Discovering functionally important sites in proteins

Matteo Cagiada[1], **Sandro Bottaro**[1], **Søren Lindemose**[1], **Signe M. Schenstrøm**[1], **Amelie Stein**[1], **Rasmus Hartmann-Petersen**[1*], **Kresten Lindorff-Larsen**[1*]

***For correspondence:**
rhpetersen@bio.ku.dk (RH-P);
lindorff@bio.ku.dk (KL-L)

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

Abstract Proteins play important roles in biology, biotechnology and pharmacology, and missense variants are a common cause of disease. Discovering functionally important sites in proteins is a central but difficult problem because of the lack of large, systematic data sets. Sequence conservation can highlight residues that are functionally important but is often convoluted with a signal for preserving structural stability. We here present a machine learning method to predict functional sites by combining statistical models for protein sequences with biophysical models of stability. We train the model using multiplexed experimental data on variant effects and validate it broadly. We show how the model can be used to discover active sites, as well as regulatory and binding sites. We illustrate the utility of the model by predicting and validating the functional consequences of missense variants in *HPRT1* which may cause Lesch-Nyhan syndrome, and pinpoint the molecular mechanisms by which they cause disease.

Introduction

Proteins carry out most functions in a cell, generally through interactions with other molecules. These molecular interactions often involve specific sites or regions whose identification plays a fundamental role in understanding biology and disease. Considerable progress has been made in the development of methods to identify functional sites (***del Sol Mesa et al., 2003***; ***Lee et al., 2007***; ***Radivojac et al., 2013***; ***Kulmanov et al., 2018***; ***Torng and Altman, 2019***) and efforts have been made to understand the relationship between sequence variability, protein function, stability and the onset of diseases (***Yue et al., 2005***; ***Adzhubei et al., 2010***; ***Kircher et al., 2014***; ***Wagih et al., 2018***; ***Livesey and Marsh, 2020***). With the entry of accurate and large-scale protein structure prediction, structure-based methods for understanding biology are becoming even more important.

Analyses of large-scale mutagenesis studies have been used to probe the role of individual residues in the stability, abundance and function of a protein (***Gray et al., 2017***; ***Dunham and Beltrao, 2021***; ***Høie et al., 2021***). Since most proteins need to be folded to function, it may, however, be difficult to deconvolute the effects of amino acid substitutions on intrinsic function from their effects on stability and cellular abundance (***Li and Lehner, 2020***). We note here that we use the term 'function' and 'functional sites' in a relatively broad sense since our goal is to examine proteins broadly. In a few, favourable cases, multiplexed assays of variant effects (MAVEs) have been used to probe the consequence of almost all individual substitutions using both a functional readout and a readout that probes cellular abundance. Analysis of such data have been used to shed light on the molecular mechanisms underlying perturbed function, and more specifically to pinpoint which functional properties an amino acid residue contributes to (***Jepsen et al., 2020***; ***Cagiada et al., 2020***; ***Chiasson et al., 2020***; ***Faure et al., 2022***). For example, variants that lose function together with loss of abundance are likely to be caused by perturbations to the overall protein fold and stability,

whereas variants that lose function while retaining wild-type-like abundance in the cell are likely to be caused by perturbing sites that directly play a role in function (***Jepsen et al., 2020***; ***Cagiada et al., 2020***; ***Faure et al., 2022***). Alternative to such multi-readout experiments, global analyses of large datasets of multi-mutant variants can be used to deconvolute effects on for example stability and binding (***Otwinowski, 2018***).

In a recent analysis of abundance and activity assays (***Cagiada et al., 2020***) we showed that approximately half of the single-point variants that show loss-of-function do so together with loss of protein abundance. This result suggests that if one uses a functional readout to detect residues that are directly involved in function, then ca. half of the variants detected as important are so simply because they cause lowered abundance. This in turn makes it difficult to separate residues that are directly involved in function (for example in catalysis, binding and signalling) from those positions that are conserved mostly due to structural constraints (***Echave and Wilke, 2017***).

The other half of loss-of-function substitutions, instead, mostly affects the protein via for example perturbing interactions with substrates or binding partners rather than stability (***Cagiada et al., 2020***; ***Nielsen et al., 2021***). We call this class of substitutions 'stable but inactive' (SBI) variants emphasising their tight and direct involvement in protein function. Pinpointing and predicting which variants are SBI is important to understand how amino acid substitutions might affect protein functions, why they possibly cause disease, and to ultimately aid the development of personalised therapeutic treatments. A very practical utility of the detection of SBI variants is that they enable the identification of amino acid residues that play a direct role in function. Indeed, much of our biochemical understanding of how proteins function relies of protein engineering studies where the effects of amino acid substitutions on various biochemical readouts are probed. Thus, positions where most substitutions affect function, but not structural stability, are often found in functional sites (***Cagiada et al., 2020***; ***Chiasson et al., 2020***; ***Faure et al., 2022***).

As an alternative to experimental measurements, computational predictions can offer a faster, cheaper and more scalable approach to pinpoint positions that are important for function. Computational prediction of SBI variants, however, is not trivial. Many existing computational protocols are based on evaluating sequence conservation to find positions and variants that cause loss of function (***Lichtarge et al., 1996***; ***Kumar et al., 2009***; ***Choi and Chan, 2015***; ***Riesselman et al., 2018***; ***Laine et al., 2019***; ***Torng and Altman, 2019***). This strategy alone, however, may not be sufficient to identify residues that are conserved due to direct functional roles because sequence evolution is subject to both functional and structural constraints, and thus the different signals cannot be easily disentangled. One possible strategy to overcome this problem is to combine evolutionary data with analyses that report directly on the effects of substitutions on protein stability (***Cheng et al., 2005***; ***Wang et al., 2008***; ***Capra et al., 2009***).

Here, we aim to create a robust prediction method to identify functional residues in proteins and provide insights into the biochemical role of these residues in the target protein. To avoid biases from annotations of functional sites, we use unbiased data generated by MAVEs that report on both function and abundance to train a supervised machine learning model. As input to the model we use a combination of sequence conservation measures, free energy changes accompanying substitutions, and physicochemical properties. We begin by showing how our model can capture a wide range of functional sites that include those in active sites, but also distributed throughout the protein structure including potential allosteric and regulatory sites. We then show how we achieve good accuracy in pinpointing functional amino acids in different validation scenarios. Across several proteins we find that roughly one in ten of the positions are functionally relevant and conserved for reasons different than structural stability. Having validated the model, we use it to provide examples of the kinds of insight that it may provide including identifying catalytic sites, regions that interact with substrates, and interfaces in complexes. Finally, we performed prospective predictions of the mechanisms of disease variants in HPRT1 (encoding the protein Hypoxanthine-Guanine Phosphoribosyltransferase 1, HPRT1) and validate these by measuring variant effects on function and abundance.

Results

Identification of functional sites via analyses of stability and conservation

Previously we have used evolutionary analyses combined with stability calculations to predict variant effects on protein function and stability and showed how these measures correlate with changes in cellular abundance or function (***Cagiada et al., 2020***; ***Høie et al., 2021***). Here, we build on these ideas to construct a model to identify functional sites in proteins via the identification of SBI variants (Fig. 1). Before detailing the model, we first provide an intuitive description of the basic idea. Our goal is to identify positions in proteins that play some role in protein function and regulation; these may for example include active sites in enzymes, but also 'second-shell' residues around the active site, protein-protein interfaces, allosteric and regulatory sites, or residues involved in recognizing ligands and substrates. Experimentally, these residues may be identified since amino acid substitutions at these sites might cause loss of 'function' (in some readout). Many variants, however, cause loss of function via loss of stability and cellular abundance, and we remove such indirect effects by requiring that the variants have close to wild-type-like abundance (i.e. the variants belong to the SBI class). Computationally, we can identify these variants by calculations of effects on protein stability and conservation, thus finding positions that are conserved during evolution, but not due to a role in protein stability. We note two effects of these choices. First, the broad definition means that we can assign a functional role to a relatively large number of residues e.g. well beyond active sites in enzymes. Second, some residues will have a direct role in function, but also be important for protein stability; our analysis will miss those residues, but as shown below most functional sites likely do not fall in this class.

Based on the ideas above, we collected the results from two complementary MAVEs that probe cellular-abundance and functional effects for three different proteins: NUDT15 (***Suiter et al., 2020***), PTEN (***Matreyek et al., 2018***; ***Mighell et al., 2018***) and CYP2C9 (***Amorosi et al., 2021***) for a total of 9945 variants at 923 positions. Based on the experimental abundance and function scores, we assigned each variant to one of the following four classes: WT-like (high abundance and high activity), total loss (low abundance and low activity), SBI (high abundance and low activity), or low abundance and high activity. We also selected input features for each variant. These were calculated from the three-dimensional structure of a protein and a multiple sequence alignment. More precisely, we included the following features: (i) the predicted change in thermodynamic protein stability ($\Delta\Delta G$) calculated using Rosetta (***Park et al., 2016***), (ii) the evolutionary sequence information scores (which we term $\Delta\Delta E$, by analogy with $\Delta\Delta G$) using GEMME (***Laine et al., 2019***), (iii) the hydrophobicity of the amino acid (***Sereda et al., 1994***; ***Monera et al., 1995***), and (iv) the weighted contact number (***Shih et al., 2012***; ***Jack et al., 2016***) (Fig. 1A).

We first examined whether one of these features on its own would be sufficient, but did not find any that individually separates SBI variants from the remainder (Fig. S1). Thus, we used the experimental labels and the input features to train a gradient boosting classifier (Fig. 1A and Fig. S2) that predicts abundance (high/low) and activity (high/low) for each variant. We determined the hyperparameters of the model using a stratified cross-validation procedure, where each validation set contained 20% of the data. The resulting model has an average validation accuracy of 58% and a Matthews correlation coefficient of 0.57. On the entire training data the model classified 9540/9945 variants (95%) correctly, of which 1638/1819 (90%) are SBI variants. Having assigned effects of the individual variants, we then used this data to pinpoint the functional residues in the target proteins. To this end, we assigned a residue to a class if at least half of the substitutions at that position belonged to that class. In particular, we focused our attention on amino acids classified as 'functional residues', where 50% or more of their variants are stable but inactive (Fig. 1B). We compared the functional positions identified from the MAVEs with the predictions from the model (Fig. 2). For the three proteins included in the training we correctly identify 116 out of 127 residues. Accuracy and true positive rates are similar for the three proteins (Fig. 2).

