Reviewed Preprint
v1 • December 16, 2024
Not revised

Reviewed Preprint
v2 • January 20, 2026
Revised by authors

Effects of residue substitutions on the cellular abundance of proteins

Thea K Schulze, Kresten Lindorff-Larsen

The Linderstøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark

For correspondence: lindorff@bio.ku.dk

Competing interests: KL-L holds stock options in and is a consultant for Peptone Ltd.

Funding: See page 42

Reviewing editor: Rosana Collepardo, University of Cambridge, United Kingdom

eLife Assessment

This **valuable** study presents a thorough analysis of protein abundance changes caused by amino acid substitutions, using structural context to improve predictive accuracy. By deriving substitution response matrices based on solvent accessibility, the authors demonstrate that simple structural features can predict abundance effects with accuracy comparable to complex methods such as free energy calculations. The strength of the evidence is **convincing**, supported by robust experimental design and comprehensive analyses.

https://doi.org/10.7554/eLife.103721.2.sa3

Abstract

Multiplexed assays of variant effects (MAVEs) make it possible to measure the functional impact of all possible single amino acid residue substitutions in a protein in a single experiment. Combination of variant effect data from several such experiments provides the opportunity to conduct large-scale analyses of variant effect scores measured across proteins, but can be complicated by variations in the phenotypes that are probed across experiments. Thus, using variant effect datasets obtained with similar MAVE techniques can help reveal general rules governing the effects of amino acid variation for a single molecular phenotype. In this work, we accordingly combined data from six individual variant abundance by massively parallel sequencing (VAMP-seq) experiments and analysed a total of 31,614 variant effect scores reporting solely on the impact of single amino acid residue substitutions on the cellular abundance of proteins. Using our combined variant effect dataset, we derived and analysed a collection of amino acid substitution matrices describing the average impact on cellular abundance of all residue substitution types in different structural environments. We found that the substitution matrices predict the cellular abundance of protein variants with surprisingly high accuracy when given structural information only in the form of whether a residue is buried or exposed. We thus propose our substitution matrix-based predictions as strong baselines for future abundance model development.

Introduction

The twenty commonly occurring amino acids have vastly different physicochemical properties, which allow them to play distinct roles for protein structure formation and function. Some amino acid residues might be particularly important for the thermodynamic and cellular stability of a certain protein and others crucial for intrinsic activity or function, with the different amino acid residue types contributing to each of these properties in unique ways (Ribeiro et al., 2020; Dunham and Beltrao, 2021; Tsuboyama et al., 2023; Cagiada et al., 2024). For certain aspects of protein functionality, the role played by each type of amino acid to maintain this functionality is well-understood. For example, the propensity of each of the twenty amino acids to form or break secondary structural elements has been studied extensively. The typical way to assess the role played by different amino acids in proteins is by using mutagenesis studies. The experimental methods available for performing such studies have developed immensely, and the field has thus

progressed from investigating single or few residue substitutions at a time, to conducting alanine scans, to now using saturation mutagenesis methods, including MAVEs, that allow the impact of thousands of residue substitutions on protein structure and function to be studied in a single experiment (Fowler and Fields, 2014).

The wealth of mutagenesis data collected in for example MAVE experiments can be used to analyse substitution effects in individual proteins comprehensively. The data also provides the opportunity to conduct large-scale studies to uncover general rules regarding the properties and functions of amino acid residue types across proteins, if data reporting on substitution effects in multiple proteins is combined. Some MAVEs assess the effects of residue substitutions on several proteins in a single experiment (Rocklin et al., 2017; Tsuboyama et al., 2023; Beltran et al., 2024), thus naturally allowing the latter type of analysis. However, most studies investigate substitution effects in a single protein at a time, and metaanalysis thus requires that data from different individual experiments are combined (Gray et al., 2017; Dunham and Beltrao, 2021; Høie et al., 2022). Such metaanalyses have proven very useful and have, among other things, revealed 100 amino acid subtypes based on substitution effect profiles (Dunham and Beltrao, 2021) and provided insight into which amino acid types that best discriminate protein-ligand binding interface from non-interface residues across proteins (Gray et al., 2017).

While combining substitution effect scores from several MAVEs allows for creation of large datasets, such an approach must take into consideration that the term MAVE covers a range of experimental methods that seek to investigate different aspects of protein function and use a variety of techniques for this purpose (Weile and Roth, 2018; Tabet et al., 2022; Geck et al., 2022). MAVE methods share a common foundation, namely setting up a system linking protein sequence or genotype to a phenotypic readout that can be selected for on large scale and traced back to genotype via sequencing (Fowler and Fields, 2014; Fowler et al., 2014). However, the molecular phenotype being studied and the selection mechanism used change from assay to assay. Some studies might take interest in variant effects on enzymatic activity or ligand binding affinity, while others might be concerned with the impact of substitutions on cellular fitness, to give a few examples. Ultimately, these assays are thus expected to report on overlapping, but non-identical variant effects.

MAVEs are additionally performed in various experimental backgrounds, for instance in different organisms and at different temperatures, which influence measured variant effect scores. It has been demonstrated that changing a single parameter related to the experimental background, for example expression level (Jiang et al., 2013; Cisneros et al., 2023) or the chemical (Mavor et al., 2016, 2018; Weile et al., 2021) or genetic (Thompson et al., 2020; Nguyen et al., 2024) background, can result in substantial differences to the MAVE readout. In an aggregated dataset consisting of a variety of MAVE scores, phenotype- or background-specific effects will thus likely be present and add noise to the analysis performed with the dataset. A final obstacle for MAVE score combination is that the effects of substitutions are quantified in different ways across experiments (Peterman and Levine, 2016; Rubin et al., 2017), meaning that score combination typically requires implementation of one (Gray et al., 2017; Munro and Singh, 2021; Høie et al., 2022) or several (Dunham and Beltrao, 2021) score transformation schemes to make scores from individual datasets compatible (or at least be on the same scale).

As the number of published MAVE datasets has grown (Esposito et al., 2019; Rubin et al., 2021), it has become increasingly feasible to perform metaanalyses of combined datasets that were obtained with similar methodologies and report on the same molecular phenotypes. On this basis, it is possible not only to decrease potential background noise, but also to discover general rules governing residue substitution effects for that specific phenotype. Thus, inspired by several previous MAVE data-based, large-scale analyses of substitution effects (Gray et al., 2017; Dunham and Beltrao, 2021) and the recent availability of homogeneous MAVE data, we here analysed a collection of six MAVE datasets that all report on the impact of single residue substitutions on protein cellular abundance. Our analysis is thus focused on a single molecular

phenotype, cellular abundance, which is known to be crucial for protein function and at the same time easily impacted by substitution of single residues, with widespread consequences for human disease (Matreyek et al., 2018; Cagiada et al., 2024).

Variant cellular abundance has been studied with different MAVE setups, including through protein fragment complementation assays that rely on monitoring cell growth (Faure et al., 2022) and by using fluorescent reporters to estimate cellular concentration (Matreyek et al., 2018). The substitution effects analysed here were all measured using the same type of MAVE experiment, specifically VAMP-seq, and the analysed effects were thus all observed in highly similar experimental backgrounds and were quantified with the same selection mechanism using almost identical methods (Matreyek et al., 2018, 2021; Amorosi et al., 2021; Suiter et al., 2020; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024). The VAMP-seq method estimates the steady-state abundance of all single residue substitution variants of a protein through expression of green fluorescent proteintagged protein variants in cultured human cells. Each cell expresses a single protein variant, and the cellular fluorescence intensity thus reports on the cellular abundance of the expressed variant. Abundance scores for thousands of variants are obtained in high-throughput by separating cells into bins of discrete fluorescence intensity intervals using fluorescence-activated cell sorting (FACS) followed by sequencing of each bin. Expression is controlled for, and changes in abundance are therefore mostly a product of changed degradation behaviour due to the substitution of residues (Matreyek et al., 2018).

The six VAMP-seq datasets have all been analysed extensively, and separately, in previous work (Matreyek et al., 2018, 2021; Amorosi et al., 2021; Suiter et al., 2020; Cagiada et al., 2021; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024). Here, we combined the six datasets and thereby obtained a total of 31,614 abundance variant effects scores, allowing us to perform a large-scale analysis of the impact of single residue substitutions on cellular abundance across proteins and thus contribute to our understanding of how different amino acid residues help maintain cellular protein levels. In our analysis, we specifically focused on analysing and predicting the impact of residue substitutions on abundance using only simple considerations of the structural environment of residues. Simple structural descriptors, such as local packing density and side-chain accessible surface area, have previously been shown to capture some substitution effects on protein folding stability, for example for cavity-creating substitutions (Serrano et al., 1992; Chakravarty and Varadarajan, 1999; Cota et al., 2000). The impact of substitutions on protein stability can to some extent also be predicted when a simple structural descriptor like residue relative solvent accessibility is used in combination with information about the physicochemical properties of wild-type and variant residues (Caldararu et al., 2021). Since substitutions that affect the free energy of folding of a protein tend to also have an impact on cellular abundance (Nielsen et al., 2017; Scheller et al., 2019; Abildgaard et al., 2019; Matreyek et al., 2018; Suiter et al., 2020; Cagiada et al., 2021; Høie et al., 2022; Gerasimavicius et al., 2023; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024), similar structural descriptors might also be useful for analysis of cellular stability sub-stitution effects. Our goal in this work is ultimately thus not to develop the most accurate possible variant abundance predictor, but rather to show how much variation in cellular abundance can be understood by simple rules that apply across different proteins.

For our structure-based analysis of the combined VAMP-seq dataset, we constructed a set of amino acid substitution matrices that for different structural environments contain the average abundance scores for all possible residue substitution types. Others have in earlier work similarly calculated substitution matrices using experimental mutagenesis data and studied the ability of the matrices to predict the impact of mutations (Yampolsky and Stoltzfus, 2005; Munro and Singh, 2021; Høie et al., 2022). The matrices presented here are, however, both cellular abundance- and structure context-specific. We provide an extensive analysis of the constructed matrices and use the matrices to show that a considerable amount of the variation in the VAMP-seq data can be explained by taking only residue solvent accessibility in the structure of the wild-type proteins into account together with wild type and variant residue types. Our results thereby highlight that simple biochemical considerations can compete with complex biophysics- and deep

learning-based models for predicting the impact of residue substitutions on cellular abundance, although we note that predictions are in both cases non-perfect and that more work is needed before the abundance of protein variants can be accurately modelled. We use the derived substitution matrices to explore the relationship between variant abundance and thermodynamic stability and show that the matrices likely contain abundance-specific signal. Finally, we also use the substitution matrices to propose a simple approach that, given a saturation mutagenesis dataset for the protein of interest, can be used to (*i*) assess structural models of the protein under the *in vivo* conditions used for the MAVE and (*ii*) discover solvent-exposed residues that play important roles for maintaining cellular protein levels.

Results and Discussion

Summary and comparison of VAMP-seq datasets

To perform our large-scale analysis of substitution effects on protein abundance, we first collected previously published VAMP-seq data for six soluble human proteins, all known to be involved in human disease or drug metabolism. Although VAMP-seq has been applied to membrane proteins (Chiasson et al., 2020; Amorosi et al., 2021), residue substitutions might affect the cellular abundance of membrane and non-membrane proteins at least partially through different mechanisms (Chiasson et al., 2020), and we therefore limited our study to only include soluble proteins. Our dataset thus consisted of PTEN (Matreyek et al., 2018, 2021), TPMT (Matreyek et al., 2018), CYP2C9 (Amorosi et al., 2021), NUDT15 (Suiter et al., 2020), ASPA (Grønbæk-Thygesen et al., 2024) and PRKN (Clausen et al., 2024) VAMP-seq scores (Table S1). We realise that CYP2C9 is in fact anchored to the endoplasmic reticulum membrane through an N-terminal, transmembrane helix, but the protein consists mostly of a large domain found in the cytoplasm (Amorosi et al., 2021), and we therefore included abundance scores for residues found in the cytoplasmic, soluble domain in our analysis.

The variant abundance scores for each of the proteins in our dataset were measured in individual VAMP-seq experiments, and the six experiments were all performed by sorting variant-expressing cells into four equally populated bins based on cellular fluorescence intensities. In all cases, the cellular abundance of a variant was quantified by taking a weighted average of the variant frequency across the four bins. The resulting scores were for each protein normalised so that a variant abundance score of 0 corresponds to nonsense-variant-like abundance and of 1 to wild-type-like abundance. We summarised and compared the six datasets by calculating a number of dataset descriptors and by mapping average abundance scores per position onto the protein structures (Fig. 1). When combined, the six VAMP-seq datasets contain in total 31,614 single residue substitution variant abundance scores. The scores span the full lengths of the protein sequences, although the mutational depth, that is, the average number of substitutions per residue position with a reported score after quality filtering, varies considerably from protein to protein (Fig. 1B). The mutational completeness of the datasets vary accordingly, with a minimum of 57% for PTEN and a maximum of 99% for PRKN (Fig. 1C). As the proteins also have different sequence lengths (Table S1), they contribute unequally to the total pool of variants (Fig. 1D).

We note that in spite of the similar experimental methodology and strategy for calculation of reported scores there are considerable differences between the shapes of the abundance score distributions obtained in the six experiments (Fig. S1). PTEN, TPMT and CYP2C9 have abundance score distributions with broad peaks at low and high abundance, whereas NUDT15, ASPA and PRKN have almost binary distributions with narrow peaks, in particular for abundance scores close to 0. All distributions appear bimodal, with the exception of the CYP2C9 score distribution, which has an additional peak around 0.5. Low abundance scores are generally distributed around 0, but for PTEN, the low abundance-peak is shifted towards higher scores. The reported abundance score standard deviations also differ in magnitude across datasets (Fig. S2). To represent the noise level in each dataset using a single number, we resampled the individual score distributions based on score standard deviations and calculated Pearson's correlation coefficient r (and Spearman's rank correlation coefficient r_s) between the original datasets and the

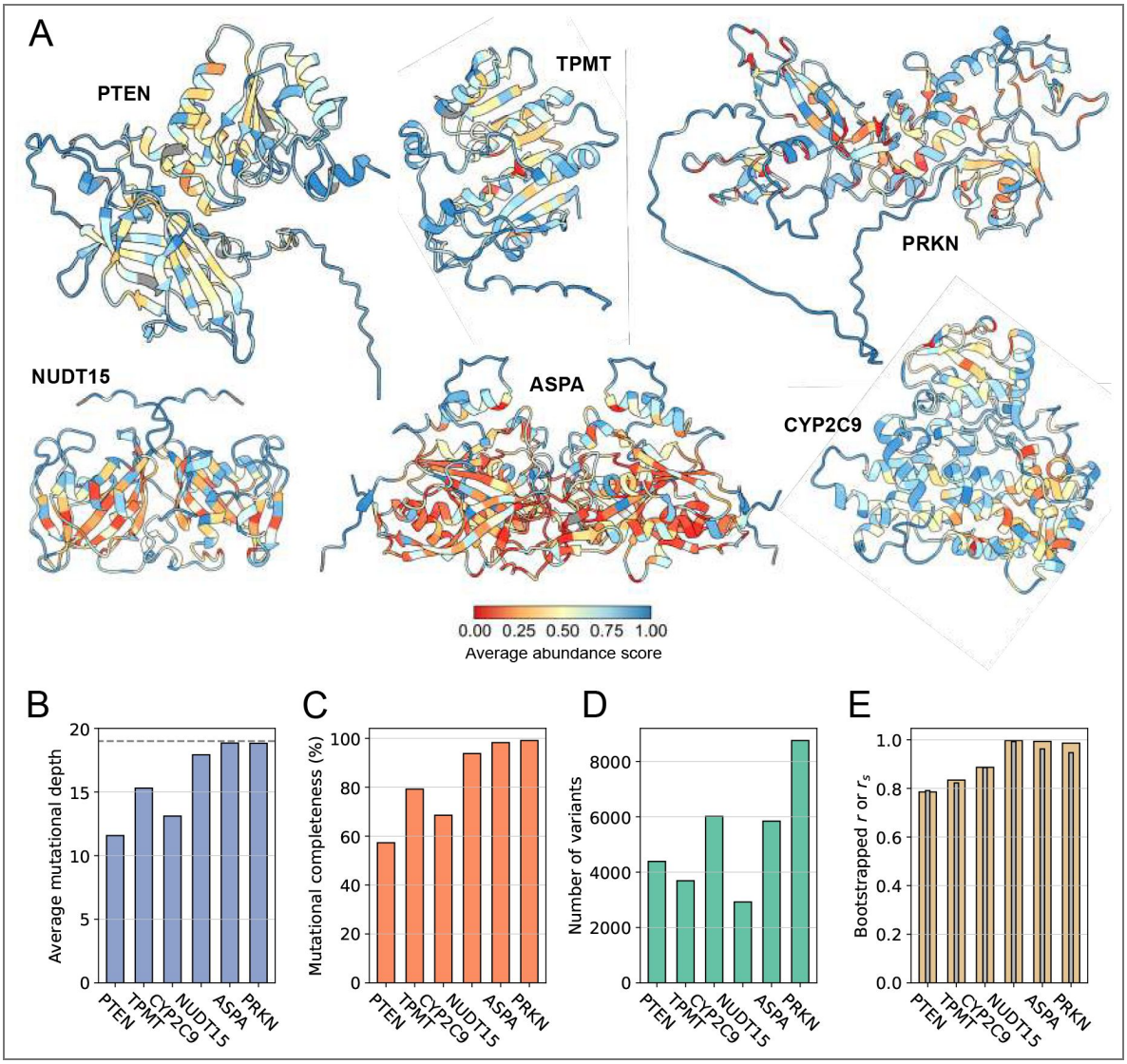


Figure 1. Overview of VAMP-seq datasets and proteins.

(A) Structures of proteins previously studied with VAMP-seq coloured by average VAMP-seq abundance score per residue. Averages are shown for all residues with abundance scores available, even if only one or few scores have been reported for the residue. Homodimer structures are shown for NUDT15 and ASPA, and residues with no reported scores are shown in gray. (B– E) Overview of VAMP-seq dataset sizes, completenesses and noise levels, showing specifically (B) the average number of variant scores per residue position, with 19 (indicated with dashed line) being the highest possible value, (C) mutational completeness, that is, the percentage of all possible single residue substitution variants with an abundance score in the dataset, (D) the total number of single residue substitution abundance scores per protein, and (E) Pearson's correlation coefficient r (yellow) and Spearman's rank correlation coefficient r_s (grey) between originally reported and resampled abundance score datasets, with resampling based on reported abundance score standard deviations.

resampled datasets. In addition to indicating the dataset noise levels, the resulting r (and r_s) values also provide an upper bound for how well abundance score predictions from a perfect variant abundance model would correlate with the experimental data (Fig. 1E). Since dataset noise levels and the broadness of peaks in the score distributions appear correlated, we assume that the distribution shape differences originate at least partly from the varying noise levels in the datasets.

We hypothesise that the dissimilar score distributions also arise from a combination of intrinsic differences between the six protein systems and from method-related, assay-specific influences on the measured scores. The PTEN and TPMT score distribution differences have for example previously been suggested to reflect dissimilar wild type protein melting temperatures (Matreyek et al., 2018). Additionally, the common VAMP-seq strategy to sort an equal number of variant-expressing cells into four bins has important implications, since variant effect scores produced in this manner are expected to depend on the composition of the variant library being sorted. Effectively, the input library composition determines the absolute values of the measured scores, and while the sorting and sequencing approach may produce an assay-independent variant ranking, the absolute values of the abundance scores are thus experiment-specific. In related work, we have expanded more on this point and explored how assay-specific influences may be accounted for when aggregating scores across datasets (Schulze et al., 2025).

In this study, we explore the six VAMP-seq datasets together without performing any additional within-dataset score transformations or across-dataset score normalisation. We show below that 'naively' combining VAMP-seq scores from the six experiments yields results that are largely in accordance with biochemical expectations and, importantly, facilitates useful structural analyses and construction of a relatively accurate prediction framework. We thus argue that direct combination of VAMP-seq scores across datasets is informative in spite of the methodological impact on absolute score values.

Abundance scores correlate with the degree of residue solvent-exposure

Our mapping of the average abundance score per residue onto the protein structures shows expected and previously described trends, most obviously that many solvent-exposed residues have high average abundance scores and that substitutions of core residues decrease abundance on average (Fig. 1A). Exceptions to these patterns exist, as some solvent-exposed residues for instance appear intolerant to substitution. These exceptions have to some extent been described and analysed in the original VAMP-seq publications. We note that the average abundance scores differ in magnitude from protein to protein, in particular in core regions.

To quantify the relation between the structural context and mutational tolerance of a residue in the context of abundance, we used the wild-type structures of the six proteins, which have relatively similar structural compositions (Fig. S3), to calculate features describing local residue environment, including relative solvent accessible surface area (rASA) and a weighted contact number (WCN). The WCN of a residue quantifies contact formation with all other residues in the protein so that a residue in a densely packed region will have a high WCN and vice versa. While rASA is sensitive to different residue environments on or close to the protein surface, rASA does not capture variation in the structure within the protein core. On the contrary, WCN has dynamic range for buried residues and is thus a complementary descriptor. Structural features were calculated with homodimer structures for NUDT15 and ASPA, in accordance with previous findings (Suiter et al., 2020; Grønbæk-Thygesen et al., 2024) and results described below. Monomeric structures were used for the four remaining proteins.

Both rASA and WCN correlate moderately well with the abundance scores, although the correlations are considerably higher for some proteins (PRKN and TPMT) than others (ASPA) (Fig. S4, S5). Generally, rASA captures the ranking of variant effects better than the WCN; Spearman's rank correlation coefficient, r_s, is on average 0.45 (0.54) between abundance scores and rASA and 0.41 (0.43) between abundance scores and WCN when including all datasets (or

when including all except for ASPA) (Fig. S4). When looking at the average abundance score per residue, we found solvent-exposed residues with reduced to strongly reduced abundance in all proteins, though in particular in ASPA and PRKN. Some residues buried in the structures on the other hand appear mutationally tolerant (Fig. S5).

Abundance score predictions from simple structural considerations

Based on the correlations between abundance scores and rASA and WCN, we hypothesised that combining the simple structural descriptors with information about the physicochemical properties of wild type and variant residue types would allow us to explain the effects of many substitutions on cellular abundance (Fig. 2). To test this hypothesis, we first established a baseline frame-work without using structural information, after which we extended our analysis to take structure into account.

For the structure-free analysis, we used the combined pool of abundance scores from all six proteins to construct a substitution matrix that, for each of the 380 possible amino acid residue sub-stitution types, contains the abundance score average (Fig. S6). We tested the ability of the substitution matrix to act as an abundance score predictor using leave-one-protein-out cross-validation; that is, we recalculated the matrix leaving out the entire VAMP-seq dataset of a single protein and then used the averages in the recalculated matrix as abundance score predictions for substitutions in the omitted protein. While the average abundance scores are not particularly good predictors of variant abundance (Fig. 2C, S7), some signal in the experimental data is indeed captured even with-out considering the structural context of the wild-type residues (Fig. S7). Previous studies have reached similar conclusions in analyses that calculate structure-independent substitution matrices and predict variant fitness by combining data from a variety of MAVE experiments (Munro and Singh, 2021; Høie et al., 2022).

We next explored whether we could improve upon this baseline by taking simple structural features into account. We used rASA values calculated for the wild-type protein structures to classify all residues as either solvent-exposed or structurally buried. We then calculated two new substitution matrices, in one matrix averaging over variant abundance scores from residues belonging to the solvent-exposed class and in the other over scores from buried residues (Fig. 2A, B). Using the same approach as above, we evaluated how well our exposure-based substitution matrices could predict unseen abundance data using leave-one-protein-out cross-validation. The abundance score prediction of a variant was thus obtained by looking up the abundance score average of the relevant substitution type in either the exposed or buried residue matrix depending on the structural class of the residue for which a prediction was made. Accounting for the structural context of the wild-type residue considerably increases the ability of the model to capture the impact of substitutions on abundance, with the average r across datasets increasing from 0.35 to 0.52 when not only the substitution type but also the degree of residue solvent-exposure is considered (Fig. 2C, S8).

We next tested if additional substitution effects could be captured by including secondary structure in our framework. To perform this analysis, we calculated two new sets of substitution matrices. In the first set, we averaged abundance scores for residues in different types of secondary structure and obtained three matrices with average effects in helix, stand and loop structure (Fig. S6). To create the other set of matrices, we split residues by both secondary structural class and degree of solvent-exposure, creating residue classes such as “solvent-exposed helix” and “buried strand”, which resulted in six matrices of average effects. We performed abundance score predictions with the two sets of matrices using the cross-validation approach described above. The substitution matrices based exclusively on residue secondary structure perform only slightly better in the prediction task than the structure-independent matrix (Fig. 2C, S9), and secondary structure is by itself thus not a suited descriptor to capture variant effects on abundance. The combination of solvent-exposure and secondary structure type similarly does not improve predictions compared to when only solvent-exposure is used (Fig. 2C, S10).

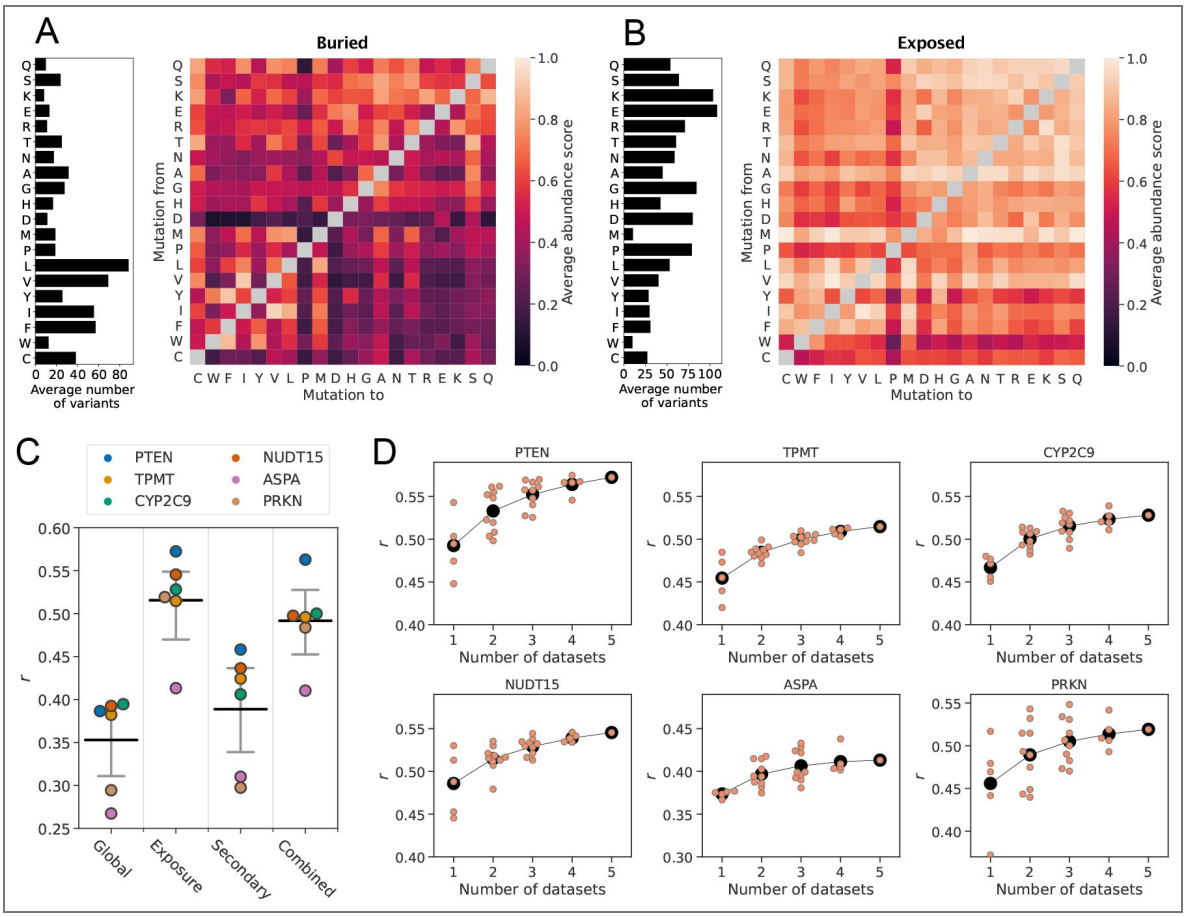


Figure 2. Average abundance score substitution matrices predict unseen abundance data.

