nature communications

Article

https://doi.org/10.1038/s41467-023-39909-0

Discovering functionally important sites in proteins

Received: 19 May 2023

Accepted: 2 July 2023

Published online: 13 July 2023

Check for updates

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

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 prospective prediction and subsequent experimental validation on the functional consequences of missense variants in *HPRT1* which may cause Lesch-Nyhan syndrome, and pinpoint the molecular mechanisms by which they cause disease.

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. Progress has been made in the development of methods to identify some types of functional sites[1–5] and efforts have been made to understand the relationship between sequence variability, protein function, stability and the onset of diseases[6–10]. 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[11–13]. 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[14]. We note here that we use the term 'function' and 'functional sites' in a relatively general sense since our goal is to examine protein function 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. Analyses 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[15–18]. 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[15,16,18]. 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[19].

In a recent analysis of abundance and activity assays[16] 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 about 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

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark. e-mail: rhpetersen@bio.ku.dk; lindorff@bio.ku.dk

signalling) from those positions that are conserved mostly due to structural constraints[20].

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[16,21]. 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[16–18].

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[5,22–26]. 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[16,27–29].

Here, we aim to create a robust and easy-to-use 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 data generated by MAVEs that report on the effects of a wide range of substitutions 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. The code for our model is freely available and we also provide access to it via a web implementation.

Results

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[13,16]. 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, our results suggest that most functional sites do not fall in this class.

Based on the ideas above, we collected the results from two complementary types of MAVEs that respectively probe cellular-abundance and functional effects for three different proteins: NUDT15[30], PTEN[31,32], and CYP2C9[33] 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[16]: 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[34], (ii) the evolutionary sequence information scores (which we term $\Delta\Delta E$, by analogy with $\Delta\Delta G$) using GEMME[26], (iii) the hydrophobicity of the amino acid[35], and (iv) the weighted contact number[36,37] (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 (Supplementary Fig. 1). Thus, we used the experimental labels and the input features to train a gradient boosting classifier (Fig. 1A and Supplementary Fig. 2) that predicts abundance (high/low) and activity (high/low) for each variant (see Methods for further details on the feature choices).

We determined model hyperparameters 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).

After feature selection and training of the final model, we tested its performance on an independent dataset (GRB2 SH3 domain,[18]) using as baseline a model using only cutoff values for $\Delta\Delta E$ and $\Delta\Delta G$ (Supplementary Fig. 3A) . We find that our model substantially outperforms the baseline model, especially in labelling functional residues. We also validated the performance of the model on a reduced

