Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance

Matteo Cagiada[1], **Kristoffer E. Johansson**[1], **Audronė Valančiūtė**[1], **Sofie V. Nielsen**[1], **Rasmus Hartmann-Petersen**[1], **Jun J. Yang**[2], **Douglas M. Fowler**[3,4,5], **Amelie Stein**[1], **Kresten Lindorff-Larsen**[1,*]

***For correspondence:**
lindorff@bio.ku.dk (KLL)

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; [2]Department of Pharmaceutical Sciences and Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, US; [3]Department of Genome Sciences, University of Washington, Seattle, WA 98195, US; [4]Department of Bioengineering, University of Washington, Seattle, WA 98195, US; [3]Genetic Networks Program, CIFAR, Toronto, ON M5G 1M1, Canada

Abstract Understanding and predicting how amino acid substitutions affect proteins is key to practical uses of proteins, and to our basic understanding of protein function and evolution. Amino acid changes may affect protein function in a number of ways including direct perturbations of activity or indirect effects on protein folding and stability. We have analysed 6749 experimentally determined variant effects from multiplexed assays on abundance and activity in two proteins (NUDT15 and PTEN) to quantify these effects, and find that a third of the variants cause loss of function, and about half of loss-of-function variants also have low cellular abundance. We analyse the structural and mechanistic origins of loss of function, and use the experimental data to find residues important for enzymatic activity. We performed computational analyses of protein stability and evolutionary conservation and show how we may predict positions where variants cause loss of activity or abundance.

Introduction

Protein engineering and mutational analysis have provided us with a wealth of information about the molecular interactions that stabilize proteins and govern their functions (***Fersht, 1999***). This information has in turn enabled us to engineer proteins with improved activities and stability, and to better understand how mutations cause disease (***Stein et al., 2019***).

Computational analyses of missense variants in genetic diseases have suggested that loss of function via loss of protein stability is a major cause of disease (***Wang and Moult, 2001***; ***Ferrer-Costa et al., 2002***; ***Steward et al., 2003***; ***Yue et al., 2005***; ***Casadio et al., 2011***; ***Gao et al., 2015***; ***Stein et al., 2019***) because unstable proteins either aggregate or become targets for the cell's protein quality control apparatus and are degraded (***Nielsen et al., 2020***). Indeed, cellular studies of disease-causing variants in a number of genes have shown that many variants are degraded in the cell (***Meacham et al., 2001***; ***Yaguchi et al., 2004***; ***Olzmann et al., 2004***; ***Ron and Horowitz, 2005***;

Yang et al., 2011, ***2013***; ***Arlow et al., 2013***; ***Nielsen et al., 2017***; ***Chen et al., 2017***; ***Matreyek et al., 2018***; ***Scheller et al., 2019***; ***Abildgaard et al., 2019***; ***Suiter et al., 2020***). For this reason, several methods for predicting and understanding disease-causing variants include predictions of changes in protein stability (***Yue et al., 2005***; ***De Baets et al., 2012***; ***Casadio et al., 2011***; ***Ancien et al., 2018***; ***Wagih et al., 2018***; ***Gerasimavicius et al., 2020***). While stability-based predictions can be relatively successful and may provide mechanistic insight into the origins of disease, it is also clear that variants can cause disease via other mechanisms such as removing key residues in an active site or perturbing interactions or regulatory mechanisms. Thus, methods used to predict the pathogenicity of missense variants often combine analysis of sequence conservation with information on protein structure and stability and other sources of information (***Kumar et al., 2009***; ***Adzhubei et al., 2010***; ***De Baets et al., 2012***; ***Kircher et al., 2014***; ***Choi and Chan, 2015***; ***Ioannidis et al., 2016***).

In order to understand better the relationship between protein stability, abundance and function we here asked the question of what fraction of single amino acid changes in a protein causes loss of function via loss of stability and cellular abundance of the proteins. Until recently, mutational analyses of proteins have mostly relied on a one-by-one approach in which individual amino acid changes are introduced and effects on various properties of a protein are tested—often using *in vitro* experiments on purified proteins. Such experiments can now be complemented by experiments that simultaneously probe the effects of thousands of variants in a single assay. Such *multiplexed assays of variant effects* (MAVEs, also often termed deep mutational scans) are based on developments in high-throughput DNA synthesis, functional assays and sequencing techniques (***Kinney and McCandlish, 2019***). Briefly, a selection procedure (e.g. for growth rate or a fluorescent reporter of a protein property) is applied to a large library of variants, each expressed in individual cells. Variants change in frequency depending on how they perform under the conditions of the selection, and the frequency of each variant before and after the selection is determined using next-generation DNA sequencing. Changes in variant frequency are used to compute a score that describes each variant's effect on the property under selection. Such data can be used as an input to protein engineering (***Araya et al., 2012***; ***Shin and Cho, 2015***), or to elucidate genotype-phenotype relationships and understand how mutations may cause disease (***Starita et al., 2015***; ***Weile and Roth, 2018***; ***Stein et al., 2019***).

Now, for the first time, we have available measurements of thousands of variant effects on two key protein properties, activity and abundance, measured in multiple proteins. Here we take advantage of these data to examine more broadly how substitutions affect activity and stability. We examine how variants may affect abundance and activity differently to find functionally important positions in proteins (***Chiasson et al., 2020***), and to understand whether different types of effects are found in different regions of a protein's structure.

To do so, we here analyse two different types of MAVEs that probe different aspects of protein function. As subjects of our study we have chosen two medically relevant human proteins, PTEN (phosphatase and tensin homolog) and NUDT15 (nucleoside diphosphate-linked to x hydrolase 15), because for both of these proteins multiplexed functional data exist from two different assays: One measuring the effect of variants on the activity of the protein via a growth rate (***Mighell et al., 2018***) or drug-sensitivity (***Suiter et al., 2020***) phenotype, and an assay that probes the effects of amino acid changes on cellular abundance (***Matreyek et al., 2018***; ***Suiter et al., 2020***). We will sometimes refer to the abundance data as reporting on 'stability' and the growth-based activity data as 'activity' or 'function', recognizing that the experiments report on a complex interplay of effects during the experimental assays. Notably, low scores in the activity-based assays might occur both due to loss of intrinsic enzymatic function, but also e.g. due to decreased protein abundance. Indeed, we use the complementary information on protein abundance to disentangle effects on abundance and intrinsic activity.

PTEN is a 403 amino-acid lipid phosphatase expressed throughout the human body and mutations have been associated with cancer and autism spectrum disorders (***Yehia et al., 2019***). In mice, PTEN has been shown to suppress tumor development via dephosphorylation of phosphatidylinositol lipids, although *in vitro* PTEN has been shown to have a broader range of substrates including proteins. PTEN is composed of two domains: a catalytic tensin-like domain (residues 14-185) and a C2 domain (residues 190-350) that mediates membrane recruitment (***Lee et al., 1999***). The C-terminal region of PTEN is disordered with a PDZ-domain binding region (residue 401-403) (***Valiente et al., 2005***). Our analysis of PTEN includes a MAVE that probes the effects of most single amino acid substitutions when assayed for lipid phosphatase activity in yeast (***Mighell et al., 2018***), whose growth had been made dependent on the ability of PTEN to catalyse the formation of essential phosphatidylinositol bisphosphate (PIP2) from its triphosphate (PIP3). We complement these data with results from a different MAVE in which variant effects on cellular abundance is determined in an experiment termed 'variant abundance by massively parallel sequencing' (VAMP-seq) (***Matreyek et al., 2018***). In VAMP-seq the steady state abundance of protein variants in cultured mammalian cells is detected by fusion to a fluorescent protein, and cells are sorted using fluorescent activated cell sorting. Our analysis here covers the 56% of all possible single amino acid variants in PTEN for which we have measurements for both the activity and abundance, and thus complements our recent analysis of a small number of disease variants in PTEN (***Jepsen et al., 2020***).

NUDT15 is a nucleotide triphosphate diphosphatase that consists of 164 amino acids in a nudix hydrolase domain featuring a conserved nudix box that coordinates the catalytic Mg^{2+}. The biologically relevant assembly is reported to be a homodimer although the monomer also has catalytic activity (***Carter et al., 2015***). NUDT15 deficiency is associated with intolerance to thiopurine drugs (***Yang et al., 2014***; ***Moriyama et al., 2016***, ***2017***; ***Nishii et al., 2018***), which are widely used in the treatments of leukemia and autoimmune diseases (***Karran and Attard, 2008***). Thiopurines are a class of anti-metabolite drugs that form the active metabolite, thio-dGTP, which competes with dGTP and causes apoptosis when incorporated extensively into DNA. NUDT15 hydrolyses thio-dGTP and thus negatively regulates the levels and cytotoxic effects of thiopurine metabolites. Therefore, NUDT15 variants that decrease function are a major cause of toxicity during thiopurine therapy, and thus the dose of the drug may be personalized to match the metabolism of these compounds (***Relling et al., 2019***). The high drug sensitivity of cells with compromised NUDT15 function has been used in a MAVE to assay 95% of all single amino acid variants for causing intolerance towards thiopurine drugs (***Suiter et al., 2020***). The same library and cells were also used in a VAMP-seq experiment to probe variant effects on cellular abundance.

Here we have analysed the effect of variants on activity and cellular abundance in both PTEN and NUDT15 to provide a global view of what fraction of variants cause substantial loss of activity in the cell, and what fraction of these variants do so via loss of protein abundance. We find that approximately one third of all variants cause loss of protein activity, and that about half of these do so most likely because of loss of protein abundance. Variants that cause loss of abundance are often found inside the protein core, while variants that cause loss of activity without affecting abundance are often found in functionally important positions including those involved in catalysis or that interact with substrates. We also find that we can predict rather accurately the positions where substitutions generally give rise to decreased abundance and activity, whereas it remains difficult to quantitatively predict the effects of individual variants.

Results and Discussion

Global analysis of variant effects

We collected data from multiplexed assays reporting on both the activity and abundance of a total of 2822 variants in NUDT15 (***Suiter et al., 2020***) and 3927 variants in PTEN (***Matreyek et al., 2018***; ***Mighell et al., 2018***) (Fig. S1). Scripts to repeat our analyses are available online at github.com/KULL-Centre/papers/tree/master/2020/mave-analysis-cagiada-et-al. Two-dimensional histograms reveal that most variants have high scores in both assays, indicating wild type-like abundance and activity under the conditions of the cellular assays (Figs. 1A and B).