Average abundance scores for all possible amino acid residue substitution types for residues sitting in either (A) structurally buried or (B) solvent-exposed environments in the wild-type protein structures. The bar plot to the left of each sub-stitution matrix indicates the average number of abundance scores that was used to calculate score averages in each row of the matrix. Amino acids are sorted according to row averages in the global substitution matrix (Fig. S6). (C) Correlations between experimental abundance scores and abundance scores predicted from average abundance score substitution matrices using leave-out-one-protein cross-validation. The matrices used for predictions were constructed using all data (Global) and for different residue categories based on residue solvent-exposure (Exposure), secondary structure (Secondary) or both solvent-exposure and secondary structure (Combined). The horizontal, black bars mark the average correlation coefficient for each category, and the grey bars mark the 95% confidence intervals of the average *r*. Confidence intervals were estimated using boot-strapping by resampling the six *r* values 10,000 times per model type. (D) Correlations between experimental abundance scores and abundance scores predicted from average abundance score substitution matrices for residues sitting in either buried or exposed environments. Scores were predicted for each protein (indicated above each plot) using average abundance score substitution matrices calculated using an increasing number of VAMP-seq datasets. Orange data points mark the prediction result from a single specific combination of datasets, and the average for each dataset count is indicated with black circles.

Since our VAMP-seq dataset is relatively small and limited to six proteins, we next assessed if our substitution matrix-based abundance models could be expected to improve with the addition of more VAMP-seq data in the future. For this assessment, we recalculated all exposure-based matrices using all possible combinations of the six datasets. We then performed predictions again, for every protein using all matrices in which the data for this protein did not occur to perform predictions. In this way, we estimated how r varied with the number of datasets included in the substitution matrix calculation. While using more datasets improves predictions on average, we observe that, depending on the exact combination of datasets, predictions can be as good or better when only one or few datasets are used to calculate the abundance score averages as when five datasets are used (Fig. 2D). Some datasets thus appear to be more related than others (Fig. S11). The average r seems to be plateauing for all six proteins, suggesting that this framework would likely only improve to a small degree with the addition of more data in the future. We observed similar trends when we performed the same analysis for the Combined matrices (Fig. S12).

The results presented above show that combining simple structural features with information about the wild-type and variant residue types enables us to capture the effects of many substitutions on the cellular abundance of a protein. It is at the same time clear that some substitution effects are not properly captured within this prediction framework. We in particular find one common group of outliers in predictions of the NUDT15, ASPA and PRKN datasets, namely substitutions that experimentally have strongly reduced abundance scores, but which the substitution matrices predict to have wild-type-like or nearly wild-type-like abundance (Fig. S7, S8, S9, S10). These prediction outliers likely occur for several reasons, including due to (*i*) biochemical and -physical features not captured in the substitution matrix framework and (*ii*) differences between VAMP-seq datasets (Fig. S1) that occur for technical rather than biochemical reasons (Schulze et al., 2025). An accurate analysis of the prediction outliers, for example to inform model refinement, would thus involve disentangling these two potential outlier causes, which, in addition to the substantial noise levels of the data, makes outlier detection and analyses challenging.

Substitution matrices and ΔΔG predictors rank variant effects equally well

We and others have previously observed a correlation between the cellular abundance of variants and (predicted) differences between variant and wild-type protein folding stabilities ($\Delta\Delta G = \Delta G_{variant} - \Delta G_{wild\ type}$), both in low-throughput cellular studies (Nielsen et al., 2017; Scheller et al., 2019; Abildgaard et al., 2019) and in high-throughput VAMP-seq studies (Matreyek et al., 2018; Suiter et al., 2020; Cagiada et al., 2021; Høie et al., 2022; Gerasimavicius et al., 2023; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024). Thermodynamically destabilised protein variants may be more likely to populate unfolded conformations than the wild-type protein, thereby exposing hydrophobic core residues that can trigger the cellular protein quality control (PQC) system to degrade the protein, for example to avoid aggregate formation (Hipp et al., 2019; Clausen et al., 2019). While ΔΔG values usually correlate with cellular abundance, we do not necessarily expect a perfect (rank) correlation between the two quantities, since substitutions of single residues can trigger degradation via a number of mechanisms that do not require complete unfolding from the native state. For example, a substitution can introduce neo-degrons at surface-exposed positions (Clausen et al., 2024), cause local unfolding that exposes a degron (Kampmeyer et al., 2022) or change degradation-relevant post-translational modification patterns (Vazquez et al., 2000). For one of several possible reasons, a variant associated with a small ΔΔG might thus be completely degraded in the cell.

To explore the relationship between variant abundance and folding stability, and to assess the performance of our substitution matrix-based prediction method, we calculated ΔΔG scores for all single residue substitution variants of the six VAMP-seq proteins. We predicted ΔΔG scores with three different models, the Rosetta energy function (Park et al., 2016) and the deep learning-based RaSP (Blaabjerg et al., 2023) and ThermoMPNN (Dieckhaus et al., 2024) models. We

calculated r_s between ΔΔG predictions and VAMP-seq scores, analysing scores from each protein separately. As expected and observed in previous work, we find that variant ΔΔG values correlate with abundance scores (Fig. 3A), although with noticeable differences in the magnitude of the rank correlation coefficient across datasets; for example, r_s ranges from −0.45 for PRKN to −0.59 for NUDT15 when predicted with Rosetta and from −0.51 for TPMT to −0.65 for NUDT15 when predicted with Ther-moMPNN (Fig. 3A). While an r_s of −1.0 is not expected due to noise in the VAMP-seq data (Fig. 1E), it is clear that predicted ΔΔG values nevertheless cannot fully explain the abundance data, likely due to a combination of limited ΔΔG prediction accuracy (Blaabjerg et al., 2023; Dieckhaus et al., 2024; Degn et al., 2025) and the fact that VAMP-seq abundance measurements contain other contributions than global changes in protein stability (ΔΔG). Moreover, variation in r_s across the six systems possibly indicates that effects of substitutions on abundance are dominated by ΔΔG effects to a higher degree in some systems than in others.

Abundance scores are ranked best by ThermoMPNN with an average r_s of −0.56 across the six proteins, in line with observations that ThermoMPNN captures experimentally measured stability changes more accurately than Rosetta and RaSP (Dieckhaus et al., 2024; Degn et al., 2025). We note that all of the models we have tested, including ThermoMPNN, are far away from predicting abundance scores perfectly, even considering the r_s upper limits given by the noise in the VAMP-seq data (Fig. 1E). Rosetta and RaSP (the latter of which was trained to predict Rosetta scores rapidly) are far more complex models than any of the abundance score substitution matrices, and it is thus noteworthy that the exposure-based matrices have a predictive performance comparable to that of Rosetta or RaSP (Fig. 3A). As an abundance score predictor, the exposure-based substitution matrices are thus competitive with RaSP, a deep learning model using a 3D convolutional neural network to process structural information, and Rosetta, which relies on a complex energy function and includes post-mutation structural sampling as part of the ΔΔG calculation pipeline. Based on this, we propose our matrix-based predictions, alongside ThermoMPNN, as a strong baseline for future abundance model development. We also emphasise the utility of simple, interpretable structure features for stability-related variant effect predictions.

Abundance score predictions and ΔΔGs contain (some) orthogonal information

Variant ΔΔG predictions and substitution matrix-based abundance score predictions correlate only moderately well (Fig. S13), suggesting that the two model types may achieve relatively similar rank correlations with the VAMP-seq data for at least somewhat different reasons. Analysis of the differences in model behaviour might shed light on the relationship between variant folding stability and cellular abundance and potentially guide improved abundance model design. We therefore next compared the exposure-based substitution matrices to the predicted ΔΔG values in detail, focusing the analysis on ThermoMPNN ΔΔG predictions. Below, we define the substitution profile of a residue as the vector of abundance scores for variants of that residue.

We compared ΔΔG predictions to predictions derived from the exposure-based substitution matrices by identifying all individual residues for which one model predicted the experimental abundance score substitution profile considerably better than the other model. For every individual residue in a protein, we specifically compared $r_{s,\Delta\Delta G}$ and $r_{s,\text{matrix}}$, which are the rank correlation coefficients between the experimental residue substitution profile and the ΔΔG-based ($r_{s,\Delta\Delta G}$) or the matrix-based ($r_{s,\text{matrix}}$) substitution profile predictions. To account for noise in the VAMP-seq data, we resampled the abundance scores of every individual substitution profile 10,000 times based on the experimental abundance score standard deviations. We then used the resampled profiles to ask how often $-r_{s,\Delta\Delta G} > r_{s,\text{matrix}}$ or $-r_{s,\Delta\Delta G} < r_{s,\text{matrix}}$ for a given residue. If one method consistently predicted the substitution profile of a residue better than the other (that is, in more than 95% of cases), we classified the residue either as “ΔΔG-favoured” (for $-r_{s,\Delta\Delta G} > r_{s,\text{matrix}}$) or “matrix-favoured” (for $-r_{s,\Delta\Delta G} < r_{s,\text{matrix}}$).

We used this approach to analyse 1816 residue substitution profiles distributed across the six proteins. Of the 1816 analysed residues, we identified 337 (19%) matrix-favoured residues and 213 (12%) ΔΔG-favoured residues. 1266 of the 1816 residues (70%) have substitution profiles that are

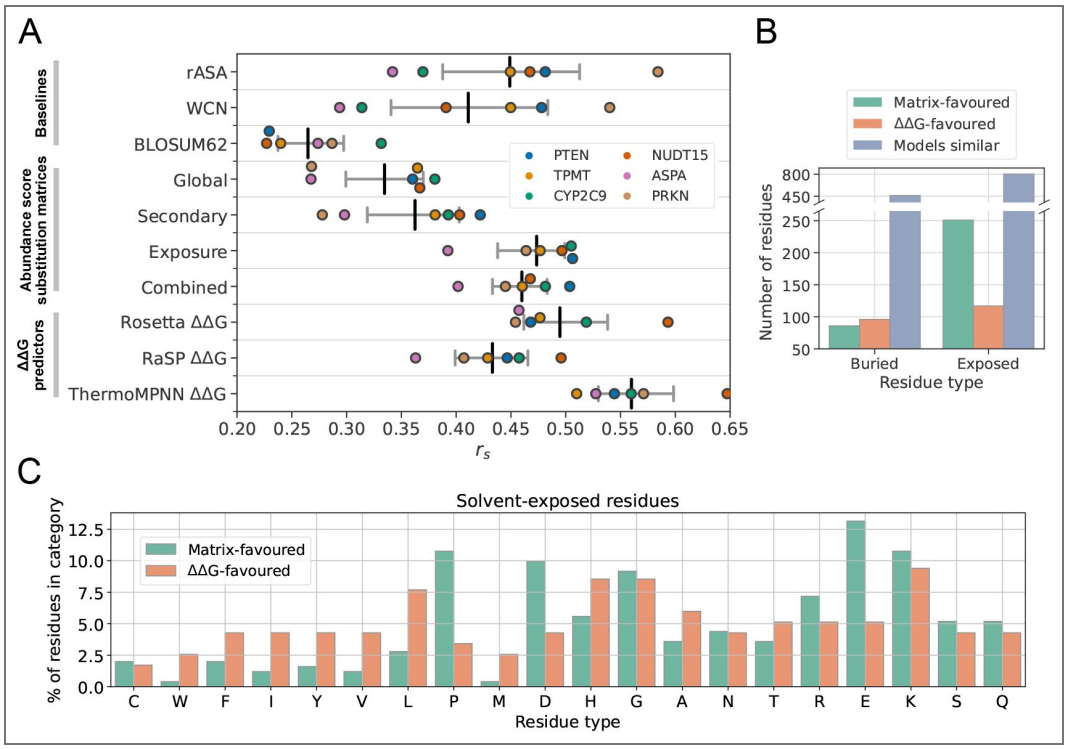


Figure 3. Ranking of variant abundance by substitution matrices and ΔΔG predictors.

(A) r_s between experimental abundance scores and variant effect score predictions from a number of models, including abundance score substitution matrices (Global, Secondary, Exposure and Combined) and ΔΔG predictors (Rosetta, RaSP and ThermoMPNN). r_s between experimental abundance scores and residue rASA or WCN values are also included, as well as variant effect scores predicted with the BLOSUM62 matrix (Henikoff and Henikoff, 1992). For the three ΔΔG predictors and WCN values, absolute r_s values are shown. Coloured points indicate r_s for individual proteins. Vertical bars in black mark the average r_s across the six proteins, and grey bars indicate 95% confidence intervals of the r_s average. Confidence intervals were estimated using bootstrapping by resampling the six (one per protein) rank correlation coefficients 10,000 times for every predictor. (B) Number of individual residue substitution profiles captured best by the exposure-based substitution matrices (green) or Ther-moMPNN ΔΔG predictions (orange), shown separately for buried and exposed residues. Variant abundance is often ranked better by the substitution matrices than ΔΔG values for exposed residues. For most residues, the substitution matrix and ΔΔG models rank residue-level variant effects with similar accuracy (blue). (C) Distributions of amino acid residue types in the groups of matrix-favoured (green) and ΔΔG-favoured (orange) solvent-exposed residues. For every amino acid residue type, the residue count in either the matrix- or ΔΔG-favoured category was normalised to the total number of residues in the category, thus revealing enrichment and depletion of certain amino acid types in one of the two groups compared to in the other.

not predicted considerably better by one method than by the other; for these residues, $r_{s,\Delta\Delta G}$ and $r_{s,matrix}$ cannot be told apart, either because the two predictors agree on how variants should be ranked or due to noise in the experimental data. When considering buried and exposed residues separately, we find 251 of 1172 (21%) exposed residues to be matrix-favoured, while only 117 of 1172 (10%) exposed residues are ΔΔG-favoured (Fig. 3B). The substitution profiles of exposed residues thus tend to be captured best by the substitution matrix model or equally well by the two model types. We also identify 182 of 644 (28%) buried residues for which either ΔΔG values or substitution matrix-based scores most accurately rank variant effects. For buried residues, the difference in the number of ΔΔG- and matrix-favoured residues is small (Fig. 3B). Since most sub-stitution profiles have indistinguishable $r_{s,\Delta\Delta G}$ and $r_{s,matrix}$ values, we do not expect a model that combines substitution matrix-based and ThermoMPNN predictions to outperform the r_s of Ther-moMPNN when evaluated globally, that is, across all variants in the individual systems.

We next set out to characterise the residues that we had classified as either matrix- or ΔΔG-favoured. To this end, we first analysed the amino acid composition of the matrix-favoured and ΔΔG-favoured residue groups, focusing our analysis on solvent-exposed positions. We specifically counted every amino acid type in the group of solvent-exposed and matrix-favoured residues (or in the group of solvent-exposed and ΔΔG-favoured residues) and then normalised the counts to the total number of residues in the group (Fig. 3C). This analysis reveals that certain amino acid types are enriched in the matrix-favoured residue group compared to in the ΔΔG-favoured group, and vice versa. Most notably, we find that aspartate, glutamate and proline are particularly enriched in the matrix-favoured residue group, meaning that for these three amino acid types, the substitution matrix model capture variant effects particularly well. The ΔΔG-favoured group is, on the contrary, enriched in aromatic and several non-polar amino acid types, and so the variant abundance of solvent-exposed residues of these amino acid types tend to be captured best by ThermoMPNN ΔΔG scores. We note that there is a relatively small number of abundance scores for substitutions from solvent-exposed aromatic and non-polar residues in our combined VAMP-seq dataset (Fig. 2B), and so the latter result might reflect that the substitution matrix for exposed residues has not seen a lot of training data for aromatic and non-polar amino acid types.

We further asked how the substitution profiles of solvent-exposed matrix-favoured and ΔΔG-favoured residues differ. To answer this question, we calculated average abundance score substitution matrices using VAMP-seq data for all residues in the two residue categories (Fig. S14A, B) and then determined the difference between the two resulting substitution matrices (Fig. S14C). We find that the average abundance scores of substitutions from several charged and polar residues to aromatic or non-polar residues tend to be smaller for matrix-favoured residues than for ΔΔG-favoured residues. Assuming that ThermoMPNN accurately models variant ΔΔG, these results suggest that non-conservative substitutions of certain polar or charged residues on the protein surface (namely those belonging to the matrix-favoured residue group) may reduce cellular abundance more than what would be expected from a purely thermodynamic perspective, and these effects seem to be captured in the substitution matrix model.

In conclusion, we have shown that predicted ΔΔG values and entries in the substitution matrices are not identical with respect to variant effect ranking, neither for buried nor exposed residues, suggesting that the substitution matrices may contain abundance-specific signal. It is interesting that exposed residues show a tendency to be matrix-favoured, as this is in line with expectations that some surface-level substitutions may affect folding stability and interactions with the PQC system in different ways. Since neither the substitution matrices nor ThermoMPNN are perfect models, it is, however, difficult to very confidently conclude that discrepancies between predictions from the two different models necessarily indicate that the effects of certain variants on abundance are driven exclusively by folding free energy changes while other variant effects are products of more complex reactions in the cell.

Substitution matrices reveal variant effect trends across protein backgrounds

We find that the average substitution effects in the VAMP-seq-derived substitution matrices are in general in accordance with biochemically expected and previously observed patterns (Munro and Singh, 2021; Høie et al., 2022). Across residue environments, substitutions conserving the physicochemical properties of the reference residue are often tolerated well, or at least better than non-conservative substitutions, and it is moreover clear that residues found in structurally buried environments tolerate substitutions less than solvent-exposed residues (Fig. 2A,B). While substitution of residues in buried environments likely decrease protein cellular abundance on average because the thermodynamic stability is compromised, substitution of exposed residues might affect cellular stability through different mechanisms (Correa Marrero and Barrio-Hernandez, 2021), for instance by interference with formation of intermolecular interactions (and thus potentially thermodynamic stability of complexes) (Suiter et al., 2020; Grønbæk-Thygesen et al., 2024) or through modification of exposed degradation motifs (Clausen et al., 2024).

We observe that, at solvent-exposed sites, substitutions from polar or charged residues to non-polar or aromatic residues are not tolerated as well as conservative substitutions. Comparing matrices across residue environments shows that substitution to proline is, as expected and previously observed, not tolerated well in any environment, and substitution from proline is also often disruptive. Mutation of cysteine to any other residue also appear to on average reduce abundance considerably in several environments. However, leaving out PRKN abundance scores from the calculation of the averages makes mutation from cysteine much less disruptive (Fig. S15), and the average effects when all proteins are included are thus dominated by the Zn^{2+}-coordinating cysteine residues that are critical for the abundance of PRKN (Clausen et al., 2024). Generally, sub-stitution matrices constructed with individual datasets display unique features (Fig. S11A,B) and relate to each other more closely for certain protein pairs than others (Fig. S11A,B), while leave-one-protein-out substitution matrices resemble each other relatively closely (Fig. S11C,D).

In line with expectations and previous results (Munro and Singh, 2021; Høie et al., 2022), the global, structure-independent matrix is furthermore asymmetric (Fig. S16), that is, for many residue pairs, substitution from residue *X* to *Y* tend to affect abundance differently than substitution from residue *Y* to *X*. In the global context, substitutions from non-polar or aromatic residues to polar or charged residues for example tend to be more disruptive than the opposite substitutions. Such asymmetry is, per construction, not present in traditional substitution matrices used for construction of sequence alignments (Dayhoff et al., 1978; Henikoff and Henikoff, 1992), although more recent work has derived asymmetric matrices using sequence data (Dang et al., 2022). The matrix with average abundance scores from buried residues is also asymmetric, in spite of the fact that substitutions from polar or charged residues to hydrophobic or aromatic residues (and not only the opposite) tend to be quite disruptive of abundance in buried environments (Fig. S16). While some asymmetric features are still present when only solvent-exposed residues are considered, the asymmetry is less pronounced, meaning that non-conservative substitutions often affect cellular stability relatively similarly irrespective of the biochemistry of the wild-type residue in this structural environment (Fig. S16).

From visual inspection of the substitution matrices, it is clear that certain pairs and groups of amino acid types share a number of substitution profile features. In a given environment, each amino acid type has two substitution profiles in form of the two vectors of nineteen abundance score averages describing effects of substitution to or from the given amino acid. We explored sub-stitution profile similarities by performing hierarchical clustering of the average score matrices, focusing our analysis on effects in buried and exposed environments (Fig. 4A). The analysis of buried residues identifies several clusters that are in accordance with residue biochemistry; for example, when considering mutational profiles for substitutions *to* a given residue, one cluster contains the aromatic residues, another cluster consists of several non-polar residues, and

residues with similar charges are identified as closely related. We also observed that, in buried environments, substitutions to glycine are in many cases as detrimental as substitutions to polar or charged residues. The clustering also highlights that substitutions from non-polar residues to the aromatic residues are not well-tolerated on average, potentially due to differences in the volume of the side chain. In fact, substitutions from non-polar buried residues to the aromatic residues share features with substitutions from non-polar buried residues to charged or polar residues. In the exposed environment, we also found clusters that are generally in accordance with residue biochemistry.

For both buried and solvent-exposed environments, several residue groups identified in the clustering of the matrix rows are also identified in the clustering of the matrix columns. However, the dendrograms along the two axes are not identical, neither in the buried nor in the exposed context. The order in which clusters form depend on the axis, and some residues appear in different clusters depending on whether they are being mutated from or to. For example, serine cluster with the majority of polar and charged residues in the *mutation from* direction in the buried matrix, but not in the *mutation to* direction. More surprisingly, aspartate and glutamate behave similarly when they occur as the variant residue in the buried environment, but have more different mutational patterns when they are the wild type residue. Generally, buried aspartate appears to play a critical role for maintaining abundance and is clearly an outlier in the *mutation from* direction, as it does not seem to cluster with any other residues. Even mutation from aspartate to glutamate is not tolerated well in the protein core, highlighting that the importance of buried aspartate for cellular stability is not only charge-related. In the exposed environment, aspartate and glutamate do cluster together as both wild-type and variant residues.

We further quantified the relations between the average substitution profiles of the 20 amino acids using principal component analysis (PCA). We concatenated the 20 *mutation to* and 20 *mutation from* abundance score averages for every amino acid residue type and used the 20 resulting 40-vectors as input for the PCA. In that way, the PCA explored amino acid similarities and differences considering simultaneously the effects of substitution from and to a given amino acid. For buried residues, the first principal component (PC1) appears to separate the polar and charged amino acids from the non-polar and aromatic amino acids (Fig. 4B), showing (as expected) that amino acid hydrophobicity is an important property for explaining variation in the substitution profiles in buried environments. Along PC1, proline and glycine are placed close to the polar and charged amino acids. PC2 for buried residues clearly separates the nonpolar and aromatic residues, with most polar residues found at a shorter distance to the nonpolar residues than the aromatics. The PCA of average abundance scores in exposed environments separates the nonpolar and aromatic amino acids from the remaining amino acids along PC2 rather than PC1. The PC1 for residues in an exposed environment, which on its own explains a relatively high amount of variation in the sub-stitution profiles, instead seems at least partially correlated with how well substitutions to polar and charged residues are tolerated (Fig. 4B).

For residues in loops, we noticed that glycine, histidine, asparagine and aspartate share certain substitution profile features, as they all appear particularly sensitive to mutation to /*3*-branched residues, including threonine, isoleucine and valine, as well as to tryptophan and proline (Fig. S6). In loop structure, many residues, and in particular exactly glycine, histidine, asparagine and aspartate, are able to adopt backbone Φ and ψ angles in the $0° < \Phi < 180°$, $-90° < \psi < 90°$ area of the Ramachandran plot (Hovmöller et al., 2002). Backbone dihedral angles in this region of the Ramachandran plot are associated with left-handed helix formation, although the vast majority of residues adopting these dihedral angles do not actually take part in forming left-handed helix (Hov-möller et al., 2002). A number of residue types, including isoleucine, threonine, proline and valine, have very little density in this region of the Ramachandran plot (Hovmöller et al., 2002). We therefore hypothesised that the ability to adopt the Φ and ψ angles of left-handed helical conformations influences the average abundance scores of residues in loop structure. To test our hypothesis, we identified all loop residues with backbone angles in the Φ,ψ area stated above. As expected, many of the identified residues were glycine and asparagine residues, whereas no isoleucine or valine residues in our six protein dataset were found to have

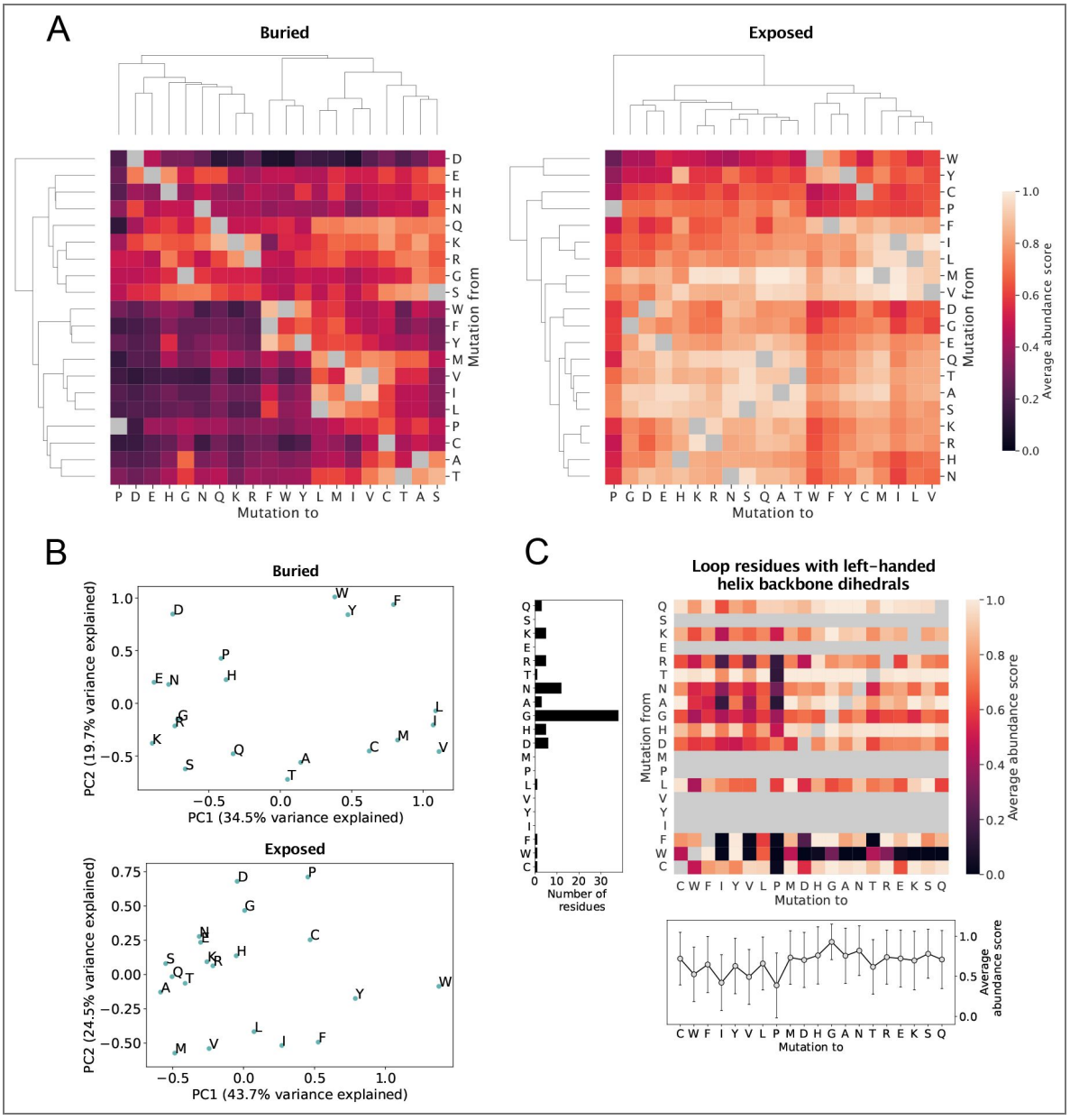


Figure 4. Analysis of average abundance score substitution matrices.