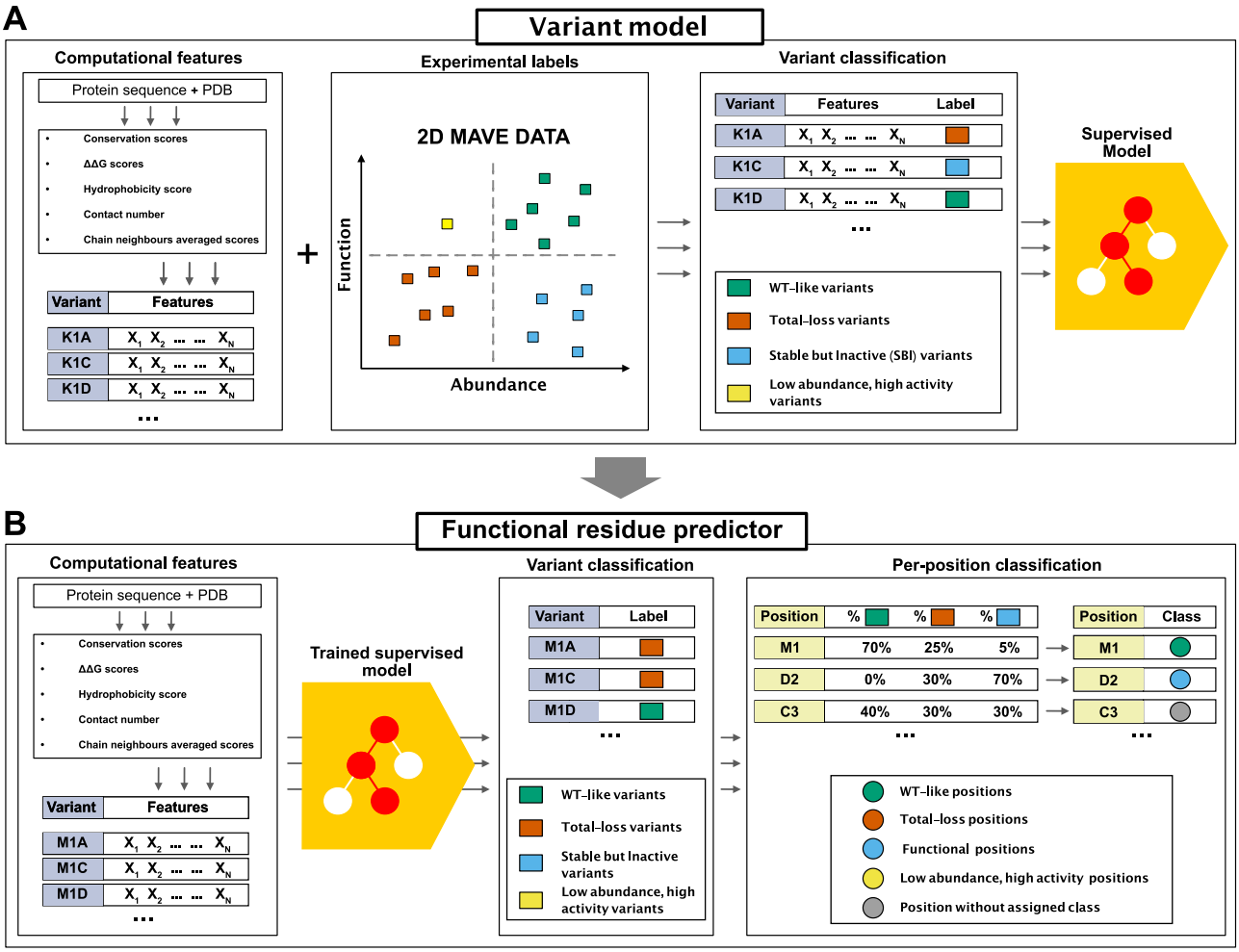


Fig. 1 | Graphical summary of our model to identify functionally important residues. A Using protein sequence and structure as input, we extract a number of features to characterise each variant. For each protein in our training set we extract all variants with MAVE measurements of abundance and function; this data is then combined with the structural 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.

variant training set (Supplementary Table 1), and by training on two of the proteins and evaluating on the third (Supplementary Table 2); in both cases we found reduced performance compared to the full model, but no evidence of substantial overfitting. We also examined other choices of features including both subsets of those used in our final model and a model with a wider set of features, and evaluated both by cross-validation and the independent dataset (Supplementary Fig. 3B). The results show that the chosen feature set outperforms the other sets, including the model with a larger number of features.

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[18]. 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 (Supplementary Fig. 4). 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 are generally found to have a large $\Delta\Delta G$ for folding in the experiments (Supplementary Fig. 4A, 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 (Supplementary Figure 4C,D). We note that for PDZ3 we found a limited number of homologues in Uniref30 for the PDZ3 domain, which might make the $\Delta\Delta E$ values more uncertain.

We then analysed 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

https://doi.org/10.1038/s41467-023-39909-0

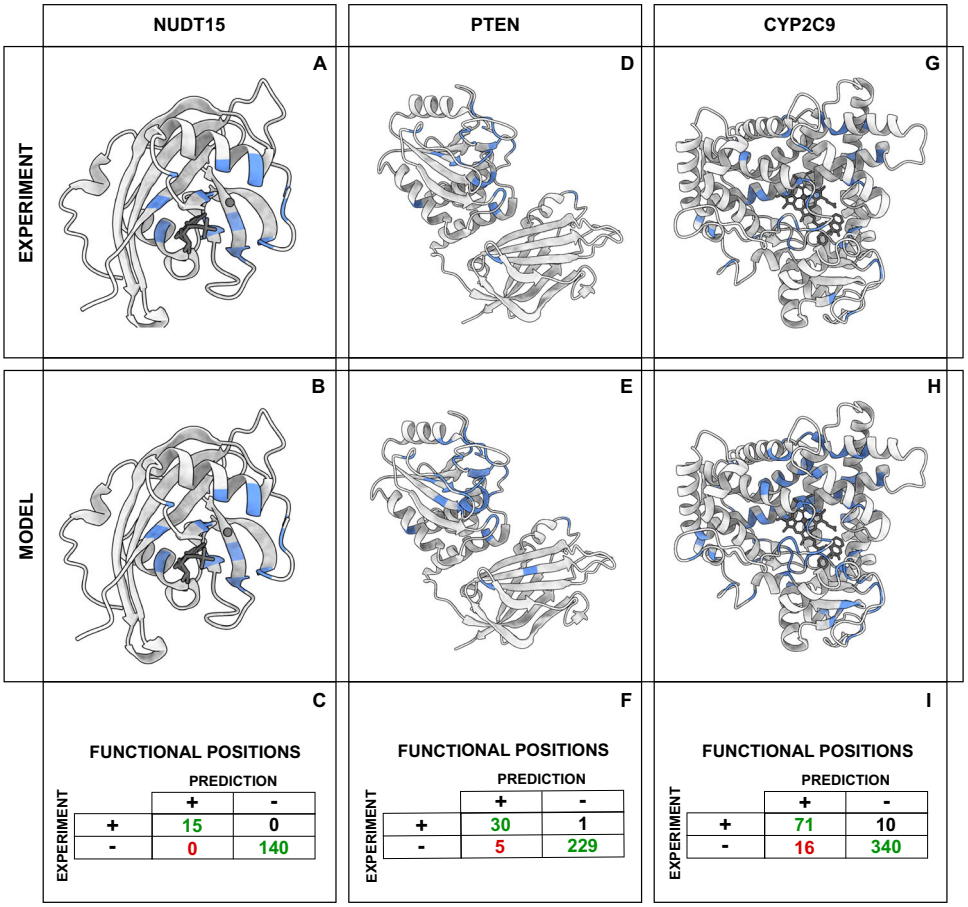


Fig. 2 | Finding functional residues in the training proteins. A, **D**, **G** Functional positions (in blue) identified using experiments. **B**, **E**, **H** Functional positions (in blue) identified using our model. **C**, **F**, **I** Statistics on the accuracy of residue level predictions of functional 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).