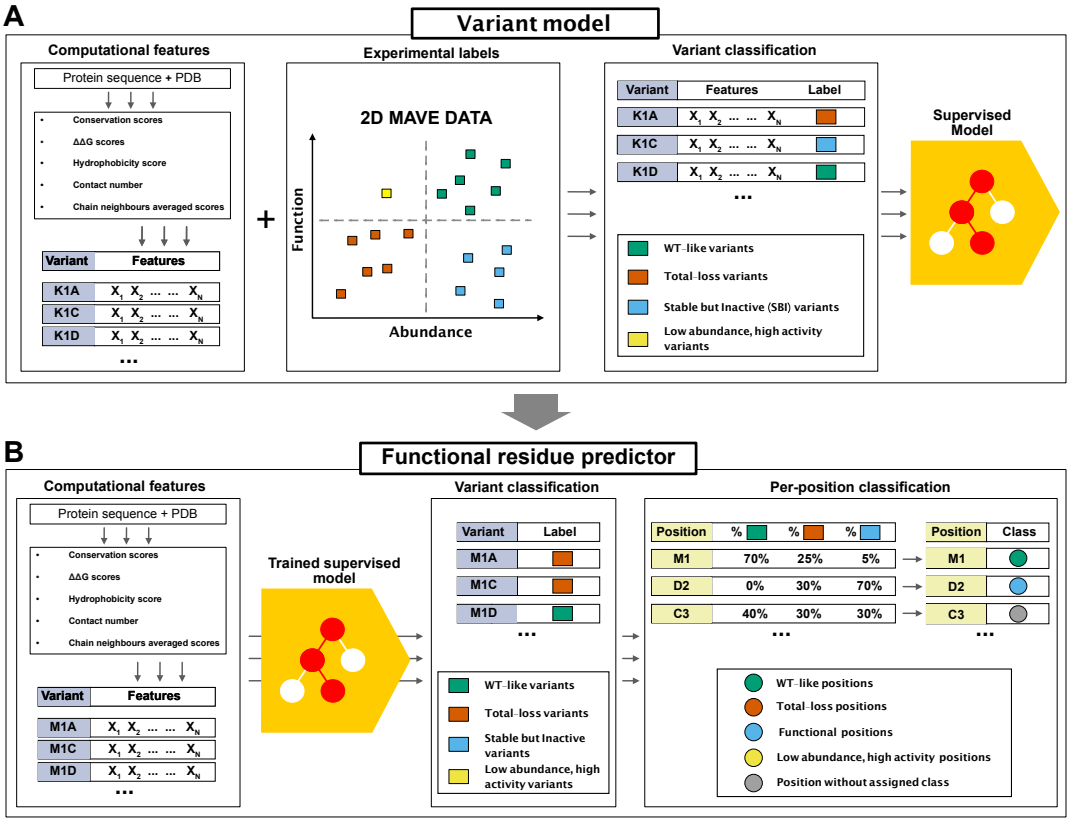


Figure 1. Graphical summary of our model to identify functionally important residues. (A) Using protein sequence and structure as input, we extract a number of features to characterise each variant. For each protein in our training set we extract all variants with MAVE measurements of abundance and function; this data is then combined with the structure and sequence features to train a gradient boosting classifier that assigns the variants to one of the four output classes. (B) The trained model takes the structure- and sequence-based features as input to classify all variants in to one of four classes. We also assign a class to each position/residue if half of the variants at that position are found in that class; remaining positions are not classified.

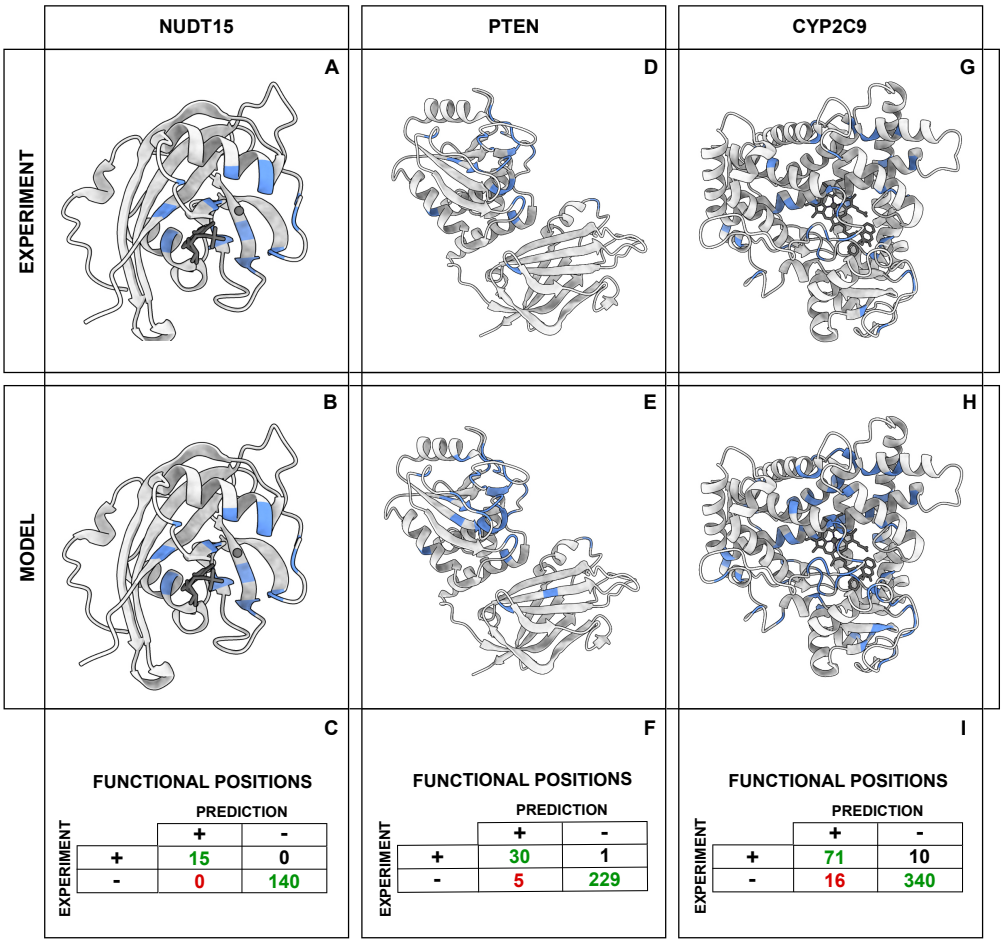


Figure 2. Finding functional residues in the training proteins. (A, D and G) Functional positions (in blue) identified using experiments. (B, E and H) Functional positions (in blue) identified using our model. (C, F, I) Statistics on the accuracy of residue level predictions of functional residues. For each of the three proteins we show the number of true positives (upper-left corner), true negatives (bottom-right corner), false positives (bottom-left corner) and false negatives (top-right).

Validation and applications

Our model predicts sites and residues that play a broad range of functional roles. This, however, makes it more difficult to validate the model as most experiments only probe the role of a small number of sites. We therefore tested the method on diverse tasks and data.

Validation using multiplexed and high-throughput data

First, we tested our model against a set of data that is relatively similar to the training data. Specifically, we applied it to two proteins (an SH3 domain from GRB2; Uniprot ID P62993 and the PDZ3 domain from PSD95; Uniprot ID P78352) that each have been assayed using two MAVEs that probe abundance and binding (***Faure et al., 2022***). A joint analysis of this data in turn enabled determining $\Delta\Delta G$ for both folding and binding. We then used our model to predict variants and residues that are important for folding and function (here binding), and compared the results to the experiments (Fig. S3). For the SH3 domain, we find a good agreement between the predictions and experiments. Specifically, we find that variants and residues that are predicted to cause loss of function due to loss of abundance generally are found to have a large $\Delta\Delta G$ for folding in the experiments (Fig. S3A,B). Similarly, the variants and residues that we predict to cause loss of function without loss of abundance, generally have a large $\Delta\Delta G$ for binding, but are experimentally found not to affect folding. Note that our model did not use information about binding interactions, but simply discovers residues important for peptide binding by the fact that they are conserved, but not for stability reasons. A similar analysis of the PDZ domain showed a more complex picture, where we find a correlation (rather than separation) between the $\Delta\Delta G$ for folding and binding, and that the predictions appear to divide the variants and residues into three categories with progressively greater $\Delta\Delta G$ values (Fig. S3C,D). We note here also that for PDZ3 we found a limited number of homologs in Uniref30 for the PDZ3 domain, resulting in a limited evolutionary characterization of the protein.

We then proceeded to analyse data from MAVEs on five different proteins to validate the performance of the variant classification step (Table 1). In contrast to the data that were used to train the model, we here only had data for a single MAVE for each protein. Four of these experiments are functional assays and the last a cellular-abundance assay. Because these experiments only probe one dimension of our two-dimensional 'function-and-abundance' landscape, we simplified the output labels from the model from four to two classes. For the four MAVEs that probe function we combined the model's output into either functional or inactive (independent of the predicted effect on abundance), and for the abundance-based MAVE we combine the predicted labels into high/low abundance (independent of the predicted effect on function). We compared these predictions to the data generated by the MAVEs which we also reduced to binary labels. On average we find an accuracy of 72% over all the variants and 68% when focusing on the variants labelled as SBI by our model (Table 1).

We also compared our predictions with the results of a high-throughput experiment on alkaline phosphatase (PafA; Uniprot ID Q9KJX5, ***Markin et al.*** (***2021***)). In these experiments, the catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$) was measured for the wild-type protein as well as variants where each residue was changed into either glycine or valine. We first analyzed how well our model was able to distinguish between variants with different levels of activity, and found a lower median activity ($k_{\text{cat}}/K_{\text{M}}$) for variants classified as total-loss and SBI compared to WT-like variants (Fig. S4A,B). The experiments also revealed different effects of valine and glycine variants, and we analysed this observation in light of our predictions. For buried positions, we find that that variants at total-loss positions generally gave rise to lowered activity, in particular when substituting to glycine (Fig. S4C). At exposed positions, we generally predicted variants to have a smaller effect (Fig. S4D). While the experiments revealed that most variants did not cause global unfolding, a number of variants appeared to affect the *in vitro* translation that was used to produce the protein variants (***Markin et al., 2021***). In particular, some variants were found to produce less active protein when expressed at 37 °C compared to 23 °C. Using a ten-fold change as a cutoff (***Markin et al., 2021***), we find most (70%) variants that

Table 1. Validation using multiplexed assays of variant effects.

Protein	PDB	MAVE	SBI accuracy	Global accuracy
Cystathionine beta-synthase	4COO	***Sun et al.*** (***2020***)	0.75	0.68
Beta-lactamase TEM	1BTL	***Stiffler et al.*** (***2015***)	0.76	0.72
Regulatory protein GAL4	3COQ	***Kitzman et al.*** (***2015***)	0.69	0.72
Small ubiquitin-related modifier 1	1WYW	***Weile et al.*** (***2017***)	0.75	0.63
Thiopurine S-methyltransferase	2H11	***Matreyek et al.*** (***2018***)	0.54	0.76

show this level of difference belong to the total-loss class (Fig. S4E,F), in line with the fact that these are predicted to cause folding defects.