(A) Hierarchical clustering of average abundance score substitution matrices for buried and solvent-exposed environments. Clustering was performed along both axes of the matrices. Grey squares indicate a synonymous substitution. (B) Principal component analysis of substitution profiles describing the effects of all substitutions from and to each of the twenty amino acid residues. The analysis was performed using average abundance scores for residues in buried and exposed environments separately. (C) Analysis of substitution profiles of loop residues with backbone dihedral angles similar to those found in left-handed helices. The heat map shows average abundance scores for all loop residues with left-handed helix-like backbone dihedral angles in the six proteins. Grey squares indicate a synonymous substitution as well as no data. The plot below the heat map shows average abundance scores for substitutions to each of the twenty amino acid residues for this particular class of loop residues, with error bars indicating the standard deviation over all abundance scores for the substitution type. The left bar plot shows the total number of loop residues that for each residue type was found across the six proteins to adopt the left-handed helix-like backbone conformation.

this backbone conformation (Fig. 4C). Moreover, for several residue types, we identified only a single residue or no residues with this conformation. We next calculated an average abundance score substitution matrix for the identified residues and averaged the abundance scores for substitutions to each of the twenty amino acids across reference residue types (Fig. 4C). Clearly, substitution to proline, isoleucine, valine and tryptophan are the least favourable substitution types on average, and substitutions to glycine are the most tolerated. Our analysis thus suggests that substitution of loop residues with left-handed helical backbone conformations typically disrupts cellular abundance if this conformation is not usual for the variant residue to adopt.

Finally, we tested whether known helix propensity scales could explain patterns in the abundance substitution matrices. To do so, we calculated the correlation between average abundance scores for residues found in *a*-helices and different experimentally and statistically derived residue helix propensity scales (Pace and Scholtz, 1998). All correlations between average abundance scores and helix propensity scales are weak (Fig. S17), indicating that the average substitution patterns in the abundance data are not dominated by helix propensities. A similar observation was made in a study aggregating data from a broader range of MAVEs (Gray et al., 2017).

Average substitution profiles identify stabilising dimer interfaces

In vitro studies have demonstrated that NUDT15 (Carter et al., 2015) and ASPA (Moore et al., 2003; Bitto et al., 2007; Le Coq et al., 2008) form homodimeric assemblies in solution and when crystallised. In a previous analysis of the NUDT15 VAMP-seq dataset, the homodimer interface present in the NUDT15 crystal structure was found to contain several residue sites at which the mutational tolerance was lower than the average tolerance across all residues and substitutions hence consistently resulted in decreased abundance (Suiter et al., 2020). Sensitivity towards substitutions has also been noticed for residues sitting in the monomer-monomer interface found in the ASPA crystal structure (Grønbæk-Thygesen et al., 2024). Since prior work thus suggests that NUDT15 and ASPA homodimerisation is important for the maintaining cellular abundance of the proteins, we have used the NUDT15 and ASPA homodimer structures for all analyses described above. However, no quantitative analysis of the mutational tolerance and hence importance for maintaining cellular stability has actually been reported for the interface residues in ASPA, and so to justify our choice, we (re)analysed the crystal structure interfaces of the two proteins using the pooled VAMP-seq dataset.

We specifically examined whether NUDT15 and ASPA homodimerisation might affect the *in vivo* stability of the proteins in the VAMP-seq experiments by quantifying the extent to which the abundance scores of the homodimer interface residues resemble the average substitution scores of buried or solvent-exposed residues. For the analysis, we recalculated the exposure-based substitution matrices using only monomeric structures for the residue classification and leaving out either NUDT15 or ASPA data from the calculation. For every interface residue in the dimer structures, we then calculated the root-mean-square-deviation (RMSD) between all variant abundance scores reported for the interface residue and the average abundance score substitution profile for the interface residue amino acid type in either buried or exposed environments. For all interface residues, we thus obtained two numbers, $RMSD_{exposed}$ and $RMSD_{buried}$, which respectively quantify how closely related the substitution profile scores of the residue are to the average scores of exposed and buried sites (Fig. 5). If the $RMSD_{buried}$ of a residue is smaller than the $RMSD_{exposed}$, we thus expect the residue to be buried in the monomer-monomer interface in the cell, and vice versa.

In NUDT15, the abundance scores for 11 out of 13 crystal structure interface residues are more similar to the average abundance scores of buried residues than exposed residues (Fig. 5A). The disruption of the interactions formed by these 11 residues thus reduces the cellular abundance of NUDT15 relatively similarly to how disruption of the interactions of core residues would, suggesting that the 11 residues are buried in the dimer interface inside the cell. Residues L153 and L156 have abundance scores that resemble those of exposed residues more closely than those of buried residues. In the crystal structure of the dimeric assembly of the protein, L153 and L156 are positioned next to each other on the border of the dimer interface (Fig. 5C). While the side

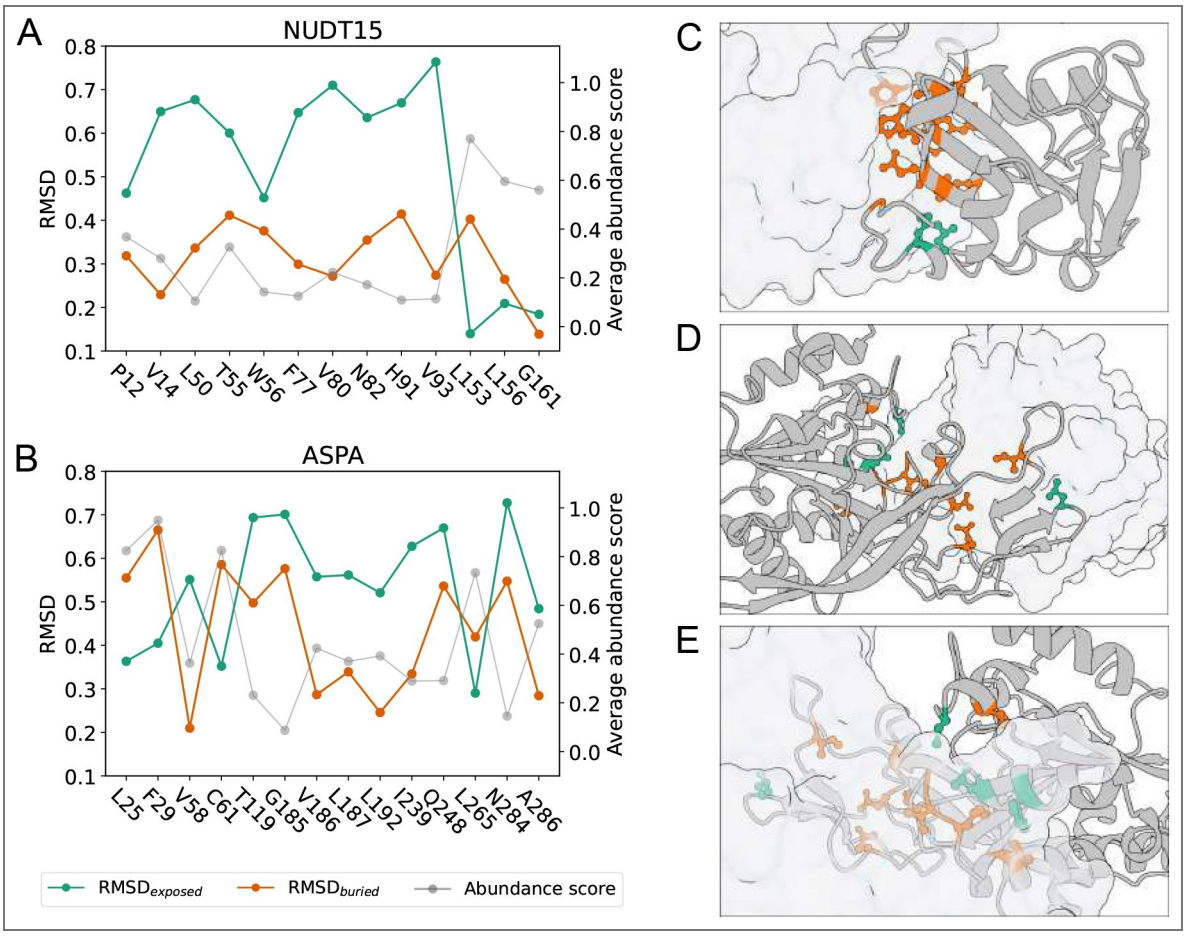


Figure 5. Homodimerisation stabilises NUDT15 and ASPA *in vivo*.

RMSD between abundance score substitution profile and average abundance scores in buried (orange) and exposed (green) environments for residues in (A) the NUDT15 dimer interface and (B) the ASPA dimer interface. For each interface residue, the average abundance score (grey) is also shown. If $RMSD_{exposed}$ is smaller than $RMSD_{buried}$, the abundance score substitution profile of the residue resembles the average profile of an exposed residue more than that of a buried residue of the same amino acid type, and vice versa. RMSD values were only calculated for residues with at least five reported abundance scores. To evaluate if $RMSD_{exposed}$ and $RMSD_{buried}$ are considerably different, the substitution profile for every residue was resampled 10,000 times based on the VAMP-seq score averages and standard deviations (see Methods), and $RMSD_{exposed}$ and $RMSD_{buried}$ values were calculated for all of the resampled profiles. For all interface residues analysed here, the standard error of the mean of the resampled RMSD values is too small to be illustrated with error bars. It is also the case that $RMSD_{buried} < RMSD_{exposed}$ or $RMSD_{exposed} < RMSD_{buried}$ for more than 95% of the resampled substitution profiles for all interface residues. (C) NUDT15 dimer with one monomer represented by cartoon of backbone and one monomer shown in transparent surface representation. Side chains are shown for all interface residues, and interface residues with $RMSD_{buried}$ smaller than $RMSD_{exposed}$ are shown in orange, while interface residues with $RMSD_{exposed}$ smaller than $RMSD_{buried}$ are shown in green. (D, E) ASPA dimer shown from two different angles represented similarly to the NUDT15 dimer.

chains of the two residues do point into the interface in the static structure, our analysis suggests that they are not more important for the cellular stability of NUDT15 than if they had been sitting in solvent-exposed positions. We note that this result is more clear for residue L153 than residue L156.

Using the same method, we found that 10 out of 14 interface residues in the ASPA dimer have a smaller $RMSD_{buried}$ than $RMSD_{exposed}$ (Fig. 5B). The majority of the ASPA dimer interface residues thus appear structurally buried, and we conclude that ASPA exists as a dimer *in vivo* and that dimerisation is important for maintaining the cellular stability of the protein. In our analysis, we found crystal structure interface residues L25, F29, C61 and L265 to have abundance scores most similar to those of solvent-exposed residues of the same amino acid types. The four residues do not appear to be as critical for ASPA dimerisation and cellular stability as the rest of the interface residues. In the crystal structure, L265 is positioned in a loop at the border of the interface (Fig. 5D), while L25, F29 and C61 are spatially close to each other at a site distant from L265 (Fig. 5E). We note that for ASPA, the $RMSD_{buried}$ for several residues is relatively high although still smaller than $RMSD_{exposed}$, in accordance with the fact that the ASPA VAMP-seq dataset contains many variants with very low abundance.

Discovering solvent-exposed residues important for cellular stability

Having analysed residues found in the monomer-monomer interfaces of the NUDT15 and ASPA structures, we next tested whether our method was able to identify other solvent-exposed residues with $RMSD_{buried} < RMSD_{exposed}$ on the non-interface surfaces of NUDT15 and ASPA as well as on the surfaces of the four other proteins in our dataset. At the same time, we tested whether any buried residues have $RMSD_{exposed} < RMSD_{buried}$. For the analysis, we used average abundance score matrices for buried and exposed residues calculated using only monomer structures and excluding data from the protein being analysed. For every residue in each of the six proteins, we calculated $RMSD_{exposed}$ and $RMSD_{buried}$ using all reported variant abundance scores for the residue (Fig. S18), as described above, allowing us to discover residues for which the structural environment is not necessarily a good indicator of the impact of substitutions on abundance.

Generally, we find many residues to behave as expected, although we identify surface-exposed residues with $RMSD_{buried} < RMSD_{exposed}$ and buried residues with $RMSD_{exposed} < RMSD_{buried}$ in all proteins. The number of residues for which the abundance score substitution profile does not match the average profile for residues in a similar structural environment depends on the protein. We for instance find 46 out of 351 (13%) exposed residues in PRKN to have $RMSD_{buried} < RMSD_{exposed}$ and 15 out of 114 (13%) buried residues in PRKN to have $RMSD_{exposed} < RMSD_{buried}$. In the ASPA monomer, we on the other hand find 86 out of 183 (47%) exposed residues with $RMSD_{buried} < RMSD_{exposed}$ and 17 out of 130 (13%) buried residues with $RMSD_{exposed} < RMSD_{buried}$. The accuracy with which environment indicates mutational tolerance is highest for PRKN and PTEN and lowest for ASPA (Fig. S18). Considering the number of non-interface, surface-exposed residues that we find to have $RMSD_{buried} < RMSD_{exposed}$ in NUDT15 and ASPA (Fig. S18, S19), we cannot exclude that the crystal structure interface residues might be important for maintaining cellular stability in ways that are not just related to the formation of a homodimer.

Given the fact that the six abundance score distributions have considerably different shapes (Fig. S1), some of the identified discrepancies between residue environments and abundance score profiles are perhaps not surprising. For example, since the PTEN abundance score distribution is shifted towards high abundance score values, it makes sense that we identify many residues in PTEN that we structurally classify as buried, but which have a smaller $RMSD_{exposed}$ than $RMSD_{buried}$ value. The opposite might be true for solvent-exposed residues in for example ASPA and PRKN that appear to have abundance score substitution profiles most similar to the average profile of buried residues. As discussed above, a number of factors, including both biochemical and technical, likely contribute to producing the different score distributions, and a combination of these factors will likely also give rise to some of the identified mismatches between structural

environment and substitution profiles. In spite of this, we have shown that our method is able to identify surface residues for which the mismatches likely are indicators of residues with functions (in this case dimerisation) important for abundance.

In line with this, we focus the rest of our analysis specifically on solvent-exposed residues for which the $RMSD_{buried}$ is smaller than $RMSD_{exposed}$, thereby generating hypotheses regarding which exposed residues might be required for maintaining cellular abundance of the proteins. We mapped this group of residues onto the protein structures (Fig. S19) and indeed found that our method works not only for analysis of residues likely involved in homodimerisation, but also for revealing exposed residues that through different types of biochemistry control abundance. In PRKN, solvent-exposed residues with $RMSD_{buried} < RMSD_{exposed}$ for example include many of the Zn^{2+}-coordinating cysteine and histidine residues that are critical for maintaining abundance (Clausen et al., 2024). Our method also identifies that the substitution profile of the solvent-exposed S109 in PRKN is more similar to that of buried than exposed serines, which is in line with the fact that certain mutations of this residue are thought to introduce a solvent-exposed quality control degron in the protein (Clausen et al., 2024). Finally, the method appears to be able to capture that certain phosphorylation sites are important for maintaining abundance. Specifically, it has been shown that residues S380, T382, T383 and S385 of PTEN are phosphorylated in the cell and that mutation of the first three residues to alanine increase the cellular degradation rate of the protein (Vazquez et al., 2000). With our method, we identify one of these sites, S385, to have $RMSD_{buried} < RMSD_{exposed}$ while being exposed in the static structure. The importance of residue S385 has also been discussed in a previous analysis of the VAMP-seq data (Matreyek et al., 2018). While the substitution profile of residue T383 does at first sight appear to have $RMSD_{buried} < RMSD_{exposed}$, the abundance scores of residue T383 are too noisy to state this with high confidence (see Methods). Since residues S380 and T382 both have less than five reported abundance scores, we did not assess whether our method picks up on these sites as being important for maintaining abundance or not.

Above, we have thus demonstrated that the $RMSD_{buried}$ and $RMSD_{exposed}$ values of a residue, which respectively measure the similarity between the substitution profile of the residue and the average substitution profiles of buried and exposed residues, can be used to (*i*) tell apart experimentally relevant from irrelevant structures and (*ii*) reveal residue sites with importance for maintaining cellular protein stability, thereby generating hypotheses for further studies. Our results highlight the usefulness of the exposure-based substitution matrices beyond abundance score predictions. The approach introduced here is related to other methods that similarly use discrepancies between structure data and single residue substitution mutagenesis data to distinguish correct or relevant structures from decoy or wrongly predicted structures (Adkar et al., 2012; Chiasson et al., 2020; Zutz et al., 2021), and to methods that aim to derive three-dimensional structures using data generated by MAVEs (Schmiedel and Lehner, 2019; Rollins et al., 2019; Drake et al., 2024). Finally, the identified discrepancies between individual and average substitution patterns of solvent-exposed residues point to several aspects of cellular biochemistry that our simple, structure-based abundance framework does not capture, including aspects of post-translational modification and interactions with the cellular degradation machinery after introduction of neodegrons. The highlighted discrepancies might thus be used to guide future abundance model development, for example via introduction of more complex biochemical input features.

Conclusions

Several previous studies have combined heterogeneous mutagenesis data from multiple MAVEs to analyse amino acid substitution effects across proteins in a general manner. In this study, we have similarly analysed a combination of six large mutagenesis datasets, however restricting our analysis to a more homogenous set of MAVE data that were all collected with the VAMP-seq method and report on the impact of substitutions on the cellular abundance of proteins. Our results therefore reveal trends in substitution effects for a single molecular phenotype. We have

analysed in total 31,614 single residue substitution variant abundance scores in a structure-based fashion through construction of amino acid substitution matrices reporting on the average abundance scores for all possible substitution types.

Importantly, we found that, by using simple structural considerations, it is possible to construct structure context-specific substitution matrices that with relatively high accuracy, at least considering the model simplicity, can predict the impact of residue substitutions on the cellular abundance of proteins. Our substitution matrices can, in fact, rank variant abundance as well as much more complex biophysics- and deep learning-based methods, highlighting the usefulness of simple, interpretable structure features for stability-related variant effect prediction.

As shown in this work and in several previous studies, the cellular abundance of variants correlates with the variant impact on protein folding stability, or ΔΔG, explaining in part why simple structural features capture abundance-related variant effects relatively well. We compared abundance score predictions from our substitution matrices to ΔΔG predictions from state-of-the-art stability change models, and while the two prediction types are correlated, we found that they also contain orthogonal information, suggesting that the substitution matrices derived from VAMP-seq data contain abundance-specific signal.

The VAMP-seq-derived substitution matrices can also be used to discriminate between experimentally, in this case cellularly, relevant and irrelevant protein structures. We have specifically shown that two of the six proteins in our combined VAMP-seq dataset, namely NUDT15 and ASPA, likely form homodimers that stabilise the proteins in cells. Using a similar method, the matrices can also point out outlier residues that have unexpected substitution profiles given their structural environment. For example, solvent-exposed residues for which variants increase degradation propensity may be captured. We propose that our substitution matrix-based approach is applicable not only for selecting between monomer and dimer structures as done here, but that it might also be useful in a broader context, for example for selecting between correctly and incorrectly predicted structures.

The average substitution effects described by our substitution matrices generally agree with biochemical expectations. In clustering analyses of the average substitution profiles of different amino acid types, amino acids tend to group together according to physicochemical properties, although with noteworthy exceptions, such as for example the extreme substitutional intolerance of aspartate in buried environments. Moreover, while structural considerations can help rationalise certain variant effects, like those of loop residues with left-handed helix-like backbone conformations, this is not always the case, as revealed by the lack of correlation between secondary structure propensity scales and average abundance scores of residues in *α*-helices.

Given the general consistency between the VAMP-seq-derived substitution matrices and biochemical expectation, we believe that the matrices may generally serve as useful references for how substitutions will likely affect abundance. While the combined analysis of the six individual VAMP-seq datasets proved useful in this work, future work developing for example prediction methods for variant abundance might benefit from more elaborate considerations regarding and treatment of the experimental and biochemical factors that influence the absolute values of the VAMP-seq abundance scores (Schulze et al., 2025).

Methods

VAMP-seq data collection and preparation

We obtained VAMP-seq abundance scores and abundance score standard deviations directly from the original VAMP-seq publications (Matreyek et al., 2018, 2021; Amorosi et al., 2021; Suiter et al., 2020; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024). In these studies, the abundance scores were calculated in a similar manner so that they should be more directly comparable; first, the abundance of each variant was expressed as a weighted average W_{var} over the frequencies with which the variant was found in each of the four FACS-sorted bins. W_{var} scores for all variants in an experiment were then min-max normalised to values respectively

representing the weighted average of nonsense variants and the weighted average of synonymous variants, so that the final abundance score of a variant would be $score_{var} = (W_{var} - W_{nonsense})/(W_{wt} - W_{nonsense})$ (Matreyek et al., 2018).

We filtered the six datasets to include only single residue substitution scores in our analysis, that is, we excluded data for synonymous and nonsense variants. We did not perform any additional quality filtering of the data beyond filtering already performed in the original publications, meaning that scores from positions with low mutational depth are included in our analysis. Moreover, since abundance scores were already normalised in the original VAMP-seq work, we used reported scores without further normalisation. PTEN VAMP-seq data has been measured multiple times to increase the mutational completeness of the variant abundance dataset (Matreyek et al., 2018, 2021). In our analysis, we used the most recently published PTEN dataset, which is based on a combination of abundance scores measured in the original and updated PTEN work (Matreyek et al., 2021). Finally, we excluded variant abundance scores for the first 28 residues of CYP2C9 from all analyses, since the CYP2C9 N-terminal is a transmembrane helix (Amorosi et al., 2021).

Quantification of dataset noise levels

We quantified the noise levels in the individual datasets using a bootstrapping approach in which we resampled each VAMP-seq dataset 100 times. Each abundance score was resampled from a Gaussian distribution with mean equal to the originally reported abundance score and standard deviation equal to the reported standard deviation of the score. We calculated r (and r_s) between the originally reported abundance scores and each resampled set of scores and reported the average of the resulting 100 r (and r_s) values in Figure 1E. We note that abundance score standard deviations are based on a different number of replicate measurements, both within and across datasets, and that our bootstrapped r values do not in all cases correspond to the correlations between replicate experiments reported in the original VAMP-seq publications, with deviations in particular for TPMT and CYP2C9, where our quantification of the noise levels gives higher numbers for r.

Structure preparation and visualisation

We used combined crystal and AlphaFold2 structures (Jumper et al., 2021) in our analysis so that our structural input preserved crystal structure coordinates where possible and at the same time had no missing residues. We created the combined structures by first aligning crystal structures and AlphaFold2 structures using the Matchmaker tool in UCSF ChimeraX (v1.3, Pettersen et al. (2021)) and then filled out gaps from missing residues in the crystal structure files with coordinates from the AlphaFold2 structures. We then ran the "alphafold2_ptm" model in ColabFold (v1.5.2, Mirdita et al. (2022)) with the combined structures as input template structures using UniProt sequences as query sequences (Table S1). We specified to not use a multiple sequence alignment for the structure prediction and otherwise used default setting, including omission of Amber relaxation. We verified that the first ranked ColabFold output structures mostly preserved the crystal structure backbone conformations and then used these structures as inputs for all of our analyses. We took this approach to include as much structural information as possible in our modelling while at the same time making sure that our input structures resembled the crystal structures in cases where crystal and AlphaFold2 structures deviated with respect to backbone conformation.

The AlphaFold2 structures that we used as input for this approach were predicted with the AlphaFold Monomer v2.0 pipeline and downloaded from the AlphaFold Protein Structure Database (Varadi et al., 2022), except for in the cases of NUDT15 and ASPA, since these proteins can be found only in their monomeric form in the database. We created the homodimeric assemblies of the proteins with the multimer pipeline "alphafold2_multimer_v3" in ColabFold using default settings, including template mode "pdb70" and MSA mode "mmseqs2_uniref_env". The top ranked dimer structures generated with this approach were then used as input AlphaFold2 structures for the procedure described above.

We processed structure data files with pdb-tools (v2.4.3, Rodrigues et al. (2018)). Our structure processing included removal of chains not noted in Table S1 and removal of all HETATM entries in the crystal structure files. All our analyses were thus performed without accounting for cofactors or ions bound to the proteins in their crystal structures. We also removed the 28 N-terminal residues of the CYP2C9 structure, as the N-terminal of CYP2C9 is known to form a transmembrane helix (Amorosi et al., 2021). In calculations that rely on the monomeric structures of NUDT15 and ASPA, we always used chain A in its conformation in the dimeric structure. All protein structure visualisations were created with ChimeraX.

Calculation of structural features

We used DSSP (Kabsch and Sander, 1983) for secondary structure assignment and calculation of solvent accessible surface area for every residue in the wild-type protein structures. Solvent accessible surface area was normalised to obtain relative solvent accessibility using a theoretically derived maximum accessibility per residue (Tien et al., 2013). As in other studies (Hovmöller et al., 2002), we used secondary structure assignments from DSSP to classify residues into three secondary structure categories, specifically helix (*a*-, 3_{10}- and *n*-helix), strand (extended stand and isolated */3*-bridge) and loop (hydrogen-bonded turn, bend and irregular/loop). These definitions were used throughout our work, except in the analysis of abundance score correlations with helix propensity scales (Fig. S17), where we only included data for *a*-helical residues.

We also evaluated contact formation between residues by calculating a WCN for every residue in the wild-type structure environment. Specifically, the WCN for every residue i was calculated as a function of the distances $r_{i,j}$ to all other residues in the protein structure according to the equation

$$\mathrm{WCN}_i = \sum_{j \neq i} s(r_{i,j}) \quad \text{with} \quad s(r) = \frac{1 - \left(\frac{r}{r_0}\right)^6}{1 - \left(\frac{r}{r_0}\right)^{12}}, \tag{1}$$

in which we set r_0 = 7 Å. We defined the distance between two residues to be the shortest distance between any pair of heavy atoms in the side chains, except if one residue involved was glycine, in which case we used the shortest distance between any interresidue atom pair. Distances were calculated with the MDTraj (v1.9.7, (McGibbon et al., 2015)) function compute_contacts. We calculated WCN according to Eq. 1, as we have previously found contact numbers calculated in this manner to correlate with variant abundance (Cagiada et al., 2023; Grønbæk-Thygesen et al., 2024; Clausen et al., 2024; Gersing et al., 2024).

