https://doi.org/10.1186/s40246-024-00605-9

RESEARCH

Open Access

FiTMuSiC: leveraging structural and (co) evolutionary data for protein fitness prediction

Matsvei Tsishyn[1,2†], Gabriel Cia[1,2†], Pauline Hermans[1,2], Jean Kwasigroch[1,2], Marianne Rooman[1,2†] and Fabrizio Pucci[1,2*†]

Abstract

Systematically predicting the effects of mutations on protein fitness is essential for the understanding of genetic diseases. Indeed, predictions complement experimental efforts in analyzing how variants lead to dysfunctional proteins that in turn can cause diseases. Here we present our new fitness predictor, FiTMuSiC, which leverages structural, evolutionary and coevolutionary information. We show that FiTMuSiC predicts fitness with high accuracy despite the simplicity of its underlying model: it was among the top predictors on the hydroxymethylbilane synthase (HMBS) target of the sixth round of the Critical Assessment of Genome Interpretation challenge (CAGI6) and performs as well as much more complex deep learning models such as AlphaMissense. To further demonstrate FiTMuSiC's robustness, we compared its predictions with *in vitro* activity data on HMBS, variant fitness data on human glucokinase (GCK), and variant deleteriousness data on HMBS and GCK. These analyses further confirm FiTMuSiC's qualities and accuracy, which compare favorably with those of other predictors. Additionally, FiTMuSiC returns two scores that separately describe the functional and structural effects of the variant, thus providing mechanistic insight into why the variant leads to fitness loss or gain. We also provide an easy-to-use webserver at https://babylone.ulb.ac.be/FiTMuSiC, which is freely available for academic use and does not require any bioinformatics expertise, which simplifies the accessibility of our tool for the entire scientific community.

Keywords Protein variants interpretation, Fitness, CAGI6, Pathogenicity

Introduction

Accurately quantifying the effect of genetic variants on the fitness of the encoded proteins is one of the open challenges in biology which, if resolved, would have a tremendous impact on the understanding and treatment of genetic diseases [1–3]. The experimental approaches commonly used to quantify variant effects include different mutagenesis experiments [4–7] and large-scale exome screening approaches [8, 9]. However, these remain expensive and time consuming, and given the ever increasing amount of genetic data that is being generated, the number of variants of unknown significance (VUS) that are waiting to be characterized keep growing [10]. Moreover, the genetic bases of the majority of rare diseases are still not deciphered [11], and this is even more true for complex diseases such as cancer [12]. New complementary approaches are thus needed to interpret and classify these VUS and, more generally, to gain novel insights into these matters.

In the last two decades, many computational tools have been developed to predict the phenotypic effect of genetic variants [13–24]. They are mainly based on evolutionary features combined using machine learning

[†] Matsvei Tsishyn, Gabriel Cia have contributed equally to this work; Marianne Rooman and Fabrizio Pucci have contributed equally to this work.

*Correspondence:
Fabrizio Pucci
Fabrizio.Pucci@ulb.be
[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles, 50 Roosevelt Ave, 1050 Brussels, Belgium
[2] Interuniversity Institute of Bioinformatics in Brussels, Triumph Bvd, 1050 Brussels, Belgium

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

techniques. The most recent predictors such as [14, 22, 24] take advantage of the advent of deep learning approaches, as enough experimental data has become available to train complex models for fitness prediction [7]. These methods could in principle help accelerate the discovery of clinically relevant variants and their molecular effect, but their low accuracy and poor generalization properties are major obstacles for having a strong impact on clinical decision. In addition, black-box machine learning models do not contribute to improve our understanding of pathogenic mechanisms.

Currently, the gold standard to assess the performance of fitness prediction methods is the blind community-wide experiment called Critical Assessment of Genome Interpretation (CAGI) [25–27], which evaluates predictors on unpublished data. CAGI allows for an unbiased assessment of the methods as well as the identification of their strengths and weaknesses. Moreover, it provides guidelines on how to translate computational predictions into clinical practice.

In this paper we present our new method, FiTMuSiC, which we used in the recent CAGI6 experiment to predict the fitness of hydroxymethylbilane synthase (HMBS) variants. We begin with a presentation and discussion of our computational approach and of its performances in CAGI6. We then showcase additional results of our method on clinically relevant variants. Our results show that FiTMuSiC achieves very good performances when applied to unseen data, which demonstrates that simple linear combination models can actually perform as well as more complex deep learning-based models such as AlphaMissense [24].