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[38]). In these experiments, the catalytic efficiency (k_{cat}/K_M) was measured for the wild-type protein as well as variants where each residue was changed into either glycine or valine. We first analysed how well our model was able to distinguish between variants with different levels of activity, and found a lower median activity (k_{cat}/K_M) for variants classified as total-loss and SBI compared to WT-like variants (Supplementary Fig. 5A, 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 (Supplementary Fig. 5C). At exposed positions, we generally predicted variants to have a smaller effect (Supplementary Fig. 5D). 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 proteins[38]. In particular, some variants resulted in less active protein when expressed at 37 °C compared to 23 °C. Using a ten-fold change as a cutoff[38], we find most (70%) variants that show this level of difference belong to the total-loss class (Supplementary Fig. 5E, F), in line with the fact that these are predicted to cause folding defects.

Finally, we compared predictions from our model with results from two previously published models for detecting functional residues[27,39]. We selected 10 enzymes that had previously been studied and for which the percentage of true positives from the previous work was available. Using our model, we find 77 out of the 109 known functional sites (true positive rate of 0.71) (Supplementary Fig. 6), which can be compared to values of 0.40 and 0.60 from the earlier models[27,39].

Table 1 | Validation using multiplexed assays of variant effects

Protein	PDB	MAVE	SBI accuracy	Global accuracy
Cystathionine beta-synthase	4COO	68	0.75	0.68
Beta-lactamase TEM	1BTL	69	0.76	0.72
Regulatory protein GAL4	3COQ	70	0.69	0.72
Small ubiquitin-related modifier 1	1WYW	71	0.75	0.63
Thiopurine S-methyltransferase	2H11	31	0.54	0.76

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[40,41]. 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[40,42]. Of these 17 amino acids, nine are located in the proximity to 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 phosphorylation reaction and Arg105 stabilises the transition state in the phosphorylation 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 to 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 (Supplementary Fig. 7). 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[43]), as well as five non-enzymes from the Protein–Protein Interaction Affinity Database 2.0[44]. 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 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[43]. 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 33% and 80%.

We examined in more detail the 25 of the 87 positions that our model did not assign as SBI. 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 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 (Supplementary Fig. 8) 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 functional roles beyond those directly involved in for example 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 vicinity 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

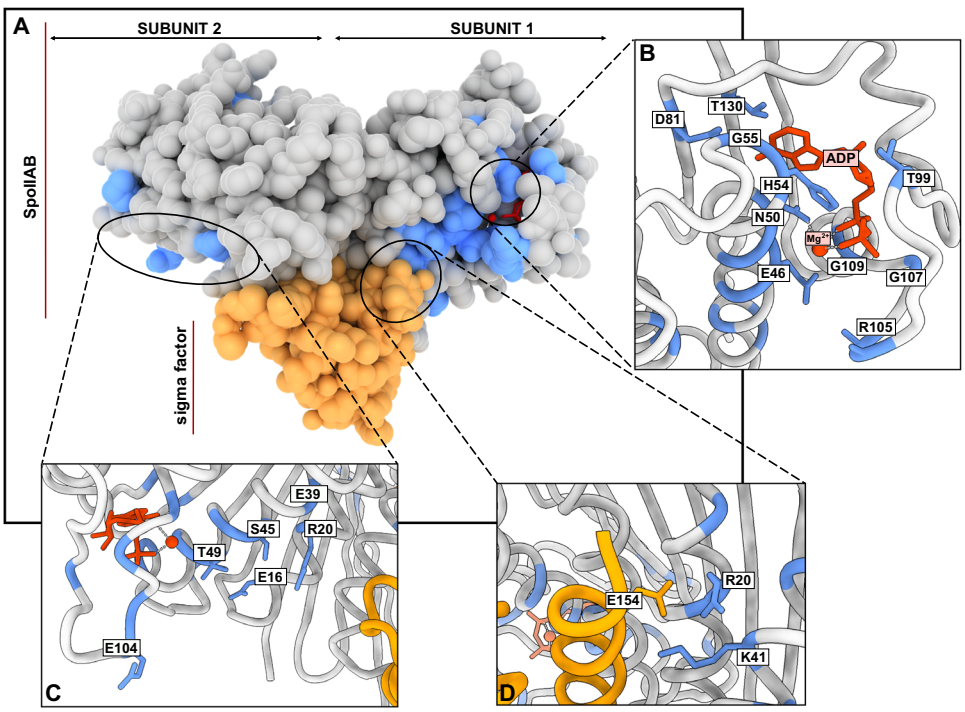
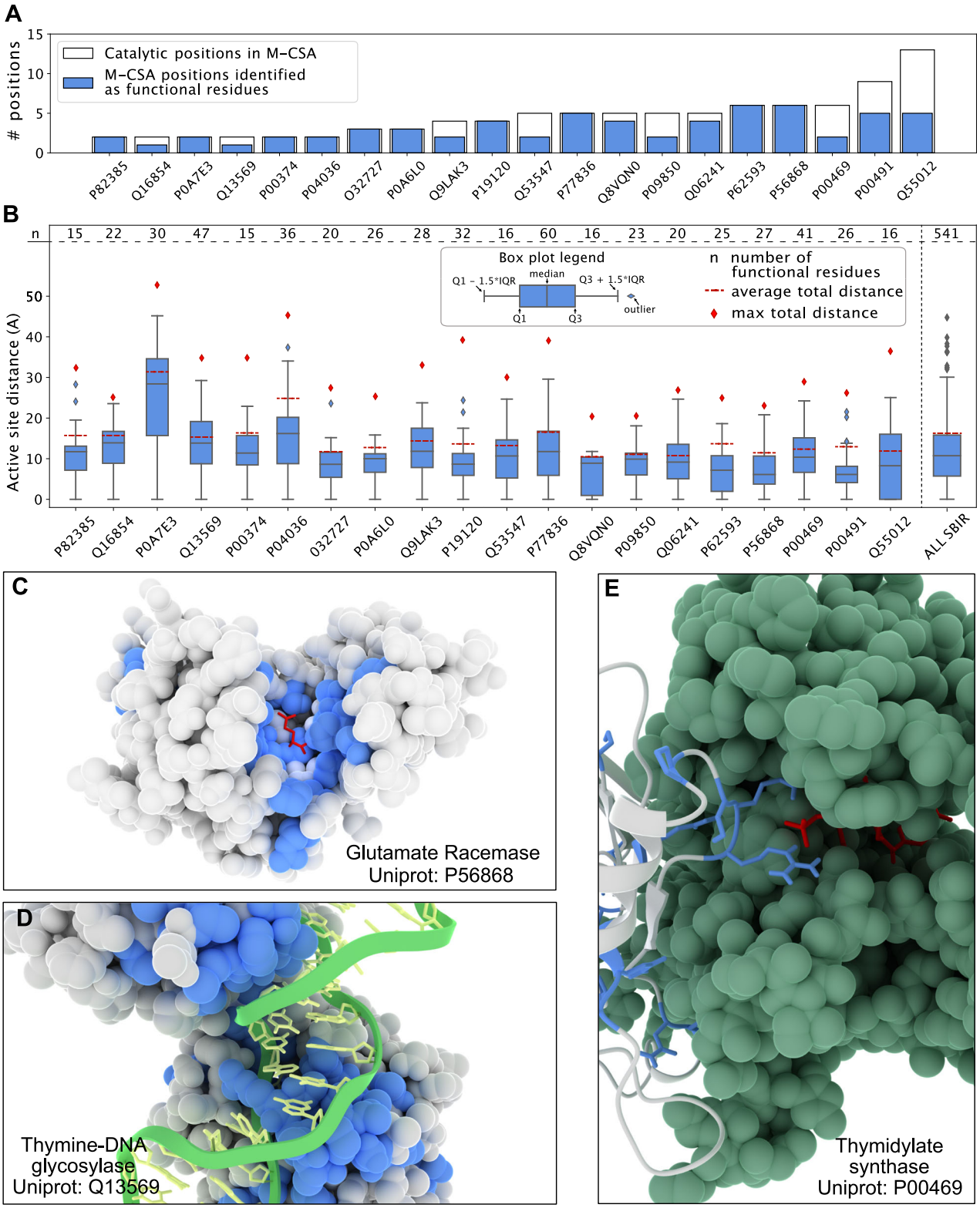