Definitions of buried and exposed residues

We classified residues with an rASA smaller than or equal to 0.1 as structurally buried and residues with an rASA larger than 0.1 as solvent-exposed. In other words, we defined the residue class (RC) as a function of rASA as follows:

$$\mathrm{RC(rASA)} = \begin{cases} \text{buried} & \text{if rASA} \leq 0.1 \\ \text{exposed} & \text{if rASA} > 0.1 \end{cases} \tag{2}$$

We assigned a residue class to all residues in the six proteins according to this definition, and to calculate, for example, the “buried” substitution matrix, we averaged VAMP-seq scores reported for all residues with RC(rASA) = buried. The definition of buried and exposed residues in Eq. 2 is used throughout the manuscript, except in the analysis we performed to choose the 0.1 cutoff value.

The rASA = 0.1 cutoff value was selected using a grid search in which we scanned combinations of rASA and WCN values to maximise the average correlation between experimental abundance scores and abundance scores predicted with the exposure-based substitution matrices. More specifically, we calculated the substitution matrices for buried and exposed residues using a range of rASA and WCN cutoff combinations to classify residues as buried or exposed. rASA cutoff values ($_{\mathrm{rASA}}$) from 0 to 0.5 were scanned together with WCN cutoff values ($_{\mathrm{WCN}}$) from 0 to 20. For this

scan, we defined buried residues as residues with a WCN larger than or equal to the WCN cutoff value *and* an rASA smaller than or equal to the rASA cutoff value. Residues not classified as buried were classified as exposed, meaning that exposed residues would have a WCN smaller than the WCN cutoff or an rASA larger than the rASA cutoff. The residue class (RC) definition as a function of both rASA and WCN was thus as follows:

$$\mathrm{RC}(\mathrm{rASA},\mathrm{WCN}) = \begin{cases} \text{buried} & \text{if } \mathrm{rASA} \leq c_{\mathrm{rASA}} \text{ and } \mathrm{WCN} \geq c_{\mathrm{WCN}} \\ \text{exposed} & \text{if } \mathrm{rASA} > c_{\mathrm{rASA}} \text{ or } \mathrm{WCN} < c_{\mathrm{WCN}} \end{cases} \quad (3)$$

For each combination of rASA and WCN cutoff values, we tested the predictive power of the substitution matrices using leave-one-protein-out cross-validation, that is, we calculated the matrices for each cutoff combination leaving out of one the six proteins and then attempted to predict the abundance scores for the omitted protein. We calculated *r* and the mean absolute error (MAE) between experimental and predicted abundance scores for each protein and then averaged results across proteins. We found that, on average, an rASA cutoff of 0.1 maximises *r* and minimises the MAE when used in combination with WCN cutoff values of 0, 5 and 10, however with no considerable difference between the three WCN cutoffs (Fig. S20A,B). A WCN cutoff value of 0 effectively corresponds to applying only the rASA cutoff. For simplicity, we therefore used an rASA cutoff of 0.1 (Eq. 2) and did not consider WCN for exposure-based residue classifications in any other analyses in this work.

Definition of substitution matrices

To calculate the substitution matrix entry $\bar{s}_{ij}$ for substitutions from amino acid type *i* to amino acid type *j*, we averaged all individual VAMP-seq scores s_{ij} for this substitution type, such that

$$\bar{s}_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} s_{ij,k}, \quad (4)$$

where N_{ij} is the total number of VAMP-seq scores for substitutions from residue type *i* to residue type *j*. To calculate $\bar{s}_{ij}$ for the Global substitution matrix, we simply used all s_{ij} values available in our combined VAMP-seq dataset. Residue environment-specific substitution matrices were calculated using s_{ij} values only for variants of residues sitting in specific environments in the wild-type protein structures. For example, entries in the two substitution matrices that depend on residue burial (Fig. 2) were calculated according to

$$\bar{s}_{ij}(\mathrm{rASA}) = \begin{cases} \bar{s}_{ij,\text{buried}} & \text{if } \mathrm{rASA} \leq 0.1 \\ \bar{s}_{ij,\text{exposed}} & \text{if } \mathrm{rASA} > 0.1 \end{cases}, \quad (5)$$

where $\bar{s}_{ij,\text{buried}}$ and $\bar{s}_{ij,\text{exposed}}$ were calculated with Eq. 4 using VAMP-seq scores s_{ij} for variants of residues that respectively sit in either buried (rASA 0.1) or exposed (rASA > 0.1) environments.

Calculation of ΔΔG values with Rosetta, RaSP and ThermoMPNN

We calculated the difference in Gibbs folding free energy between wild-type proteins and all possible single residue substitution variants ($\Delta\Delta G = \Delta G_{\text{variant}} - \Delta G_{\text{wild type}}$) with Rosetta using the cartesian-_ddg protocol and *opt-nov15* energy function (Park et al., 2016). We used an in-house pipeline (*https://github.com/KULL-Centre/PRISM/tree/main/software/rosetta_ddG_pipeline*, v0.2.1) to first prepare and relax input structures and then calculate energies, which we converted from Rosetta energy units (REU) to kcal/mol with the conversion factor 2.9 REU/(kcal/mol) (Park et al., 2016). For evaluation of the effect of a given substitution in dimeric structures, we made the residue substitution in both monomers simultaneously and obtained a ΔΔG for the dimer, which we divided by 2 to get the stability change per residue.

We also calculated variant ΔΔG values using the deep learning-based RaSP (Blaabjerg et al., 2023) and ThermoMPNN (Dieckhaus et al., 2024) models, both of which estimate single residue substitution ΔΔG values from static protein structures. We used the six wild-type protein structures as inputs to the models, representing NUDT15 and ASPA with their homodimer structures. Since ThermoMPNN was trained to predict monomer stability, we compared the rank correlations between abundance scores and ΔΔG values predicted using either monomer or homodimer structures of NUDT15 and ASPA as input. We found r_s to decrease slightly (smaller is better) from −0.62 to −0.65 when ΔΔG values were predicted with the NUDT15 homodimer rather than the monomer structure. When only variants of NUDT15 interface residues were considered, r_s between ΔΔG values and abundance scores decreased from −0.23 for monomer-based predictions to −0.36 for dimer-based predictions. ΔΔG values calculated with ASPA monomer and homodimer structures correlated equally well (r_s = −0.53) with abundance scores when all variants of ASPA were taken into account. Interface variants of ASPA had r_s values of −0.25 and −0.21 for monomer- and dimer-based evaluations, respectively. Given these results and for consistency with the rest of our analysis, we used the dimer-based ΔΔG predictions for NUDT15 and ASPA from ThermoMPNN through-out our work.

Methods to compare pairs of substitution profiles

We compared how well ThermoMPNN ΔΔG scores and our exposure-based substitution matrices ranked variant effects at a residue-level in the following way. For every individual residue with VAMP-seq scores reported for at least five variants, we calculated r_s between experimental abundance scores and predicted scores in form of either ThermoMPNN ΔΔGs or substitution matrix values, thus obtaining $r_{s,\Delta\Delta G}$ and $r_{s,\text{matrix}}$, respectively. If $-r_{s,\Delta\Delta G}$ and $r_{s,\text{matrix}}$ for a given residue were both positive numbers, we proceeded to resample the experimental residue substitution profile 10,000 times, randomly drawing new abundance scores from variant-specific Gaussian distributions with μ equal to the experimental VAMP-seq score of a variant and σ equal to the VAMP-seq score standard deviation of the same variant, assuming independent errors across variants of a single residue. For every residue, we recalculated $r_{s,\Delta\Delta G}$ and $r_{s,\text{matrix}}$ using the 10,000 resampled substitution profiles and then checked how often $-r_{s,\Delta\Delta G} > r_{s,\text{matrix}}$ or $-r_{s,\Delta\Delta G} < r_{s,\text{matrix}}$. Residues for which at least 95% of resampled substitution profiles satisfied $-r_{s,\Delta\Delta G} > r_{s,\text{matrix}}$ were classified as ΔΔG-favoured. On the other hand, we classified residues as matrix-favoured if $-r_{s,\Delta\Delta G} < r_{s,\text{matrix}}$ for more than 95% of resampled profiles. For the remaining residues, we report that the two predictors capture variant effects on a residue-level equally well. We only classify residues with at least five variant effect scores and for which $-r_{s,\Delta\Delta G}$ and $r_{s,\text{matrix}}$ are both positive, and the sum of residues in the the three categories (ΔΔG-favoured, matrix-favoured and "models similar") is thus smaller than the total number of residues in the six proteins.

We compared the experimental substitution profile of every individual residue to the average profiles of buried and exposed residues in a similar fashion. For every residue, we resampled the residue's substitution profile 10,000 times, drawing new abundance scores from variant-specific Gaussian distributions as described above. We then used the resampled substitution profiles to calculate 10,000 residue-specific $\text{RMSD}_{\text{exposed}}$ and $\text{RMSD}_{\text{buried}}$ values. Only when either $\text{RMSD}_{\text{exposed}} < \text{RMSD}_{\text{buried}}$ or $\text{RMSD}_{\text{exposed}} > \text{RMSD}_{\text{buried}}$ for more than 95% of the resampled substitution profiles did we consider a given residue to have an exposed-like or a buried-like substitution behaviour. We only performed this analysis for residues with substitution profile consisting of at least 5 abundance scores. In the text and figures, we only report results for residues that live up to our 95% criterium, unless otherwise noted, such as in the case of residue T383 in PTEN.

Hierarchical clustering and principal component analysis

Hierarchical clustering of substitution matrix rows and columns was performed and visualised using the seaborn function clustermap. We set function parameters method='average' and metric='euclidean', meaning that distances between clusters were evaluated by calculating the

average distance between all pairs of elements in the clusters using the Euclidean distance as distance metric. Prior to performing the clustering, we set the average abundance scores for synonymous substitutions equal to one.

We performed PCA of abundance scores averages for residues in buried and exposed environments by constructing, for each environment type, a 20×40 matrix that for each of the 20 amino acid types specified the 40 abundance score averages for all substitutions to and from that amino acid residue type. In the reformatted matrices, we set scores for synonymous substitutions equal to one. We then used the scikit-learn class sklearn.decomposition. PCA with default parameters to perform the PCA. The input features were not scaled prior to the analysis.

Analysis of substitution matrices for helix- and loop-forming residues

We tested the correlation between average abundance score matrices for residues forming *a*-helix using a number of different helix propensity scales (Pace and Scholtz, 1998; Rohl et al., 1996; Muñoz and Serrano, 1995; Chou and Fasman, 1978; Williams et al., 1987; Luque et al., 1996). The scales all report on the residue ΔΔG relative to alanine for helix formation. All numbers for our analysis were taken from a single paper which compared six different helix propensity scales (Pace and Scholtz, 1998), and we have here followed the naming of the scales presented in that paper. To test the correlation with our average abundance score matrices, we calculated a helix propensity substitution matrix for each helix propensity scale by subtracting the ΔΔG of the variant residue from the ΔΔG of the reference residue. We then calculated the correlation coefficient between the helix propensity and the average abundance score substitution matrices. For this analysis, we used average abundance score matrices constructed with abundance scores only for residues forming *a*-helix, unlike in all our other analyses in this paper, where our helix matrices includes data from residues forming *a*-helix as well as 3_{10}- and *n*-helix

To analyse the substitution patterns of loop residues adopting backbone conformations similar to those found in left-handed helices, we used the MDTraj functions compute_phi and compute_psi to identify all loop residues for which the backbone Φ and ψ angles simultaneously fell within the ranges $0° < \Phi < 180°$ and $-90° < \psi < 90°$ (Hovmöller et al., 2002). We excluded residues that were not present in the crystal structures of the proteins from this analysis to make sure that wrongly predicted loop residue conformations from AlphaFold2 did not influence the results.

Analysis of solvent-exposed residues important for cellular stability

For our analysis of NUDT15 and ASPA homodimerisation, we defined homodimer interface residues as residues that were classified as exposed in the monomer structures and as buried in the dimer structures. To filter out residues which experience minor changes in rASA upon dimerisation, but still change category with the strict 0.1 rASA cutoff used to distinguish between buried and exposed residues, we only included residues in the interface residue category if the change in rASA was more than 0.01 when going from monomer to dimer. The results are not particularly sensitive to the exact cutoff value of 0.01, which was selected based on visual inspection of monomer and dimer structures.

For the analysis of the homodimer interfaces as well as for the analysis of the substitution profiles for all remaining residues in the six proteins, we calculated $RMSD_{buried}$ and $RMSD_{exposed}$ values only for residue positions for which at least five abundance scores were reported in the experimental datasets. In our results, we thus show no $RMSD_{buried}$ and $RMSD_{exposed}$ data for any residues with fewer than five abundance scores. We note that all NUDT15 interface residues have at least five abundance scores and that all ASPA interface residues except one (residue 242) have at least five abundance scores.

Protein name	Abbreviation	UniProt ID	PDB ID	PDB chain	Residues
Phosphatase and Tensin Homolog	PTEN	P60484-1	1d5r	A	1-403
Thiopurine S-methyltransferase	TPMT	P51580	2h11	A	1-245
Cytochrome P450 2C9	CYP2C9	P11712-1	1r9o	A	29-490
Nucleotide triphosphate diphosphatase	NUDT15	Q9NV35	5bon	A+B	1-164
Aspartoacylase	ASPA	P45381	2o53	A+B	1-313
E3 ubiquitin-protein ligase parkin	PRKN	O60260-1	5c1z	A	1-465

Table S1. Overview of structure and sequence data used for analysis. Protein sequences are given by their UniProt identifiers and crystal structure input files (Lee et al., 1999; Wu et al., 2007; Wester et al., 2004; Carter et al., 2015; Le Coq et al., 2008; Kumar et al., 2015) by their PDB identifiers. We used the PDB file chains and residues noted in the last two columns.

Figure S1. VAMP-seq abundance score distributions shown separately for each of the six datasets included in our analysis.

The datasets only include scores for single residue substitution variants at residue sites in soluble protein regions.

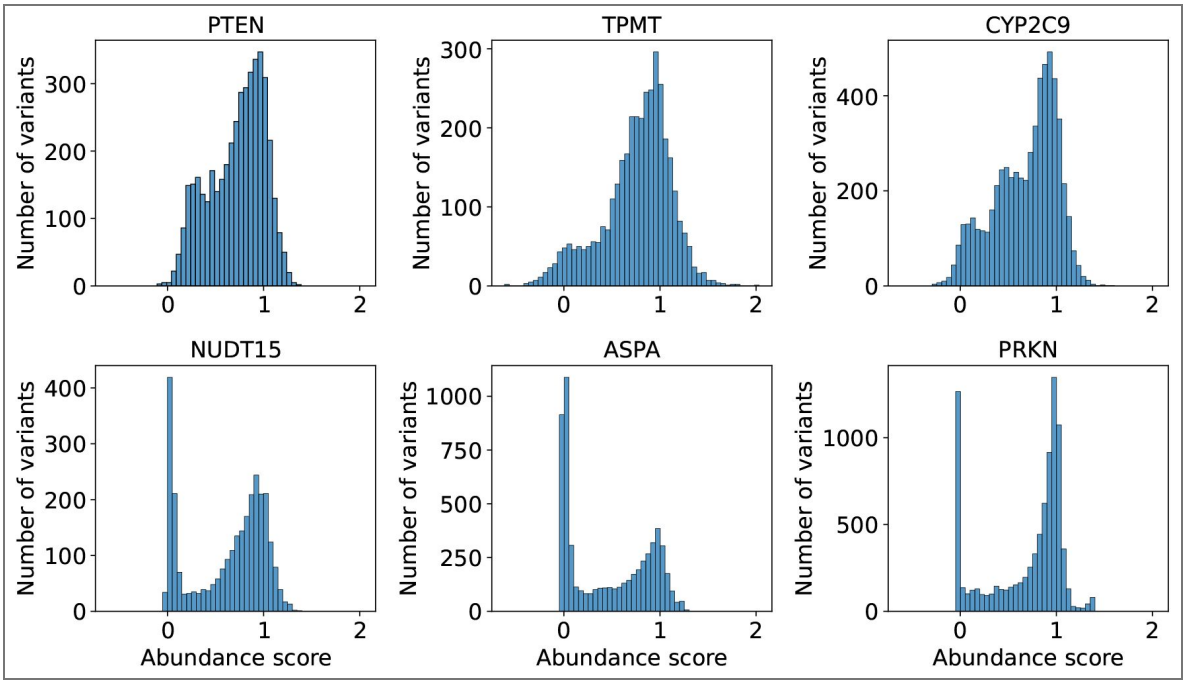
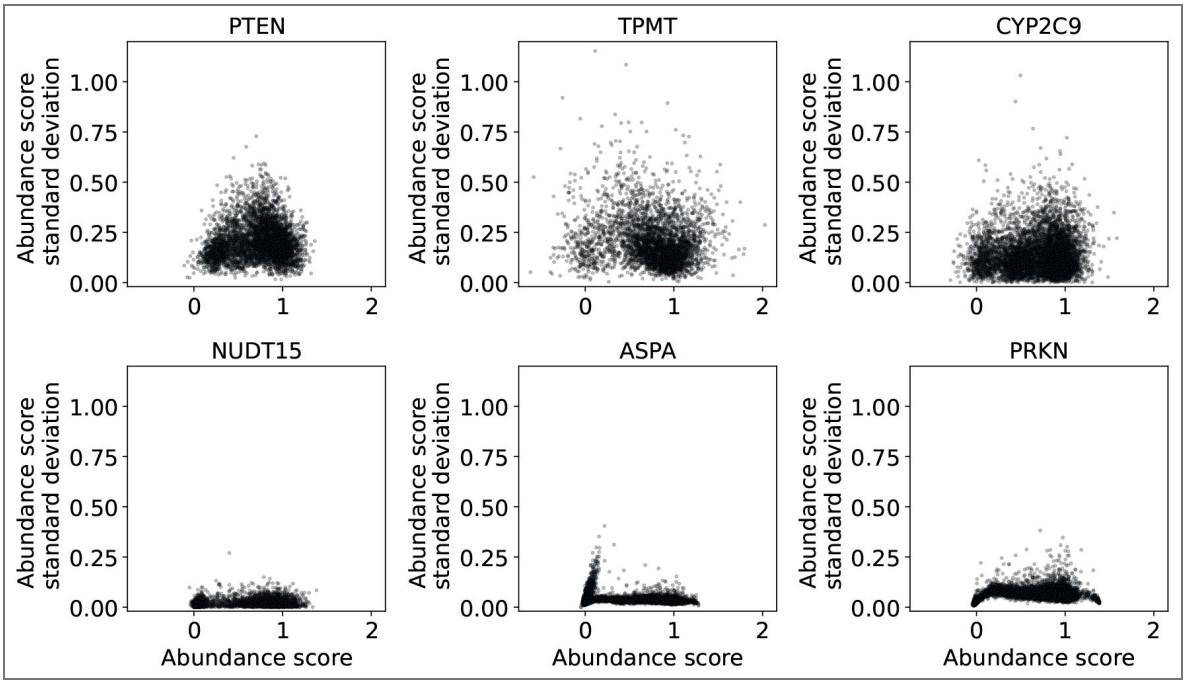


Figure S2. VAMP-seq abundance score standard deviations plotted as function of abundance scores separately for each VAMP-seq dataset.

The standard deviations were calculated using a varying number of experimental replicates in the original VAMP-seq publications.



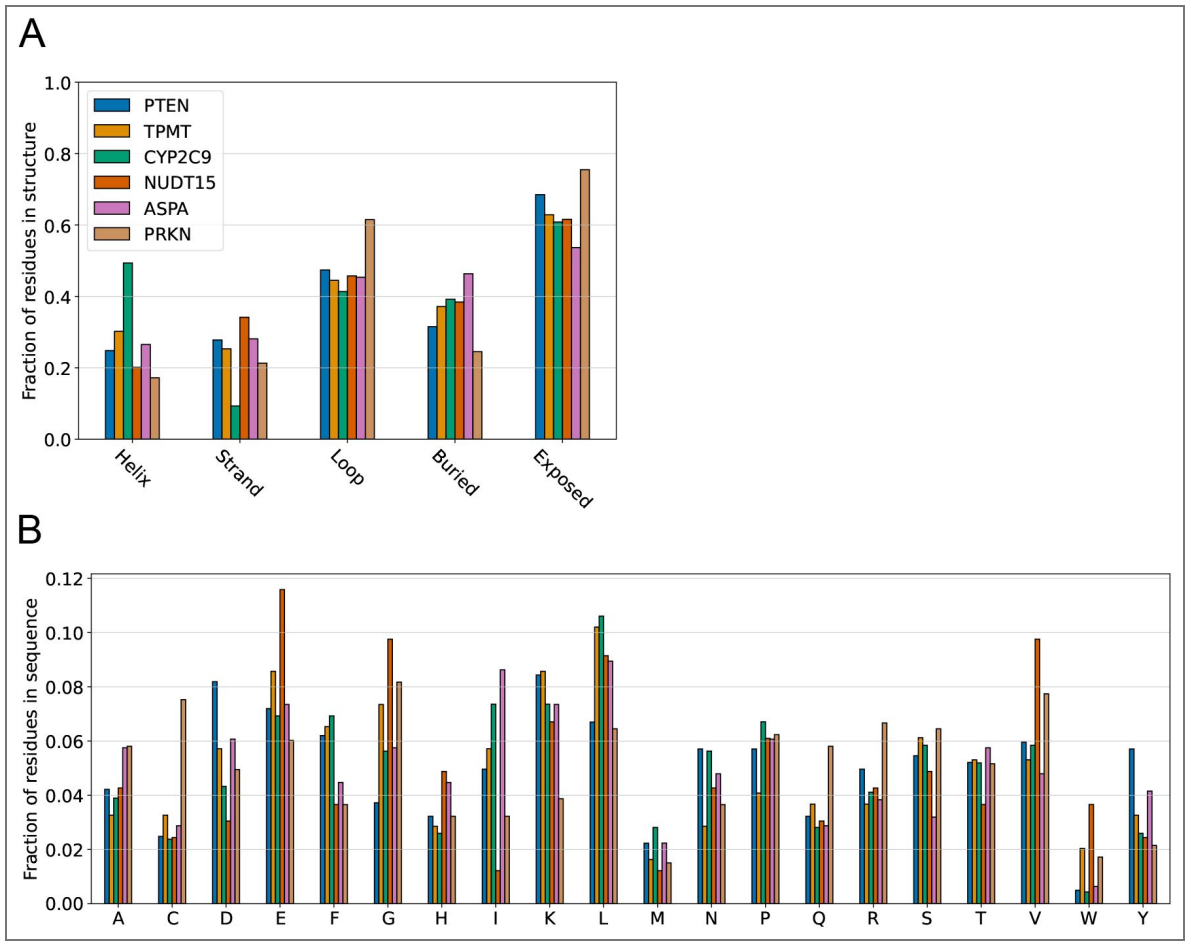


Figure S3. Summary of protein structure and sequence compositions.

(A) The per protein fractions of residues forming different types of secondary structure are shown alongside with the fractions of residues found at structurally buried and solvent-exposed sites. For NUDT15 and ASPA, fractions were calculated using homodimer structures. The six proteins have relatively similar structural compositions, however with exceptions such as the relative helical enrichment in CYP2C9 and the loop and thereby also exposed site enrichment in PRKN. (B) Protein sequence compositions are similarly shown as residue type fractions with respect to the total number of residues in each protein.

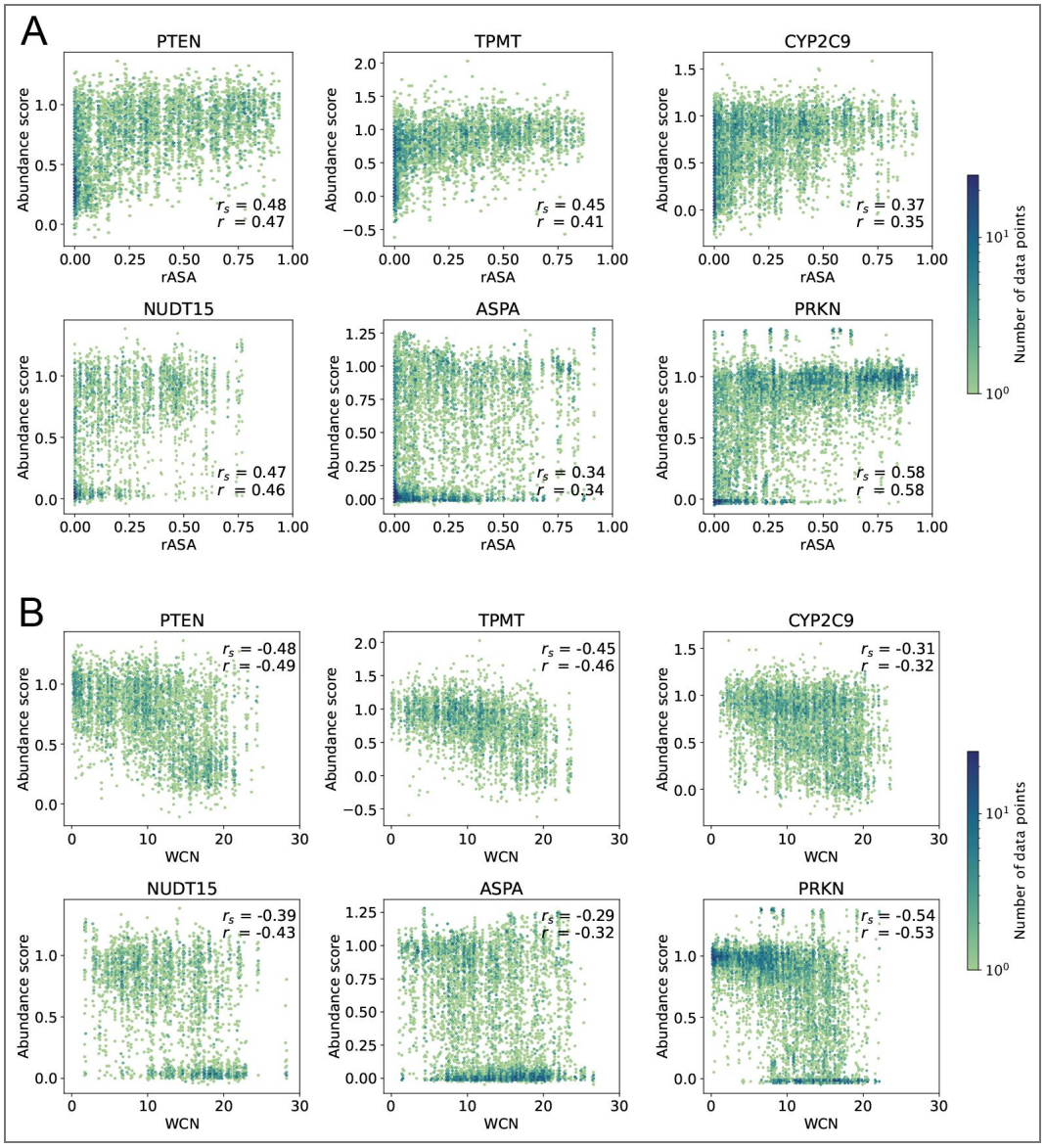


Figure S4. rASA and WCN correlate with single residue substitution abundance scores.

Abundance scores for all single residue substitution variants are plotted against residue (A) rASA and (B) WCN values calculated using the wild-type protein structures. Pearson's correlation coefficient r and Spearman's rank correlation coefficient r_s between abundance scores and rASA or WCN values were calculated for each protein and are shown in the indvidual plots.