Predicting functional sites in enzymes and protein interactions

As an example of the kinds of insights one might gain on specific proteins, we applied the model to the Anti-sigma F factor (Uniprot: O32727, Fig. 3A), for which previous studies have shown the importance of a number of sites (***Campbell et al., 2002***; ***Fu et al., 2019***). Our model predicted 509 SBI variants (20% of the total number of variants) and we used these to label 21 positions as predicted functional sites. When comparing these to previous biochemical studies, we found reported functional roles for 17 of the 21 predicted positions (***Campbell et al., 2002***; ***Masuda et al., 2004***). Of these 17 amino acids, nine are located in the proximity of the active site (Fig 3B). His54, Gly55, Thr99, Gly107, Gly109 and Thr130 interact and stabilise the ADP/ATP in the active site, Asn50 is involved in the chelation of a Mg^{+2} ion, Glu46 acts as the catalytic base in the kinase reaction and Arg105 stabilises the transition state in the kinase reaction. In addition to these active site residues, we found another cluster of functional residues (Glu104, Thr49, Glu16, Ser45, Glu39 and Arg20) in the proximity of the binding site for the Anti-sigma F factor antagonist (Uniprot O32723, Fig. 3B), while predicted functional positions Arg20 and Lys41 have been described as mediators of the interaction of the enzyme with the sigma factor. The roles of the remaining four positions that our model highlighted (Asn3, Asn15, Gly129 and Pro95) have, to our knowledge, not been analysed; they could either be false positives or residues with functional roles that have not yet been characterized. We observed that the functional sites in the Anti-sigma F factor could be grouped in two structurally compact clusters (Fig. S5). The first cluster includes all of the residues that have a role in the catalytic process, while the second cluster contains positions in the interaction network with Anti-sigma F factor antagonist and the sigma-factor.

Having shown the model performance using the Anti-sigma F factor as an example, we extended this analysis of functional positions to a larger data set, consisting of 20 enzymes in the Mechanism and Catalytic Site Atlas database (M-CSA, ***Ribeiro et al.*** (***2018***)), as well as five non-enzymes from the Protein-Protein Interaction Affinity Database 2.0 (***Vreven et al., 2015***). We applied our model to these 25 proteins and identified 16167 SBI variants (15.2% of the total) and assigned 588 residues to be important for function (12.7% of the total available positions). We found a small difference in the fraction of predicted functional sites between the enzymes (13.9% of the total positions) and the other five proteins (11.6%). For the 20 proteins in M-CSA, we collected a curated list of 87 residues known to be involved in catalysis (***Ribeiro et al., 2018***). Given the well-defined functional roles we would expect that most substitutions at these sites would affect the enzymatic function. Our model assigned 62 of these 87 positions (71%) as functional positions (Fig. 4A). In 9 out of 20 enzymes the entire set of catalytic residues were classified as functional residues, while in the rest of the dataset the fraction of matching sites ranged between 20% and 80%.

We examined the 25 of the 87 positions that our model did not assign as SBI in more detail. Most of these (17/25 residues; 68%) have a median $\Delta\Delta G$ greater 2 kcal/mol, and our model labelled them as 'total-loss' positions. This finding highlights one of the limitations of our approach. While we can assign conserved residues that do not affect stability to have a likely functional role beyond

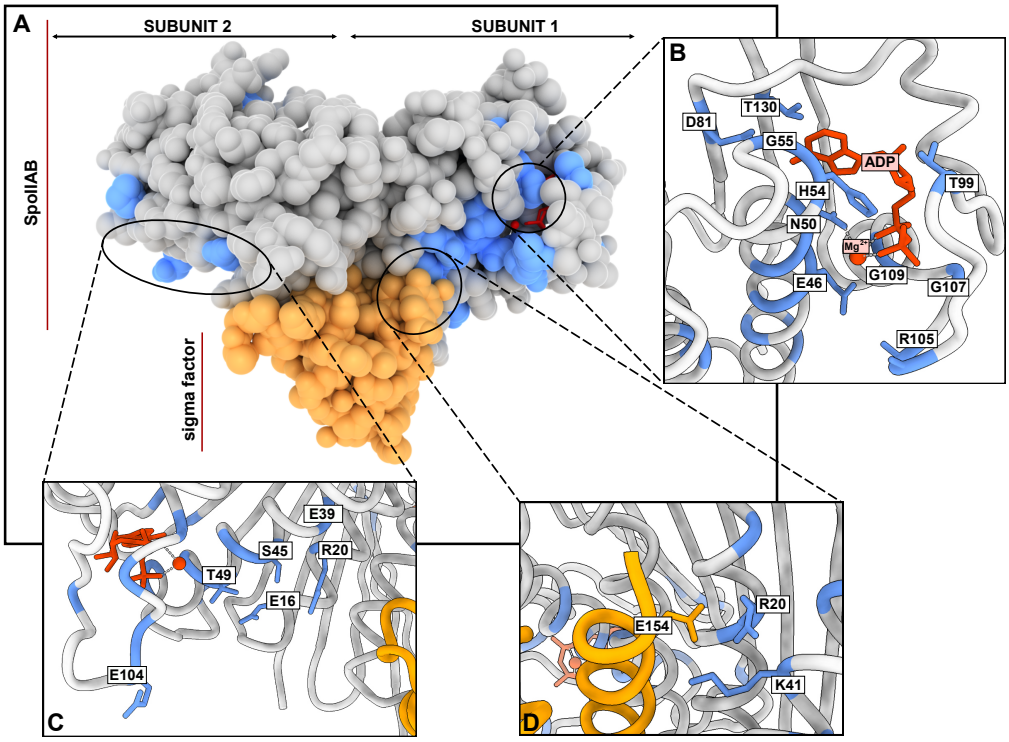


Figure 3. Identifying functional residues in an Anti-sigma F factor (UNIPROT O32727, PDB: 1I0O). (A) The dimeric Anti-sigma F factor is shown using Van der Walls surfaces (grey). Predicted functional sites are coloured in blue, Mg^{+2}:ADP is shown in red and the bound sigma factor is shown in yellow. (B, C and D) Functional sites identified by our model and previously described in the literature (***Campbell et al., 2002***; ***Masuda et al., 2004***) and labelled. (B) Predicted functional sites in the proximity of the active site and reported to influence activity. CD) Predicted functional sites involved in interactions with the σ factor. (D) Predicted sites that interact with the Anti-sigma F factor antagonist SpoIIAA.

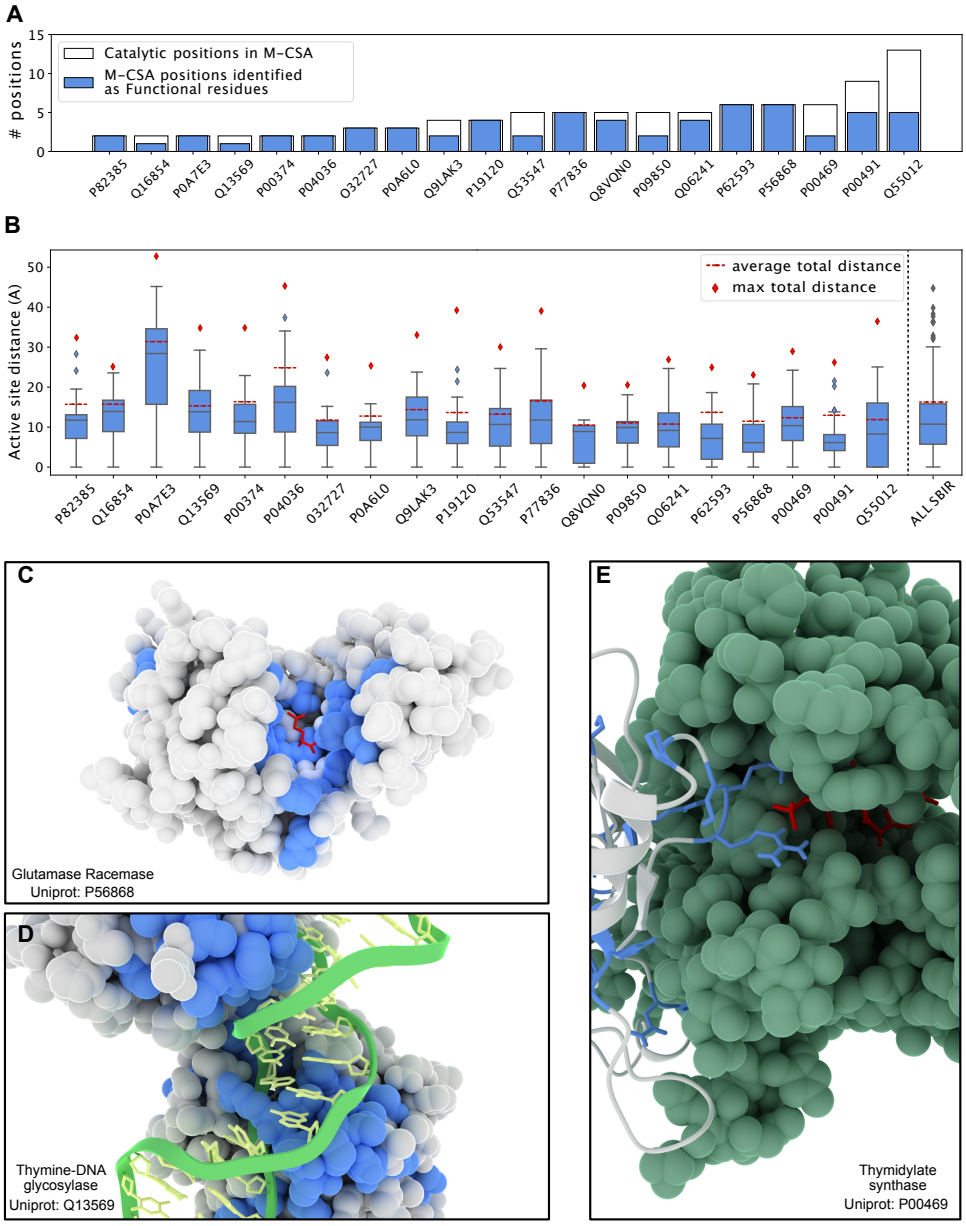


Figure 4. Predicting functional sites in enzymes. We used our model to study functional residues in a set of 20 enzymes from M-CSA, which also provides annotations of residues in catalytic sites. (A) Number of known catalytic residues for each protein in M-CSA (white) and the number of these catalytic residues classified as functional positions by our model (blue). (B) Functional sites are generally close to the active site. For each protein, we show the distribution of distances (using a boxplot) between the predicted functional sites and the (nearest) active site residue (from M-CSA). For comparison we show the average and maximum pairwise residue-residue distances (red dotted line and a red squared dot, respectively). The rightmost boxplot shows the cumulative data. (C, D and E) Examples of predicted functional sites (blue) in three proteins from the M-CSA set. (C) Functional sites in glutamase racemase are found in a single cluster close to the active site. (D) Functional sites in thymine-DNA glycosylase are both located in the active site, but also in the region needed to bind the target DNA chain. (E) Predicted functional sites are also found in protein-protein interfaces, such as in the interface of homo-dimeric thymidylate synthase.

structural stability, the reverse is not true. In these particular cases, ca. 20% (17/85) of the amino acids in the active sites appeared to be important both for function and structural stability. As an example we show the results for an Endo-1,4-beta-xylanase (Fig. S7) where three of the five catalytic positions were assigned as functional residues by our model, and the remaining two have a median $\Delta\Delta G > 2$ kcal/mol.

