAVE was performed. If only a part of a protein is covered by mutagenesis data, we only searched for structures of that part. We chose structures by balancing highest available coverage and resolution, selecting structures solved using X-ray crystallography when available. All calculations were carried out using the Rosetta version with GitHub SHA 28f338acfb3bfd87048b38a04772486975dc83fa from July 2, 2020. We first relaxed the structures using the relax application and the following flags:

- fa_max_dis 9
- relax:constrain_relax_to_start_coords
- ignore_unrecognized_res

- missing_density_to_jump
- nstruct 1
- relax:coord_constrain_sidechains
- relax:cartesian
- beta
- score:weights beta_nov16_cart
- ex1
- ex2
- relax:min_type lbfgs_armijo_nonmonotone
- flip_hnq
- no_opth false

Subsequently, we carried out saturation mutagenesis to calculate for each single amino acid substitution using the Cartesian ΔΔG protocol and the beta_nov16_cart energy function with three iterations as previously described (Park et al., 2016; Frenz et al., 2020). Flags for the ΔΔG calculations were:

- fa_max_dis 9.0
- ddg::dump_pdbs false
- ddg:iterations 3
- score:weights beta_nov16_cart
- missing_density_to_jump
- ddg:mut_only
- ddg:bbnbrs 1
- beta_cart
- ex1
- ex2
- ddg::legacy true
- optimize_proline true

Scores from the three iterations were averaged. Values of ΔΔG in Rosetta Energy Units were divided by 2.9 to bring them onto a scale corresponding to kcal/mol (Park et al., 2016).

Evolutionary conservation analysis using GEMME

We calculated evolutionary conservation scores (ΔΔE) using GEMME, a global epistatic model for predicting mutational effects (Laine et al., 2019), based on a multiple sequence alignment of homologs of each protein of interest. We used the sequence used in the MAVE experiment as input to HHblits (version 2.0.15 and settings -e 1e-10 -i 1 -p 40 -b 1 -B 20000) to search UniRef30_2020_03_hhsuite.tar.gz (Mirdita et al., 2017; UniProt Consortium, 2019; Steinegger et al., 2019). We de-gapped the alignment with respect to the MAVE sequence, removed sequences with 50% gaps and used the output alignment as input to GEMME, with default settings. Finally, we rank-normalized the output ΔΔE scores and scaled them to a [0,1] scale.

Collecting and normalizing data from MAVEs

We downloaded 39 data sets that had been generated by MAVEs from publicly available repositories including MAVEdb (Esposito et al., 2019) and a recent compilation of data (Livesey and Marsh, 2020) (see Table S1). We used the variant fitness scores as presented in the original publication including possible normalization. Three data sets (P53, MAPK1 and Src) showed a reverse relationship between variant scores and ΔΔG and ΔΔE and were therefore reversed to match our convention of high scores for wild-type-like activity and low scores for low activity. We then rank-normalised and scaled these scores to a [0,1] range. We term these normalized scores s_{exp}.

We next merged each data set from the MAVEs with the corresponding ΔΔE and ΔΔG scores aligning the sequences based on using Biopython, pairwise2.align.globalds(target_seq.upper(), seq.upper(), MatrixInfo.blosum62, -3, -1).

Baseline substitution model

For each MAVE data set, we calculated a 20×20 amino-acid matrix containing the average of each the 400 possible Wild-type → Mutant variant scores (e.g. the average s_{exp} for all valine to alanine substitutions). We then calculated a global substitution score matrix ($\tilde{s}_{exp}$) by averaging these matrices.

Features and random forest models

For each variant, we extracted in total 47 computational features. Depending on the model, each s_{exp} was matched with its variant ΔΔG and ΔΔE score, as well as the 20 ΔΔG and 20 ΔΔE values corresponding to all available scores at the respective position, plus the mean ΔΔG and ΔΔE scores for the position. We note that this results in some redundancy in the use of the data, but simplifies the

data structure in the model. We also included three global features from the baseline substitution model: (i) The mean MAVE score $\tilde{s}_{WT\rightarrow Mut}$, (ii) $\tilde{s}_{WT\rightarrow Any}$ was calculated as the mean for any substitution from the given wild-type amino acid, and (iii) $\tilde{s}_{Any\rightarrow Mut}$ as the mean for mutating to the specified amino-acid from any wild-type. Thus, these three features correspond to (i) an entry in $\tilde{s}_{exp}$ as well as the (ii) row-average, and (iii) column-average of $\tilde{s}_{exp}$ (Figure S2).

We trained the Random Forest models using the RandomForestRegressor in Scikit-Learn (Pedregosa et al., 2011), with a mean-squared-error loss function, 150 trees and a minimum of 15 samples per leaf.

We first trained two models, using the variant ΔΔG or ΔΔE score only (1 feature). Next we trained a combined model using both features. Training of the position-context model was performed with all 47 available features. We iteratively trained the model and evaluated validation set performance in a leave-one-protein-out cross-validation, with the removal of all validation data sets for the selected protein from the training data set for each training round.

Model analysis

To extract feature importance from a globally-trained model, we trained a new random forest model using all 47 features and all available data sets. The feature co-correlation dendrogram was constructed by first calculating the Spearman correlation between each pair of features, which in turn was converted into a distance as $1 - r_S$, and used as input to build the dendogram using SciPy (Virtanen et al., 2020). We measured partial dependence using PDPbox (Jiangchun, 2018).

Relative solvent accessibility

We used DSSP (Kabsch and Sander, 1983) to calculate the relative solvent accessibility (RSA) using the structures that we also used for the Rosetta calculations. The RSA was used in analysis but not in training.

UMAP projection of MAVE positions

To map variant scores to positions, we removed positions with fewer than 15 experimental s_{exp} scores from the data set, and calculated position means for s_{exp}, ΔΔE, ΔΔG, $\tilde{s}_{WT\rightarrow Mut}$ and RSA. Any missing s_{exp} scores were then imputed using sklearn.impute.SimpleImputer from the amino-acid mean across all extracted positions, and all scores were then normalized to a 0–1 range. For computational efficiency we reduced the data set to 20 features using PCA before projecting the data into two dimensions (UMAP1 and UMAP2), using UMAP-learn with default settings (McInnes et al., 2018).

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantitative analyses were performed with Python and SciPy (Virtanen et al., 2020). Stability calculations were performed using Rosetta (Leman et al., 2020) and sequence analyses using GEMME (Laine et al., 2019). All analyses and details are described in STAR Methods and linked code.