In order to separate wild-type like variants from those with decreased activity and/or abundance,

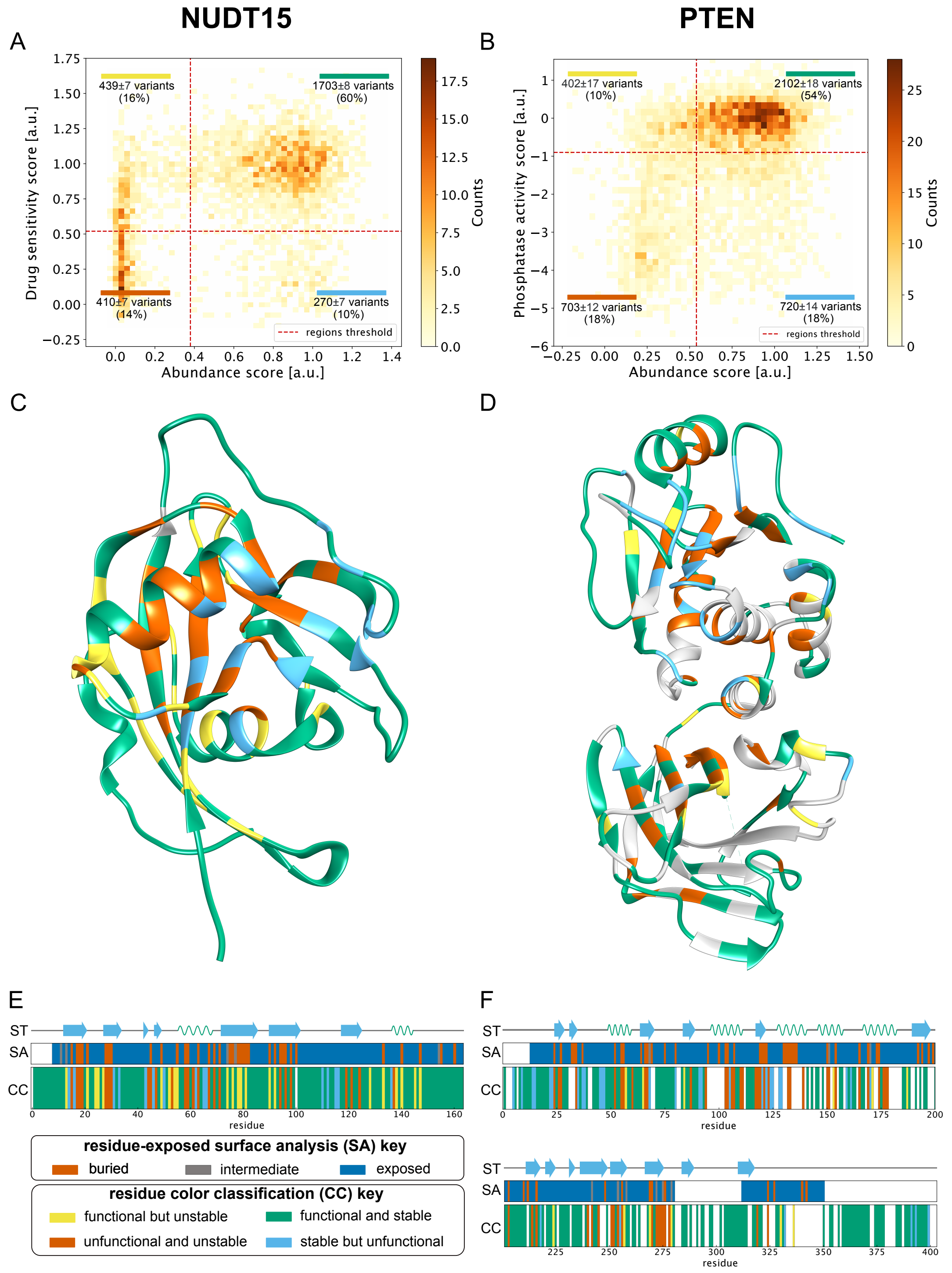


Figure 1. Overview of the NUDT15 and PTEN multiplexed data analysed in this work. A and B show 2D histograms that combine the data from the activity-based MAVE on the y-axis with the results from the VAMP-seq experiment on the x-axis. Variants are categorised based on the region of the 2D histogram (dashed lines) they belong to. The fractions of variants falling in each of the four quadrants are indicated, with errors of the mean estimated by bootstrapping using the uncertainties of the experimental scores. Panels C and D show a per-position consensus category (CC) coloured onto the structure of the proteins (PDB entry 5LPG for NUDT15 and 1D5R for PTEN). Panels E and F show the positional colour categories together with the secondary structure (ST) and solvent accessibility (SA).

we define a threshold value for all scores (Fig. S2). These thresholds define four classes of variants according to whether the variant showed high or low scores in the activity-based and abundance MAVEs. For simplicity each class is also associated with a colour. 'WT-like' variants had wild-type-like activity and abundance and are shown in green. 'Low activity, high abundance' variants had WT-like abundance but low activity in the assays, and are shown in blue. 'Low abundance, high activity' variants had WT-like activity but low abundance in the assays and are shown in yellow. 'Total loss' variants had low activity and low abundance and are shown in red.

For both proteins, the majority of variants are wild-type-like (60% for NUDT15 and 54% for PTEN; Figs. 1A and B; green). The total-loss category represents variants that both show loss of activity and low cellular abundance (14% for NUDT15 and 18% for PTEN; Figs. 1A and B; red), and as discussed further below we expect that most of these variants lose activity because of their low abundance. Of the total of 680 and 1403 variants with low activity in NUDT15 and PTEN respectively, 60% and 50% lose activity together with loss of abundance. The low activity, high abundance variants are still abundant in the cell but inactivated by other means, e.g. by changes in amino acids in the active site (Figs. 1A and B; blue). The low abundance, high activity class, which contains 16% of NUDT15 and 10% of PTEN variants (Figs. 1A and B; yellow), show low abundance levels, but high levels of activity in the activity-based assay and are not as easily explained by a single mechanism.

To focus our analysis on different types of variant effects in different parts of the protein structure we converted the variant data into positional categories that represent the most frequent class (also represented with the same names and colours as for the variants) among the variants at that position. We performed this classification procedure at all positions with more than five tested variants (99% for NUDT15 and 88% for PTEN). This results in 62% and 60% positions classified as WT-like for NUDT15 and PTEN respectively (Figs. S1C and D; green). On the other hand, at 16% and 22% of the positions most variants cause loss of activity together with loss of abundance (Figs. S1C and D; red), whereas the most common outcome at 9% and 12% of the positions are loss of activity without loss of abundance (Figs. S1C and D; blue). Finally, at 14% of the positions in NUDT15 and 9% in PTEN the variants most often have low abundance, but high levels of activity (Figs. S1C and D; yellow).

We validated the classifications using a clustering method that does not depend on defining cutoffs for the experimental scores. We grouped together positions with similar variant profiles in the two MAVEs (see Methods), and find overall very good agreement with the cutoff-based method in particular for the WT-like, total-loss and loss of activity, high abundance categories (Figs. S3 and S4). For NUDT15 we find that 133/188 positions are classified in the same way using the two different methods, with the most variable results occurring in the category with low abundance but sufficient activity to sustain growth (Fig. S3). For PTEN, we analysed the data using either three or four clusters, with the former appearing to be the more natural classification. In that case, 246/310 positions are classified in the same way using the two methods, with the 12 positions in the low abundance, high activity (yellow) category ending either as WT-like or total-loss. This indicates that three of the four categories of position effects are identified more robustly, corresponding to substitutions generally resulting in (i) WT-like activity in both assays, (ii) loss of activity and abundance or (iii) loss of activity, while retaining WT-like abundance. The low abundance, high activity positions are, however, less robustly classified and we do not analyse them further.

As expected, amino acids at buried positions are in general sensitive to mutations. In NUDT15, 35 out of the 163 amino acids are fully buried, and half of these (49%) are classified as sensitive to mutations in both the activity- and abundance-based assays (red label) with the remaining buried positions mainly classified as low abundance, high activity (34%; Figs. S1E and F). Because the variant coverage is lower in PTEN, only 355 of 403 positions can be classified in this way, and only 34 of these 355 are fully buried. Among these 34, 80% are classified as 'unstable' positions (low abundance, high activity and total-loss categories). Thus, loss of abundance is the typical reason for loss of activity for variants at buried positions.

Low activity, high abundance positions are defined as having the majority of the tested variants that have lost activity, but are still abundant in the cell, and previously such positions have been found to map to functionally important sites in the membrane protein VKOR (***Chiasson et al., 2020***). In PTEN, these variants and positions are mainly found in the catalytic phosphatase domain (Fig. S5) and include the active site (Figs. 2A and B). In NUDT15 we find the low activity, high abundance positions in three different regions. The first is located in proximity of the substrate binding site, the second includes the residues that coordinate a magnesium ion (***Suiter et al., 2020***) and the third region consists of four positions: Asn111, Asn117, Gln44 and Arg34, where most variants cause loss of activity, but not loss of abundance (Figs. 2C and D). Arg34 is directly involved in the hydrolysis of the thiopurine drugs (***Carter et al., 2015***). Moreover, these four residues form a connected hydrogen-bond network that positions a loop (residues 111–117) to enable binding of the substrate. In particular, the presence of Asn111 and Asn117 appears to fix the position of the Glu113 and preserves its coordinating function with the magnesium ion (***Suiter et al., 2020***).

Using Gly47 in NUDT15 and Arg130 in PTEN as reference points in the active sites in these two proteins we find that the low activity, high abundance positions, where variants typically show loss of activity, but not loss of abundance, are clustered around the active sites. Specifically, we find all of these positions in NUDT15 are within 14 Å of Gly47 (C_α-distances). The average distance between low activity, high abundance positions and Gly47 is 9Å, a value that can be compared to the average (15Å) over all positions in NUDT15. In PTEN, we find that 29 of 32 low activity, high abundance positions are found in the catalytic domain. All of these 29 positions are within 22Å of Arg130, with the average distance to Arg130 being 14Å (compared to 21Å over all positions).

Computational predictions of multiplexed data from MAVEs

As described previously and demonstrated above, MAVEs provide a wealth of data not only for use in medical applications (***Weile and Roth, 2018***; ***Stein et al., 2019***) but also for understanding basic properties of proteins (***Dunham and Beltrao, 2020***). Despite recent advances in proteome-wide experiments (***Després et al., 2020***), it is still not possible to probe all possible variants in all proteins experimentally, and thus computational methods remain an important supplement to predict and understand variant effects. Experimental data from MAVEs are thus increasingly used to benchmark prediction methods, as they provide a broad view of the effect of amino acid substitutions in proteins (***Hopf et al., 2017***; ***Jepsen et al., 2020***; ***Livesey and Marsh, 2020***; ***Reeb et al., 2020***).