A strength of our model is that it identifies residues with likely function roles beyond those directly involved e.g. in catalysis. We examined the positions that our model predicted to be functional sites in the 20 enzymes from M-CSA, and found that a substantial fraction of these are localised in the proximity of the catalytic site (Fig. 4B); on average 48% of the predicted functional residues are located less than 10 Å from the closest catalytic site residue. We expect that many of these residues are important for the catalytic process, as we show for glutamate racemase (Uniprot: P56868, Fig. 4C), where 19 out of 26 of the predicted functional residues (including M-CSA catalytic positions) are less than 10 Å from the D-glutamine in the catalytic site (***Hwang et al., 1999***).

In some cases we also found that the predicted functional sites were located both close to as well as further away (20Å or more) from the catalytic site, as illustrated by OPRTase (Uniprot: P0A7E3, Fig. S8), where predicted functional residues are located near several distinct interaction sites. Most proteins in the M-CSA set have clusters of predicted functional sites further away from catalytic sites. We found that these positions, for example, may be involved in interactions with substrates, but not directly involved in the catalytic activity. As an example we show a thymine-DNA glycosylase (Uniprot Q13569, Fig 4D), where we identified a subset of functional residues far (14.5 Å) from the enzymatic active site. These include Arg275 and the residues in loop 274-277 (***Maiti et al., 2009***; ***Fu et al., 2019***). These amino acids are not directly involved in the catalysis, but their role is to push the target DNA base into the catalytic pocket, which consists of two residues: Arg140 and His151 (***Maiti et al., 2009***; ***Kanaan et al., 2015***). Our model identified the former as a functional position, while the latter was predicted to be a total-loss position.

In addition to finding positions that are important to catalysis and binding of substrates, our model can also help identify positions that play a role in forming in protein-protein interactions as we show for thymidylate synthase (P00469, Fig. 4E), a homo-dimeric enzyme with a catalytic pocket in both of the subunits. Of the 41 residues identified by our model, 20 are located close to the protein-protein interface. We note here that all of our stability calculations are performed on monomeric structures, and so we identify these residues as being conserved, but not due to the structural stability of the individual subunits. Looking at the predicted functional sites at the interface, we find that some are involved in forming the active site which includes residues from both subunits (Arg178,Cys198, ***Pookanjanatavip et al.*** (***1992***)). We, however, also found a number of other residues, which we suggest are conserved as they stabilize the (obligatory) dimer structure.

Predicting and understanding disease variants in HPRT1

Having validated and exemplified our model using previously published data, we then used our model in a prospective study to predict the impact and mechanism of human missense variants. We and others have previously shown that many, but not all, disease-causing missense variants cause loss of function by loss of abundance. The ability to assign a functional status to so-called variants of uncertain significance is one of the major outstanding challenges in clinical genetics. In order to understand the molecular origin of disease and develop potential treatments it is, however, also important to be able to predict why variants cause loss of function.

We therefore selected hypoxanthine-guanine phosphoribosyltransferase-1 (HPRT1, Uniprot: P00492), an enzyme involved in the onset of Lesch-Nyhan disease and its attenuated variants, for a prospective study of the accuracy and utility of our model (***Fu and Jinnah, 2012***; ***Fu et al., 2014***). We estimated structural and sequence features for 190 of 210 residues in HPRT1; of the 3610 total possible single amino acid variants, our model predicts 471 to be SBI and 1046 to be total-loss variants.

We then selected 17 variants for experimental characterization. These variants were selected

either from gnomAD (***Karczewski et al., 2020***), ClinVar (***Landrum et al., 2018***) or variants that we selected to test our model more broadly. Five of the variants were predicted to have wild-type-like activity, six variants were predicted to be total-loss (i.e. loss of activity and loss of abundance), and six variants were predicted to be SBI (i.e. loss of activity without loss of abundance) (Fig 5).

To test our predictions, we established a yeast-based growth assay for HPRT1 function. HPRT1 catalyses the formation of inosine and guanosine monophosphate (IMP and GMP) from hypoxanthine and guanine, respectively. Previous studies have shown that mycophenolic acid (MPA) acts as an inhibitor of IMP dehydrogenase, which is responsible for the conversion of IMP to xanthosine monophosphate (XMP, ***Woods et al.*** (***1983***); ***Escobar-Henriques and Daignan-Fornier*** (***2001***)). Thus, in the presence of MPA, yeast cells can only generate GMP through the salvage pathway, and the yeast HPRT1 orthologue, Hpt1, therefore becomes essential (Fig. S9). We introduced each of the 17 variants as well as a wild-type and vector controls into a yeast strain lacking Hpt1 to test their effects on function and abundance. We found that 11/17 variants grew worse that the wild-type control in the presence of MPA, showing that they cause loss of function (Fig 5B). We also measured the abundance of the wild-type and 17 variants using western blots and found that 3/17 variants had substantially reduced levels (Fig 5C).

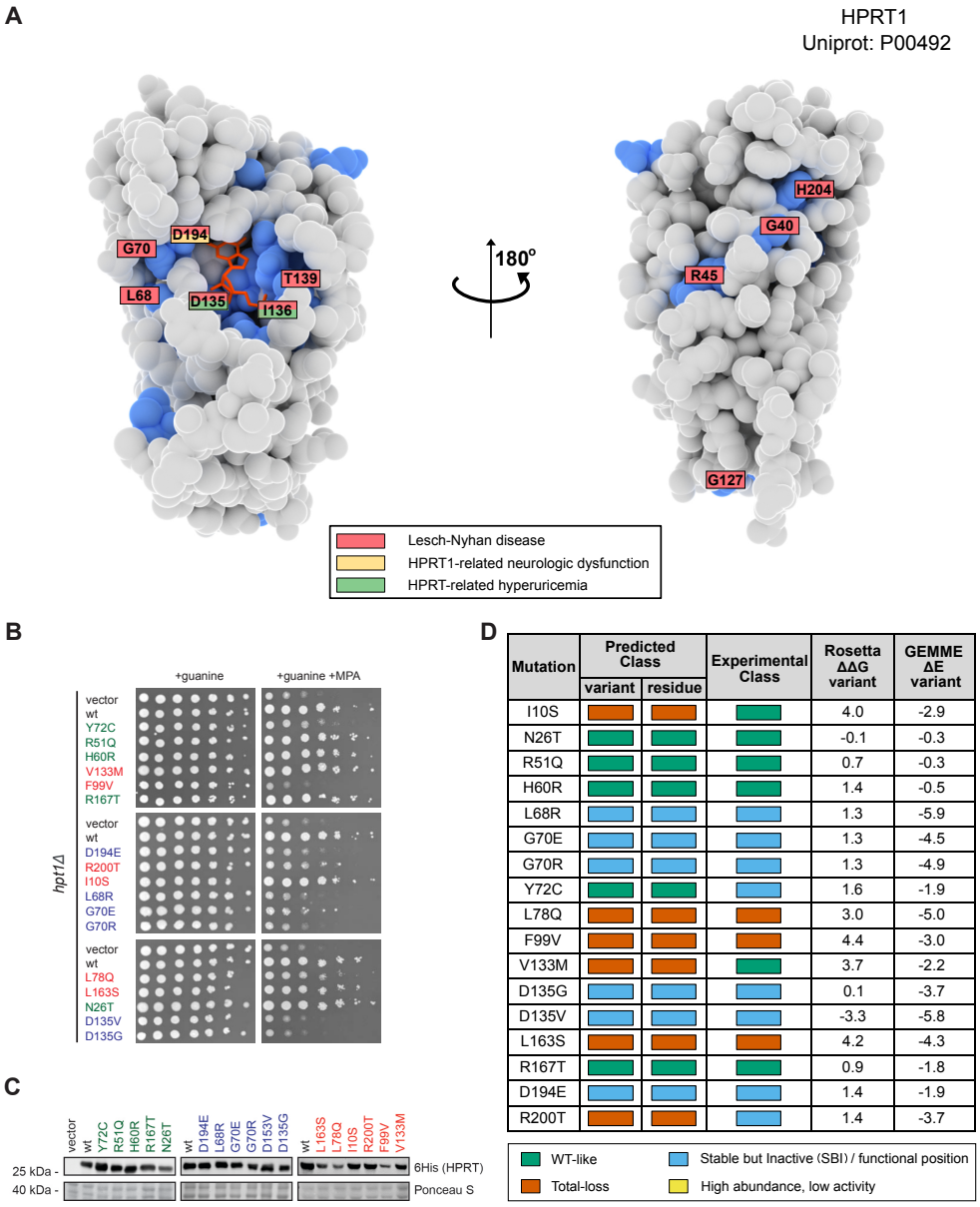
We used the experimental data to classify the variants into wild-type-like, SBI and total-loss and compared the results to the computational predictions (Fig 5D). Overall we find that 13/17 (76%) of the variants are predicted correctly including all (6/6) of the SBI residues. This result highlights that our model can predict variant effects relatively well, and in particular can help separate loss-of-function variants that lose function due to intrinsic function from those due to structural stability. We note that some of these variants have also been characterized biochemically (***Fu and Jinnah, 2012***), with overall good agreement between our predictions, the yeast assays and the biochemical experiments.

Discussion

There is a long history of studying protein function through analyses of sequence conservation, and conserved residues are generally important for the protein. Because most proteins need to be folded to function it is, however, difficult to disentangle the role of individual amino acids on the stability of the overall fold from more direct roles of individual amino acids for protein function. Clearly, it may not always be possible to separate the two, and we have shown for example that substitutions at some catalytic residues may result in loss of stability. In many cases, however, substitutions at functionally important sites either have a small negative effect on stability, or may even give rise to increased stability (***Shoichet et al., 1995***; ***Bloom et al., 2006***). Building on these ideas we have here presented a method that finds functionally important sites as those that are conserved, but not immediately due to a role in protein stability.

To construct and train our model we leverage (i) large-scale experimental assays reporting on protein activity and abundance and (ii) computational methods to estimate variant effects on protein stability as well as general functional effects using conservation in multiple sequence alignments. We have validated our model using both detailed biochemical experiments on individual proteins (via M-CSA), as well as larger scale data generated by high-throughput experiments. We also validated the method through prospective predictions of the mechanism of disease variants in HPRT1.