Fig. 3 | Identifying functional residues in an Anti-sigma F factor (Uniprot O32727, PDB: 1I0O). A The dimeric Anti-sigma F factor is shown using Van der Walls surfaces (grey). Predicted functional sites are coloured in blue, Mg^{2+}:ADP is shown in red and the bound sigma factor is shown in yellow. **B**–**D** Functional sites identified by our model (in blue) and previously described in the literature[40,42] are labelled. **B** Predicted functional sites in the proximity to 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.

https://doi.org/10.1038/s41467-023-39909-0



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[45].

In some cases we also found that the predicted functional sites were located both close to as well as further away (20 Å or more) from the catalytic site, as illustrated by orotate phosphoribosyltransferase (OPRTase, Uniprot: P0A7E3, Supplementary Fig. 9), where predicted functional residues are located in several distinct regions. 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[41,46].

Fig. 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) with the total number of functional residues reported (as *n*) in the top of the distribution. 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. The composition of each boxplot (boundaries and elements) is reported in the Figure legend. **C**–**E** Examples of predicted functional sites (blue) in three proteins from the M-CSA set. **C** Functional sites in glutamate racemase are found in a single cluster close to the active site (substrate in red). **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 (green and yellow structure). **E** Predicted functional sites are also found in protein–protein interfaces, such as in the interface of homo-dimeric thymidylate synthase. The second dimeric sub-unit is coloured in green with the substrate in the active site in red.

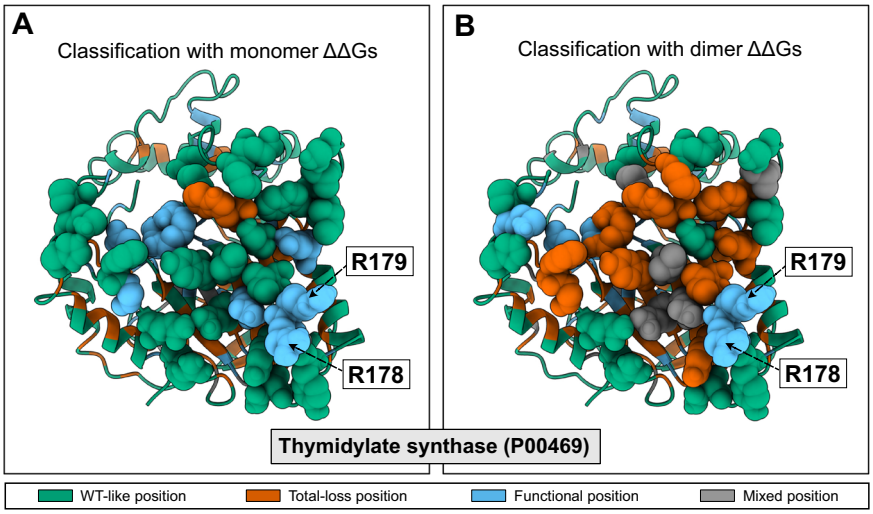


Fig. 5 | Analysing functional sites at a protein-protein interface. We show the results from predicting functional sites in thymidylate synthase using either the structure of **A** the monomer or **B** the dimer as input to Rosetta for calculating $\Delta\Delta G$. Residues at the interface are shown with Van der Walls atomic representation. Residues with known catalytic activity are labelled.