Recently we exploited the two different MAVEs for PTEN to analyse a small number of pathogenic variants together with variants that have been observed in a broader analysis of the human population (***Jepsen et al., 2020***). Specifically, we compared the abundance-based (VAMP-seq) and activity-based multiplexed data to two computational methods aimed at capturing either (i) specifically protein stability or (ii) function more broadly. Here we build on this work, by (i) applying computational modelling to predict changes in thermodynamic protein stability using Rosetta (***Park et al., 2016***) and (ii) using evolutionary conservation as a more general view of which amino acid changes would be tolerated while maintaining function (***Ekeberg et al., 2014***). The former uses as input the structure of NUDT15 or PTEN to predict the change in protein stability ($\Delta\Delta G$), while the latter uses a sequence alignment of homologuous proteins as input to a computational assessment of conservation, taking both site and pair-conservation (co-evolution) into account, quantified by a score (which we by analogy to $\Delta\Delta G$ term $\Delta\Delta E$) that estimates how likely a substitution would be. As previously argued (***Jepsen et al., 2020***), the $\Delta\Delta G$ calculations are more akin to the results of a abundance-based MAVE (both capturing aspects of protein stability), while the $\Delta\Delta E$ values capture a broader range of effects as would also be expected from an activity-based MAVE.

We thus compared the computational predictions of $\Delta\Delta G$ and $\Delta\Delta E$ with each of the two multiplexed assays for NUDT15 and PTEN. As expected, we find that stability predictions correlate better with the abundance-based MAVE than with the activity-based MAVE, while for the evolutionary analysis the situation is reversed (Fig. S6). This supports the notion that analysis of conservation is a better predictor of general aspects of protein function, while the Rosetta calculations support the expected relationship between cellular protein abundance and thermodynamic stability (***Matreyek***

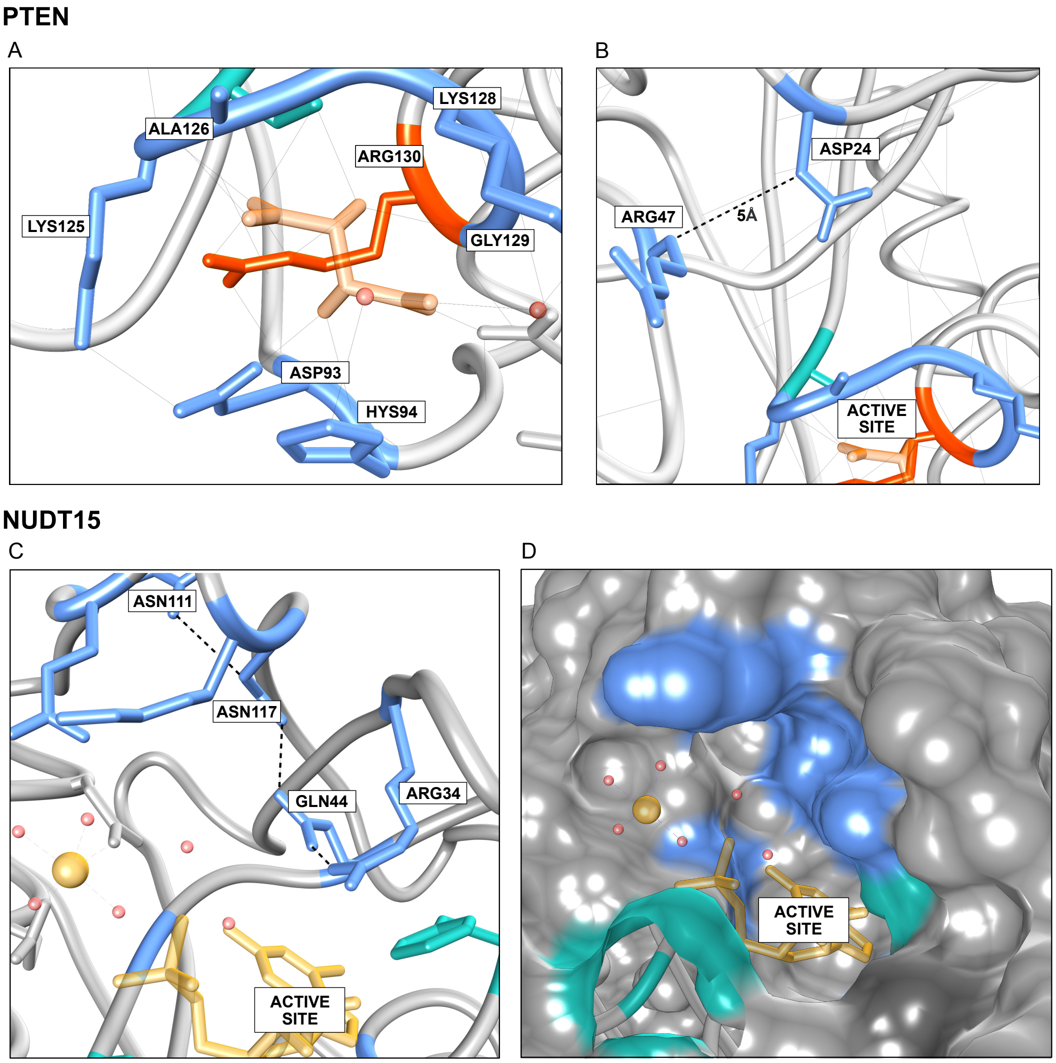


Figure 2. Examples of variants that lose activity but not abundance. A: Residues in PTEN in the low activity, high abundance category (blue) include residues in and surrounding the catalytic phosphatase site including some that directly interact with the substrate (here mimicked by the inhibitor tartrate (***Lee et al., 1999***)). B: Other residues that are more distant to the active site also fall in this category, and variants in this region likely perturb the integrity of the active site. C-D: Examples of functionally important residues in NUDT15 that are close to, but outside of the active site. In particular, we identified four conserved residues (Asn111, Asn117, Gln44, Arg44) that are connected by a hydrogen bond network and likely involved in the hydrolysis of the thiopurines.

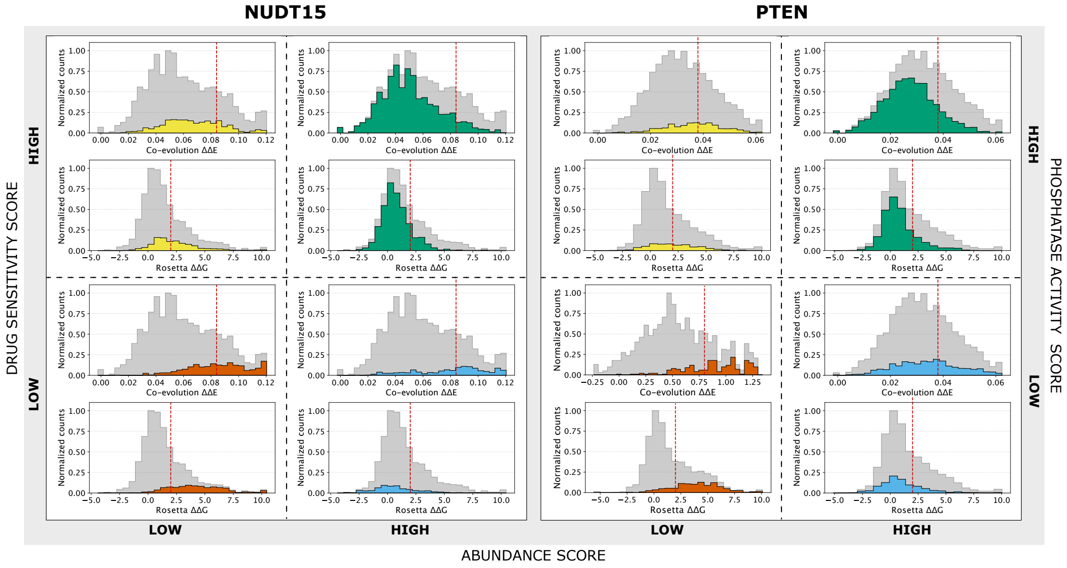


Figure 3. Histograms of the two computational scores ($\Delta\Delta G$ and $\Delta\Delta E$) in NUDT15 and PTEN. $\Delta\Delta G$ aims to capture effects purely on the thermodynamic stability, with high values indicating destablized variants. $\Delta\Delta E$ captures evolutionary conservation, as calculated by a model that takes both site and pairwise co-evolution into account, and with high values indicating non-conservative substitutions. Thus, for both $\Delta\Delta G$ and $\Delta\Delta E$ positive values indicate detrimental substitutions, whereas in the experiments low values indicate substitutions that cause loss of activity or abundance. For both proteins we split the histograms up according to the four categories of variants determined from the experiments, as indicated by the axes with high and low experimental scores for abundance and activity. Thus, for example, the two green histograms for NUDT15 indicate the distributions of $\Delta\Delta G$ and $\Delta\Delta E$ values for those variants that are classified as stable and active by the MAVEs, and indeed it is clear that most of these variants have scores that are below the cutoff (red dashed lines). In addition to the coloured histograms we also show the full histogram of all analysed variants (grey) to ease comparison between the subsets and the full set of variants.

et al., 2018; ***Abildgaard et al., 2019***; ***Jepsen et al., 2020***).

We define threshold values for the computational scores (Fig. S7) to separate wild-type-like from deleterious variants and construct four categories that we label with colours as above. Using a threshold of 2 kcal/mol for the $\Delta\Delta G$ for both proteins results in 69% (NUDT15) and 65% (PTEN) of the variants being predicted stable. Similarly, from the evolutionary conservation analysis 78% and 58% of all variants for NUDT15 and PTEN, respectively, have scores that indicated that the substitutions are tolerated. Note that, by convention, positive $\Delta\Delta G$ and $\Delta\Delta E$ scores indicate loss of stability or sequence tolerance, respectively, and hence the scales are inverted compared to the scores from the MAVEs.

To enable a more direct comparison between the experimental and computational scores, we show histograms of the two computational scores ($\Delta\Delta G$ and $\Delta\Delta E$) for each of the four classes based on the experimental scores (Fig. 3). We find that the variants that experimentally were classified as WT-like (stable and active) generally have low computational values; thus the computational predictions suggest that these substitutions have a mild effect on stability (low $\Delta\Delta G$) and are compatible with substitutions observed in homologous proteins (low $\Delta\Delta E$). We make similar observations for the total loss category, where the computational scores are generally above the cutoff, and for the low activity, high abundance category where the computational analysis finds low values of $\Delta\Delta G$ but higher values of $\Delta\Delta E$. Despite these general trends, we find variable agreement in the classification of individual variants by experiments and computation (Fig. S8), with the best agreement in the WT-like and total-loss categories.