Our model for functionally important sites, in some sense, lies between two extremes that have previously been used to analyse proteins. At one end of the scale, activity-based MAVEs and sequence conservation analyses provide a global view of residues that play functional roles, but often do not give direct mechanistic insights. At the other end of the scale, detailed biochemical, biophysical and structural experiments are often needed to pinpoint how individual amino acids affect function, but are (with exceptions such as work by ***Markin et al.*** (***2021***)) difficult to scale to a large number of variants and proteins. To test our model we have thus both compared it to data generated by MAVEs, reducing to a one-dimensional view when only a single assay is available, as



Mutation	Predicted Class		Experimental Class	Rosetta ΔΔG variant	GEMME ΔE variant
	variant	residue			
I10S				4.0	-2.9
N26T				-0.1	-0.3
R51Q				0.7	-0.3
H60R				1.4	-0.5
L68R				1.3	-5.9
G70E				1.3	-4.5
G70R				1.3	-4.9
Y72C				1.6	-1.9
L78Q				3.0	-5.0
F99V				4.4	-3.0
V133M				3.7	-2.2
D135G				0.1	-3.7
D135V				-3.3	-5.8
L163S				4.2	-4.3
R167T				0.9	-1.8
D194E				1.4	-1.9
R200T				1.4	-3.7

Figure 5. Predicting consequences of missense variants in HPRT1. (A) Structure of HPRT1 shown as Van der Walls surfaces with the predicted functional residues coloured in blue. Residues which contain at least one disease-causing variant are labelled and coloured according to which diseases it has been associated with (see legend). (B) Results of a yeast-based assay for HPRT1 function. Yeast strains carrying a vector control, wild-type HPRT1 or one of the 17 HPRT1 variants all grow on medium containing guanine. In the presence of the inosine monophosphate dehydrogenase inhibitor mycophenolic acid (MPA), yeast cells cannot grow in the absence of a functional HPRT1 protein. (C) Assessment of protein abundance using western blots of the 6His-tagged HPRT1 variants. (D) Comparing predictions of the effects of the variants with the experimental measurements. Rosetta $\Delta\Delta G$ values are in units of kcal/mol.

well as results from detailed assessments of the role of individual amino acids in enzymes and protein-protein interactions.

Technological and methodological developments have led to rapid increases in the number of known protein sequences, as well as our ability to assign overall biological functions to many of these. The ability to identify functionally important sites in proteins is important for our understanding of how proteins work, our ability to engineer enzymes and to understand the mechanisms that underlie diseases. Our model may, for example, be used to select residues for protein engineering experiments, or for generating focused libraries in enzyme optimization. As we have illustrated, the model can help pinpoint the molecular mechanisms of disease variants, which may both guide further experiments and be used as starting point for developing mechanism-based therapies. We hope that the method that we have described here will help researchers make progress in these and other areas.

Methods

Sequence features

For each protein in the training (Table S1) and validation (Tables S2, S3 and S4) data we performed a statistical analysis of multiple sequence alignments (MSA). We extracted the sequence of the first isoform from UniProt and used HHBlits (***Remmert et al., 2012***) to build an MSA, on which we applied additional filters before evaluating the substitution effects and conservation scores. The first filter keeps only the positions (columns) that are present in the target sequence, and the second filter removes the sequences (rows) where the number of total gaps exceeds 50% of the total number of positions. We used GEMME (***Laine et al., 2019***) to calculate an evolutionary conservation score. GEMME explicitly models the evolutionary history of the protein, returning a score ($\Delta\Delta E$) which estimates an 'evolutionary distance' of a variant from the query wild-type sequence. The $\Delta\Delta E$ scores for the variants range between 0 (conservative substitution) and -7 (substitutions that appear incompatible based on the MSA). In addition to the effect of the individual variants, we also calculated a 'neighbour score' as the average of the $\Delta\Delta E$ values for the previous and proceeding residue. In addition to the $\Delta\Delta E$ scores, we used the hydrophobicity (***Monera et al., 1995***) of the target amino acid as a feature in our model.

Structural features

We predicted changes in thermodynamic stability using Rosetta (GitHub SHA1 99d33ec59ce9fcecc5e4f3800c778a54afdf8504). We used the Cartesian ddG protocol (***Park et al., 2016***) on crystal structures corresponding to the Uniprot sequence listed in Tables S1,S2 and S3. The $\Delta\Delta G$ values obtained from Rosetta were divided by 2.9 to bring them from Rosetta energy units onto a scale corresponding to kcal/mol (***Park et al., 2016***). We used these $\Delta\Delta G$ values as a feature for the classifier, setting all values below/above a range of 0 – 5 kcal/mol, to 0 or 5 kcal/mol, respectively. We also calculated a neighbour score, averaging over the two nearest neighbours, as for the $\Delta\Delta E$ values.

Finally, we used MDTraj v.1.9.3 (***McGibbon et al., 2015***) to calculate the weighted contact number (WCN) for each residue:

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

where $r_{i,j}$ is the C_α distance between residue i and j, r_0 is a switching parameter (set to 7.0Å).

We determined solvent exposed and buried regions using GetArea (***Fraczkiewicz and Braun, 1998***) using 20% as a threshold to divide the data in to two classes.

Training the classifier

We used eight different features as input to our classifier. Each variant was pairs with (1, 2) variant $\Delta\Delta G$ and $\Delta\Delta E$, (3–6) residue average and neighbour average $\Delta\Delta G$ and $\Delta\Delta E$ scores, (7) the hydrophobicity of the target amino acid, and (8) the WCN.

We used a gradient descent machine to classify the variants, as implemented in Catboost (***Prokhorenkova et al., 2017***) v.0.26.1, using a Multinomial/Multiclass Cross Entropy Loss and L2 regularization. We first set the number of gradient descent iterations (Fig. S2A) on the vanilla model and then used a grid scan to find optimal value for the remaining hyperparameters. We measured precision and recall using five-fold cross validation on the training data to select the optimal of hyperparameters (Fig. S2B). We compared the model performance with a null model from the sklearn python package (*sklearn.dummy.DummyClassifier*) using the 'prior' strategy, which returns the most frequent class label in the observed argument passed to it. We also used a random forest classifier model (*sklearn.ensemble.RandomForestClassifier*) in the comparison, optimizing the hyperparameters using a grid scan protocol and k-fold cross validation.

Cloning

Full-length human HPRT1 was expressed in yeast from the pYES2 vector (Invitrogen). An N-terminal RGS6His-tag was inserted upstream of an SRS linker peptide, before HPRT Met1. Point mutations were generated by Genscript.

Yeast strains and techniques

The *hpt1*Δ (MatA, *his3*Δ*1, leu2*Δ*0, met15*Δ*0, ura3*Δ*0, hpt1::KanMX*) *Saccharomyces cerevisiae* yeast strain was from Invitrogen. The cells were cultured in synthetic complete (SC) medium (2% glucose, 6.7 g/L yeast nitrogen base without amino acids and with ammonium sulfate (Sigma)) and supplemented for selection with 1.92 g/L uracil drop-out mix (US Biological). For expression, the glucose was replaced with 2% galactose.

Yeast transformations were performed with lithium acetate (***Gietz and Schiestl, 2007***). For growth assays, the cells were cultured at 30 °C to exponential phase and diluted to an OD(600nm) of 0.40. From this, dilution series (5-fold) were prepared and 5 μL of each dilution was applied as droplets on agar plates. Colonies formed after 2–3 days of incubation at 30 °C. The agar contained 50 μg/mL guanine (Sigma) and 0.1 mg/mL MPA (Sigma).

Whole cell lysates for western blotting were prepared from exponential phase cultures using glass beads and trichloroacetic acid (TCA) as described before (***Kampmeyer et al., 2022***). Briefly, $1.2 \cdot 10^8$ cells in exponential phase were harvested and washed in water by centrifugation (3000 g, 5 min.). The cells were resuspended in 1 mL of 20% TCA and centrifuged (3000 g, 5 min.). The supernatant was discarded and the pellet was resuspended in 200 μL 20% TCA and transferred to 2 mL screwcap tubes containing 0.5 mL 400-600 micron glass beads (Sigma). The tubes were then applied to a Mini-BeadBeater machine (BioSpec Products Inc.) set at 3 cycles of 10 s. With a needle, a hole was made in the bottom of the tube and the tube was placed inside a 15 mL tube. Then, 400 μL 5% TCA was added and the material was eluted by centrifugation (1000 g, 5 min.) into the 15 mL tube. The eluted material was centrifuged (10000 g, 5 min.) at 4 °C. The pellet washed twice with ice-cold 80% acetone. Finally, the pellet was resuspended in 100 μL sample buffer for SDS-PAGE (62.5 mM Tris/HCl pH 6.8, 2% SDS, 25% glycerol, 0.01% bromphenol blue, 5% β-mercaptoethanol) and incubated for 5 min. at 100 °C.

SDS-PAGE and western blotting

SDS-PAGE was performed using 12.5% acrylamide gels. After electrophoresis, the proteins were transferred to 0.2 μm nitrocellulose membranes (Advantec) by electro-blotting. After transfer, the blots were stained with Ponceau S (0.1% Ponceau S in 5% acetic acid) and blocked in PBS (10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 137 mM NaCl, 3 mM KCl, pH 7.4) with 5% skimmed milk powder. The

antibodies were anti-RGSHis (Qiagen, Cat. No. 34650) and peroxidase-conjugated anti-mouse antibody (Dako, Cat. No. P0260).

Data and code availability

Data available at https://github.com/KULL-Centre/papers/tree/main/2022/functional-sites-cagiada-et-al

Acknowledgments

We acknowledge access to computing resources from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen. The research was supported by the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (NNF18OC0033950, to A.S., R.H.P. and K.L.-L.).

References

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.

Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, Kowalski JP, Rettie AE, Fowler DM, Dunham MJ. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. bioRxiv. 2021; .

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences. 2006; 103(15):5869–5874.

Cagiada M, Johansson KE, Valančiūtė A, Nielsen SV, Hartmann-Petersen R, Yang JJ, Fowler DM, Stein A, Lindorff-Larsen K. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. bioRxiv. 2020; .

Campbell EA, Masuda S, Sun JL, Muzzin O, Olson CA, Wang S, Darst SA. Crystal structure of the Bacillus stearothermophilus anti-σ factor SpoIIAB with the sporulation σ factor σF. Cell. 2002; 108(6):795–807.

Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology. 2009; 5(12):e1000585.

Cheng G, Qian B, Samudrala R, Baker D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Research. 2005; 33(18):5861–5867.

Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, Sun S, Roth FP, DeSloover D, Marks DS, et al. 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.

Dunham AS, Beltrao P. Exploring amino acid functions in a deep mutational landscape. Molecular systems biology. 2021; 17(7):e10305.

Echave J, Wilke CO. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. Annual review of biophysics. 2017; 46:85–103.

Escobar-Henriques M, Daignan-Fornier B. Transcriptional regulation of the yeast gmp synthesis pathway by its end products. Journal of Biological Chemistry. 2001; 276(2):1523–1530.

Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022; 604(7904):175–183.

Fraczkiewicz R, Braun W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. Journal of computational chemistry. 1998; 19(3):319–333.

Fu R, Ceballos-Picot I, Torres RJ, Larovere LE, Yamada Y, Nguyen KV, Hegde M, Visser JE, Schretlen DJ, Nyhan WL, et al. Genotype–phenotype correlations in neurogenetics: Lesch-Nyhan disease as a model disorder. Brain. 2014; 137(5):1282–1303.