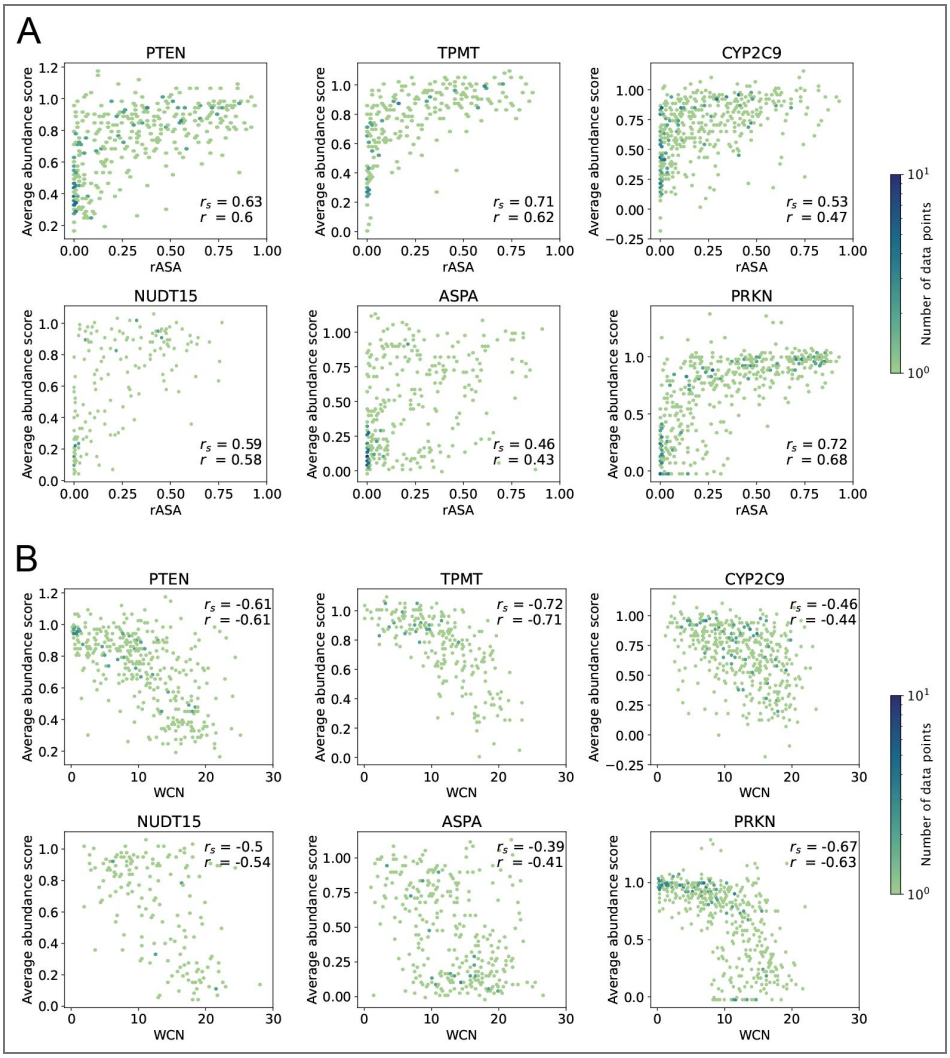


Figure S5. rASA and WCN correlate with residue-averaged abundance scores.

Residue-averaged abundance scores are plotted against residue (A) rASA and (B) WCN values calculated using the wild-type protein structures. Pearson's correlation coefficient r and Spearman's rank correlation coefficient r_s between residue-averaged abundance scores and rASA or WCN were calculated for each protein and are shown in the individual plots. Averages are shown for all residues with abundance scores available, even if only one or few abundance scores have been reported for a given residue.

Figure S6. Average abundance score substitution matrices calculated using the six VAMP-seq datasets.

The average score matrices were calculated by averaging over all scores (Global), scores for residues forming helix (Helix), scores for residues forming strand (Strand) and scores for residues forming loop structure (Loop), with structural context evaluated on wild-type structures. In each panel, the bar plot to the left of the average score matrix shows the average number of data points used to calculate the averages in a matrix row.

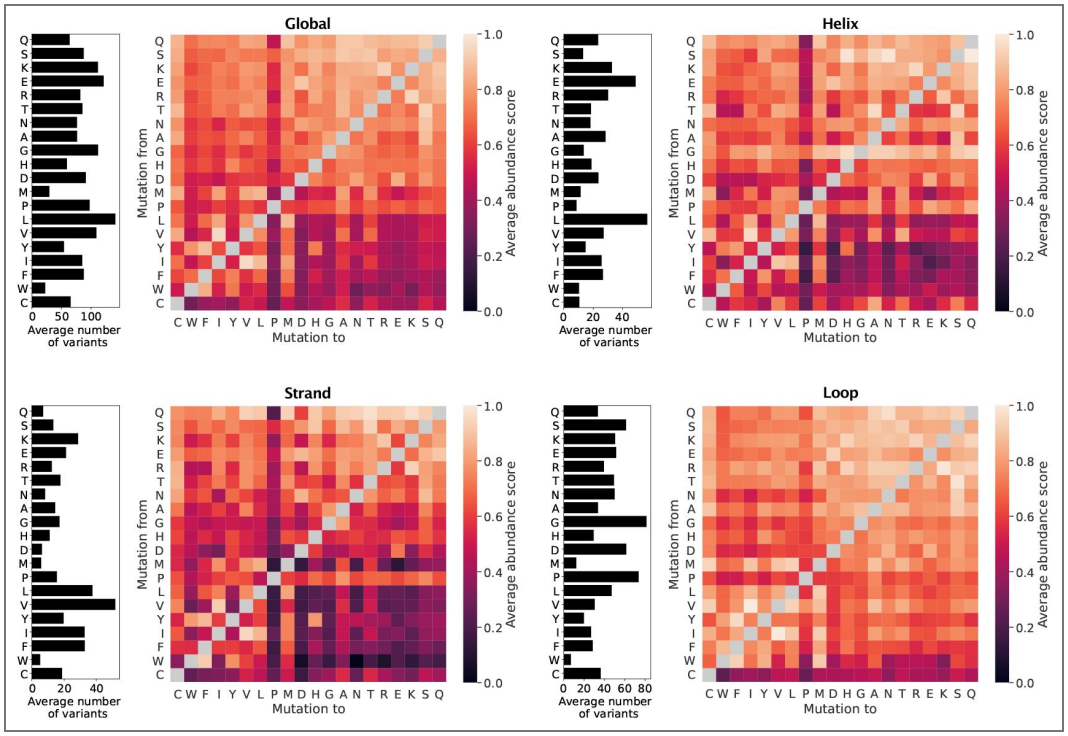


Figure S7. VAMP-seq abundance scores plotted against abundance score predictions obtained from average abundance score substitution matrices constructed without consideration of the structural context of the wild-type residue (Global in Fig. 2).

All predictions were performed using leave-out-one-protein cross-validation. Data point density is shown using an upper limit of 50 on the colour scale, meaning that a bin will be coloured red if it contains 50 data points or more.

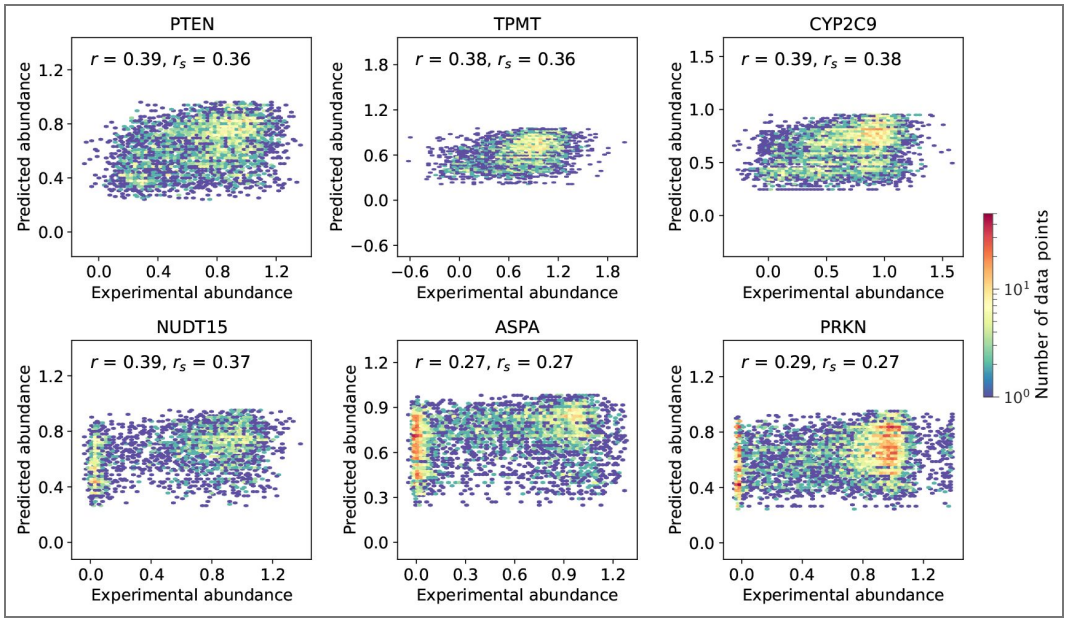


Figure S8. VAMP-seq abundance scores plotted against abundance score predictions obtained from average abundance score substitution matrices constructed by considering the degree of solvent-exposure of the wild-type residue (Exposure in Fig. 2).

All predictions were performed using leave-out-one-protein cross-validation. Data point density is shown using an upper limit of 50 on the colour scale, meaning that a bin will be coloured red if it contains 50 data points or more.

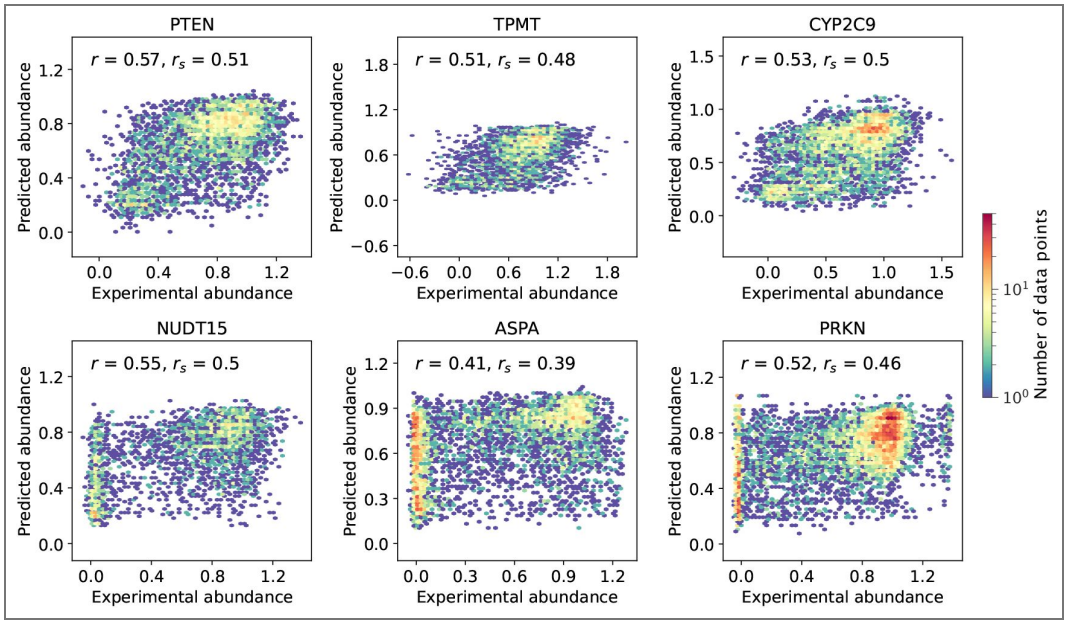
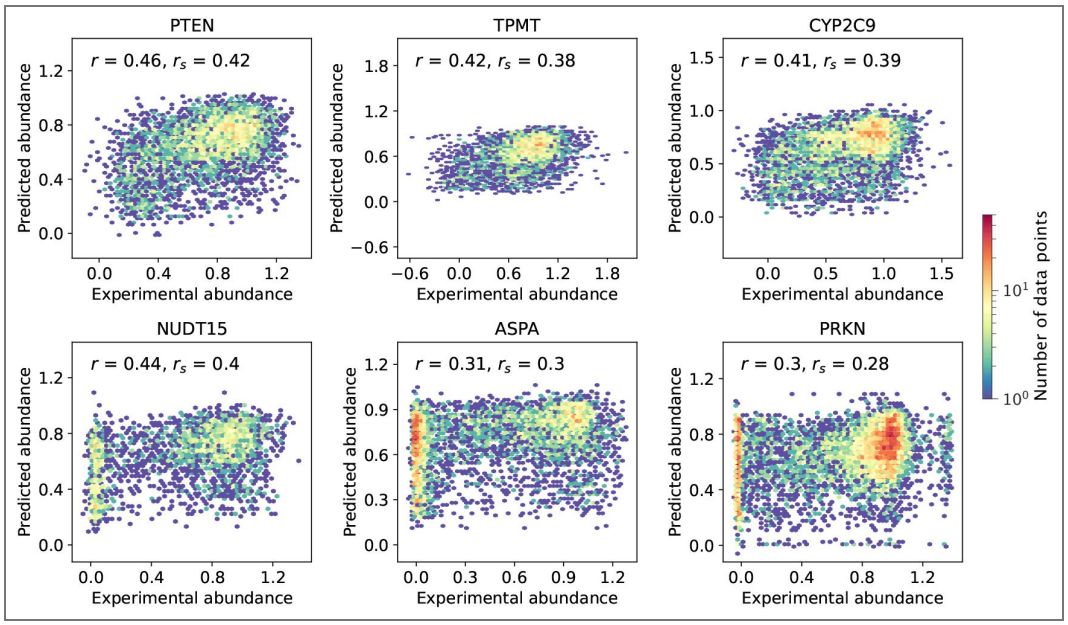


Figure S9. VAMP-seq abundance scores plotted against abundance score predictions obtained from average abundance score substitution matrices constructed by considering the secondary structure context of the wild-type residue (Secondary in Fig. 2).

All predictions were performed using leave-out-one-protein cross-validation. Data point density is shown using an upper limit of 50 on the colour scale, meaning that a bin will be coloured red if it contains 50 data points or more.



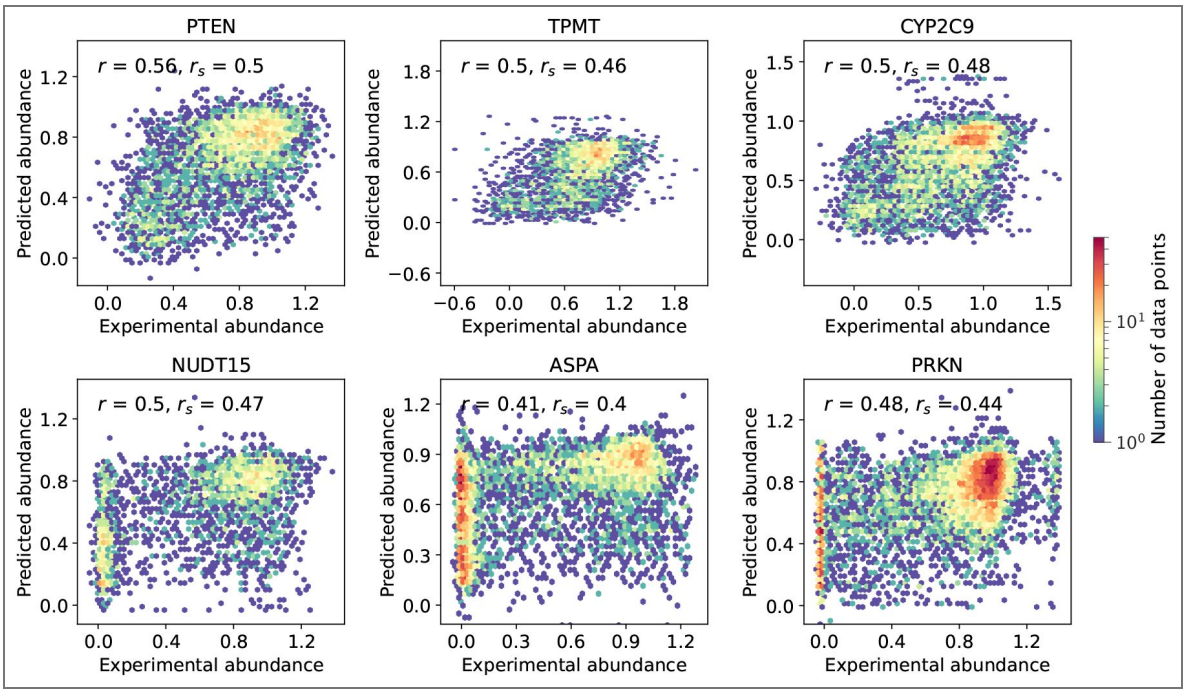


Figure S10. VAMP-seq abundance scores plotted against abundance score predictions obtained from average abundance score substitution matrices constructed by considering both the degree of solvent-exposure and the secondary structure context of the wild-type residue (Combined in Fig. 2).

All predictions were performed using leave-out-one-protein cross-validation. Data point density is shown using an upper limit of 50 on the colour scale, meaning that a bin will be coloured red if it contains 50 data points or more.

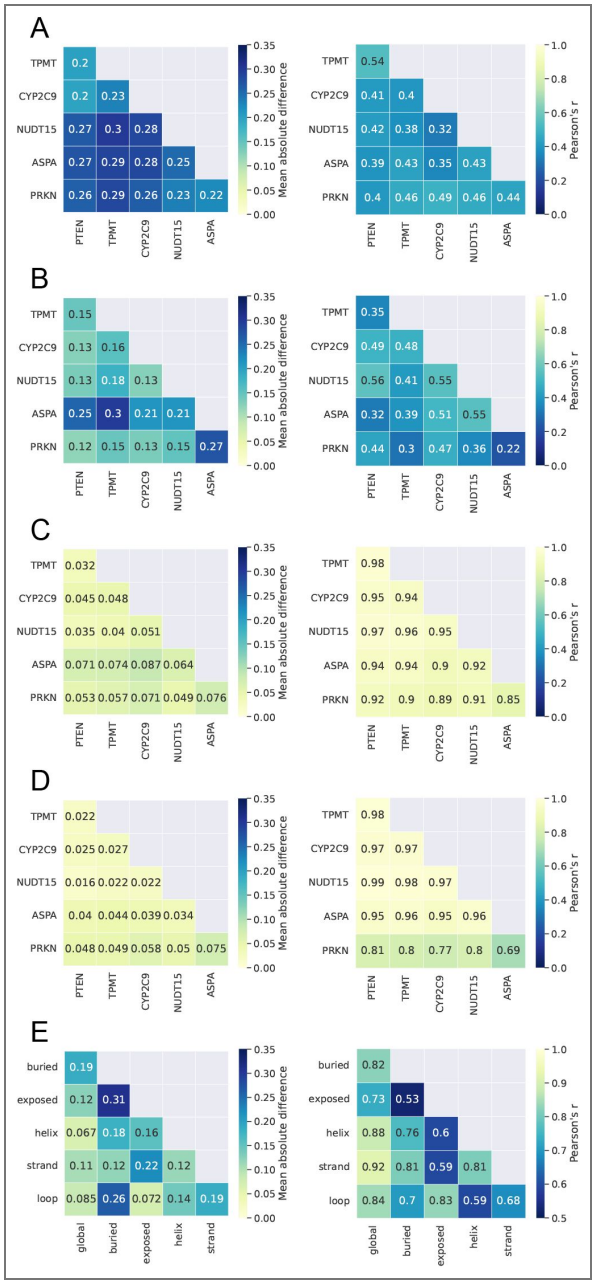


Figure S11. Comparison of average abundance score substitution matrices across proteins and residue environments.

A mean absolute error and *r* were calculated between pairs of substitution matrices considering all abundance score averages in the two matrices to evaluate matrix similarity. (A) Comparison of buried environment matrices that were each calculated using data from only a single protein. Protein names on the axes indicate the single dataset used for the substitution matrix calculation. (B) Comparison of exposed environment matrices each calculated with data from single a protein. (C) Comparison of buried environment matrices, with matrices calculated using five out of six VAMP-seq datasets. Protein names on the axes indicate which dataset was omitted from the substitution matrix calculation. (D) Comparing matrices for exposed environments, with matrices calculated using five out of six VAMP-seq datasets. (E) Comparing matrices for different structural environments using matrices based on all data from all six proteins.

Figure S12. Correlations between experimental abundance scores and abundance scores predicted from the six average abundance score substitution matrices that consider both residue burial and secondary structure context of the wild-type residue.

Scores were predicted for each protein (plot title) using average abundance score substitution matrices calculated with an increasing number of VAMP-seq datasets, though always leaving out the dataset for which predictions are performed. Orange data points mark prediction results from a single specific combination of datasets, and black data points are the average of the orange ones.

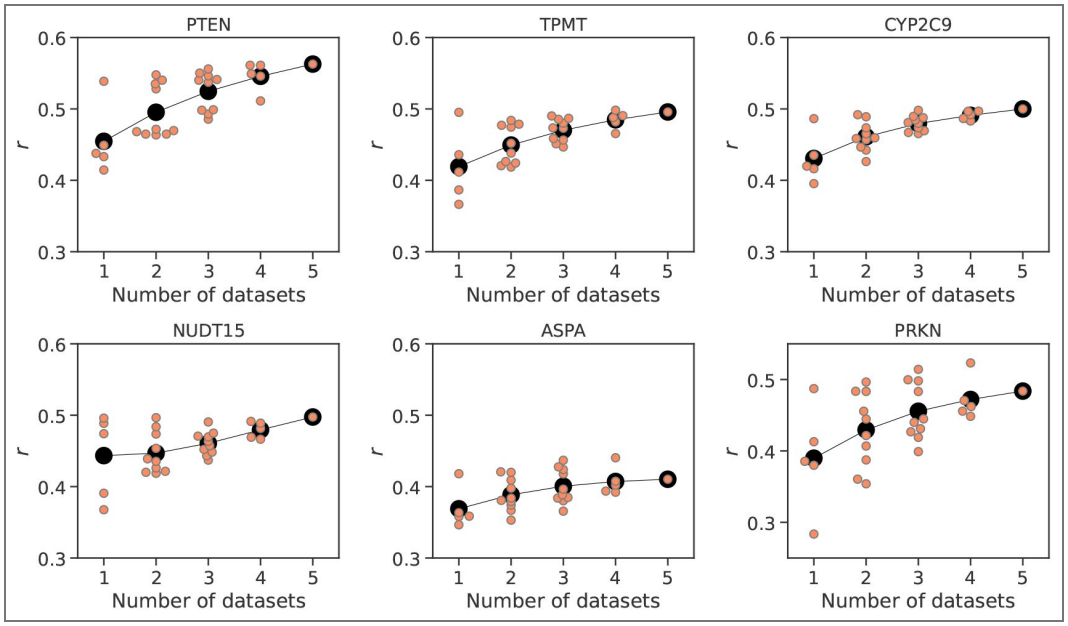
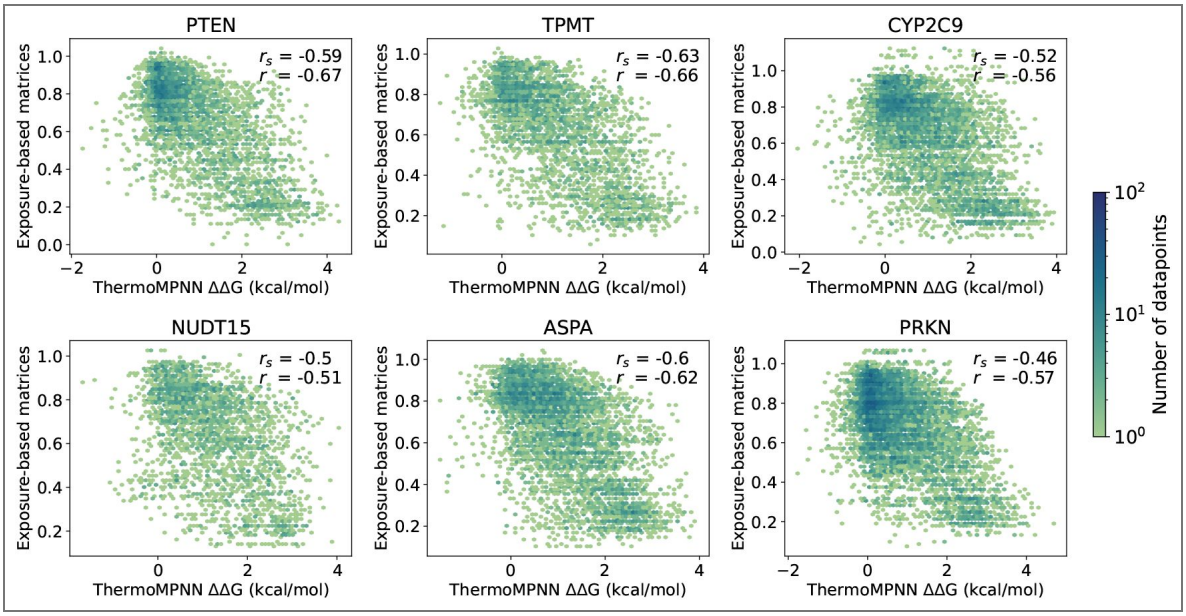


Figure S13. Predicted ΔΔG values correlate with predicted abundance scores.

ΔΔG values for all single residue substitution variants in the six proteins were estimated using the ThermoMPNN model and correlate moderately well with variant abundance scores predicted with the exposure-based substitution matrices, as indicated with the *r* and r_s values per dataset in the individual plots. Data point density is shown using an upper limit of 100 on the colour scale.



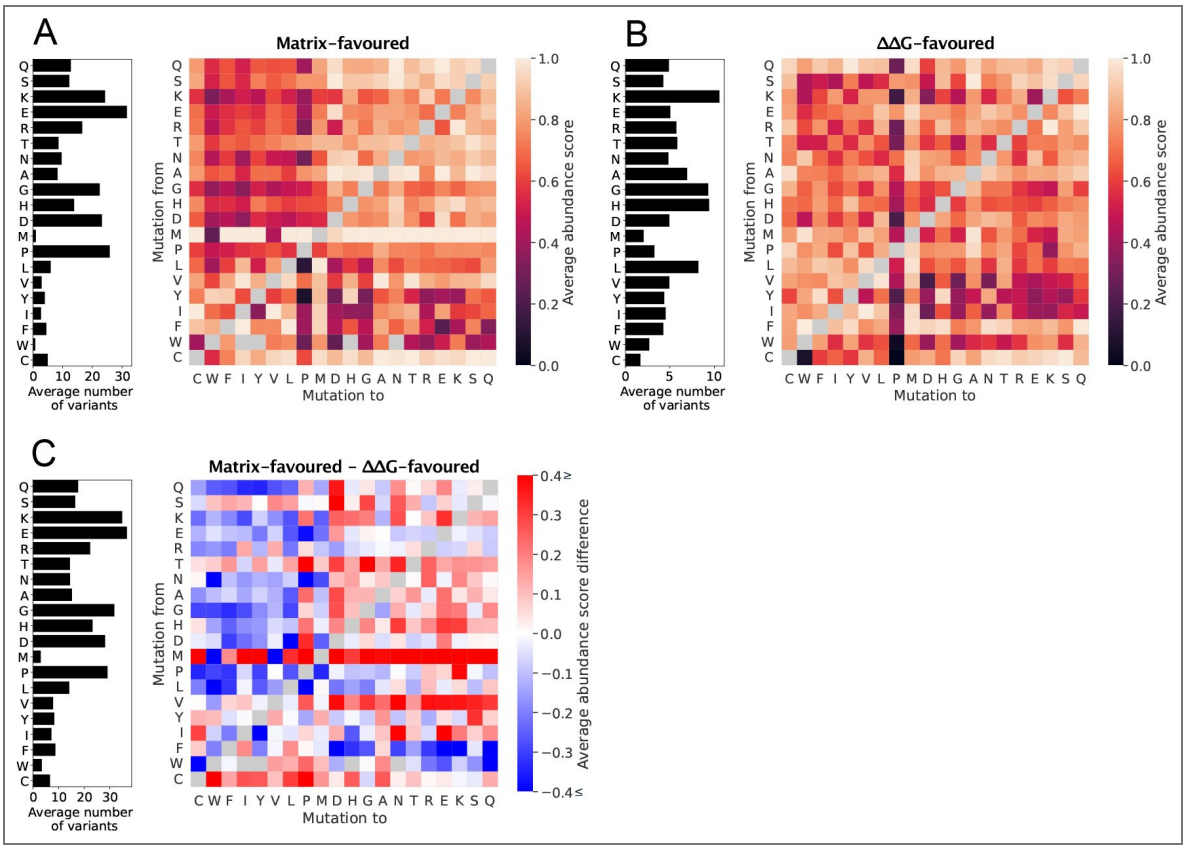


Figure S14. Abundance score substitution matrices for variants of solvent-exposed residues classified as either (A) matrix-favoured or (B) ΔΔG-favoured, as well as (C) the difference between these two substitution matrices.