We proceeded by generating and examining the structure-function relationships that we extracted from the computational analyses (Fig. S9). We used the computational results to group the positions into four categories and found a high overlap to those found in experiments (Fig. S10), in particular for the WT-like and total-loss categories. This result suggests that the computational analyses better capture general effects at positions compared to individual variants as discussed above (Fig. 3). We again used a clustering procedure as an alternative approach to classify positions, and find good agreement both with the cutoff-based classification of the computational data as well as with the experiment-based classifications (Fig. S11). Thus, together these results show that a joint computational analysis of stability and conservation can be used to find positions in the protein where substitutions are likely to disrupt thermodynamic stability, and other positions where they will cause loss of activity via removing functionally important residues.

Conclusions

Large-scale analyses of proteins using multiplexed assays provide opportunities to obtain a global view of variant effects (***Gray et al., 2017***; ***Dunham and Beltrao, 2020***). By combining different assays to read out different properties of a protein it becomes possible to dissect which positions contribute most to which property (***Jepsen et al., 2020***). Most proteins need to be folded to be active, and thus amino acid substitutions that lead to loss of stability will often lead to loss of function, and indeed loss of stability appears to be an important driver for disease (***Yue et al., 2005***; ***Stein et al., 2019***).

We have here exploited the availability of data generated by MAVEs for two proteins, with one experiment probing general effects on protein activity and another directly assessing cellular abundance. We show that a global analysis of these experiments can provide insight into how proteins function and how activity may be perturbed. With the assays considered here, we find that most variants have at most a modest effect on protein activity. Of the ca. 30% of the variants that cause substantial loss of activity we find that ca. 50% also cause loss of abundance. Interestingly, the latter number can be compared to our previous analysis of 42 disease-causing variants in PTEN, where we found a comparable fraction (∼60%) of the disease-causing variants appear to cause loss of function via loss of stability and thereby cellular protein abundance (***Jepsen et al., 2020***). Similarly, in our studies of pathogenic missense variants in the MLH1 gene we found low (<50% of wild type) steady state protein levels in 7 out 16 pathogenic variants (***Abildgaard et al., 2019***). Thus, at least in these cases, it appears that the fraction of variants that cause disease via this mechanism reflects the overall fraction of 'total loss' variants in the protein. An interesting question for future experiments is how many of these variants would be active if protein levels could be restored for example by chemical chaperones or modulating the protein quality control apparatus (***Arlow et al., 2013***; ***Kampmeyer et al., 2017***).

Building on previous work (***Chiasson et al., 2020***) we also show how we can use variant effects on protein activity and abundance/stability to find functionally important residues both by experiments and computation. For several surface-exposed residues many variants cause loss of activity, but without substantial loss of abundance. We find that these include the active sites in NUDT15 and PTEN, but also discover functionally important sites adjacent to these active sites. The importance of second shell positions for modulating the structure or dynamics of active site residues has for example also emerged in studies of ligand binding (***Tinberg et al., 2013***) and enzyme evolution and design (***Campbell et al., 2016***; ***Broom et al., 2020***).

The relatively tight confinement of these low-activity/high-abundance positions may also explain why predictions of changes in protein stability can be used to predict a substantial number of disease variants: At least in NUDT15 and PTEN the number of positions where substitutions typically cause loss of abundance (and thereby activity) is greater than the number of positions where substitutions cause loss of activity while retaining protein abundance. Indeed, while functional sites induce substantial constraints on amino acid variation during evolution, the strongest effects are those closest to the active sites (***Jack et al., 2016***; ***Mayorov et al., 2019***). Our ability to predict these sites by combining evolutionary analysis and stability calculations also suggest an approach for discovering new functionally-important sites using combined analyses of protein structure and sequences.

We find that ca. 12% of variants in NUDT15 and PTEN appear to be able to support wild-type like growth in the cellular assays even at substantially reduced protein levels. Clearly, there can be a non-linear relationship between a growth phenotype and protein abundance (***Jiang et al., 2013***), and this may help explain some of these variants. Future experiments that probe the relationship between expression levels and variant effects in NUDT15 and PTEN may shed further light on these variants. Further, the abundance-based MAVE for PTEN was performed in a cultured mammalian cell line (***Matreyek et al., 2018***) and the activity-based MAVE was performed in yeast (***Mighell et al., 2018***), leading to potential differences due to the differences in the quality control and proteostasis machinery in these cells.

In summary, we demonstrate how multiplexed assays and computational analyses are beginning to provide a coherent and comprehensive view of the global effects of variants in proteins. The results highlight that many effects are correctly predicted and thus computation can be used not only to predict whether a variant will cause loss of activity or not, but also provide some mechanistic insight. Clearly, there is room for improvement, and additional experiments on more proteins and covering more aspects of the complicated relationship between protein sequence and functions will help further our ability to predict these effects computationally.

Methods

Conservation analysis of variant effects

We used a statistical analysis of multiple sequence alignments (MSAs) of the two proteins to estimate the tolerance towards specific substitutions. In line with previous work, we use a method that includes both site and pairwise conservation (co-evolution). We used HHBlits (***Remmert et al., 2012***) to build initial MSAs, which we filtered before calculating the variant effects. The first filter removes sequences (rows) in the MSA with more than 50% gaps. The second filter keeps only positions (columns) that are present in the human target sequences of NUDT15 or PTEN. Finally, we apply a similarity filter (***Ekeberg et al., 2013***) to remove redundant sequences. We use a modified version of the lbsDCA algorithm (***Ekeberg et al., 2014***), based on l_2-regularized maximization with pseudo-counts to predict the likelihood of every variant of the protein. We use the energy potential generated by the algorithm to evaluate the log-likelihood difference between the wild type and the variant sequences ($\Delta\Delta E$). The results of this analysis, the stability calculations and scripts to reproduce our analyses are available github.com/KULL-Centre/papers/tree/master/2020/mave-analysis-cagiada-et-al.

Structural analyses

We used Rosetta (GitHub SHA1 99d33ec59ce9fcecc5e4f3800c778a54afdf8504) to predict changes in thermodynamic stability ($\Delta\Delta G$) from the structure of NUDT15 and PTEN using the Cartesian ddG protocol (***Park et al., 2016***). As starting points we used the crystal structures of NUDT15 (***Valerie et al., 2016***) (PDB ID: 5LPG) and PTEN (***Lee et al., 1999***) (PDB ID: 1D5R). The values obtained from Rosetta were divided by 2.9 to bring them from Rosetta energy units onto a scale corresponding to kcal/mol (Frank DiMaio, University of Washington; personal correspondence) (***Jepsen et al., 2020***). We used DSSP-2.28 (***Kabsch and Sander, 1983***; ***Touw et al., 2015***) and the same crystal structures as above to classify the burial with a three state model (***Rost and Sander, 1994***) (buried, intermediate, or exposed).

Defining thresholds for classifying variants

We defined thresholds for the scores from both MAVEs (Fig. S2), by fitting the variant score distributions using the minimal number of Gaussians (three) needed to obtain a reasonable fit. We then used the intersection of the first and last Gaussian as cutoff for our classifications. We use a cutoff of 2 kcal/mol (similar to the value used in our previous study (***Jepsen et al., 2020***)) for $\Delta\Delta G$ and varied the cutoff for $\Delta\Delta E$ to maximize the overlap in the classification of positions (Fig. S10).

To examine the threshold-based classifications, we used a hierarchical clustering algorithm (***Virtanen et al., 2020***) to group positions with similar responses to amino acid substitutions. Each position was represented by a 40-dimensional vector that contains the scores for each of the 20 possible amino acids in the two MAVEs. Missing values were replaced by the average score over that position. We use the Euclidean distance between these vectors as similarity score in the hierarchical clustering (***Ward Jr, 1963***). To compare with the threshold-based classification we analysed this using four clusters, though in the case of PTEN we also show the results using only three clusters.

Acknowledgement

This work is a contribution from the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (to R.H.-P., D.M.F., A.S. and K.L.-L.). J.J.Y. acknowledges funding from the NIH (R01GM118578).

References

Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, Hoffmann ER, Bernstein I, Gerdes AM, Takahashi M, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. Elife. 2019; 8:e49138.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4):248–249.

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

Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the National Academy of Sciences. 2012; 109(42):16858–16863.

Arlow T, Scott K, Wagenseller A, Gammie A. Proteasome inhibition rescues clinically significant unstable variants of the mismatch repair protein Msh2. Proceedings of the National Academy of Sciences. 2013; 110(1):246–251.

Broom A, Rakotoharisoa RV, Thompson MC, Zarifi N, Nguyen E, Mukhametzhanov N, Liu L, Fraser JS, Chica RA. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. Nature Communications. 2020 Sep; 11(1):4808. https://doi.org/10.1038/s41467-020-18619-x, doi: 10.1038/s41467-020-18619-x.

Campbell E, Kaltenbach M, Correy GJ, Carr PD, Porebski BT, Livingstone EK, Afriat-Jurnou L, Buckle AM, Weik M, Hollfelder F, et al. The role of protein dynamics in the evolution of new enzyme function. Nature chemical biology. 2016; 12(11):944–950.

Carter M, Jemth AS, Hagenkort A, Page BDG, Gustafsson R, Griese JJ, Gad H, Valerie NCK, Desroses M, Boström J, Warpman Berglund U, Helleday T, Stenmark P. Crystal structure, biochemical and cellular activities demonstrate separate functions of MTH1 and MTH2. Nature Communications. 2015; 6(1):7871. https://doi.org/10.1038/ncomms8871, doi: 10.1038/ncomms8871.

Casadio R, Vassura M, Tiwari S, Fariselli P, Luigi Martelli P. Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. Human Mutation. 2011; 32(10):1161–1170.

Chen L, Brewer MD, Guo L, Wang R, Jiang P, Yang X. Enhanced degradation of misfolded proteins promotes tumorigenesis. Cell reports. 2017; 18(13):3143–3154.

Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, Sun S, Roth F, DeSloover D, Marks DS, Rettie AE, Fowler DM. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. eLIFE. 2020; 9(e58026).

Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015; 31(16):2745–2747.

De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, Schymkowitz J, Rousseau F. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic acids research. 2012; 40(D1):D935–D939.

Després PC, Dubé AK, Seki M, Yachie N, Landry CR. Perturbing proteomes at single residue resolution using base editing. Nature communications. 2020; 11(1):1–13.

Dunham A, Beltrao P. Exploring amino acid functions in a deep mutational landscape. BioRxiv. 2020; p. 2020.05.26.116756.

Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. Journal of Computational Physics. 2014; 276:341–356. http://dx.doi.org/10.1016/j.jcp.2014.07.024, doi: 10.1016/j.jcp.2014.07.024.

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707.

Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. Journal of molecular biology. 2002; 315(4):771–786.

Fersht A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. Macmillan; 1999.

Gao M, Zhou H, Skolnick J. Insights into disease-associated mutations in the human proteome through protein structural analysis. Structure. 2015; 23(7):1362–1369.

Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. Scientific Reports. 2020 Sep; 10(1):15387. https://doi.org/10.1038/s41598-020-72404-w, doi: 10.1038/s41598-020-72404-w.

Gray VE, Hause RJ, Fowler DM. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. Genetics. 2017; 207(1):53–61.

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nature biotechnology. 2017; 35(2):128–135.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics. 2016; 99(4):877–885.

Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biology. 2016; 14(5):e1002452.

Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Classifying disease-associated variants using measures of protein activity and stability. In: *Protein Homeostasis Diseases* Elsevier; 2020.p. 91–107.

Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DN. Latent effects of Hsp90 mutants revealed at reduced expression levels. PLoS Genet. 2013; 9(6):e1003600.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules. 1983; 22(12):2577–2637.

Kampmeyer C, Nielsen SV, Clausen L, Stein A, Gerdes AM, Lindorff-Larsen K, Hartmann-Petersen R. Blocking protein quality control to counter hereditary cancers. Genes, Chromosomes and Cancer. 2017; 56(12):823–831.

Karran P, Attard N. Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer. Nature Reviews Cancer. 2008; 8(1):24–36.

Kinney JB, McCandlish DM. Massively parallel assays and quantitative sequence–function relationships. Annual review of genomics and human genetics. 2019; .

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014; 46(3):310–315.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009; 4(7):1073.

Lee JO, Yang H, Georgescu MM, Di Cristofano A, Maehama T, Shi Y, Dixon JE, Pandolfi P, Pavletich NP. Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. Cell. 1999; 99(3):323–334.

Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Molecular Systems Biology. 2020; p. e9380.

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, Bhatia S, Evans WE, Relling MV, Yang W, Shendure J, Fowler DM. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature Genetics. 2018; 50(6):874–882. http://dx.doi.org/10.1038/s41588-018-0122-z, doi: 10.1038/s41588-018-0122-z.

Mayorov A, Dal Peraro M, Abriata LA. Active site-induced evolutionary constraints follow fold polarity principles in soluble globular enzymes. Molecular biology and evolution. 2019; 36(8):1728–1733.

Meacham GC, Patterson C, Zhang W, Younger JM, Cyr DM. The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. Nature cell biology. 2001; 3(1):100–105.

Mighell TL, Evans-Dutson S, O'Roak BJ. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. American Journal of Human Genetics. 2018; 102(5):943–955. https://doi.org/10.1016/j.ajhg.2018.03.018, doi: 10.1016/j.ajhg.2018.03.018.

Moriyama T, Nishii R, Lin TN, Kihira K, Toyoda H, Nersting J, Kato M, Koh K, Inaba H, Manabe A, et al. The effects of inherited NUDT15 polymorphisms on thiopurine active metabolites in Japanese children with acute lymphoblastic leukemia. Pharmacogenetics and genomics. 2017; 27(6):236.

Moriyama T, Nishii R, Perez-Andreu V, Yang W, Klussmann FA, Zhao X, Lin TN, Hoshitsuki K, Nersting J, Kihira K, et al. NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. Nature genetics. 2016; 48(4):367–373.

Nielsen SV, Schenstrøm SM, Christensen CE, Stein A, Lindorff-Larsen K, Hartmann-Petersen R. Protein destabilization and degradation as a mechanism for hereditary disease. In: *Protein Homeostasis Diseases* Elsevier; 2020.p. 111–125.

Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, Kassem MM, Rasmussen LJ, Lindorff-Larsen K, Hartmann-Petersen R. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genetics. 2017; 13(4):e1006739.

Nishii R, Moriyama T, Janke LJ, Yang W, Suiter CC, Lin TN, Li L, Kihira K, Toyoda H, Hofmann U, et al. Preclinical evaluation of NUDT15-guided thiopurine therapy and its effects on toxicity and antileukemic efficacy. Blood. 2018; 131(22):2466–2474.

Olzmann JA, Brown K, Wilkinson KD, Rees HD, Huai Q, Ke H, Levey AI, Li L, Chin LS. Familial Parkinson's disease-associated L166P mutation disrupts DJ-1 protein folding and function. Journal of Biological Chemistry. 2004; 279(9):8506–8515.

Park H, Bradley P, Greisen Jr P, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. Journal of chemical theory and computation. 2016; 12(12):6201–6212.

Reeb J, Wirth T, Rost B. Variant effect predictions capture some aspects of deep mutational scanning experiments. BMC bioinformatics. 2020; 21(1):1–12.

Relling MV, Schwab M, Whirl-Carrillo M, Suarez-Kurtz G, Pui CH, Stein CM, Moyer AM, Evans WE, Klein TE, Antillon-Klussmann FG, et al. Clinical pharmacogenetics implementation consortium guideline for thiopurine dosing based on TPMT and NUDT 15 genotypes: 2018 update. Clinical Pharmacology & Therapeutics. 2019; 105(5):1095–1105.

Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods. 2012; 9(2):173.

Ron I, Horowitz M. ER retention and degradation as the molecular basis underlying Gaucher disease heterogeneity. Human molecular genetics. 2005; 14(16):2387–2398.

Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins: Structure, Function, and Bioinformatics. 1994; 20(3):216–226.

Scheller R, Stein A, Nielsen SV, Marin FI, Gerdes AM, Di Marco M, Papaleo E, Lindorff-Larsen K, Hartmann-Petersen R. Toward mechanistic models for genotype–phenotype correlations in phenylketonuria using protein stability calculations. Human Mutation. 2019; 40(4):444–457.

Shin H, Cho BK. Rational protein engineering guided by deep mutational scanning. International journal of molecular sciences. 2015; 16(9):23094–23110.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015; 200(2):413–422.

Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. Trends in biochemical sciences. 2019; 44(7):575–588.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. TRENDS in Genetics. 2003; 19(9):505–513.

Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, Yang W, Hoshitsuki K, Singh M, Trehan A, Parish C, Smith C, Li L, Bhojwani D, Yuen LYP, Li Ck, Li Ch, Yang Yl, Walker GJ, Goodhand JR, et al. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. Proceedings of the National Academy of Sciences. 2020; p. 201915680. http://www.pnas.org/lookup/doi/10.1073/pnas.1915680117, doi: 10.1073/pnas.1915680117.

Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013; 501(7466):212–216.

Touw WG, Baakman C, Black J, Te Beek TA, Krieger E, Joosten RP, Vriend G. A series of PDB-related databanks for everyday needs. Nucleic acids research. 2015; 43(D1):D364–D368.

Valerie NCK, Hagenkort A, Page BDG, Masuyer G, Rehling D, Carter M, Bevc L, Herr P, Homan E, Sheppard NG, al Stenmark P, Jemth AS, Helleday T. NUDT15 Hydrolyzes 6-Thio-DeoxyGTP to Mediate the Anticancer Efficacy of 6-Thioguanine. Cancer Research. 2016; 76(18):5501–5511. https://cancerres.aacrjournals.org/content/76/18/5501, doi: 10.1158/0008-5472.CAN-16-0584.

Valiente M, Andrés-Pons A, Gomar B, Torres J, Gil A, Tapparel C, Antonarakis SE, Pulido R. Binding of PTEN to specific PDZ domains contributes to PTEN protein stability and phosphorylation by microtubule-associated serine/threonine kinases. Journal of Biological Chemistry. 2005; 280(32):28936–28943.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Jarrod Millman K, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey C, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020; 17:261–272. doi: https://doi.org/10.1038/s41592-019-0686-2.

Wagih O, Galardini M, Busby BP, Memon D, Typas A, Beltrao P. A resource of variant effect predictions of single nucleotide variants in model organisms. Molecular systems biology. 2018; 14(12):e8430.

Wang Z, Moult J. SNPs, protein structure, and disease. Human mutation. 2001; 17(4):263–270.

Ward Jr JH. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963; 58(301):236–244.

Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. Human genetics. 2018; 137(9):665–678.

Yaguchi H, Ohkura N, Takahashi M, Nagamura Y, Kitabayashi I, Tsukada T. Menin missense mutants associated with multiple endocrine neoplasia type 1 are rapidly degraded via the ubiquitin-proteasome pathway. Molecular and cellular biology. 2004; 24(15):6569–6580.

Yang C, Asthagiri AR, Iyer RR, Lu J, Xu DS, Ksendzovsky A, Brady RO, Zhuang Z, Lonser RR. Missense mutations in the NF2 gene result in the quantitative loss of merlin protein and minimally affect protein intrinsic function. Proceedings of the National Academy of Sciences. 2011; 108(12):4980–4985.

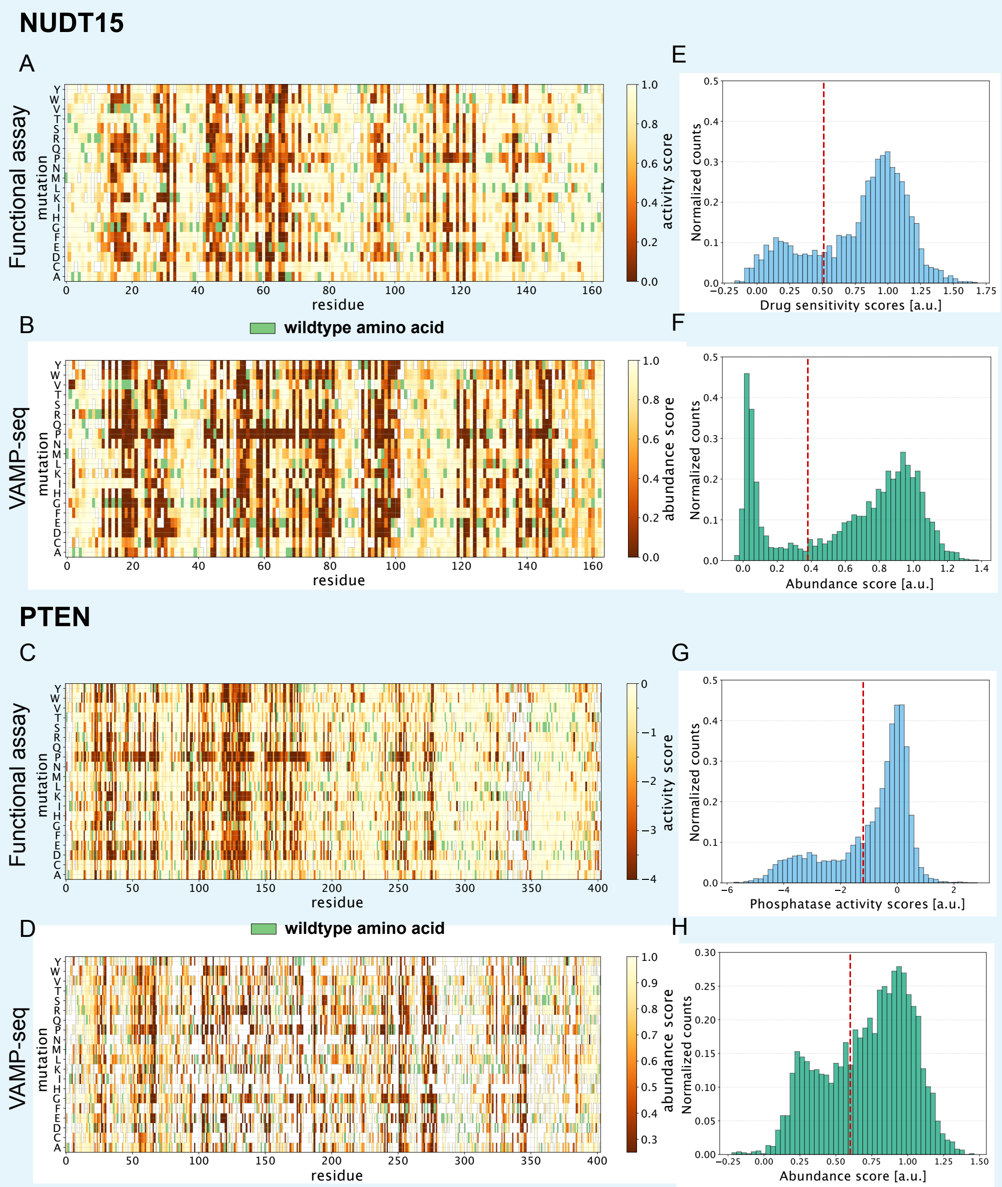
Yang C, Huntoon K, Ksendzovsky A, Zhuang Z, Lonser RR. Proteostasis modulators prolong missense VHL protein activity and halt tumor progression. Cell reports. 2013; 3(1):52–59.