Fu R, Jinnah HA. Genotype-phenotype correlations in Lesch-Nyhan disease. Journal of Biological Chemistry. 2012; 287(5):2997–3008.

Fu T, Liu L, Yang QL, Wang Y, Xu P, Zhang L, Liu S, Dai Q, Ji Q, Xu GL, et al. Thymine DNA glycosylase recognizes the geometry alteration of minor grooves induced by 5-formylcytosine and 5-carboxylcytosine. Chemical science. 2019; 10(31):7407–7417.

Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nature protocols. 2007; 2(1):31–34.

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.

Høie MH, Cagiada M, Frederiksen AHB, Stein A, Lindorff-Larsen K. Predicting and interpreting large scale mutagenesis data using analyses of protein stability and conservation. bioRxiv. 2021; .

Hwang KY, Cho CS, Kim SS, Sung HC, Yu YG, Cho Y. Structure and mechanism of glutamate racemase from Aquifex pyrophilus. Nature structural biology. 1999; 6(5):422–426.

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.

Kampmeyer C, Larsen-Ledet S, Wagnkilde MR, Michelsen M, Iversen HK, Nielsen SV, Lindemose S, Caregnato A, Ravid T, Stein A, et al. Disease-linked mutations cause exposure of a protein quality control degron. Structure. 2022; .

Kanaan N, Crehuet R, Imhof P. Mechanism of the glycosidic bond cleavage of mismatched thymine in human thymine DNA glycosylase revealed by classical molecular dynamics and quantum mechanical/molecular mechanical calculations. The Journal of Physical Chemistry B. 2015; 119(38):12365–12380.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434–443.

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.

Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. Nature methods. 2015; 12(3):203–206.

Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018; 34(4):660–668.

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):1073–1081.

Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. Molecular biology and evolution. 2019; 36(11):2604–2619.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2018; 46(D1):D1062–D1067.

Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nature reviews molecular cell biology. 2007; 8(12):995–1005.

Li X, Lehner B. Biophysical ambiguities prevent accurate genetic prediction. Nature communications. 2020; 11(1):1–11.

Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. Journal of molecular biology. 1996; 257(2):342–358.

Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Molecular systems biology. 2020; 16(7):e9380.

Maiti A, Morgan MT, Drohat AC. Role of two strictly conserved residues in nucleotide flipping and N-glycosylic bond cleavage by human thymine DNA glycosylase. Journal of Biological Chemistry. 2009; 284(52):36680–36688.

Markin C, Mokhtari D, Sunden F, Appel M, Akiva E, Longwell S, Sabatti C, Herschlag D, Fordyce P. Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. Science. 2021; 373(6553):eabf8761.

Masuda S, Murakami KS, Wang S, Olson CA, Donigian J, Leon F, Darst SA, Campbell EA. Crystal structures of the ADP and ATP bound forms of the Bacillus anti-σ factor SpoIIAB in complex with the anti-anti-σ SpoIIAA. Journal of molecular biology. 2004; 340(5):941–956.

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.

McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophysical Journal. 2015; 109(8):1528 – 1532. doi: 10.1016/j.bpj.2015.08.015.

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.

Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodges RS. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. Journal of peptide science: an official publication of the European Peptide Society. 1995; 1(5):319–329.

Nielsen SV, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Multiplexed assays reveal effects of missense variants in MSH2 and cancer predisposition. PLoS Genetics. 2021; 17(4):e1009496.

Otwinowski J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. Molecular biology and evolution. 2018; 35(10):2345–2354.

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.

Pookanjanatavip M, Yuthavong Y, Greene PJ, Santi DV. Subunit complementation of thymidylate synthase. Biochemistry. 1992; 31(42):10303–10309.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:170609516. 2017; .

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al. A large-scale evaluation of computational protein function prediction. Nature methods. 2013; 10(3):221–227.

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.

Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic acids research. 2018; 46(D1):D618–D623.

Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature methods. 2018; 15(10):816–822.

Sereda TJ, Mant CT, Sönnichsen FD, Hodges RS. Reversed-phase chromatography of synthetic amphipathic α-helical peptides as a model for ligand/receptor interactions Effect of changing hydrophobic environment on the relative hydrophilicity/hydrophobicity of amino acid side-chains. Journal of Chromatography A. 1994; 676(1):139–153.

Shih CH, Chang CM, Lin YS, Lo WC, Hwang JK. Evolutionary information hidden in a single protein structure. Proteins: Structure, Function, and Bioinformatics. 2012; 80(6):1647–1657.

Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. Proceedings of the National Academy of Sciences. 1995; 92(2):452–456.

del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. Journal of molecular biology. 2003; 326(4):1289–1302.

Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell. 2015; 160(5):882–892.

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.

Sun S, Weile J, Verby M, Wu Y, Wang Y, Cote AG, Fotiadou I, Kitaygorodsky J, Vidal M, Rine J, et al. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. Genome medicine. 2020; 12(1):1–18.

Torng W, Altman RB. High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics. 2019; 35(9):1503–1512.

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. Journal of molecular biology. 2015; 427(19):3031–3041.

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 K, Horst JA, Cheng G, Nickle DC, Samudrala R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS computational biology. 2008; 4(9):e1000181.

Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, Wu Y, Pons C, Wong C, van Lieshout N, et al. A framework for exhaustively mapping functional missense variants. Molecular systems biology. 2017; 13(12):957.

Woods RA, Roberts DG, Friedman T, Jolly D, Filpula D. Hypoxanthine: guanine phosphoribosyltransferase mutants in Saccharomyces cerevisiae. Molecular and General Genetics MGG. 1983; 191(3):407–412.

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.

Supplementary Material

Supplementary Figures

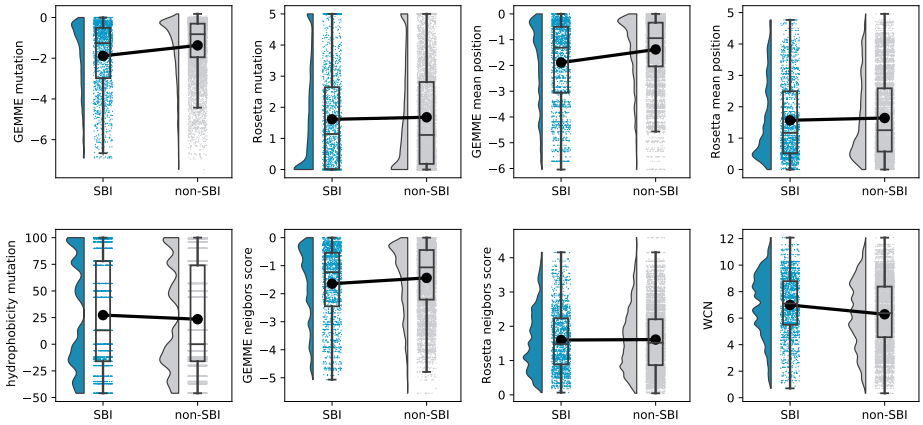


Figure S1. Comparing values of the individual features for SBI and non-SBI variants. Each raincloud plot shows, for each feature used in the model, the distribution of feature values in the training proteins for variants belonging to the SBI class (as assigned by experiments) or in one of the other three classes (WT-like, total loss and 'low abundance, high activity'). Each plot also shows the median, quartiles with a box plot and comparison between median values with a black line. Raw data are also displayed as points under the box plot.

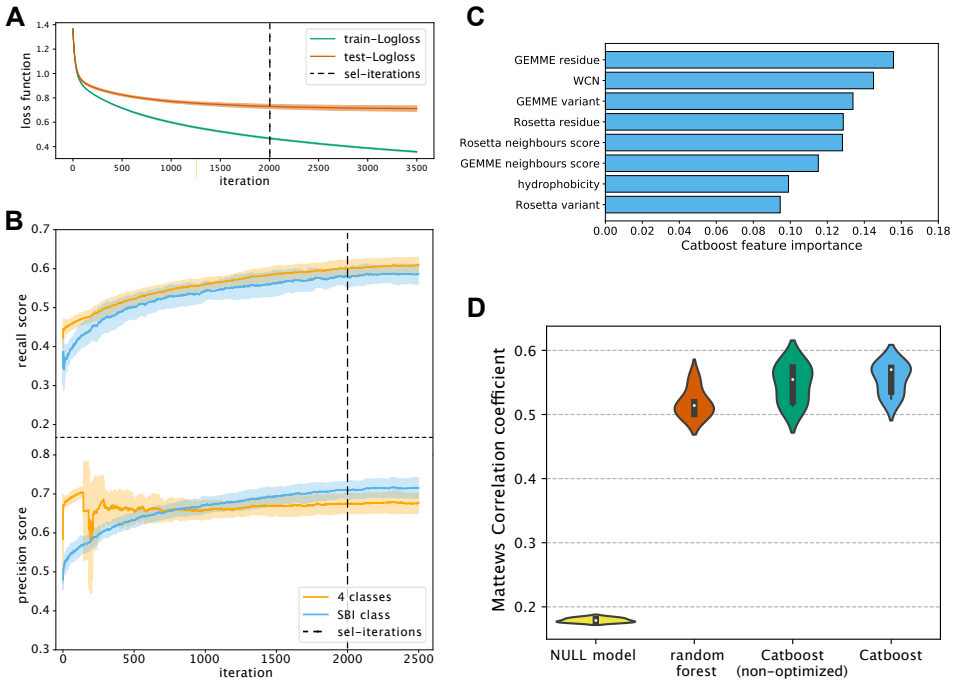


Figure S2. Details of the variant classifier training process. (A) Changes in the loss function for the (green) training data and (red) test data during model training. (B) Progress of performance (measuring recall and precision on the test set) of the Catboost classifier during training. The yellow lines represent the recall and precision for correctly classifing the variant in one of the four classes, while the blue lines show the performance of the model for classifing the SBI variants. The shaded areas in panels A and B represent the standard deviation from the k-cross validation procedure on the training dataset, and the selected number of iterations is shown as a black vertical line. (C) Feature importance for each of the features used to train the model. (D) Comparison of the performance of the Catboost model (using the Matthews' correlation coefficient), a Catboost version before optimizing hyper parameters, an optimized Random Forest model, and a Null model (which always returns the most frequent class label). The white points inside the 'violin plots' represent the median values and the black squared areas represent one standard deviation.