These amino acids are not directly involved in catalysis, but their role is to push the target DNA base into the catalytic pocket, which consists of two residues: Arg140 and His151[46,47]. 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 for catalysis and binding of substrates, our model can also help identify positions that play a role in forming 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 functional residues identified by our model, 20 are located close to the protein-protein interface. We note here that, unless otherwise noted, all the 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, Arg179, Cys198,[48]). We, however, also found a number of other residues, which we suggest are conserved as they stabilize the (obligatory) dimer structure.

To examine whether we could extract more information on interaction interfaces we also calculated $\Delta\Delta G$ values using the structure of the dimer of thymidylate synthase, introducing each amino acid substitution in both chains. Together with contact numbers calculated using the dimer structure we used these new scores as input to our model. We compared the resulting classification with the results obtained when using using the monomer structure as input to our model (Fig. 5). The analysis shows that nine residues at the interface are classified as total-loss when the calculations are based on the dimer structure (compared to just one when the monomer structure is used). Of these nine total-loss positions, three were labelled as functional residues in calculations based on the monomer structures. Residues at the interface that are known to be important for the catalytic activity (Arg178, Arg179,[48]) remained labelled as functional sites. We performed a similar set of calculations comparing the classification based on structures of the monomer or dimer structure for OPRTase (Supplementary Fig. 10A,B). As for thymidylate synthase, we found differences including an increased number of total-loss positions at the interface in the classification made using the dimer; also in this case the model correctly identified the catalytic residues located at the protein-protein interface.

The results described above show that comparing the predictions based on structures of monomers and oligomers can help disentangle functional residues that are key for protein–protein interactions from those with other roles. Specifically, when the structure of the oligomer is known and used as input, our model focuses on residues that are important for functions other than protein-protein interactions. To illustrate this point further, we compared the classification for (monomeric) myoglobin with that for the $\alpha_2\beta_2$ tetramer haemoglobin. For haemoglobin, we made predictions for both the α and β subunits, and using both the monomer and tetramer structures as input. We first compared the classification results for all three proteins when calculations are based on the monomer structure (Supplementary Fig. 10C). For all three proteins we found a comparable classification of the residues surrounding the heme group, with 7–9 residues classified as functional. We, however, found differences between myoglobin and the two haemoglobin chains at the residues that form interfaces in the haemoglobin tetramer (Supplementary Fig. 10C). For these residues,

the model assigned the WT-like label for most of the positions in myoglobin, while several interface residues in haemoglobin are classified as functional. This difference likely arises because the calculated $\Delta\Delta G$ values at the interface residues are small (because we used the monomer structures as input), but the residues in haemoglobin are more conserved than those in myoglobin. In line with this hypothesis, many of the interface residues in haemoglobin are classified as total-loss when we use the tetramer structure as input to our model (Supplementary Fig. 10C). In this case, substitutions at the interface are more destabilizing, and so the model assigns these residues to the class that corresponds to destabilization of the functional protein (the tetramer). Together, the results on thymidylate synthase (Fig. 5), OPRTase (Supplementary Fig. 10A, B), and myoglobin/haemoglobin (Supplementary Fig. 10C) show that residues at key protein-protein interaction sites behave differently depending on whether one considers destabilization of the monomer or oligomer form. In cases where the oligomer structure is known, this difference can help to distinguish residues that play functional roles due to protein-protein interactions from those that are for example directly involved in catalysis. Similarly, when using the monomer structure as input, the model can help shed light on key interaction interfaces, which appear as surface exposed patches of conserved residues.

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[49,50], for a prospective study of the accuracy and utility of our model. 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[51], ClinVar[52] 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. 6).

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,[53,54]). 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 (Supplementary Figure 11). 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. 6B). 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. 6C).

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. 6D). Overall we find that 13/17 (76%) of the variants are predicted correctly including all (6/6) of the SBI residues. This result shows 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[49], with overall good agreement between our predictions, the yeast assays and the biochemical experiments.

Making the model more easily accessible

Evaluating $\Delta\Delta G$ values with Rosetta is relatively slow, making the widespread application of our model less straightforward. We have, however, recently developed a method, called RaSP, for rapid stability predictions[55], which uses a deep-learning representation to approximate Rosetta $\Delta\Delta G$ values orders of magnitude faster than Rosetta. We first tested the results when using $\Delta\Delta G$ values from RaSP as input to the functional model described above (which was trained with $\Delta\Delta G$ values generated by Rosetta). Despite the relatively high correlation between the Rosetta and RaSP $\Delta\Delta G$ values (average Spearman correlation coefficient of 0.78), we found that the performance of the model was lower during the cross-validation when we used $\Delta\Delta G$ generated by RaSP (Supplementary Fig. 12A). This result suggests that differences in the $\Delta\Delta G$ values generated by RaSP and Rosetta combined with the threshold-based structure of gradient boosting machines can shift the prediction for variants with feature values close to the threshold values used inside the model. We therefore retrained our model using instead $\Delta\Delta G$ values generated by RaSP. We found that the RaSP-based model performs as well as the model trained with Rosetta on our validation sets (Supplementary Fig. 12B) and therefore decided to test the performance for many of the tasks we used to test our Rosetta-based model on. We found that the RaSP-trained model identifies the same number of functional sites as the Rosetta-trained model (62/87) when we examine the enzymes in the M-CSA set, with 57/62 being the same (Supplementary Fig. 12C). We also found that the RaSP-based model predicts the effects of the HPRT1 variants as well as the Rosetta-trained model (Supplementary Fig. 12D).