Yang SK, Hong M, Baek J, Choi H, Zhao W, Jung Y, Haritunians T, Ye BD, Kim KJ, Park SH, et al. A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. Nature genetics. 2014; 46(9):1017.

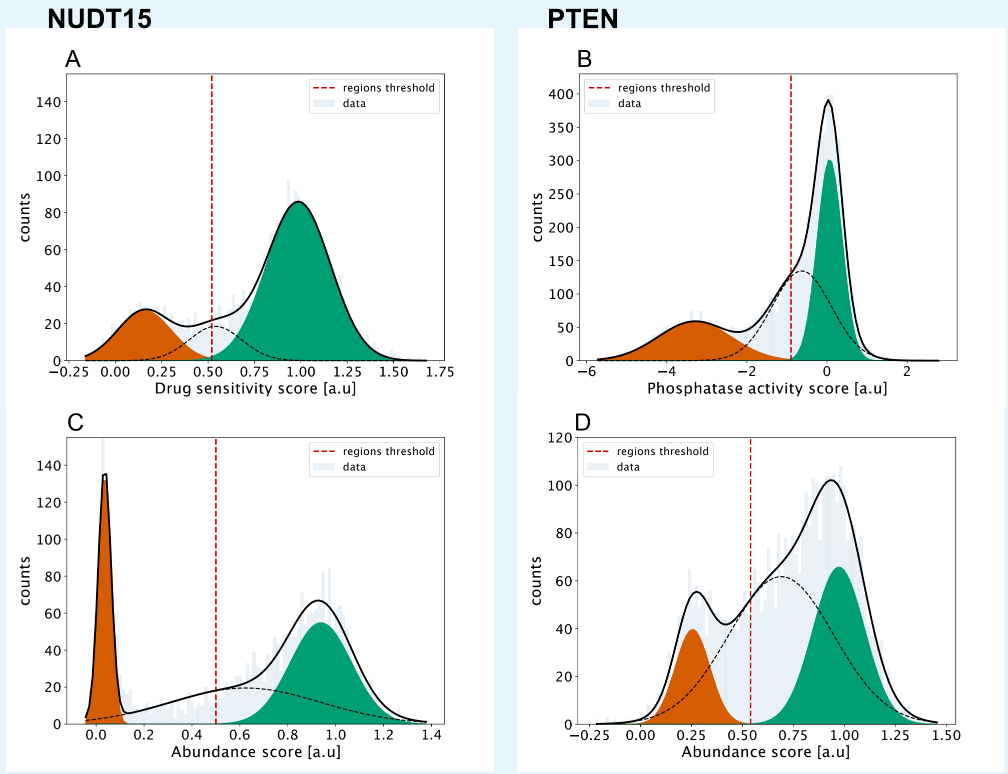
Yehia L, Ngeow J, Eng C. PTEN-opathies: from biological insights to evidence-based precision medicine. The Journal of clinical investigation. 2019; 129(2):452–464.

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. Journal of molecular biology. 2005; 353(2):459–473.

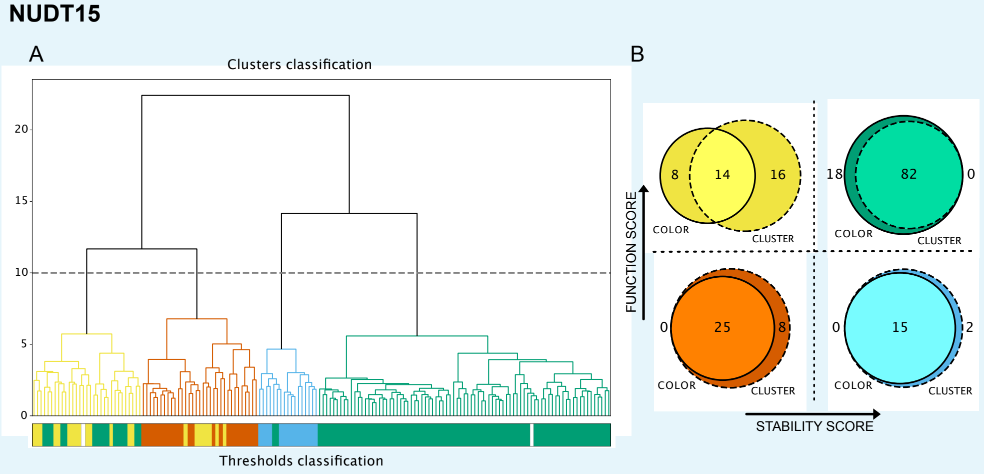
Supporting Material



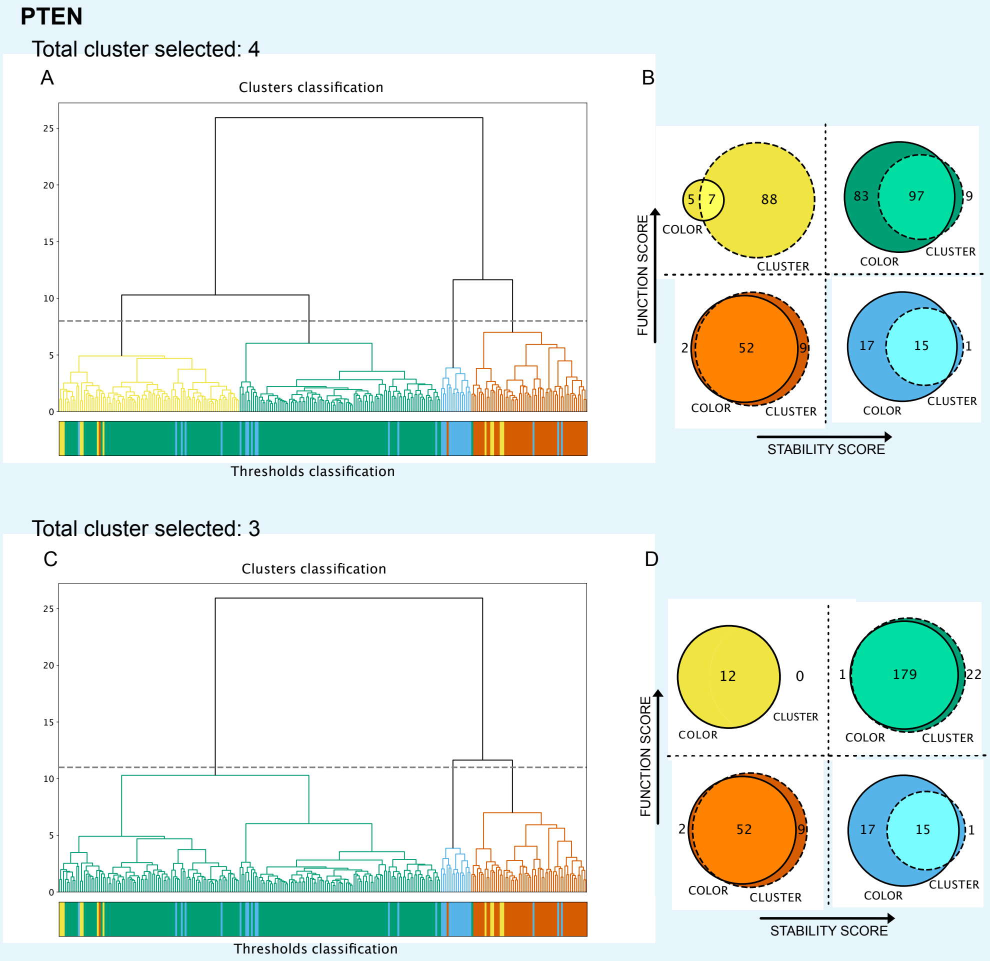
Supporting Figure S1. Experimental data for NUDT15 and PTEN. Panels A–D show the multiplexed data as heat maps. The experimental scores are indicated by colours and the wild type sequence is represented by a light green block at each position. Variants with wild-type-like behaviour have higher scores, and variants with lower scores indicate those of either low abundance (VAMP-seq) or activity (activity-based assays). Panels E–H show the distribution of values in each of the four assays with the red dashed lines indicating the cutoffs used to separate low and high scoring variants.



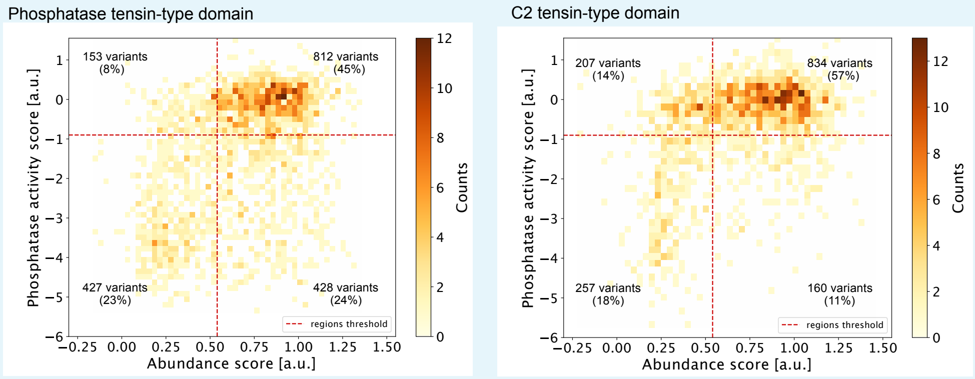
Supporting Figure S2. Protocol used to define thresholds for the multiplexed data generated by MAVEs. In each panel we fit the distribution of scores (light blue) to a mixture model with three Gaussian distributions (black line). The first and last Gaussian are shown in red and green, respectively, and we used the intersection between these to define cutoffs for classifying the variants (red dotted line).