In (A) and (B), the bar plot to the left of the substitution matrix shows the average number of data points used to calculate the averages in a matrix row. In (C), the bar plot was constructed by summing the number of variants informing the abundance score averages in the matrix-favoured and ΔΔG-favoured substitution matrices before the average of number of variants in each row was determined. We note that for several reference residue types, such as methionine, tryptophan and other aromatic and non-polar residues, the number of VAMP-seq scores used to calculate the substitution matrices is very small, since these reference residue types are depleted on the protein surfaces. The substitution matrices and the difference between them are thus relatively poorly determined for these reference residue types.

Figure S15. Average abundance score substitution matrices for (A) buried, (B) exposed and (C) loop environments calculated without abundance scores from the PRKN VAMP-seq dataset, but including all data from the five remaining datasets.

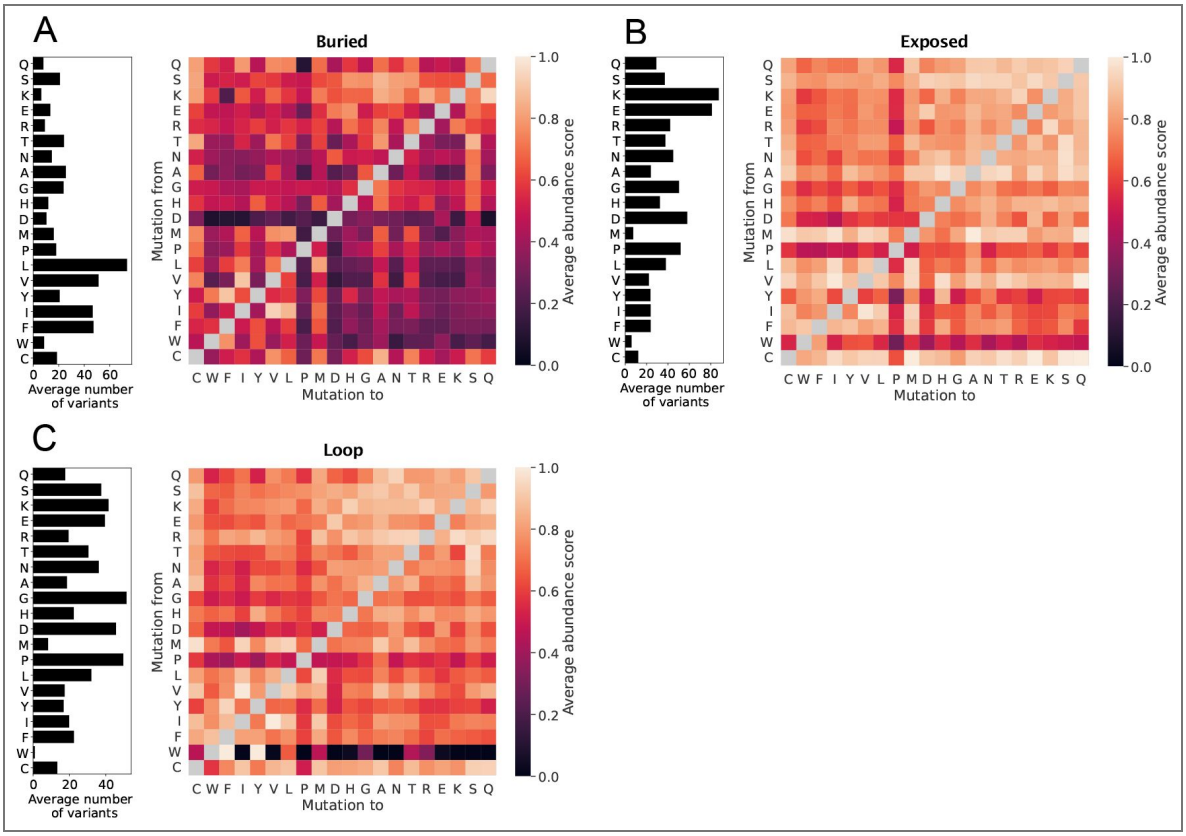
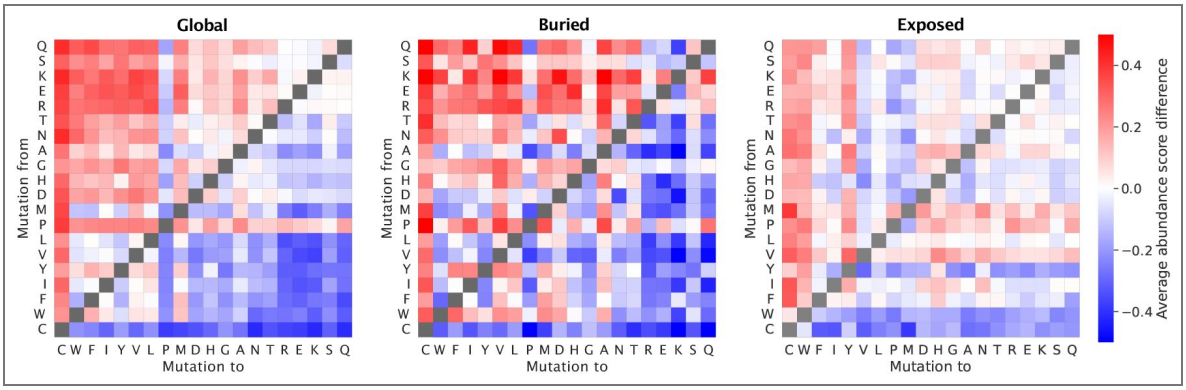


Figure S16. Heatmaps indicating the degree of symmetry across the diagonal in average abundance score sub-stitution matrices.

The symmetry was for each of the three residue environments (global, buried and exposed) evaluated for a given substitution type by subtracting the average abundance score for the reverse substitution from the average abundance score of the forward substitution. Hence, white indicates that a given residue pair behaves symmetrically, red indicates that the substitution is less disruptive in the forward than in the reverse direction, and blue means that the substitution is more disruptive in the forward than in the reverse direction.



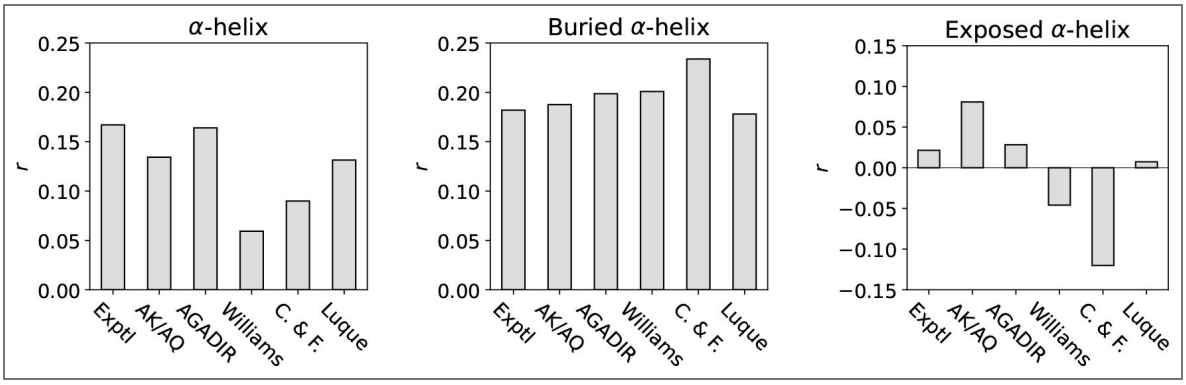


Figure S17. Analysis of the impact of amino acid helix propensity on average abundance scores.

Pearson's correlation coefficient, *r*, between the average abundance scores for residues sitting in *α*-helical structure (either all *α*-helical structure or in only buried or exposed *α*-helical structure) and substitution matrices constructed from different helix propensity scales reporting on the ΔΔG of helix residue substitutions. Helix propensity scales were obtained from and are named as in previous work comparing the different scales (Pace and Scholtz, 1998). Explt scale reports on helix propensities for solvent-exposed, non-capping residues (Pace and Scholtz, 1998), AK/AQ (Rohl et al., 1996) and AGADIR (Muñoz and Serrano, 1995) were constructed from experimental data on peptides in solution, and C. & F. (Chou and Fasman, 1978) and Williams (Williams et al., 1987) scales are based on the frequency of amino acid occurrence in *α*-helices, buried as well as exposed. Finally, the Luque scale was originally derived from structure-based thermodynamical calculations and reports on helix formation propensities for residues in solvent-exposed helices (Luque et al., 1996).

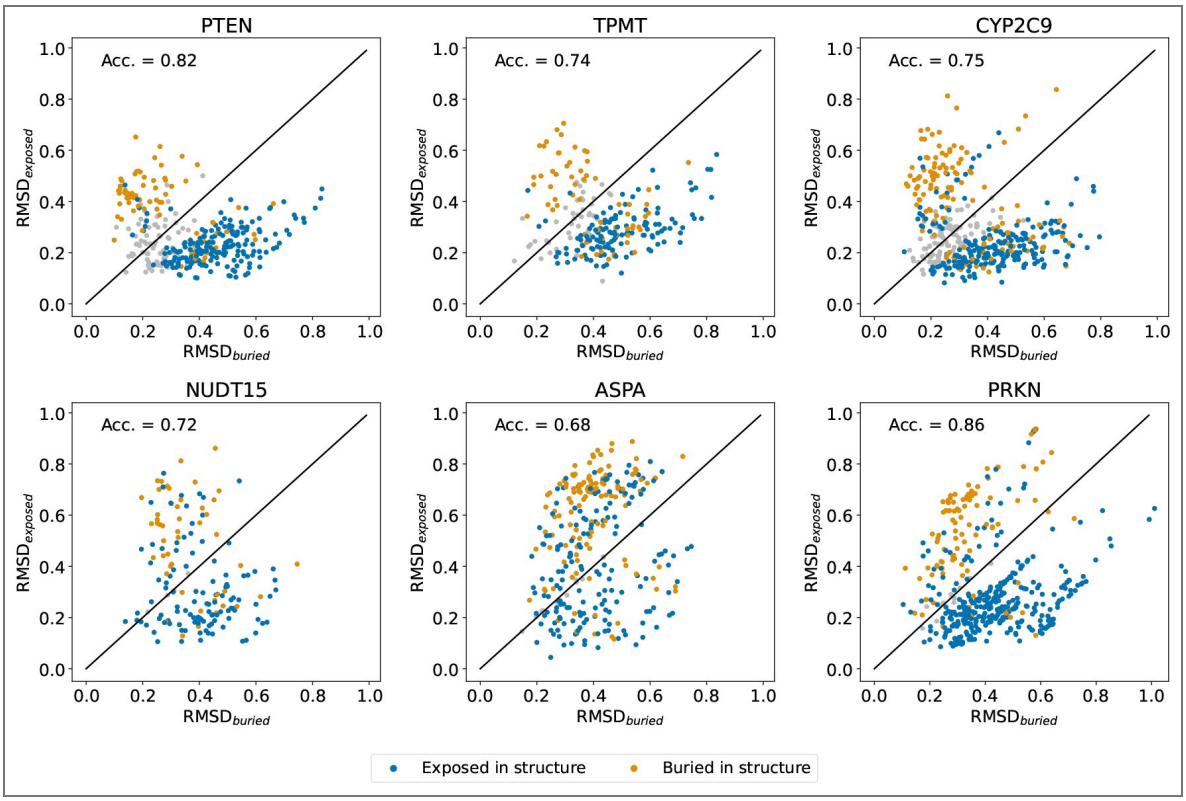


Figure S18. Discovering residues with abundance score substitution profiles non-typical for their structural environment.

Each datapoint in the plots corresponds to a single residue coloured by its wild-type structure environment type and for which we calculated $RMSD_{exposed}$ and $RMSD_{buried}$. Solvent-exposed residues (blue) are expected to have $RMSD_{exposed} < RMSD_{buried}$ and thus appear below the plot diagonals, while the opposite is true for buried residues (yellow). Blue datapoints falling above the diagonal thus correspond to solvent-exposed residues with substitution profiles most similar to the average profile of buried residue, and yellow datapoints below the diagonal represent buried residues with substitution profiles relatively similar to the average profile of exposed residues. Grey datapoints indicate residues for which we could not with confidence determine if $RMSD_{exposed}$ is larger or smaller than $RMSD_{buried}$ (see Methods). We quantified the extent to which buried and exposed residues respectively satisfy $RMSD_{buried} < RMSD_{exposed}$ and $RMSD_{exposed} < RMSD_{buried}$ by calculating the balanced classification accuracy (Acc.) per dataset, including only non-grey datapoints in the calculation. Data is only shown for residues with a least five measured abundance scores.

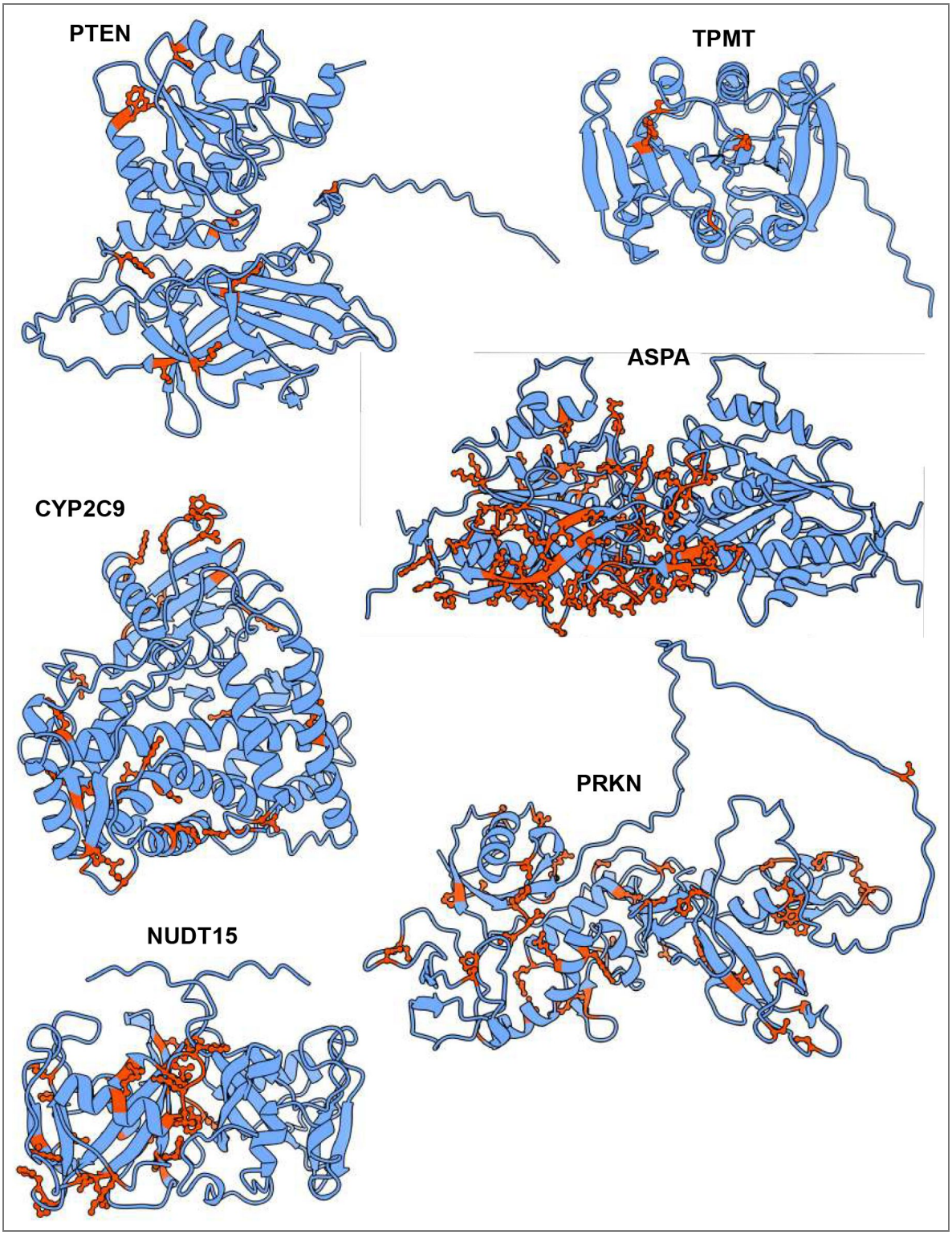


Figure S19. Protein structures with solvent-exposed residues found to have a smaller $RMSD_{buried}$ than $RMSD_{exposed}$ shown in dark orange.

We only show exposed residues for which $RMSD_{buried}$ was found to be smaller than $RMSD_{buried}$ for more than 95% of resampled residue substitution profiles, where resampling was based on experimental errors of VAMP-seq scores (see Methods). NUDT15 and ASPA are shown as homodimers, but the exposed residues discovered in the analysis are shown for only one of the monomers in the complexes.

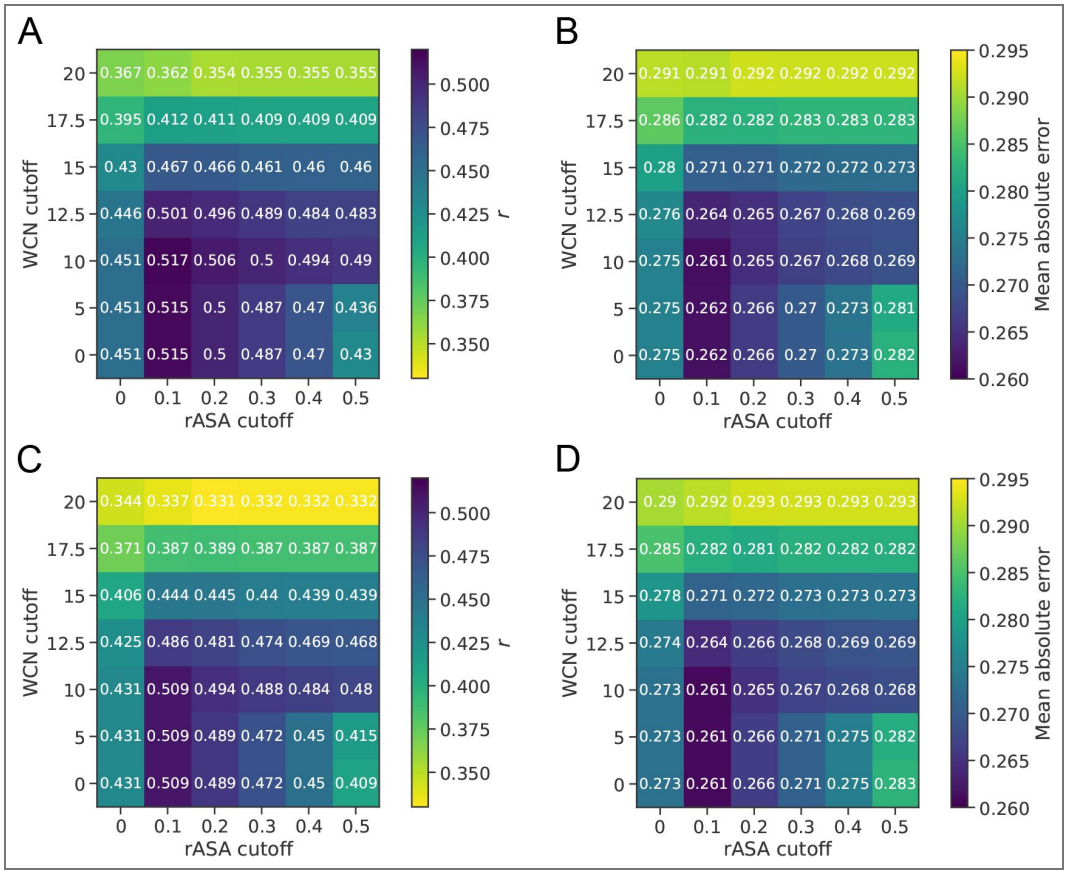


Figure S20. Grid search for selection of structure feature cutoffs to classify residues as solvent-exposed or structurally buried.

(A) Pearson's correlation coefficient *r* between experimental and predicted abundance scores, with predictions based on exposure-based substitution matrices of average abundance scores calculated using different rASA and WCN cutoff values to classify residues as buried or exposed. Predictions were made using leave-one-protein-out cross-validation, and *r* is an average of results obtained for each of the proteins in our analysis. (B) Mean absolute error between between experimental and predicted abundance scores for the same predictions that results are shown for in panel A. Based on the results shown in panels A and B, we have used an rASA cutoff value of 0.1 to classify residues as solvent-exposed or buried in all other analyses in this work. (C,D) Results for grid search analysis similar to the analysis presented in panel A and B, but using median rather than average abundance scores to construct substitution matrices. We explored if using exposure-based substitution matrices of median abundance scores rather than average abundance scores would result in higher prediction accuracy. We found that, for our selected cutoffs, both the *r* and the MAE between experimental and predicted abundance scores are highly similar for the median- and average-based matrices.

Data availability

Data and code for reproducing the results presented in this work are available at *https://github.com/KULL-Centre/_2024_Schulze_abundance-analysis*.

Acknowledgements

This work was funded by the Novo Nordisk Foundation challenge program PRISM (Protein Interactions and Stability in Medicine and Genomics, NNF18OC0033950, to K.L.-L.). We thank Lene Clausen, Martin Grønbæk-Thygesen, Vasileios Voutsinos, Kristoffer E. Johansson, Matteo Cagiada, Amelie Stein, and Rasmus Hartmann-Petersen and other members of the PRISM centre for productive discussions and suggestions on the topic of protein abundance. We acknowledge access to computational resources at the Biocomputing Core Facility at the Department of Biology, University of Copenhagen.

Additional information

Funding

Funder	Grant reference number	Author
Novo Nordisk Fonden (NNF)	NNF18OC0033950	Kresten Lindorff-Larsen

Author ORCID iDs

Thea K Schulze: https://orcid.org/0000-0001-6587-6749
Kresten Lindorff-Larsen: https://orcid.org/0000-0002-4750-6039

References

1. **Abildgaard AB,** Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, Hoffmann ER, Bernstein I, Gerdes AM, Takahashi M, *et al.* (2019) Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. *eLife* **8**:e49138 https://doi.org/10.7554/eLife.49138
2. **Adkar BV,** Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, Swarnkar MK, Gokhale RS, Varadarajan R. (2012) Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure* **20**:371-381
3. **Amorosi CJ,** Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, Kowalski JP, Rettie AE, Fowler DM, Dunham MJ (2021) Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *The American Journal of Human Genetics* **108**:1735-1751
4. **Beltran T,** Jiang X, Shen Y, Lehner B. (2024) Site saturation mutagenesis of 500 human protein domains reveals the contribution of protein destabilization to genetic disease. *bioRxiv*
5. **Bitto E,** Bingman CA, Wesenberg GE, McCoy JG, Phillips GN (2007) Structure of aspartoacylase, the brain enzyme impaired in Canavan disease. *Proceedings of the National Academy of Sciences* **104**:456-461
6. **Blaabjerg LM,** Kassem MM, Good LL, Jonsson N, Cagiada M, Johansson KE, Boomsma W, Stein A, Lindorff-Larsen K. (2023) Rapid protein stability prediction using deep learning representations. *eLife* **12**:e82593 https://doi.org/10.7554/eLife.82593
7. **Cagiada M,** Bottaro S, Lindemose S, Schenstrøm SM, Stein A, Hartmann-Petersen R, Lindorff-Larsen K. (2023) Discovering functionally important sites in proteins. *Nature Communications* **14**:4175

8. **Cagiada M,** Johansson KE, Valanciute A, Nielsen SV, Hartmann-Petersen R, Yang JJ, Fowler DM, Stein A, Lindorff-Larsen K. (2021) Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Molecular Biology and Evolution* **38**:3235-3246
9. **Cagiada M,** Jonsson N, Lindorff-Larsen K. (2024) Decoding molecular mechanisms for loss of function variants in the human proteome. *bioRxiv*
10. **Caldararu O,** Blundell TL, Kepp KP (2021) Three Simple Properties Explain Protein Stability Change upon Mutation. *Journal of Chemical Information and Modeling* **61**:1981-1988
11. **Carter M,** Jemth AS, Hagenkort A, Page BDG, Gustafsson R, Griese JJ, Gad H, Valerie NCK, Desroses M, Boström J, *et al.* (2015) Crystal structure, biochemical and cellular activities demonstrate separate functions of MTH1 and MTH2. *Nature Communications* **6**:7871
12. **Chakravarty S,** Varadarajan R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **7**:723-732
13. **Chiasson MA,** Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, Sun S, Roth FP, DeSloover D, Marks DS, *et al.* (2020) Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* **9**:e58026 https://doi.org/10.7554/eLife.58026
14. **Chou PY,** Fasman GD (1978) Empirical Predictions of Protein Conformation. *Annual Review of Biochemistry* **47**:251-276
15. **Cisneros AF,** Gagnon-Arsenault I, Dubé AK, Després PC, Kumar P, Lafontaine K, Pelletier JN, Landry CR (2023) Epistasis between promoter activity and coding mutations shapes gene evolvability. *Science Advances* **9**:eadd9109
16. **Clausen L,** Abildgaard AB, Gersing SK, Stein A, Lindorff-Larsen K, Hartmann-Petersen R. (2019) Protein stability and degradation in health and disease. In: Donev R (Ed). *Advances in Protein Chemistry and Structural Biology* **114** Elsevier. pp. 61-83
17. **Clausen L,** Voutsinos V, Cagiada M, Johansson KE, Grønbæk-Thygesen M, Nariya S, Powell RL, Have MKN, Oestergaard VH, Stein A, *et al.* (2024) A mutational atlas for Parkin proteostasis. *Nature Communications* **15**:1541
18. **Correa Marrero M,** Barrio-Hernandez I. (2021) Toward Understanding the Biochemical Determinants of Protein Degradation Rates. *ACS Omega* **6**:5091-5100
19. **Cota E,** Hamill SJ, Fowler SB, Clarke J. (2000) Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *Journal of Molecular Biology* **302**:713-725
20. **Dang CC,** Minh BQ, Mcshea H, Masel J, James JE, Sy L, Lanfear R. (2022) nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models. *Systematic Biology* **71**
21. **Dayhoff M,** Schwartz R, Orcutt B. (1978) A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**:345-352
22. **Degn K,** Utichi M, Besora PSI, Tiberti M. (2025) In-Silico Stability Predictors: Investigation of Performance Towards balanced Experimental Data. *bioRxiv*
23. **Dieckhaus H,** Brocidiacono M, Randolph NZ, Kuhlman B. (2024) Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences* **121**:e2314853121
24. **Drake ZC,** Day E, Toth P, Lindert S. (2024) Deep-Learning Structure Elucidation from Single-Mutant Deep Mutational Scanning. *bioRxiv*
25. **Dunham AS,** Beltrao P. (2021) Exploring amino acid functions in a deep mutational landscape. *Molecular Systems Biology* **17**:e10305
26. **Esposito D,** Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF (2019) MaveDB: an opensource platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology* **20**:223