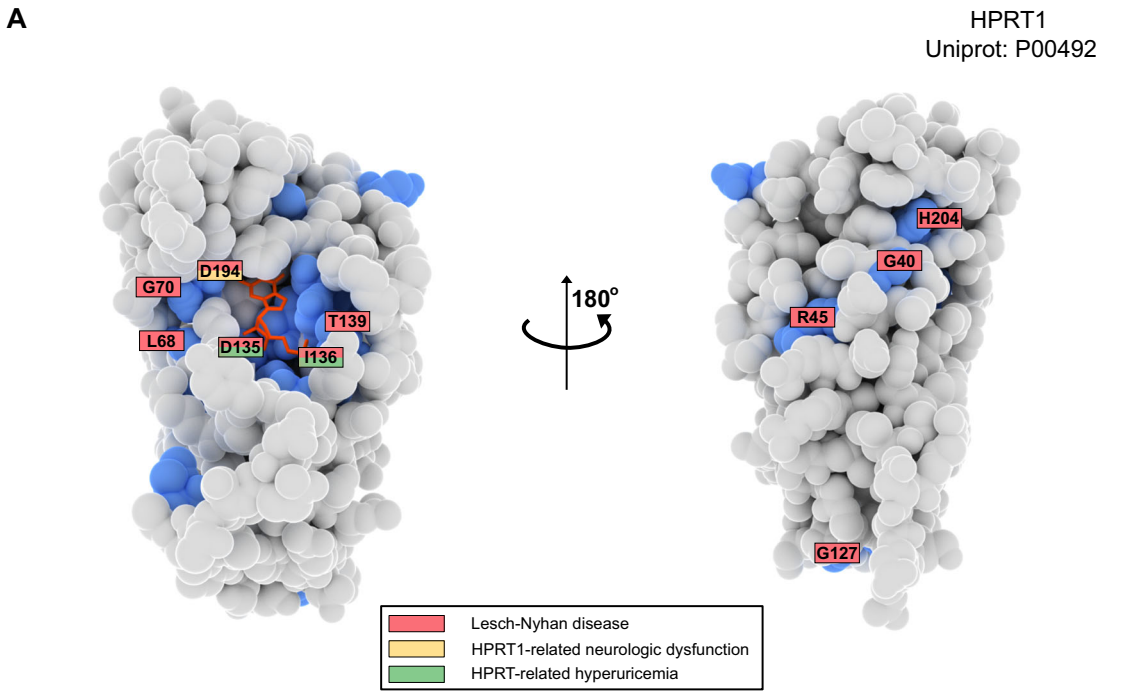
We make our model available via a notebook that can be run using Google Colaboratory (available via https://github.com/KULL-Centre/_2022_functional-sites-cagiada). The notebook guides the user with a step-by-step procedure to generate input data, using for example the GEMME and RaSP web implementations, and to generate the predictions of functional sites.

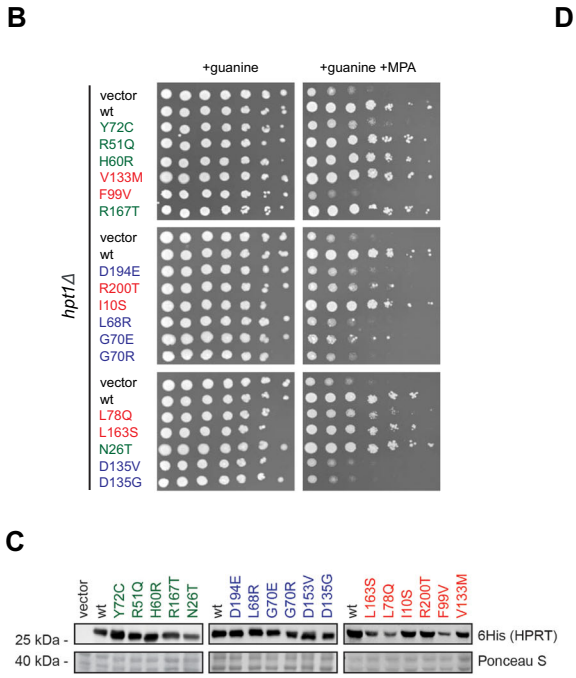
Discussion

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[56,57]. 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

https://doi.org/10.1038/s41467-023-39909-0





D

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

WT-like
Stable but Inactive (SBI) / functional position
Total-loss
Low abundance, high activity

Fig. 6 | 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** 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. Panels **B** and **C** show representative results of three replicates. Source data for this figure are provided as a Source Data File.

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 is freely available. For users that want to avoid slower, more costly Rosetta calculations, we have also trained a model that uses our deep-learning-based method for stability predictions[55]. Since both Rosetta[58] and RaSP[55] give relatively accurate results using structures predicted using AlphaFold, our model can also be applied to predicted protein structures. Very recently[59], it has been shown that one can also use large-scale measurements of stability changes as input to an approach inspired by our work.