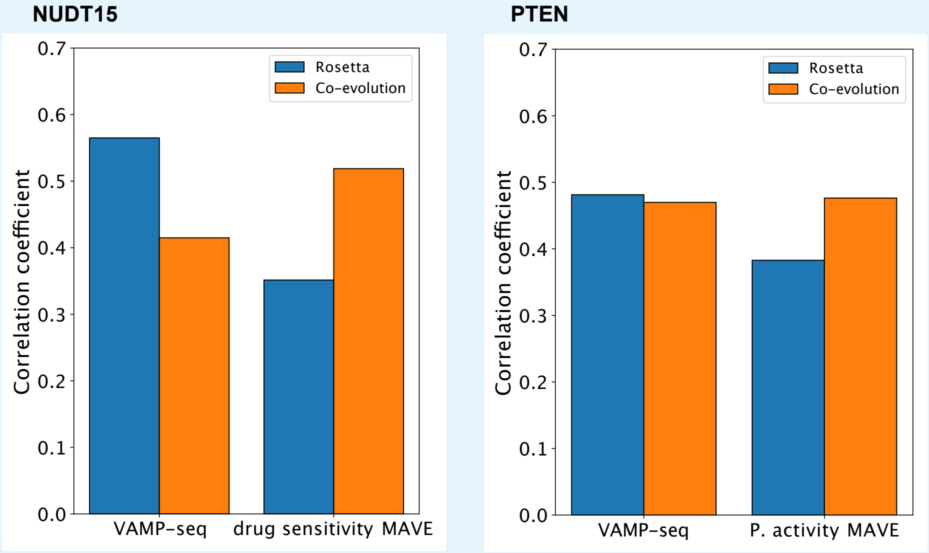
Supporting Figure S3. Assessment of the quality of a cutoff-based classification of positions in NUDT15. A: A hierarchical cluster analysis grouping together positions with similar responses to amino acid substitutions. Using the horizontal line to define the number of clusters, we colour the four clusters analogous to the results of the threshold-based classification. The bar plot under the cluster figure shows the colour assigned to each position in the threshold-based classification. B: Agreement between the two classifications represented using Venn diagrams.



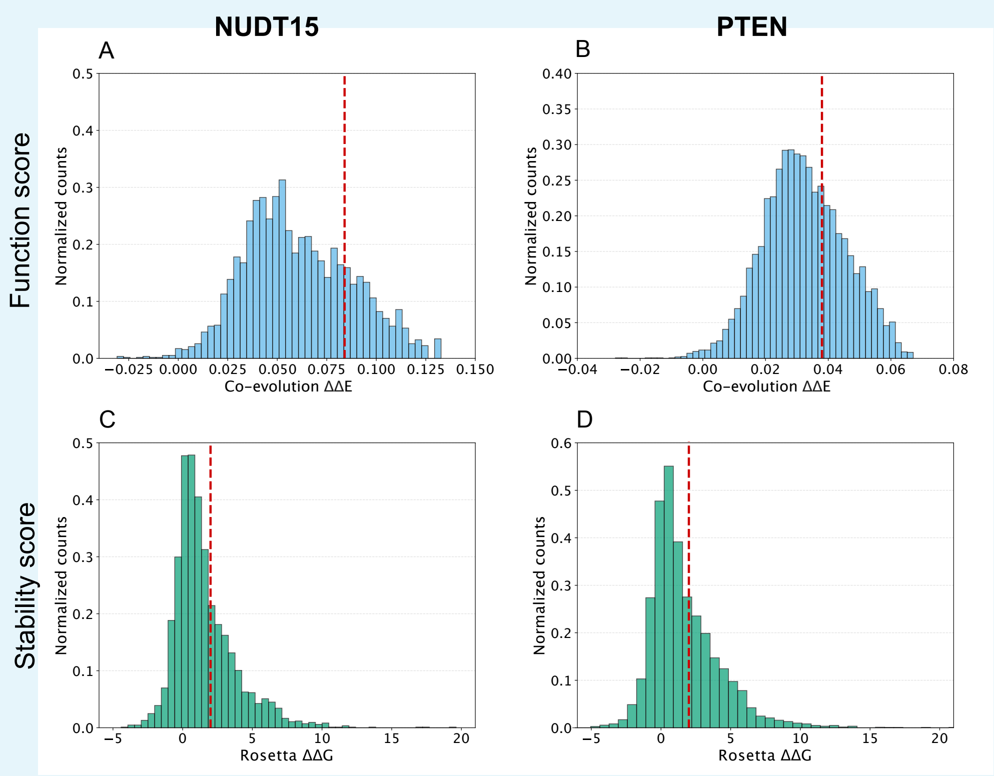
Supporting Figure S4. Assessment of the quality of a cutoff-based classification of positions in PTEN. A: A hierarchical cluster analysis grouping together positions with similar responses to amino acid substitutions. Using the horizontal line to define the number of clusters, we colour the four clusters analogous to the results of the threshold-based classification. The bar plot under the cluster figure shows the colour assigned to each position in the threshold-based classification. B: Agreement between the two classifications represented using Venn diagrams. Panels C and D use the same clustering as in A, but with the cutoff set so as to obtain only three clusters. In this case, the yellow group disappears and is merged with the green cluster.



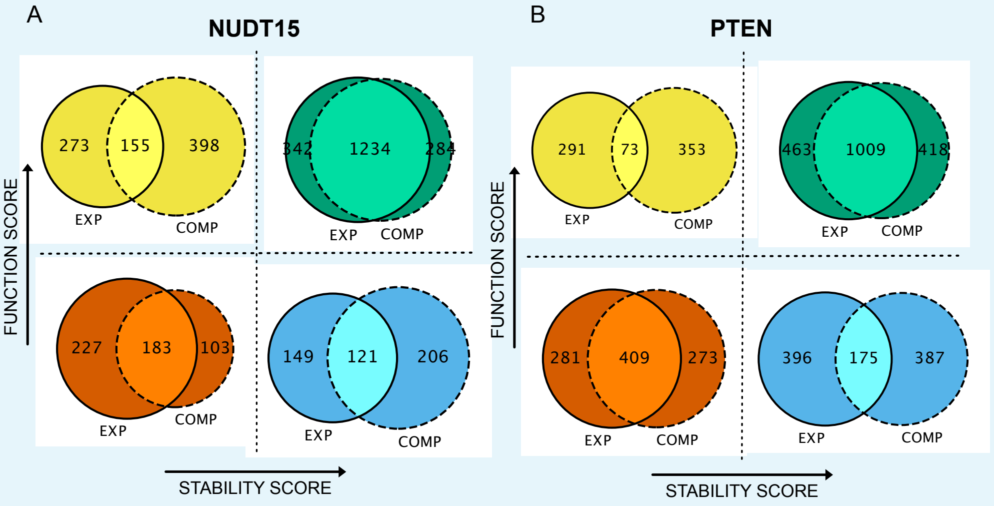
Supporting Figure S5. Analysis of PTEN variants by structural domain. The figures show the 2D histograms of the MAVE scores in PTEN for each of the two structural domains. While the overall picture in the two domains is similar, it is clear that a greater fraction of variants (47%) in the catalytic phosphatase domains causes loss of activity compared to the C2 domain (29%).



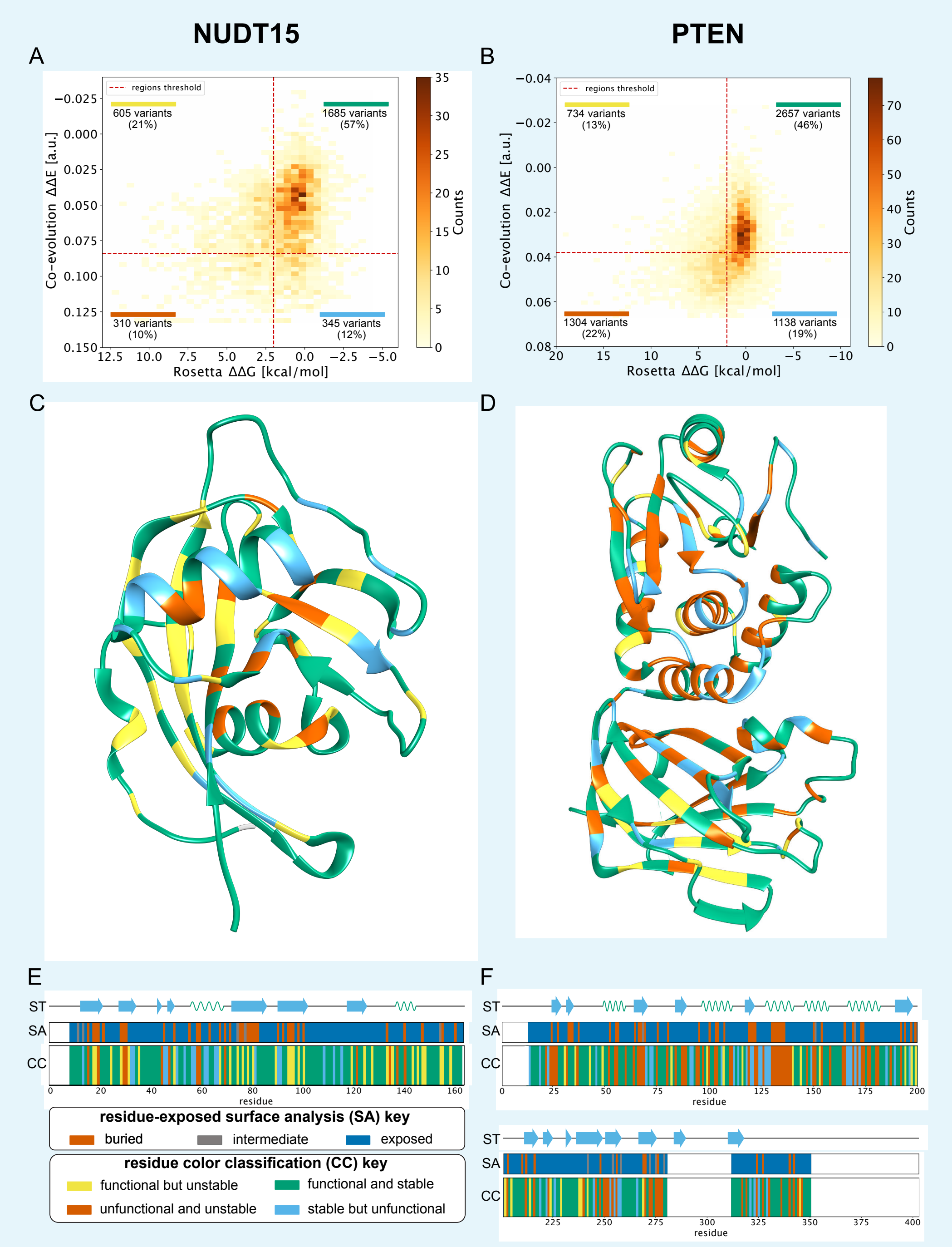
Supporting Figure S6. Correlation between the experimental data and two computational scores. For each of the two MAVEs in the two proteins we calculated the Pearson's correlation coefficient to either (blue bars) the Rosetta stability predictions or (orange bars) an assessment of tolerance towards substitutions using an evolutionary model (co-evolution).