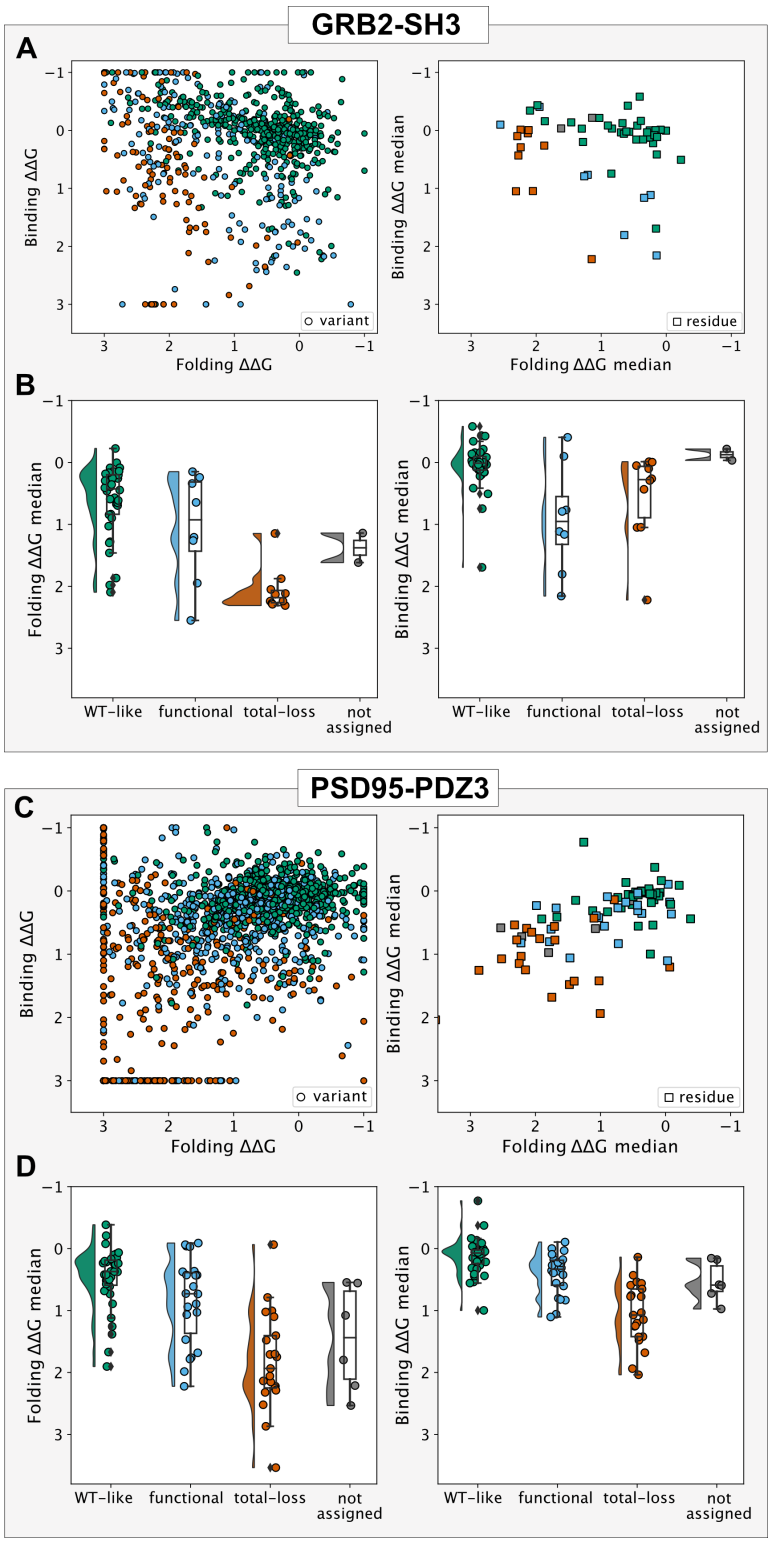


Figure S3. Comparison of experiments and predictions for the GRB2-SH3 and PSD95-PDZ3 domains. (A, C) For each of the two domains we used our model to classify variants and residues and compared to $\Delta\Delta G$ values for domain stability (Folding $\Delta\Delta G$) and peptide binding (Binding $\Delta\Delta G$) inferred from experiments. Variants and residues are coloured according to the predicted class (WT-like: green; SBI/Functional site: blue; Total-loss: red; Not assigned: grey). (B, D) Rain-cloud plots of the median value of the folding or stability $\Delta\Delta G$ values for different positions divided into the different classifications. The box plots show the median and quartiles (of the residue median values).

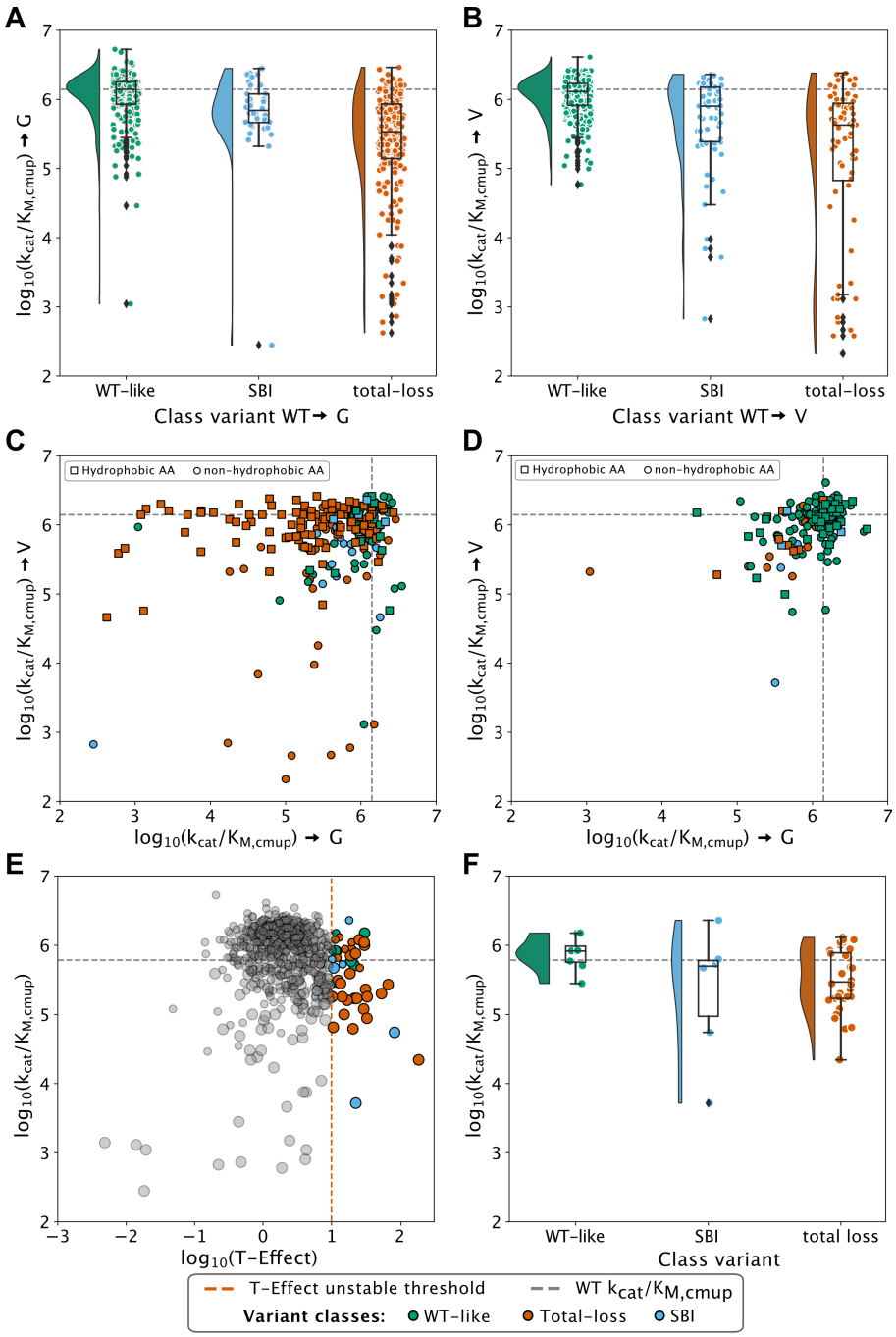


Figure S4. Comparison of experiments and predictions for PafA (Uniprot: Q9KJX5). Raincloud plots for experimentally determined $k_{cat}/K_{M,cMUP}$ values when the wild-type amino acid was was substituted for (A) glycine or (B) valine. In both A and B, the values are shown separately for the different classes predicted by our model. The box plots shows the median and quartiles for each set of variants. The dashed horizontal lines show $k_{cat}/K_{M,cMUP}$ for WT PafA. (C, D) Scatter plot of $k_{cat}/K_{M,cMUP}$ values when substituting a residue with either glycine or valine. (C) Shows the comparison for buried residues (with an exposed surface area of less than 20%) and (D) shows exposed residues. In both C and D squared markers represent positions where the WT residue is hydrophobic and circles indicate non-hydrophobic WT residues. (E) Comparison of experimental $k_{cat}/K_{M,cMUP}$ values and temperature effects during expression. In particular, the T-effect value represents the change in measured catalytic efficiency when the protein was expressed at 23 °C or 37 °C (but with the enzymatic assay performed at 23 °C in both cases). Variants that show a substantial (greater than 10-fold) change nearly all belong to the total-loss category. Points with larger markers represent data for which the T-effect might be underestimated due to experimental limitations. (F) Raincloud plots for the coloured points in panel E (large T-effect).

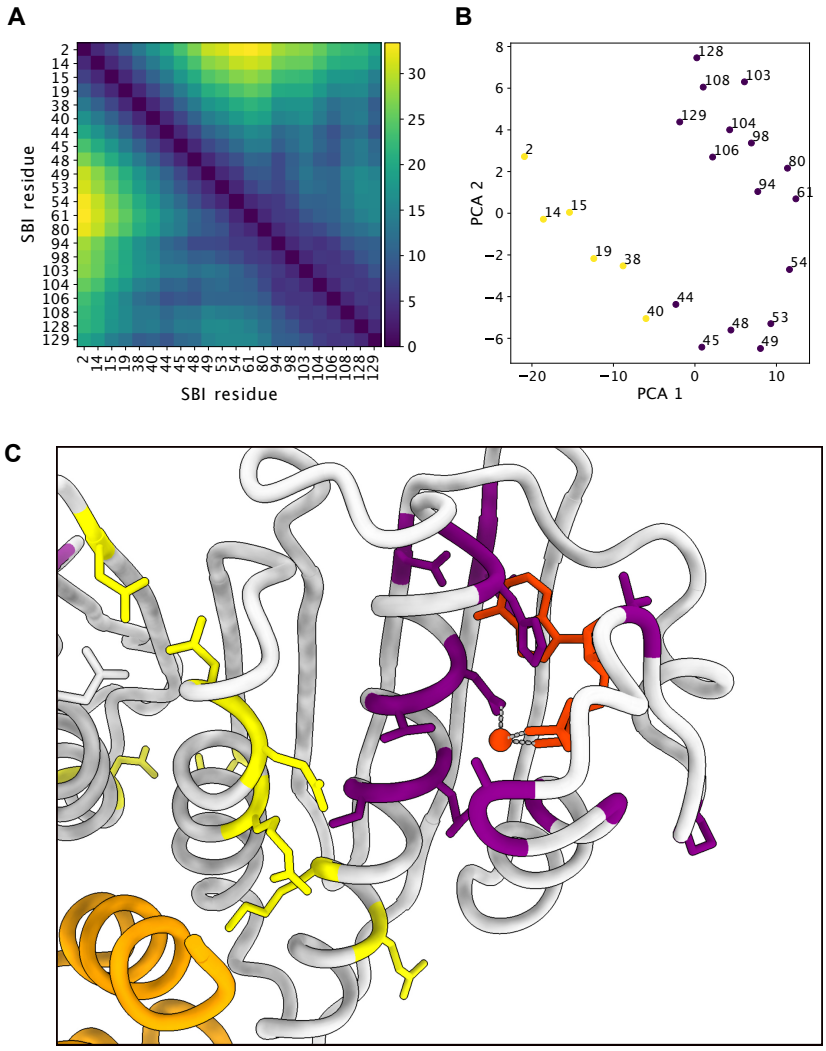


Figure S5. Clusters of functional positions in Anti-Sigma factor. (A) C_α distance matrix for the positions classified as functional by our model. (B) A principal component analysis of the distance matrix suggests that the predicted functional sites can roughly be separated into two spatial clusters (coloured in yellow and purple along the first two principal components). (C) Functional sites coloured using the clustering.