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[38]) 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 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 additional data is generated by MAVEs reporting on both function and abundance, the model can be improved further. 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 such as finding variant that may or may not be rescued by pharmacological chaperones. Recent work shows how it is possible to annotate enzyme function at larger scale[60] and our model can help pinpoint functionally important sites in newly discovered enzymes. We thus anticipate that the method that we have described here will help researchers make progress in these and many other areas.

Methods

Preprocessing of the training data

The training data include paired MAVE data for three different proteins: NUDT15, PTEN, and CYP2C9. For each protein, we selected thresholds for the scores generated by each MAVE as outlined previously[16]. Briefly, we fit the variant score distributions to three Gaussians and use the intersection of the first and last Gaussian as the cutoff. We use such binary classification for each of the two MAVEs to classify the variants into four categories[16].

Sequence features

For each protein in the training (Supplementary Table 3) and validation (Supplementary Tables 4,5 and 6) data we performed a statistical analysis of multiple sequence alignments (MSA). We extracted the sequence of the first isoform from Uniprot and used HHBlits[61] 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[26] 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[35] 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[34] on crystal structures corresponding to the Uniprot sequence listed in Supplementary Tables 3,4 and 5. 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[34]. Stability changes using RaSP were performed as described[55]. 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[62] 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 7.0 Å).

We determined solvent exposed and buried regions using GetArea[63] using 20% as a threshold to divide the data in to two classes.

Training the classifier

We began by including up to eleven features as input to our classifier. After testing on the three proteins (NUDT15,PTEN and CYP2C9), we proceeded with only eight of them. The features used for each variant were: (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. The three features discarded during model building were: wild-type amino acid type, binary exposure classes based on solvent exposure and hydrophobicity of the wild-type amino acid.

We used a gradient descent machine to classify the variants, as implemented in Catboost[64] v.0.26.1, using a Multinomial/Multiclass Cross Entropy Loss and L2 regularization. We first set the number of gradient descent iterations (Supplementary Figure 2A) 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 optimise the hyperparameters (Supplementary Figure 2B). 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 *hpt1Δ* (MatA, *his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, 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[65]. 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[66]. Briefly, 1.2×10^8 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 mouse IgG1 anti-RGSHis (Qiagen, Cat. No. 34650) diluted 1:2000 and peroxidase-conjugated polyclonal rabbit anti-mouse antibody (Dako, Cat. No. P0260) diluted 1:5000.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The input data for the model and the predictions used in this study are available at https://github.com/KULL-Centre/_2022_functional-sites-cagiada which also contains the final trained model. The datasets are also available on Zenodo[67], and can be accessed via https://doi.org/10.5281/zenodo.8046585. Uniprot and PDB ID codes for the proteins studied are available in Supplementary Tables 4, 5 and 6. Source data (full images) for Fig. 6 are provided as a Source Data File in zip format. Source data are provided with this paper.

Code availability

The code used to generate the main text figures, to recreate the predictions, and to generate new predictions is available at https://github.com/KULL-Centre/_2022_functional-sites-cagiada. An implementation of the model is also available via Google Colaboratory via the same link. The code is also available on Zenodo[67], and can be accessed via https://doi.org/10.5281/zenodo.8046585.

References

1. del Sol Mesa, A., Pazos, F. & Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302 (2003).
2. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995–1005 (2007).
3. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
4. Kulmanov, M., Khan, M. A. & Hoehndorf, R. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2018).
5. Torng, W. & Altman, R. B. High precision protein functional site detection using 3d convolutional neural networks. *Bioinformatics* **35**, 1503–1512 (2019).
6. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).
7. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. methods* **7**, 248–249 (2010).
8. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
9. Wagih, O. et al. A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* **14**, e8430 (2018).
10. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
11. Gray, V. E., Hause, R. J. & Fowler, D. M. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* **207**, 53–61 (2017).
12. Dunham, A. S. & Beltrao, P. Exploring amino acid functions in a deep mutational landscape. *Mol. Syst. Biol.* **17**, e10305 (2021).
13. Høie, M. H., Cagiada, M., Frederiksen, A. H. B., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022).
14. Li, X. & Lehner, B. Biophysical ambiguities prevent accurate genetic prediction. *Nat. Commun.* **11**, 1–11 (2020).
15. Jepsen, M. M., Fowler, D. M., Hartmann-Petersen, R., Stein, A. & Lindorff-Larsen, K. in *Chapter 5 - classifying disease-associated variants using measures of protein activity and stability* (ed. Pey, A. L.) *Protein Homeostasis Diseases* 91–107 (Academic Press, 2020).
16. Cagiada, M. et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol. Biol. Evolution* **38**, 3235–3246 (2021).
17. Chiasson, M. A. et al. Multiplexed measurement of variant abundance and activity reveals vkor topology, active site and human variant impact. *elife* **9**, e58026 (2020).
18. Faure, A. J. et al. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
19. Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. evolution* **35**, 2345–2354 (2018).
20. Echave, J. & Wilke, C. O. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annu. Rev. Biophys.* **46**, 85–103 (2017).
21. Nielsen, S. V., Hartmann-Petersen, R., Stein, A. & Lindorff-Larsen, K. Multiplexed assays reveal effects of missense variants in msh2 and cancer predisposition. *PLoS Genet.* **17**, e1009496 (2021).
22. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
23. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
24. Choi, Y. & Chan, A. P. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
25. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