27. **Faure AJ,** Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. (2022) Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**:175-183

28. **Fowler DM,** Fields S. (2014) Deep mutational scanning: a new style of protein science. *Nature Methods* **11**:801-807

29. **Fowler DM,** Stephany JJ, Fields S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols* **9**:2267-2284

30. **Geck RC,** Boyle G, Amorosi CJ, Fowler DM, Dunham MJ (2022) Measuring Pharmacogene Variant Function at Scale Using Multiplexed Assays. *Annual Review of Pharmacology and Toxicology* **62**:531-550

31. **Gerasimavicius L,** Livesey BJ, Marsh JA (2023) Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Science* **32**:e4688

32. **Gersing S,** Schulze TK, Cagiada M, Stein A, Roth FP, Lindorff-Larsen K, Hartmann-Petersen R. (2024) Characterizing glucokinase variant mechanisms using a multiplexed abundance assay. *Genome Biology* **25**:98

33. **Gray VE,** Hause RJ, Fowler DM (2017) Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* **207**:53-61

34. **Grønbæk-Thygesen M,** Voutsinos V, Johansson KE, Schulze TK, Cagiada M, Pedersen L, Clausen L, Nariya S, Powell RL, Stein A, *et al.* (2024) Deep mutational scanning reveals a correlation between degradation and toxicity of thousands of aspartoacylase variants. *Nature Communications* **15**:4026

35. **Henikoff S,** Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**:10915-10919

36. **Hipp MS,** Kasturi P, Hartl FU (2019) The proteostasis network and its decline in ageing. *Nature Reviews Molecular Cell Biology* **20**:421-435

37. **Hovmöller S,** Zhou T, Ohlson T. (2002) Conformations of amino acids in proteins. *Acta Crystallographica Section D Biological Crystallography* **58**:768-776

38. **Høie MH,** Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. (2022) Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports* **38**:110207

39. **Jiang L,** Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA (2013) Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLoS Genetics* **9**:e1003600

40. **Jumper J,** Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583-589

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

42. **Kampmeyer C,** Larsen-Ledet S, Wagnkilde MR, Michelsen M, Iversen HKM, Nielsen SV, Lindemose S, Caregnato A, Ravid T, Stein A, *et al.* (2022) Disease-linked mutations cause exposure of a protein quality control degron. *Structure* **30**:1245-1253.e5

43. **Kumar A,** Aguirre JD, Condos TE, Martinez-Torres RJ, Chaugule VK, Toth R, Sundaramoorthy R, Mercier P, Knebel A, Spratt DE, *et al.* (2015) Disruption of the autoinhibited state primes the E3 ligase parkin for activation and catalysis. *The EMBO Journal* **34**:2506-2521

44. **Le Coq J,** Pavlovsky A, Malik R, Sanishvili R, Xu C, Viola RE (2008) Examination of the Mechanism of Human Brain Aspartoacylase through the Binding of an Intermediate Analogue. *Biochemistry* **47**:3484-3492

45. **Lee JO,** Yang H, Georgescu MM, Cristofano AD, Maehama T, Shi Y, Dixon JE, Pandolfi P, Pavletich NP (1999) Crystal Structure of the PTEN Tumor Suppressor: Implications for Its Phosphoinositide Phosphatase Activity and Membrane Association. *Cell*

46. **Luque I,** Mayorga OL, Freire E. (1996) Structure-Based Thermodynamic Scale of a-Helix Propensities in Amino Acids. *Biochemistry* **35**:13681-13688

47. **Matreyek KA,** Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50**:874-882

48. **Matreyek KA,** Stephany JJ, Ahler E, Fowler DM (2021) Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Medicine* **13**:165

49. **Mavor D,** Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, Gaskins G, Liu Z, Deming L, Axen SD, *et al.* (2016) Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife* **5**:e15802 https://doi.org/10.7554/eLife.15802

50. **Mavor D,** Barlow KA, Asarnow D, Birman Y, Britain D, Chen W, Green EM, Kenner LR, Mensa B, Morinishi LS, *et al.* (2018) Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biology Open* **7**:bio036103

51. **McGibbon RT,** Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**:1528-1532

52. **Mirdita M,** Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods* **19**:679-682

53. **Moore RA,** Le Coq J, Faehnle CR, Viola RE (2003) Purification and preliminary characterization of brain aspartoacylase. *Archives of Biochemistry and Biophysics* **413**:1-8

54. **Munro D,** Singh M. (2021) DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36**:5322-5329

55. **Muñoz V,** Serrano L. (1995) Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. II†. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides. *Journal of Molecular Biology* **245**:275-296

56. **Nguyen TN,** Ingle C, Thompson S, Reynolds KA (2024) The genetic landscape of a metabolic interaction. *Nature Communications* **15**:3351

57. **Nielsen SV,** Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, Kassem MM, Rasmussen LJ, Lindorff-Larsen K, Hartmann-Petersen R. (2017) Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLOS Genetics* **13**:e1006739

58. **Pace CN,** Scholtz JM (1998) A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical Journal* **75**:422-427

59. **Park H,** Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. (2016) Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation* **12**:6201-6212

60. **Peterman N,** Levine E. (2016) Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**:206

61. **Pettersen EF,** Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**:70-82

62. **Ribeiro AJM,** Tyzack JD, Borkakoti N, Holliday GL, Thornton JM (2020) A global analysis of function and conservation of catalytic residues in enzymes. *Journal of Biological Chemistry* **295**:314-324

63. **Rocklin GJ,** Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, *et al.* (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**:168-175

64. **Rodrigues J,** Teixeira J, Trellet M, Bonvin A. (2018) pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. *F1000Research* **7**

65. **Rohl CA,** Chakrabartty A, Baldwin RL (1996) Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Protein Science* **5**:2623-2637

66. **Rollins NJ,** Brock KP, Poelwijk FJ, Stiffer MA, Gauthier NP, Sander C, Marks DS (2019) Inferring protein 3D structure from deep mutation scans. *Nature Genetics* **51**:1170-1176

67. **Rubin AF,** Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM (2017) A statistical framework for analyzing deep mutational scanning data. *Genome Biology* **18**:150

68. **Rubin AF,** Min JK, Rollins NJ, Esposito D, Harrington M, Stone J, Bianchi AH, Dias M, Frazer J, Fu Y, *et al.* (2021) MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. *bioRxiv*

69. **Scheller R,** Stein A, Nielsen SV, Marin FI, Gerdes A, Marco M, Papaleo E, Lindorff-Larsen K, Hartmann-Petersen R. (2019) Toward mechanistic models for genotype–phenotype correlations in phenylketonuria using protein stability calculations. *Human Mutation* **40**:444-457

70. **Schmiedel JM,** Lehner B. (2019) Determining protein structures using deep mutagenesis. *Nature Genetics* **51**:1177-1186

71. **Schulze TK,** Blaabjerg LM, Cagiada M, Lindorff-Larsen K. (2025) Supervised learning of protein variant effects across large-scale mutagenesis datasets. *bioRxiv* https://doi.org/10.1101/2025.04.02.646878

72. **Serrano L,** Jr JTK, Cann P, Matouschek A, Fersht AR (1992) The Folding of an Enzyme II. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability. *Journal of Molecular Biology* **224**:783-804

73. **Suiter CC,** Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, Yang W, Hoshitsuki K, Singh M, Trehan A, *et al.* (2020) Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proceedings of the National Academy of Sciences* **117**:5394-5401

74. **Tabet D,** Parikh V, Mali P, Roth FP, Claussnitzer M. (2022) Scalable Functional Assays for the Interpretation of Human Genetic Variation. *Annual Review of Genetics* **56**:441-465

75. **Thompson S,** Zhang Y, Ingle C, Reynolds KA, Kortemme T. (2020) Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *eLife* **9**:e53476

76. **Tien MZ,** Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE* **8**:e80635

77. **Tsuboyama K,** Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, Mangan NM, Ovchinnikov S, Rocklin GJ (2023) Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**:434-444

78. **Varadi M,** Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**:D439-D444

79. **Vazquez F,** Ramaswamy S, Nakamura N, Sellers WR (2000) Phosphorylation of the PTEN Tail Regulates Protein Stability and Function. *Molecular and Cellular Biology* **20**:5010-5018

80. **Weile J,** Kishore N, Sun S, Maaieh R, Verby M, Li R, Fotiadou I, Kitaygorodsky J, Wu Y, Holenstein A, *et al.* (2021) Shifting landscapes of human MTHFR missense-variant effects. *The American Journal of Human Genetics* **108**:1283-1300

81. **Weile J,** Roth FP (2018) Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics* **137**:665-678

82. **Wester MR,** Yano JK, Schoch GA, Yang C, Griffin KJ, Stout CD, Johnson EF (2004) The Structure of Human Cytochrome P450 2C9 Complexed with Flurbiprofen at 2.0-Å Resolution. *Journal of Biological Chemistry* **279**:35630-35637

83. **Williams RW,** Chang A, Juretić D, Loughran S. (1987) Secondary structure predictions and medium range interactions. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **916**:200-204

84. **Wu H,** Horton JR, Battaile K, Allali-Hassani A, Martin F, Zeng H, Loppnau P, Vedadi M, Bochkarev A, Plotnikov AN, *et al.* (2007) Structural basis of allele variation of human thiopurine-S-methyltransferase. *Proteins: Structure, Function, and Bioinformatics* **67**:198-208

85. Yampolsky LY, Stoltzfus A. (2005) The Exchangeability of Amino Acids in Proteins. *Genetics* **170**:1459-1472

86. Zutz A, Hamborg L, Pedersen LE, Kassem MM, Papaleo E, Koza A, Herrgård MJ, Jensen SI, Teilum K, Lindorff-Larsen K, *et al.* (2021) A dual-reporter system for investigating and optimizing protein translation and folding in E. coli. *Nature Communications* **12**:6093

Peer reviews

Reviewer #1 (Public review):

Significance:

While most MAVEs measure overall function (which is a complex integration of biochemical properties, including stability), VAMP-seq-type measurements more strongly isolate stability effects in a cellular context. This work seeks to create a simple model for predicting the response for a mutation on the "abundance" measurement of VAMP-seq.

Public Review:

Of course, there is always another layer of the onion, VAMP-seq measures contributions from isolated thermodynamic stability, stability conferred by binding partners (small molecule and protein), synthesis/degradation balance (especially important in "degron" motifs), etc. Here the authors' goal is to create simple models that can act as a baseline for two main reasons:

(1) how to tell when adding more information would be helpful for a global model;

(2) how to detect when a residue/mutation has an unusual profile indicative of an unbalanced contribution from one of the factors listed above.

As such, the authors state that this manuscript is not intended to be a state-of-the-art method in variant effect prediction, but rather a direction towards considering static structural information for the VAMP-seq effects. At its core, the method is a fairly traditional asymmetric substitution matrix (I was surprised not to see a comparison to BLOSUM in the manuscript) - and shows that a subdivision by burial makes the model much more predictive. Despite only having 6 datasets, they show predictive power even when the matrices are based on a smaller number. Another success is rationalizing the VAMPseq results on relevant oligomeric states.

Comments on revision:

We have no further comments on this manscript.

https://doi.org/10.7554/eLife.103721.2.sa2

Reviewer #3 (Public review):

"Effects of residue substitutions on the cellular abundance of proteins" by Schulze and Lindorff-Larsen revisits the classical concept of structure-aware protein substitution matrices through the scope of modern protein structure modelling approaches and comprehensive phenotypic readouts from multiplex assays of variant effects (MAVEs). The authors explore 6 unique protein MAVE datasets based on protein abundance through the lens of protein structural information (residue solvent accessibility, secondary structure type) to derive combinations of context-specific substitution matrices that predict variant impact on protein abundance. They are clear to outline that the aim of the study is not to produce a new best abundance predictor, but to showcase the degree of prediction afforded simply by utilizing structural information.

Both the derived matrices and the underlying 'training' data are comprehensively evaluated. The authors convincingly demonstrate that taking structural solvent accessibility contexts into account leads to more accurate performance than either a structure-unaware matrix, secondary structure-based matrix, or matrices combining both solvent accessibility and secondary structure. The capacity for the approach to produce generalizable matrices is explored through training data combinations, highlighting factors such as the variable quality of the experimental MAVE data and the biochemical differences between the protein targets themselves, which can lead to limitations. Despite this, the authors demonstrate their simple matrix approach is generally on par with dedicated protein stability predictors in abundance effect evaluation, and even outperforms them in a niche of solvent accessible surface mutations, revealing their matrices provide orthogonal abundance-specific signal. More importantly, the authors further develop this concept to creatively show their matrices can be used to identify surface residues that have buried-like substitution profiles, which are shown to correspond to protein interface residues, post-translational modification sites, functional residues or putative degrons.

The paper makes a strong and well-supported main point, demonstrating the widespread utility of the authors' approach, empowered through protein structural information and cutting edge MAVE datasets. This work creatively utilizes a simple concept to produce a highly interpretable tool for protein abundance prediction (and beyond), which is inspiring in the age of impenetrable machine learning models.

https://doi.org/10.7554/eLife.103721.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer # 1 (Public review):***
>
> *Significance:*
>
> *While most MAVEs measure overall function (which is a complex integration of biochemical properties, including stability), VAMP-seqtype measurements more strongly isolate stability effects in a cellular context. This work seeks to create a simple model for predicting the response for a mutation on the "abundance" measurement of VAMPseq.*

We thank the reviewer for their evaluation of our work and for their comments and feedback below.

> *Of course, there is always another layer of the onion, VAMP-seq measures contributions from isolated thermodynamic stability, stability conferred by binding partners (small molecule and protein), synthesis/degradation balance (especially important in "degron" motifs), etc. Here the authors' goal is to create simple models that can act as a baseline for two main reasons:*
>
> *(1) how to tell when adding more information would be helpful for a global model;*
>
> *(2) how to detect when a residue/mutation has an unusual profile indicative of an unbalanced contribution from one of the factors listed above.*
>
> *As such, the authors state that this manuscript is not intended to be a state-of-the-art method in variant effect prediction, but rather a direction towards considering static structural information for the VAMP-seq effects. At its core, the method is a fairly traditional asymmetric substitution matrix (I was surprised not to see a comparison to*

> *BLOSUM in the manuscript) - and shows that a subdivision by burial makes the model much more predictive. Despite only having 6 datasets, they show predictive power even when the matrices are based on a smaller number. Another success is rationalizing the VAMPseq results on relevant oligomeric states.*

We thank the reviewer for their summary of the main points of our work. Based on the suggestion by the reviewer, we have added a comparison to predictions with BLOSUM62 to our revised manuscript, noting that we have previously compared the BLOSUM62 matrix to a broader and more heterogeneous set of scores generated by MAVEs (Høie et al, 2022).

> *Specific Feedback:*
>
> *Major points:*
>
> *The authors spend a good amount of space discussing how the six datasets have different distributions in abundance scores. After the development of their model is there more to say about why? Is there something that can be leveraged here to design maximally informative experiments?*

We believe that these effects arise from a combination of intrinsic differences between the systems and assay-specific effects. For example, biophysical differences between the systems, such as differences in absolute folding stabilities or melting temperatures, will play a role, as will the fact that some proteins contain multiple domains.

Also, the sequencing-based score for an individual variant in a sort-seq experiment (such as VAMP-seq) depends both on the properties of that variant and on the composition of the entire FACS-sorted cell library. This is because cells are sorted into bins depending on the composition of the entire library, which means that library-to-library composition differences can contribute to the differences between VAMP-seq score distributions.

From our developed models and outliers in predictions from these, it is difficult to tell which of the several possible underlying reasons cause the differences. We have briefly expanded the discussion of these points in the manuscript, and we have moreover elaborated on this in subsequent work (Schulze et al., 2025).

> *They compare to one more "sophisticated model" - RosettaddG - which should be more correlated with thermodynamic stability than other factors measured by VAMP-seq. However, the direct head-tohead comparison between their matrices and ddG is underdeveloped. How can this be used to dissect cases where thermodynamics are not contributing to specific substitution patterns OR in specific residues/regions that are predicted by one method better than the other? This would naturally dovetail into whether there is orthogonal information between these two that could be leveraged to create better predictions.*

We thank the reviewer for this suggestion and indeed had spent substantial effort trying to gain additional biological insights from variants for which MAVE scores or MAVE predictions do not match predicted ΔΔG values. One major caveat in this analysis is that the experimental MAVE scores, MAVE predictions and the predicted ΔΔG values are rather noisy, making it difficult to draw conclusions based on individual variants or even small subsets of variants.

In our revised manuscript, we have added an analysis to discover residue substitution profiles that are predicted most accurately either by a ΔΔG model or by our substitution matrix model, thereby avoiding analysis of individual variant effect scores.

We find that many substitution profiles are predicted equally well by the two model types, but also that there are residues for which one method predicts substitution effects better than the other method. We have added an analysis of the characteristics of the residues and

variants for which either the ΔΔG model or the substitution matrix model is most useful to rank variants. Since we only find relatively few residues for which this is the case, we do not expect a model that leverages predicted scores from both methods to perform better than ThermoMPNN across variants.

> *Perhaps beyond the scope of this baseline method, there is also ThermoMPNN and the work from Gabe Rocklin to consider as other approaches that should be more correlated only with thermodynamics.*

We acknowledge that there are other approaches to predict ΔΔG beyond Rosetta including for example ThermoMPNN and our own method called RaSP (Blaabjerg et al, eLIFE, 2023), and we have added comparisons to ThermoMPNN and RaSP in the revised manuscript. We are unsure how one would use the data from Rocklin and colleagues directly, but we note that e.g. RaSP has been benchmarked on this data and other methods have been trained on this data. We originally used Rosetta since the Rosetta model is known to be relatively robust and because it has never seen large databases during training (though we do not think that training of ThermoMPNN and RaSP would be biased towards the VAMP-seq data). We note also that we have previously compared both Rosetta calculations and RaSP with VAMP-seq data for TPMT, PTEN and NUDT15 (Blaabjerg et al, eLIFE, 2023)

> *I find myself drawn to the hints of a larger idea that outliers to this model can be helpful in identifying specific aspects of proteostasis. The discussion of S109 is great in this respect, but I can't help but feel there is more to be mined from Figure S9 or other analyses of outlier higher than predicted abundance along linear or tertiary motifs.*

We agree with these points and have previously spent substantial time trying to make sense of outliers in Figure S9 and Figure S18 (Figure S8 and Figure S18 of revised manuscript). The outlier analysis was challenging, in part due to the relatively high noise levels in both experimental data and predictions, and we did not find any clear signals. Some outliers in e.g. Figure S9 are very likely the result of dataset-specific abundance score distributions, which further complicates the outlier analysis. We now note this in the revised paper and hope others will use the data to gain additional insights on proteostasis-specific effects.

> ***Reviewer # 2 (Public review):***
>
> *Summary:*
>
> *This study analyzes protein abundance data from six VAMP-seq experiments, comprising over 31,000 single amino acid substitutions, to understand how different amino acids contribute to maintaining cellular protein levels. The authors develop substitution matrices that capture the average effect of amino acid changes on protein abundance in different structural contexts (buried vs. exposed residues). Their key finding is that these simple structure-based matrices can predict mutational effects on abundance with accuracy comparable to more complex physics-based stability calculations (ΔΔG).*
>
> *Major strengths:*
>
> *(1) The analysis focuses on a single molecular phenotype (abundance) measured using the same experimental approach (VAMP-seq), avoiding confounding factors present when combining data from different phenotypes (e.g., mixing stability, activity, and fitness data) or different experimental methods.*
>
> *(2) The demonstration that simple structural features (particularly solvent accessibility) can capture a significant portion of mutational effects on abundance.*
>
> *(3) The practical utility of the matrices for analyzing protein interfaces and identifying functionally important surface residues.*

We thank the reviewer for the comments above and the detailed assessment of our work.

> *Major weaknesses:*
>
> *(1) The statistical rigor of the analysis could be improved. For example, when comparing exposed vs. buried classification of interface residues, or when assessing whether differences between prediction methods are significant.*

We agree with the reviewer that it is useful to determine if interface residues (or any of the residues in the six proteins) can confidently be classified as buried- or exposed-like in terms of their substitution profiles. Thus, we have expanded our approach to compare individual substitution profiles to the average profiles of buried and exposed residues to now account for the noise in the VAMP-seq data. In our updated approach, we resample the abundance score substitution profile for every residue several thousand times based on the experimental VAMP-seq scores and score standard deviations, and we then compare every resampled profile to the average profiles for buried and exposed residues, thereby obtaining residue-specific distributions of $RMSD_{buried}$ and $RMSD_{exposed}$ values. These RMSD distributions are typically narrow, since many variants in several datasets have small standard deviations. In the revised manuscript, we report a residue to have e.g. a buried-like substitution profile if $RMSD_{buried} < RMSD_{exposed}$ for at least 95% of the resampled profiles. We do not recalculate average scores in substitution matrices for this analysis.

Moreover, to illustrate potential overlap in predictive performance between prediction methods more clearly than in our preprint, we have added confidence intervals in Fig. 2 and Fig. 3 of the revised manuscript. We note that the analysis in Fig. 2 is performed using a leave-one-protein-out approach, which we believe provides the cleanest assessment of how well the different models perform.

> *(2) The mechanistic connection between stability and abundance is assumed rather than explained or investigated. For instance, destabilizing mutations might decrease abundance through protein quality control, but other mechanisms like degron exposure could also be at play.*

We agree that we have not provided much description of the relation between stability and abundance in our original preprint. In the revised manuscript, we provide some more detail as well as references to previous literature explaining the ways in which destabilising mutations can cause degradation. We have moreover performed and added additional analyses of the relationship between thermodynamic stability and abundance through comparisons of stability predictions and predictions performed with our substitution matrix models.

> *(3) The similar performance of simple matrix-based and complex physics-based predictions calls for deeper analysis. A systematic comparison of where these approaches agree or differ could illuminate the relationship between stability and abundance. For instance, buried sites showing exposed-like behavior might indicate regions of structural plasticity, while the link between destabilization and degradation might involve partial unfolding exposing typically buried residues. The authors have all the necessary data for such analysis but don't fully exploit this opportunity.*

This is similar to a point made by reviewer 1, and our answer is similar. We were indeed hoping that our analyses would have revealed clearer differences between effects on thermodynamic protein stability and cellular abundance and have tried to find clear signals. One major caveat in performing the suggested analysis is that both the experimental MAVE scores, ΔΔG predictions and our simple matrix-based predictions are rather noisy, making it difficult to make conclusions based on individual variants or even small subsets of variants.

To address this point, we have added an analysis to discover residue substitution profiles that are predicted most accurately either by a ΔΔG model or by our substitution matrix model, thereby avoiding analysis of individual variant effect scores. We find that many substitution profiles are predicted equally well by the two model types, but we also, in particular, find solvent-exposed residues for which the substitution matrix model is the better predictor. These residues are often aspartate, glutamate and proline, suggesting that surface-level substitutions of these amino acid types often can have effects that are not captured well by a thermodynamical model, either because this model does not describe thermodynamic effects perfectly, or because in-cell effects are necessary to account for to provide an accurate description.

> *(4) The pooling of data across proteins to construct the matrices needs better justification, given the observed differences in score distributions between proteins (for example, PTEN's distribution is shifted towards high abundance scores while ASPA and PRKN show more binary distributions).*

We agree with the reviewer that the differences between the score distributions are important to investigate further and keep in mind when analysing e.g. prediction outliers. However, our results show that the pooling of VAMP-seq scores across proteins does result in substitution matrices that make sense biochemically and can identify outlier residues with proteostatic functions. As we also respond to a related point by reviewer 1, the differences in score distributions likely have complex origins. In that sense, we also hope that our results can inspire experimentalists to design methods to generate data that are more comparable across proteins.

For example, biophysical differences between the systems, such as differences in absolute folding stabilities or melting temperatures will play a role, as will the fact that some proteins contain multiple domains. Also, the sequence-based score for an individual variant in a sort-seq experiment (such as VAMP-seq) depends both on the properties of that variant and from the composition of the entire FACS-sorted cell library. This is because cells are sorted into bins depending on the composition of the entire library, which means that library-to-library composition can contribute to the differences between VAMP-seq score distributions. From our developed models and outliers in predictions from these, it is difficult to tell which of the several possible underlying reasons cause the differences.

Thus, even when experiments on different proteins are performed using the same technique (VAMP-seq), quantifying the same phenomenon (cellular abundance) and done in similar ways (saturation mutagenesis, sort-seq using four FACS bins), there can still be substantial differences in the results across different systems. An interesting side result of our work is to highlight this including how such variation makes it difficult to learn across experiments. We now elaborate on these points in the revised manuscript.

> *(5) Some key methodological choices require better justification. For example, combining "to" and "from" mutation profiles for PCA despite their different behaviors, or using arbitrary thresholds (like 0.05) for residue classification.*

We hope we have explained our methodological choices clearer in the revised paper.

We removed the dependency of the threshold of 0.05 used for residue classification in Fig. S19 of the original manuscript; in the revised manuscript we only report a residue to have e.g. a buried-like substitution profile if $RMSD_{buried} < RMSD_{exposed}$ for at least 95% of the abundance score profiles that we resampled according to VAMP-seq score noise levels, as explained above.

With respect to combining “to” and “from” mutational profiles for PCA, we could have also chosen to analyse these two sets of profiles separately to take potentially different behaviours

along the two mutational axes into account. We do not think that there should be anything wrong with concatenating the two sets of profiles in a single analysis, since the analysis on the concatenated profiles simply expresses amino acid similarities and differences in a more general manner.

> *The authors largely achieve their primary aim of showing that simple structural features can predict abundance changes. However, their secondary goal of using the matrices to identify functionally important residues would benefit from more rigorous statistical validation. While the matrices provide a useful baseline for abundance prediction, the paper could offer deeper biological insights by investigating cases where simple structure-based predictions differ from physics-based stability calculations.*
>
> *This work provides a valuable resource for the protein science community in the form of easily applicable substitution matrices. The finding that such simple features can match more complex calculations is significant for the field. However, the work's impact would be enhanced by a deeper investigation of the mechanistic implications of the observed patterns, particularly in cases where abundance changes appear decoupled from stability effects.*

We agree that disentangling stability and other effects on cellular abundance is one of the goals of this work. As discussed above, it has been difficult to find clear cases where amino acid substitutions affect abundance without stability beyond for example the (rare) effects of creating surface exposed degrons. Our new analysis, in which we compare substitution matrix-based predictions to stability predictions, does offer deeper insight into the relationship between the two predictor types and hence possibly between folding stability and abundance.