Methods

Features

We briefly describe the features used by our method, which are of two kinds: structural and evolutionary. Structural features use the 3-dimensional (3D) structure of the wild-type protein as input. They include:

- Relative solvent accessibility (RSA). It is defined as the ratio (in %) between the solvent accessible surface area of a residue in its given 3D structure and in a Gly-X-Gly tripeptide extended conformation; it is computed by an in-house program [28].
- PoPMuSiC (PoP) [29]. This computational tool predicts the change in protein thermodynamic stability upon point mutations ($\Delta\Delta G$) using the 3D structure of the target protein as input. It is based on the formalism of statistical potentials [30], with the energy values and RSA used as features in an artificial neural network.
- MAESTRO (MAE) [31]. This tool also predicts the $\Delta\Delta G$ based on the protein 3D structure. It uses contact potentials as features, as well as some biophysical properties of the mutated and wild-type residues such as hydrophobicity and isoelectric point.
- SNPMuSiC (SNP) [16]. It is a predictor of variant deleteriousness based on structural and evolutionary features. Its evolutionary part is the PROVEAN algorithm [17], and its structural part consists of statistical potentials and RSA appropriately combined with artificial neural networks (ANN). We used here the structural part only, since PROVEAN is used by FiTMuSiC as a separate feature.

FiTMuSiC also includes four evolutionary features. To compute them, we generated a multiple sequence alignment (MSA) of the target sequence using JackHMMER [32] (with database UniRef90 [33], one iteration and an E-value threshold of 0.01). The evolutionary features are:

- PROVEAN score (PVS) [17]. It is a pure evolutionary tool that predicts the functional effect of variants. We used a re-implemented in-house version of the program which has some small differences with respect to the original version. Namely, it uses the pairwise alignment of the wild-type with the homologous sequences to calculate the alignment scores of the variants, rather than realigning them for each variant.
- Conservation Index (CI) [34]. It is calculated from $f_i(a)$ and $f(a)$, the regularized frequencies of amino acid a at position i in the MSA and in the full MSA, respectively, which are computed as:

$$f_i(a) = \frac{c_i(a)}{m}(1-\theta) + \frac{\theta}{21}$$
$$f(a) = \frac{c(a)}{m \times N}(1-\theta) + \frac{\theta}{21}, \tag{1}$$

where $c_i(a)$ and $c(a)$ are the number of occurrences of a at position i and in the full MSA, respectively, m is the depth of the MSA and N its length. The pseudocount parameter θ is set to 0.01 and defines the strength of the regularization; 21 is the number of possible states (20 amino acids and 1 gap). The CI score is calculated as:

$$\mathrm{CI}(i) = \left[\sum_{a \in A}(f_i(a) - f(a))^2\right]^{1/2}, \tag{2}$$

where A is the set of 20 standard amino acids.
- Log-odd ratio score (LOR) [35]. The log-odd ratio of observing the wild-type amino acid wt with respect

to the mutated amino acid *mt* at position *i* is defined as:

$$\mathrm{LOR}(i) = \log \frac{f_i(mt)}{1 - f_i(mt)} - \log \frac{f_i(wt)}{1 - f_i(wt)}. \quad (3)$$

- pyCoFitness score (PYF) [36]. This score is obtained through a method that infers a coevolutionary model from the MSA using a pseudo-likelihood maximization direct coupling analysis approach [37], and employs the inferred model to compute the change in fitness due to the variant.

Model structure and training

The FiTMuSiC model is a simple linear combination of the eight features listed above. The mathematical expression of the model is:

$$\begin{aligned}\mathrm{FiTMuSiC} =&\alpha_1 \mathrm{RSA} + \alpha_2 \mathrm{PoP} + \alpha_3 \mathrm{MAE} + \alpha_4 \mathrm{SNP} + \\ &\alpha_5 \mathrm{PVS} + \alpha_6 \mathrm{CI} + \alpha_7 \mathrm{LOR} + \alpha_8 \mathrm{PYF} + \alpha_9,\end{aligned} \quad (4)$$

where α_i $(i = 1, \ldots, 9)$ are free parameters that were identified based on a training set of deep-mutagenesis scanning data on three proteins: SUMO-conjugating enzyme UBC9 (UBE2I), small ubiquitin-related modifier 1 (SUMO1) and thiamin pyrophosphokinase 1 (TPK1) [38]. Structural features were computed using models from the AlphaFold Protein Structure Database [39].