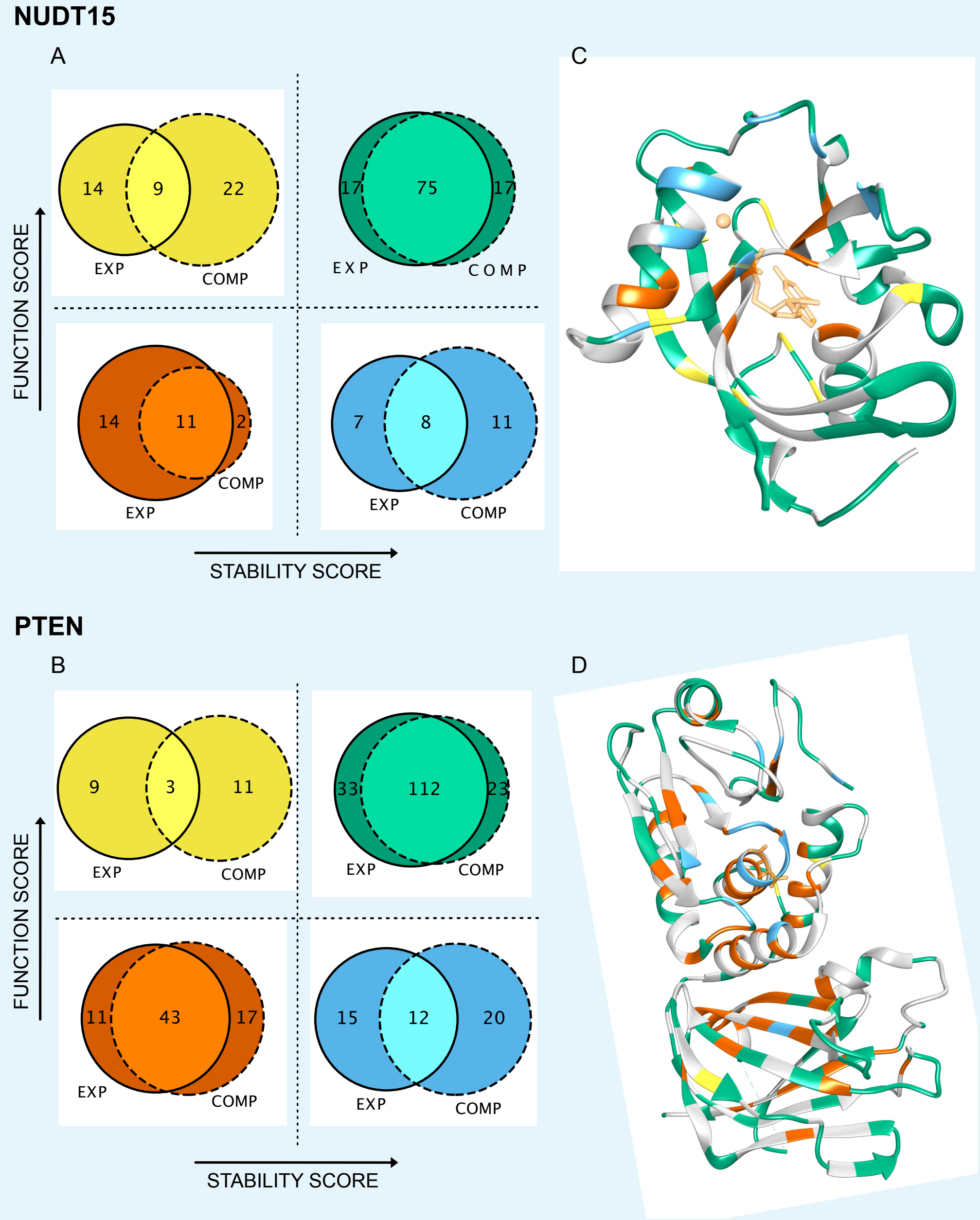
Supporting Figure S7. Distributions of scores and thresholds used in the computational analyses. Panels A and B show the distribution of values in the co-evolution analysis for both the target proteins, with the red dashed lines indicating the cutoffs used for classifying the variants. The Panels C and D show the distributions and the thresholds for the Rosetta $\Delta\Delta G$ values.



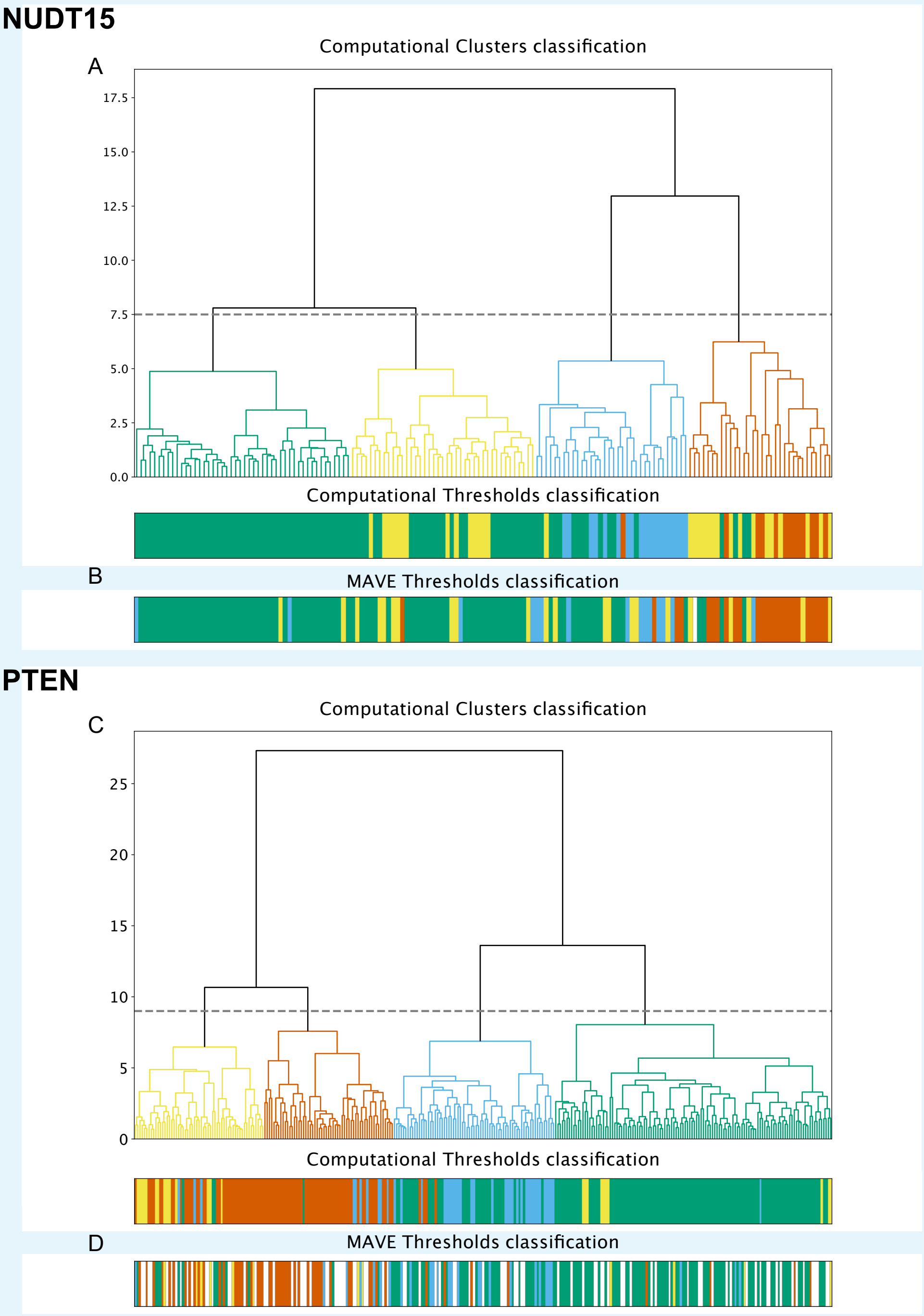
Supporting Figure S8. Comparison of the classification of variants by analysing the computational and experimental data analysis. The Venn diagrams show the agreement between the classification of the individual variants in (A) NUDT15 and (B) PTEN with ‘EXP’ and ‘COMP’ representing the cutoff-based classification using either the data generated by the two MAVEs or the $\Delta\Delta G/\Delta\Delta E$ analysis, respectively.



Supporting Figure S9. Overview of the NUDT15 and PTEN computational data generated in this work. Panels A and B show 2D histograms that combine the evolutionary conservation analysis $\Delta\Delta E$ on the y-axis with the Rosetta $\Delta\Delta G$ values on the x-axis. Variants are categorised based on the region of the 2D histogram they belong to. Panels C and D show a per-position consensus category (CC) coloured onto the structure of the proteins (PDB entry 5LPG for NUDT15 and 1D5R for PTEN). Panels E and F show the positional consensus categories together with the secondary structure (ST) and solvent accessibility (SA).



Supporting Figure S10. Comparing the classification of positions obtained using experimental data and the computational analysis. Panels A and B show the agreement between the classification from the analysis of experiments and computation (on the common subset of tested positions) using Venn diagrams. The protein structures in panels C and D are coloured at those positions where the experimental and computational classification is the same with the remaining positions shown as grey.



624

Supporting Figure S11. Cluster analysis of computational $\Delta\Delta E$ and $\Delta\Delta G$ data. The panels show the result of a hierarchical clustering of (A) NUDT15 and (C) PTEN. The clusters are shown with the same colours used to define the classes of position in the cutoff-based classification. The two bar plots (B and D) show the class assigned to each position in the cutoff-based classification using either computational or experimental data.