26. Laine, E., Karami, Y. & Carbone, A. Gemme: a simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.* **36**, 2604–2619 (2019).
27. 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 Res.* **33**, 5861–5867 (2005).
28. Wang, K., Horst, J. A., Cheng, G., Nickle, D. C. & Samudrala, R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. *PLoS Comput. Biol.* **4**, e1000181 (2008).
29. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. *PLoS Comput. Biol.* **5**, e1000585 (2009).
30. Suiter, C. C. et al. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl Acad. Sci. USA* **117**, 5394–5401 (2020).
31. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
32. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
33. Amorosi, C. J. et al. Massively parallel characterization of cyp2c9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).
34. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. theory Comput.* **12**, 6201–6212 (2016).
35. Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M. & Hodges, R. S. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *J. Pept. Sci.* **1**, 319–329 (1995).
36. Shih, C.-H., Chang, C.-M., Lin, Y.-S., Lo, W.-C. & Hwang, J.-K. Evolutionary information hidden in a single protein structure. *Proteins: Struct. Funct. Bioinf.* **80**, 1647–1657 (2012).
37. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol.* **14**, e1002452 (2016).
38. Markin, C. et al. Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, eabf8761 (2021).
39. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504 (2004).
40. Campbell, E. A. et al. Crystal structure of the bacillus stearothermophilus anti-σ factor spoiiab with the sporulation σ factor σf. *Cell* **108**, 795–807 (2002).
41. Fu, T. et al. Thymine dna glycosylase recognizes the geometry alteration of minor grooves induced by 5-formylcytosine and 5-carboxylcytosine. *Chem. Sci.* **10**, 7407–7417 (2019).
42. Masuda, S. et al. Crystal structures of the adp and atp bound forms of the bacillus anti-σ factor spoiiab in complex with the anti-anti-σ spoiiaa. *J. Mol. Biol.* **340**, 941–956 (2004).
43. Ribeiro, A. J. M. et al. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
44. Vreven, T. et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* **427**, 3031–3041 (2015).
45. Hwang, K. Y. et al. Structure and mechanism of glutamate racemase from aquifex pyrophilus. *Nat. Struct. Biol.* **6**, 422–426 (1999).
46. Maiti, A., Morgan, M. T. & Drohat, A. C. Role of two strictly conserved residues in nucleotide flipping and n-glycosylic bond cleavage by human thymine dna glycosylase. *J. Biol. Chem.* **284**, 36680–36688 (2009).
47. 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. *J. Phys. Chem. B* **119**, 12365–12380 (2015).
48. Pookanjanatavip, M., Yuthavong, Y., Greene, P. J. & Santi, D. V. Subunit complementation of thymidylate synthase. *Biochemistry* **31**, 10303–10309 (1992).
49. Fu, R. & Jinnah, H. A. Genotype-phenotype correlations in lesch-nyhan disease. *J. Biol. Chem.* **287**, 2997–3008 (2012).
50. Fu, R. et al. Genotype–phenotype correlations in neurogenetics: Lesch-nyhan disease as a model disorder. *Brain* **137**, 1282–1303 (2014).
51. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
52. Landrum, M. J. et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids Res.* **46**, D1062–D1067 (2018).
53. Woods, R. A., Roberts, D. G., Friedman, T., Jolly, D. & Filpula, D. Hypoxanthine: guanine phosphoribosyltransferase mutants in saccharomyces cerevisiae. *Mol. Gen. Genet. MGG* **191**, 407–412 (1983).
54. Escobar-Henriques, M. & Daignan-Fornier, B. Transcriptional regulation of the yeast gmp synthesis pathway by its end products. *J. Biol. Chem.* **276**, 1523–1530 (2001).
55. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *Elife* **12**, e82593 (2023).
56. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA* **92**, 452–456 (1995).
57. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).
58. Akdel, M. et al. A structural biology community assessment of alphafold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
59. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and protein design. *bioRxiv* 2022–12 (2022).
60. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
61. Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. Methods* **9**, 173 (2012).
62. McGibbon, R. T. et al. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528 – 1532 (2015).
63. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19**, 319–333 (1998).
64. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516* (2017).
65. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the liac/ss carrier dna/peg method. *Nat. Protoc.* **2**, 31–34 (2007).
66. Kampmeyer, C. et al. Disease-linked mutations cause exposure of a protein quality control degron. *Structure* (2022).
67. Cagiada, M. et al. _2022_functional-sites-cagiada: v.1.0-publication *Zendo*https://doi.org/10.5281/zenodo.8046585 (2023).

68. Sun, S. et al. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med.* **12**, 1–18 (2020).
69. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell* **160**, 882–892 (2015).
70. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
71. Weile, J. et al. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).

Acknowledgements

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.). We acknowledge access to computing resources from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen.

Author contributions

M.C., S.B., and K.L.-L. conceived the overall study, building on preliminary work by A.S., K.L.-L., and R.H.P. M.C. and S.B. wrote the first draft of the manuscript with input from K.L.-L. M.C. performed the computational analyses and experiments with input from S.B. and K.L.-L. R.H.P, K.L.-L., and A.S. conceived the experiments on HPRT, which were performed by S.L. and S.M.S. under the supervision of R.H.P. All authors provided input to the study, and edited and provided input to the manuscript.

Competing interests

K.L.-L. holds stock options in and is a consultant for Peptone Ltd. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-39909-0.

Correspondence and requests for materials should be addressed to Rasmus Hartmann-Petersen or Kresten Lindorff-Larsen.

Peer review information *Nature Communications* thanks Gabriel Rocklin, Guillaume Lamoureux and the other anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.