> ***Reviewer #3 (Public review):***
>
> *"Effects of residue substitutions on the cellular abundance of proteins" by Schulze and Lindorff-Larsen revisits the classical concept of structure-aware protein substitution matrices through the scope of modern protein structure modelling approaches and comprehensive phenotypic readouts from multiplex assays of variant effects (MAVEs). The authors explore 6 unique protein MAVE datasets based on protein abundance (and thus stability) by utilizing structural information, specifically residue solvent accessibility and secondary structure type, to derive combinations of context-specific substitution matrices predicting variant abundance. They are clear to outline that the aim of the study is not to produce a new best abundance predictor but to showcase the degree of prediction afforded simply by utilizing information on residue accessibility. The performance of their matrices is robustly evaluated using a leave-one-out approach, where the abundance effects for a single protein are predicted using the remaining datasets. Using a simple classification of buried and solvent-exposed residues, and substitution matrices derived respectively for each residue group, the authors convincingly demonstrate that taking structural solvent accessibility contexts into account leads to more accurate performance than either a structureunaware matrix, secondary structure-based matrix, or matrices combining both solvent accessibility or secondary structure. Interestingly, it is shown that the performance of the simple buried and exposed residue substitution matrices for predicting protein abundance is on par with Rosetta, an established and specialized protein variant stability predictor. More importantly, the authors finish off the paper by demonstrating the utility of the two matrices to identify surface residues that have buried-like substitution profiles, that are shown to correspond to protein interface residues, posttranslational modification sites, functional residues, or putative degrons.*
>
> *Strengths:*

> *The paper makes a strong and well-supported main point, demonstrating the utility of the authors' approach through performance comparisons with alternative substitution matrices and specialized methods alike. The matrices are rigorously evaluated without introducing bias, exploring various combinations of protein datasets. Supplemental analyses are extremely comprehensive and detailed. The applicability of the substitution matrices is explored beyond abundance prediction and could have important implications in the future for identifying functionally relevant sites.*

We thank the reviewer for the supportive comments on our work.

> *Comments:*
>
> *(1) A wider discussion of the possible reasons why matrices for certain proteins seem to correlate better than others would be extremely interesting, touching upon possible points like differences or similarities in local environments, degradation pathways, posttranslation modifications, and regulation. While the initial data structure differences provide a possible explanation, Figure S17A, B correlations show a more complicated picture.*

We agree with the reviewer that biochemical and biophysical differences between the proteins might contribute to the fact that some matrices correlate better than others. We also agree that it would be very interesting to understand these differences better. While it might be possible to examine some of the suggested causes of the differences, like differences or similarities in local environments, we have generally found that noise and differences in score distributions make such analyses difficult (see also responses to reviewers 1 and 2). For now, we will defer additional analyses to future work.

> *(2) The performance analysis in Figure 2D seems to show that for particular proteins "less is more" when it comes to which datasets are best to derive the matrix from (CYP2C9, ASPA, PRKN). Are there any features (direct or proxy), that would allow to group proteins to maximize accuracy? Do the authors think on top of the buried vs exposed paradigm, another grouping dimension at the protein/domain level could improve performance?*

We don't currently know if any protein- or domain-level features could be used to further split residues into useful categories for constructing new substitution matrices, but it is an interesting suggestion. We note that every substitution matrix consists of 380 averages, and creating too many residue groupings will cause some matrix entries to be averaged over very few abundance scores, at least with the current number of scores in the pooled VAMP-seq dataset. For example, while previous work has shown different mutational effects e.g. in helices and sheets (as one would expect), we find that a model with six matrices ({buried,exposed}x{helix,sheet,other}) does not lead to improved predictions (Fig. 2C), presumably because of an unfavourable balance between parameters and data.

> *(3) While the matrices and Rosetta seem to show similar degrees of correlation, do the methods both fail and succeed on the same variants? Or do they show a degree of orthogonality and could potentially be synergistic?*

These are good questions and are related to similar questions from reviewers 1 and 2. In the revised manuscript, we have added additional analyses of differences between predictions from our substitution matrix model and a stability model, and we indeed find that the two methods show a degree of orthogonality. However, since we identify only relatively few residues for which one method performs better than the other, we don't expect a synergistic model to outperform the stability predictor across all variants in any of the six proteins.

> *Overall, this work presents a valuable contribution by creatively utilizing a simple concept through cutting-edge datasets, which could be useful in various.*

> ***Reviewing Editor:***
>
> *As discussed in more detail below, to strengthen the assessment, the authors are encouraged to:*
>
> *(1) Include more thorough statistical analyses, such as confidence intervals or standard errors, to better validate key claims (e.g., RMSD comparisons).*
>
> *(2) Perform a deeper comparison between substitution response matrices and ΔΔG-based predictions to uncover areas of agreement or orthogonality*
>
> *(3) Clarify the relationship between structural features, stability, and abundance to provide more mechanistic insights.*

As discussed above and below, we have added new analyses and clarifications to the revised manuscript.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *Minor points:*
>
> *Why is a continuous version of the contact number used here, instead of a discrete count of neighbouring residues? WCN values of the residues in the core domain can be affected by residues far away (small contribution but not strictly zero; if there are many of them, it adds up).*

We have previously found WCN, which quantifies residue contact numbers in a continuous manner, to be a useful input feature for a classifier that determines whether individual residues are important for maintaining protein abundance or function (Cagiada et al, 2023). We have also found WCN and the cellular abundance of single substitution variants to correlate well in individual analyses of different proteins (Grønbæk-Thygesen et al., 2024; Gersing et al., 2024; Clausen et al., 2024).

We have calculated the WCN as well as a contact number based on discrete counts of neighbouring residues for the six proteins in our dataset. When distances between residues are evaluated in the same way (i.e. using the shortest distance between any pair of heavy atoms in the side chains), and when the cutoff value used for the discrete count is equal to the r_0 of the WCN function, the continuous and discrete evaluations of residue contact numbers are highly and linearly correlated, and their rank correlation with the VAMP-seq data are very similar. We only observe minor contributions from residues far away in the structure on the WCN.

> *Typos in SI figure captions e.g. Figure S8-11 "All predictions were performed using using...."*

Thank you for pointing this out. We have corrected the typos in Figure S8-11 (Figure S7-S10 in the revised manuscript).

> *Personally, I'd appreciate a definition of these new substitution matrices under the constraints of rASA/WCN values. It was unclear to me until I read the code but we think that the definition is averaging the substitution matrix based on the clusters they are assigned to. If so, this could be straightforwardly defined in the method section with a heaviside step function.*

We have added a definition of the “buried” and “exposed” substitution matrices as a function of rASA in the methods section (“Definitions of buried and exposed residues” and “Definition of substitution matrices”) of the manuscript, as well as a definition of how we classified residues as either buried or exposed using both rASA and WCN as input. Our final

substitution matrices, as shown in e.g. Fig. 2, do not depend on the WCN; only the substitution matrix results in Figure S6 (Figure S20 in the revised manuscript) depend on both WCN and rASA.

> ***Reviewer #2 (Recommendations for the authors):***
>
> *The following suggestions aim to strengthen the analysis and clarify the presentation of your findings:*
>
> *(1) Specific analyses to consider:*
>
> *(1.1) Analyze buried positions where the exposed matrix performs better. Understanding these cases might reveal properties of protein core regions that show unexpected mutational tolerance.*

We agree with the reviewer that a more detailed analysis of buried residues with exposed-like substitution profiles would be very interesting.

We note that for proteins where the VAMP-seq score distribution is shifted towards high values (as it is the case for PTEN, TPMT and CYP2C9), our identification of such residues may be a result of the score distribution differences between the six datasets. To confidently identify mutationally tolerant core regions, it would be best to (a) correct for the distribution differences prior to the analysis or (b) focus the analysis on residues that fall far below the diagonal in Figure S18.

In additional data (which can be found at *https://github.com/KULL-Centre/_2024_Schulze_abundance-analysis)* ,we provide, for each of the proteins, a list of buried residues for which $RMSD_{exposed}$ <$RMSD_{buried}$ (for more than 95% of resampled substitution profiles, as described under 1.6). We have not analysed these residues further.

> *(1.2) A systematic comparison of matrix-based vs. ΔΔG-based predictions could help understand both exposed sites that behave as buried (as analyzed in the paper) and buried sites that behave as exposed (1.1), potentially revealing mechanisms underlying abundance changes.*

In our revised manuscript, we have added additional analyses to compare matrixbased and ΔΔG-based predictions, focusing on exposed sites for which one prediction method captures variant effects on abundance considerably better the other prediction method. We have not investigated buried sites with exposed-like behaviour any further in this work.

> *(1.3) Explore different normalization approaches when pooling data across proteins. In particular, consider using log(abundance score): if the experimental error in abundance measurements is multiplicative (which can be checked from the reported standard errors), then log transformation would convert this into a constant additive error, making the analysis more statistically sound.*

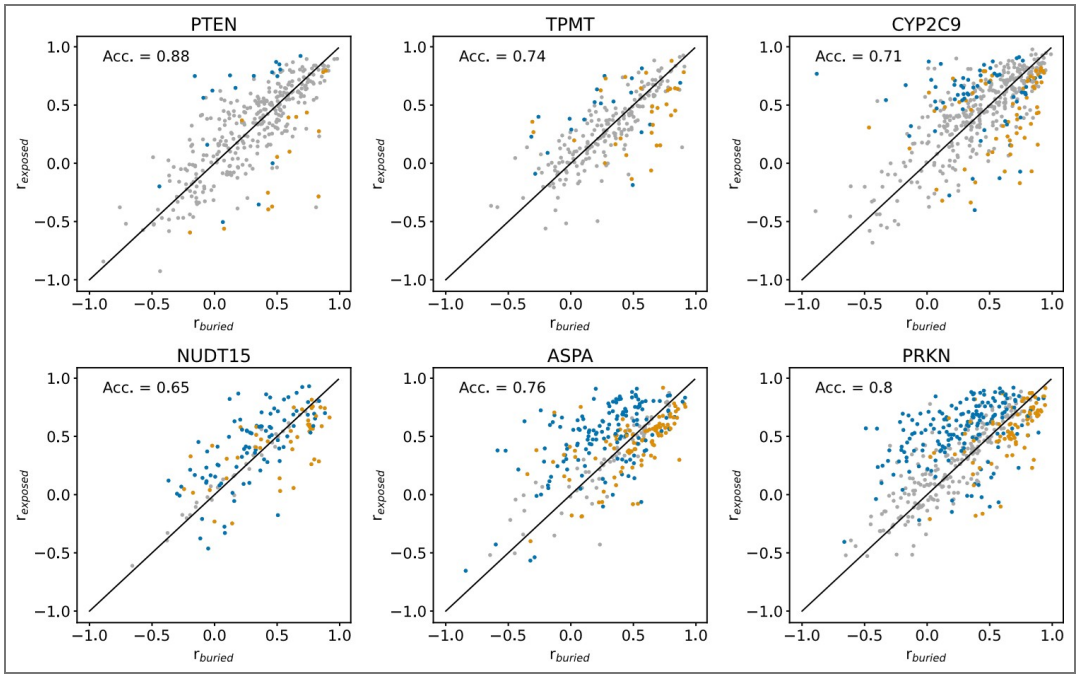
As we answer below to point 2.2, the abundance scores are, within each dataset, min-max normalised to nonsense and synonymous variant scores, and the score scale is thus in this way consistent across the six datasets. We have explained above and in the revised manuscript that abundance score distribution differences across datasets are likely partially a result of the FACS binning of assay-specific variant libraries. Using only the VAMP-seq scores (that is, without further information about the individual experiments), we cannot correct for the influence of the sorting strategy on the reported scores. A score normalisation across datasets that places all data points on a single scale would require inter-dataset references variant scores, which we do not have. We note that in a subsequent manuscript (Schulze et al, bioRxiv, 2025) we have attempted to take system- and experimentspecific score distributions into account. We now refer to this work in the revised manuscript.

> *(1.4) Consider using correlation coefficients between predicted and observed abundance profiles as an alternative to RMSD, which is sensitive to the absolute values of the scores.*

We agree with the reviewer that using correlation coefficients to compare substitution profiles might also be useful, in particular for datasets with relatively unique VAMP-seq score distributions, such as the ASPA dataset. To explore this idea, we have repeated the analysis presented in Fig. S18 using the Pearson correlation coefficient r rather than the RMSD.

As in Fig. S18, we derive r_{buried} and $r_{exposed}$ for every residue in the six proteins, specifically by calculating r between the abundance score substitution profile of every individual residue and the average abundance score substitution profiles of buried and exposed residues. VAMP-seq data for the protein for which r_{buried} and $r_{exposed}$ are evaluated is omitted from the calculation of average abundance score substitution profiles, and we use only monomer structures to determine whether residues are buried or exposed.

We show the results of this analysis in an Author response image 1 below. In each panel of the figure, r_{buried} and $r_{exposed}$ are shown for individual residues of a single protein. Blue datapoints indicate residues that are solvent-exposed in the wild-type protein structures, and yellow datapoints indicate residues that are buried in the wild-type structures. Residues for which it is not the case that $r_{buried} < r_{exposed}$ or $r_{exposed} < r_{buried}$ in more than 95% of 1000 resampled residue substitution profiles (see explanation of resampling method above) are coloured grey. "Acc." is the balanced classification accuracy, calculated using all non-grey datapoints, indicating how many buried residues have buried-like substitution profiles ($r_{exposed} < r_{buried}$) and how many solvent-exposed residues have exposed-like substitution profiles ($r_{buried} < r_{exposed}$). The classification accuracy per protein in this figure cannot be compared to the classification accuracy of the same protein in Fig. S18, since the number of datapoints used in the accuracy calculation differ between the r- and RMSD-based analyses.



Author response image 1.

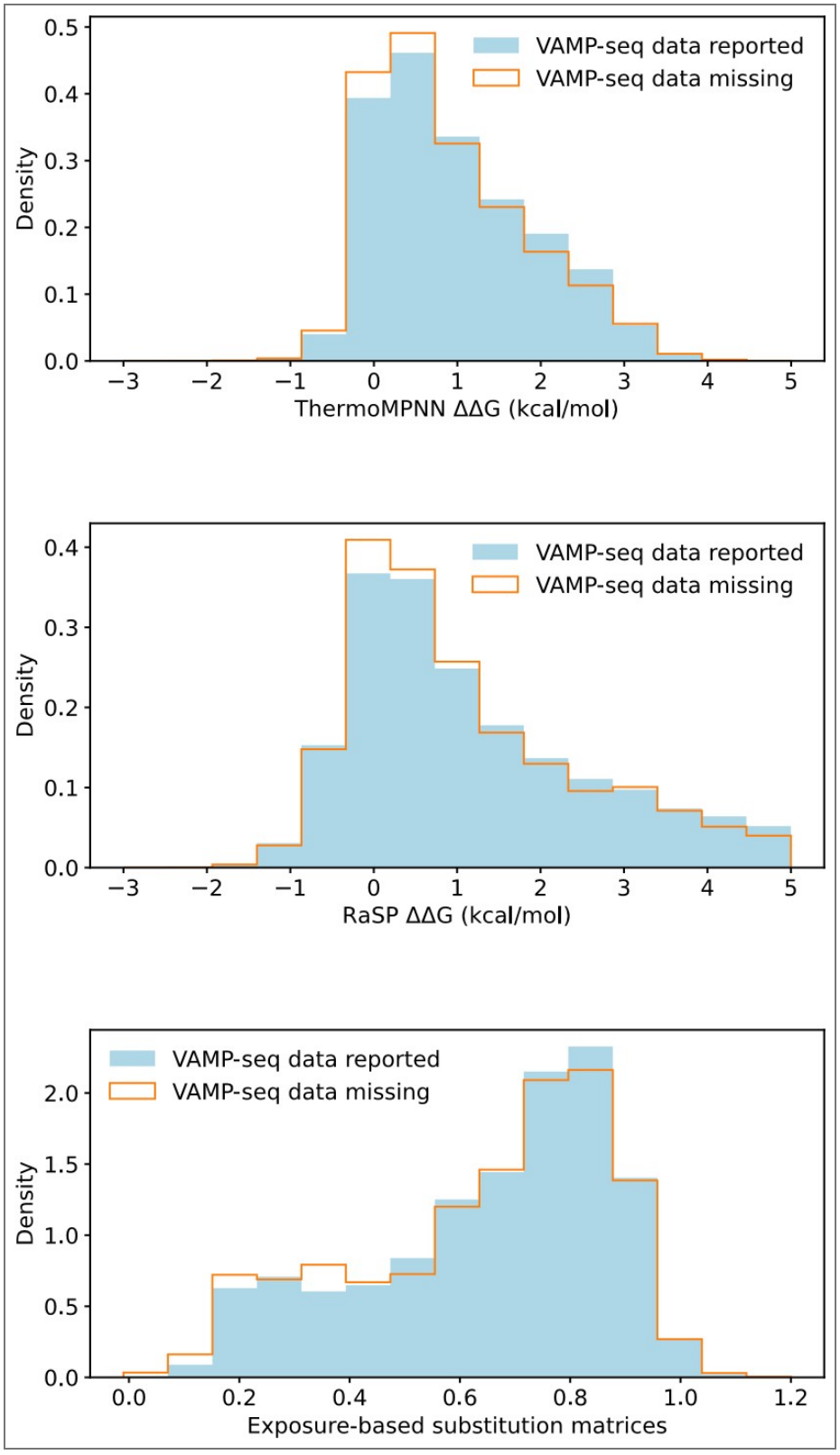
Comparing the r-based approach to the RMSD-based approach (Fig. S18), it is clear that the r-based method is less robust than the RMSD-based method for noisy and incomplete datasets. For the noisiest and most mutationally incomplete VAMP-seq datasets (i.e., PTEN, TPMT and

CYP2C9) (Fig. 1), there are relatively few residues for which we with high confidence can determine if the substitution profile is more buried- or more exposed-like. When the VAMP-seq data is less noisy and has high mutational completeness, the r-based method becomes more robust and may thus be relevant in potential future work on new VAMP-seq data with small error bars.

In conclusion, we find that RMSD-based approach to compare substitution profiles is more robust than an r-based approach for several of the VAMP-seq datasets that are included in our analysis. We do believe than an approach based on the correlation coefficient, or potentially several metrics, could be relevant to use, since abundance score distributions from VAMP-seq datasets can differ significantly across datasets. So as not to increase the length of the main text of our manuscript, we have not added this analysis to the revised manuscript.

> *(1.5) Consider treating missing abundance scores as zero values, as they might indicate variants with very low abundance, rather than omitting them from the analysis.*

This suggestion would be most relevant for the PTEN, TPMT and CYP2C9 datasets, which all have a relatively small average mutational depth and completeness, as shown in Fig. 1B and 1C. To assess if setting missing abundance scores as zero values would be reasonable, we have compared the distributions of predicted ΔΔG values (from RaSP and ThermoMPNN) and of predicted abundance scores (from our exposure-based substitution matrices) for variants with reported and missing VAMP-seq data. We show the result in Author response image 2, with data aggregated across the six protein systems:



Author response image 2.

We find that variants with and without VAMP-seq data have similar ΔΔG score distributions and similar predicted abundance score distributions, and there is thus no clear enrichment of predicted loss of abundance for variants with missing VAMP-seq scores. This suggests that

missing abundance scores do not necessarily indicate very low abundance. One cause of missing data might instead be problems with library generation (Matreyek et al, 2018, 2021).

We show in Fig. S9 (Fig. S8 of the revised manuscript) that predicted scores for variants with experimental abundance scores of 0 are often overestimated for NUDT15, ASPA and PRKN, but this is not so much a problem for PTEN, TMPT and CYP2C9, the datasets with most missing scores. The lack of an enrichment of low abundance variants from the various predictors would thus still support that missing scores do not necessarily indicate low abundance.

> *(1.6) Develop a proper statistical framework for comparing buried vs exposed predictions (whether using RMSD or correlations), including confidence intervals, rather than using arbitrary thresholds.*

As explained above and in the methods section of our revised manuscript, we have expanded our approach to compare the substitution profile of a residue to the average profiles of buried and exposed residues, and our method now accounts for the noise in the VAMP-seq data, making the analysis more statistically rigorous. In our expanded approach, we compare the substitution profiles of individual residues to the average profiles for buried and exposed residues 10,000 times per residue to get a residue-specific distribution of $RMSD_{buried}$ and $RMSD_{exposed}$ values. Individual $RMSD_{buried}$ and $RMSD_{exposed}$ values are calculated by resampling abundance scores from a Gaussian distribution defined by the experimentally reported abundance score and abundance score standard deviation per variant. We now only report a residue to have e.g. a buried-like substitution profile if $RMSD_{buried} < RMSD_{exposed}$ in at least 95% of our samples. We do not recalculate average scores in substitution matrices for this analysis. We have updated the plots in our manuscript, e.g. in Fig. S18 and S19 of the revised version, to indicate which residues are confidently classified as buried- or exposed-like.

> *(2) Presentation improvements:*
>
> *(2.1) In Figure 4, consider removing the average abundance scores, which are not directly related to the RMSD comparison being shown.*

We have decided to keep the average abundance scores in Fig. 4 (now Fig. 5), as we find the average abundance scores useful for guiding interpretation of the RMSD values. For example, an unusually small average abundance score with a relatively small standard deviation may explain a case where $RMSD_{buried}$ and $RMSD_{exposed}$ are both large. This is for example the case for residue G185 in ASPA.

In our preprint, the error bars on the average abundance scores in Fig. 4 (now Fig. 5) indicated the standard deviation across the abundance scores that were used to calculate the average per position. We have removed these error bars in the revised manuscript, as we realised that these were not necessarily helpful to the reader.

> *(2.2) I am assuming that abundance scores are defined as the ratio abundance_variant/abundance_wt throughout the analysis, but I don't think this has been explicitly defined. If this is correct, please state it explicitly. In such case, log(abundance_score) would have a simple interpretation as the difference in abundance between variant and wild-type.*

Abundance scores are defined throughout the manuscript as sequence-based scores that have been min-max normalised to the abundance of nonsense and synonymous variants, i.e. abundance_score = (abundance_variant abundance_nonsense)/(abundance_wt–abundance_nonsense). We have described the normalisation of scores to wild-type and nonsense variant abundance in lines 164-166 of the original manuscript. We have now added additional information about the normalisation scheme in the methods section. We note that

we did not ourselves apply this normalisation to the data; the scores were reported in this manner in the original publications that reported the VAMP-seq experiments for the six proteins.

> *(2.3) Consider renaming "rASA" to the more commonly used "RSA" for relative solvent accessibility.*

We have decided to keep using “rASA” throughout the manuscript.

> *(2.4) The weighted contact number function used differs from the established WCN measure (Σ1/rij²) introduced by Lin et al. (2008, Proteins). This should be acknowledged and the choice of alternative weighting scheme justified.*

As we have also responded to the first minor point of reviewer 1, we have previously found WCN, as it is defined in our manuscript, to be a useful input feature for a classifier that determines whether individual residues are important for maintaining protein abundance or function (Cagiada et al, 2023). We have also previously found this type of WCN to correlate well with variant abundance of individual proteins, as measured with VAMP-seq or protein fragment complementation assays (Grønbæk-Thygesen et al., 2024; Clausen et al., 2024; Gersing et al., 2024). We acknowledge that residue contact numbers or weighted contact numbers could also be expressed in other ways and that alternative contact number definitions would likely also produce values that correlate well with VAMP-seq data. Since the WCN, as defined in our manuscript, already correlates relatively well with abundance scores, we have not explored whether alternative definitions produce better correlations.

> *(2.5) Replace the phrase "in the above" with specific references to sections or simply "above" where appropriate. Also, consider replacing many instances of "moreover" with simpler alternatives such as "also" or "in addition" to improve readability.*

We have changed several sentences according to this suggestion and hope that we have improved the readability of our manuscript.

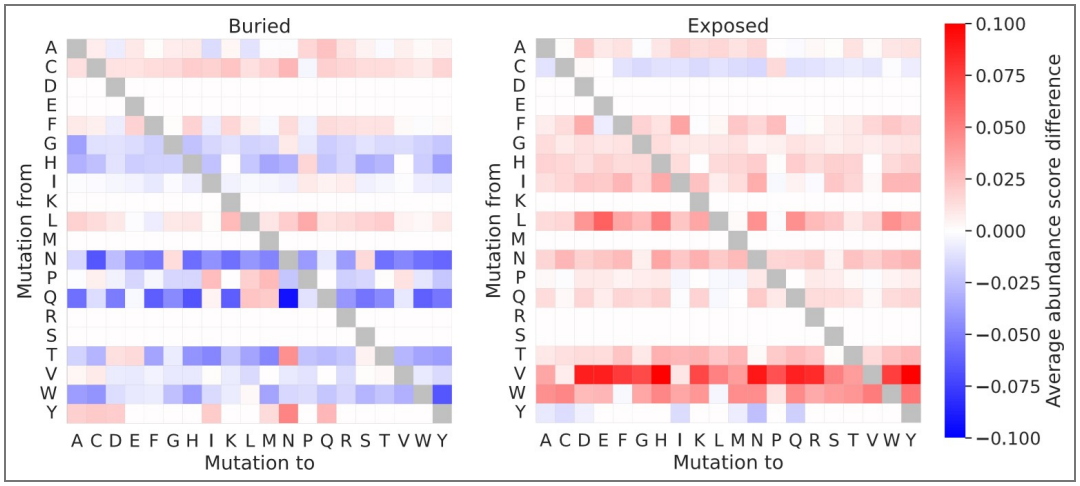
> ***Reviewer #3 (Recommendations for the authors):***
>
> *(1) It should be explicitly confirmed earlier that complex structures are used for NUDT15 and ASPA when assessing rASA/WCN. Additionally, it would be interesting to see the effect that deriving the matrices using NUDT15 and ASPA monomers would have.*

We have commented on the use of NUDT15 and ASPA homodimer structures earlier in the revised manuscript (specifically already in the subsection Abundance scores correlate with the degree of residue solvent-exposure section).

When residues are classified using monomer rather than dimer structures of NUDT15 and ASPA, there is a small effect on the resulting “buried” and “exposed” substitution matrices. Entries in this set of substitution matrices calculated using either monomer or dimer structures typically differ by less than 0.05, and only a single entry differ by more than 0.1. As expected, the “exposed” matrix tend to contain slightly larger numbers when derived from dimer structures than when derived from monomer structures, meaning that when the interface residues are included in the exposed residue category, the average abundance scores of the “exposed” matrix are lowered. For buried residues, the picture is more mixed, although the overall tendency is that the interface residues make the “buried” matrix contain smaller average abundance scores for dimer compared to monomer structures. These results generally support the use of dimer structures for the residue classification.

We here show the differences between the substitution matrices calculated with dimer or monomer structures of NUDT15 and ASPA and using data for all six proteins in our combined

VAMP-seq dataset (average_abundance_score_differece = average_abundance_score_dimers – average_abundance_score _monomers):



Author response image 3.

We have not explored these alternative matrices further.

> *(2) While the supplemental analyses are rigorous, the abundance of various metrics being presented can be confusing, especially when they seem to differ in their result. For instance, the discussion of Figure S17 (paragraph starting 428) contains mentions of mean differences but then switches to correlations, while both are presented for all panels. The claim "The datasets thus mainly differ due to differences in substitution effects in buried environments. " is well supported by the observed mean differences, but for Pearson's correlations the average panel A ,B values of buried 0.421 vs exposed 0.427 are hardly different. Which of the metrics is more meaningful, and are both needed?*

We agree with the reviewer that the claim that “The datasets thus mainly differ due to differences in substitution effects in buried environments” is not well-supported by the r between the substitution matrices, and we have removed this claim from the text.

Since some datasets share VAMP-seq score distribution features, while others do not, the absolute difference between scores or matrices may be relevant to check for some dataset pairs, while the r may be more relevant to check for other dataset pairs. Hence, we have included both metrics in Fig S17 (Fig S11 in the revised manuscript).

> *(3) Lines 337-340 - does not feel like S7 is the topic, perhaps the authors meant Figure 2A, B? In general, the supplemental figure references are out of order and panel combinations are sometimes confusing.*

We have corrected figures references to now be correct and changed the arrangement of supplemental figures so that they now occur in the correct order. We have looked through the panel combinations with clarity in mind, and hope that the current set of main and supplementary figures balances overview and detail.

> *(4) Line 363 "are also are also".*

We have corrected this typo.

https://doi.org/10.7554/eLife.103721.2.sa0