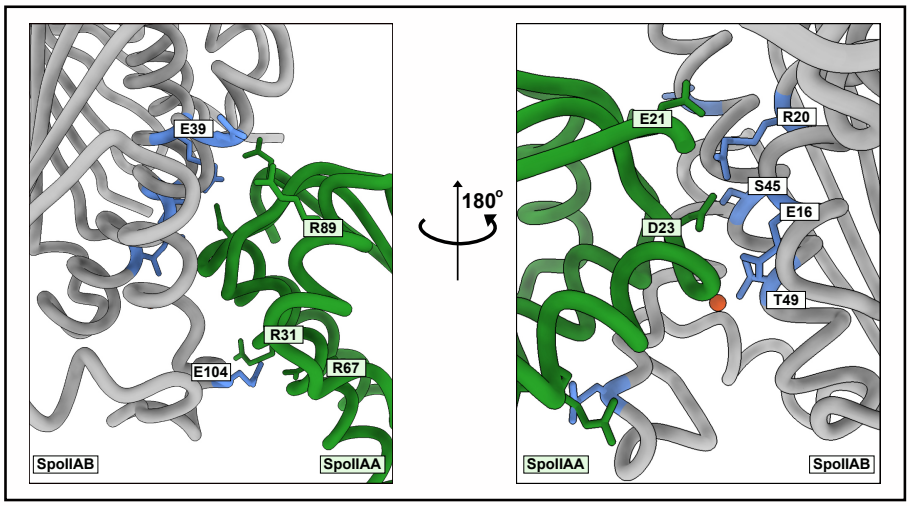


Figure S6. Map of detected functional residues at the interface between the Anti-sigma factor (SpoIIAB in grey) and its antagonist (SpoIIAA in green). Predicted functional residues in SpoIIAB involved in interface interactions are labelled and coloured in blue.

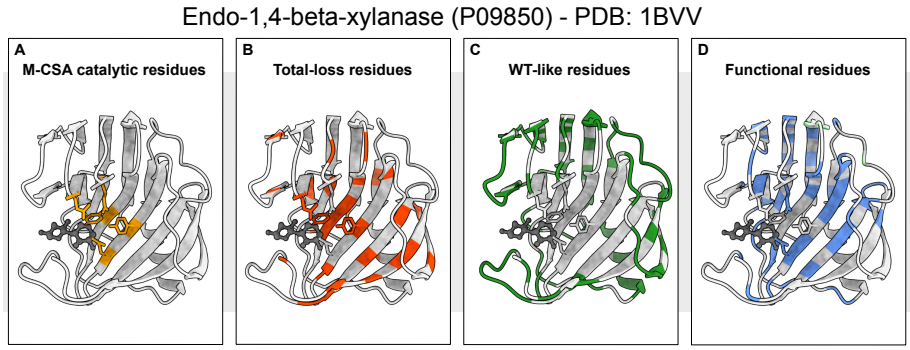


Figure S7. Model predictions for Endo-1,4-beta-xylanase. (A) Catalytic residues in Endo-1,4-beta-xylanase as assigned by M-CSA. (B, C and D) show the residues which are classified by our model as (B) total loss, (C) wild-type like, and (D) functional residues. Our model predicts that substitutions at all five catalytic residues affect the protein function rendering it inactive. Three of the five positions are predicted to be important for function, but not for protein stability, whereas two positions are predicted to be important for both function and stability.

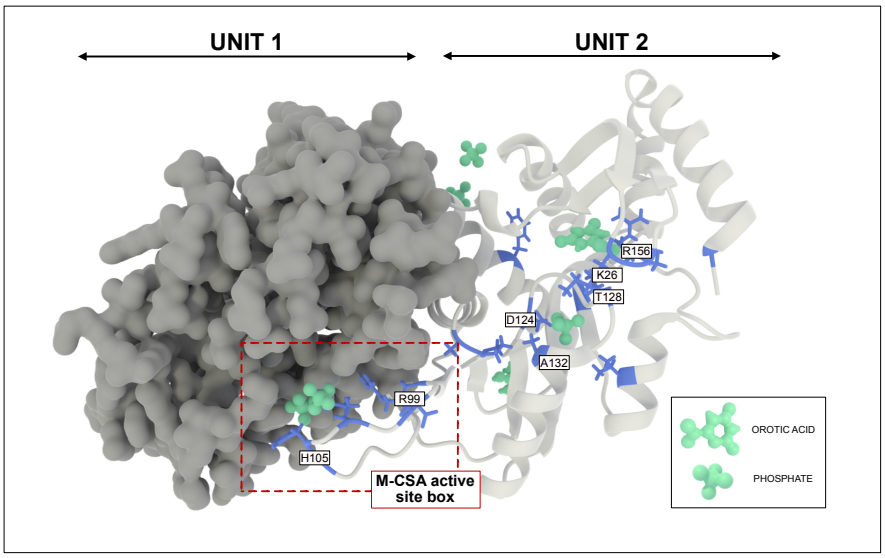


Figure S8. Location of predicted functional sites in OPRTase. The figure shows the dimer structure of OPRTase, with the two sub-units coloured in grey and white, respectively. The residues classified as functional by our model are coloured in blue and labelled on the second sub-unit. The region surrounded by a dotted red box contains catalytic residues reported in the M-CSA database (again shown for the second sub-unit). Orotic acid and phosphate ions bound to the OPRTase complex are coloured in green.

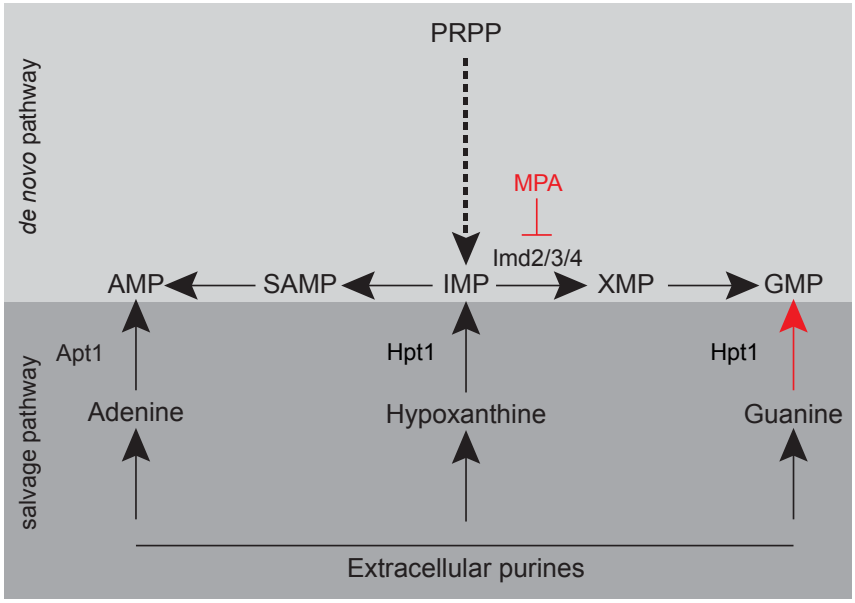


Figure S9. *Hpt1* is essential in the presence of MPA. The figure shows relevant parts of yeast nucleotide metabolism including the salvage pathway where *Hpt1* is essential for generating GMP when XMP synthesis is blocked by MPA. A *hpt1*Δ yeast strain can therefore not grow in the presence of MPA, but can be rescued by human HPRT1 or functional HPRT1 variants.

Supplementary Tables

Table S1. Number of variants per class predicted by our model on the training set

Protein	WT-like	SBI	Total loss	Low abundance, high activity
NUDT15	1824 (68%)	350 (13%)	509 (19%)	1 (<1%)
PTEN	1689 (54%)	662 (21%)	755 (24%)	0 (0%)
CYP2C9	2396 (57%)	1111 (26%)	657 (17%)	0 (0%)

Table S2. Enzymes analysed from the Mechanism and Catalytic Site Atlas database

UNIPROT ID	PDB	Chain	# Analysed residues	# Catalytic residues	# Functional residues
Q9LAK3	1RO7	A	291	1	16
P82385	1DO6	A	124	2	11
P09850	1C5H	A	213	3	39
P56868	1B73	A	254	6	17
P0A7E3	1ORO	A	213	2	17
Q13569	3UFJ	A	410	1	26
P62593	1BTL	A	286	5	10
P00374	1DHF	A	187	1	9
P00469	1LCB	A	316	3	25
P00491	1RR6	A	289	3	16
Q06241	1R44	A	202	5	11
P04036	1ARZ	A	273	2	25
P0A6L0	1P1X	A	259	3	19
P19120	1KAZ	A	650	3	19
P77836	1BRW	A	433	5	35
Q16854	2OCP	A	277	2	17
Q53547	1AUO	A	219	3	9
Q55012	1PS1	A	337	6	8
Q8VQN0	1JC5	A	148	3	10

Table S3. Proteins analysed from the Protein-Protein Interaction Affinity Database 2.0

UNIPROT ID	PDB	Chain	# Analysed residues	# SBI residues
P0AE67	1FFW	A	129	6
P01588	1EER	A	193	17
P61972	1A2K	A	127	19
P0A0L5	3BZD	B	237	7
O77044	1KSD	A	433	12

Table S4. Proteins used in the manuscript which are not included in the above databases

Protein	UNIPROT ID	PDB	Chain
Alkaline phosphatase PafA	Q9KJX5	5tj3	A
Growth factor receptor-bound protein 2 (GRB2)	P62993	AlphaFold	A
Disks large homolog 4 (DLH4)	P78352	6QJJ	A
Hypoxanthine-guanine phosphoribosyltransferase (HPRT1)	P00492	3BZD	A
Anti-signa F factor	O32727	1L0O	A