The scale convention of FiTMuSiC values is the following: a value of 1 means equal fitness for wild-type and mutant; a value of 0 or below means the mutant is not fit at all; a value larger than 1 means that the mutant is fitter than the wild-type.

Additional models submitted to CAGI6

In addition to FiTMuSiC, we submitted the predictions of two other models to the CAGI6 challenge. The first is a simple rescaling of the SNPMuSiC score (SNP):

$$\mathcal{SNP} = \beta_1 \,\mathrm{SNP} + \beta_2, \quad (5)$$

where the numerical factors β_1 and β_2 were chosen to rescale the SNPMuSiC values and were identified on the fitness training set described in the previous subsection.

Although stability and fitness are imperfectly correlated [40], we also submitted a prediction model based on a rescaling of the score of the thermodynamic stability predictor PoPMuSiC (POP):

$$\mathcal{POP} = -\mathrm{ReLU}[-\mathrm{ReLU}[\gamma_1 \mathrm{POP} + \gamma_2] + 1] + 1, \quad (6)$$

where the parameters γ_1 and γ_2 were identified on the same training set as the other models. The ReLU functions bound the output between 0 and 1.

Model interpretation

To give information about the molecular effect of variants, FiTMuSiC provides four scores in addition to the global fitness of the variants. The first is the RSA of the mutated residue, which provides information on its spatial location in the 3D structure. The second is the z-score $\mathcal{Z}$ defined as:

$$\mathcal{Z} = \frac{\mathrm{FiTMuSiC} - \mu[\mathrm{FiTMuSiC}]}{\sigma[\mathrm{FiTMuSiC}]}, \quad (7)$$

where μ and σ represent the mean and standard deviation over all mutations on the given protein, respectively. Negative z-scores correspond to mutants that are less fit than average mutants; positive z-scores indicate mutants that are fitter than average mutants, with very positive values corresponding to mutants fitter than the wild-type.

The last two scores, $\mathcal{Z}_{\mathrm{str}}$ and $\mathcal{Z}_{\mathrm{evo}}$, give information about the extent to which the structural features (SNP, POP, MAE) and evolutionary features (CI, LOR, PVS, PYF) contribute to the global fitness of the considered variant. Defining the structural (STR) and evolutionary (EVO) contributions to the fitness as:

$$\mathrm{STR} =\alpha_2 \mathrm{PoP} + \alpha_3 \mathrm{MAE} + \alpha_4 \mathrm{SNP}, \quad (8)$$

$$\mathrm{EVO} =\alpha_5 \mathrm{PVS} + \alpha_6 \mathrm{CI} + \alpha_7 \mathrm{LOR} + \alpha_8 \mathrm{PYF}, \quad (9)$$

their z-scores $\mathcal{Z}_{\mathrm{str}}$ and $\mathcal{Z}_{\mathrm{evo}}$ are expressed as:

$$\mathcal{Z}_{\mathrm{str}} =\frac{\mathrm{STR} - \mu[\mathrm{STR}]}{\sigma[\mathrm{STR}]}, \quad (10)$$

$$\mathcal{Z}_{\mathrm{evo}} =\frac{\mathrm{EVO} - \mu[\mathrm{EVO}]}{\sigma[\mathrm{EVO}]}. \quad (11)$$

Negative $\mathcal{Z}_{\mathrm{str}}$ values correspond to mutations that destabilize the structure more than average mutations; positive $\mathcal{Z}_{\mathrm{str}}$ values indicate mutations that are less destabilizing than average mutations or are even stabilizing. Negative $\mathcal{Z}_{\mathrm{evo}}$ values correspond to mutations into residues that are rarely to never observed at that position across evolution or, more precisely, that are evolutionary unfavorable in the sequence context; positive $\mathcal{Z}_{\mathrm{evo}}$ values indicate mutations into residues that are evolutionary favorable.

Results

Predicting fitness of HMBS variants

HMBS, also known as porphobilinogen deaminase, is an enzyme involved in the heme biosynthesis pathway, and more specifically in the conversion of porphobilinogen into heme precursor hydroxymethylbilane [41]. Mutations in this gene have been associated with acute intermittent porphyria (AIP), which is a rare metabolic disease with life-threatening neurovisceral attacks that require frequent hospitalization of patients [42]. As almost one third of HMBS variants annotated in the ClinVar database [43] are VUS, saturation mutagenesis experiments using high-throughput yeast complementation assays have recently been performed to estimate the fitness of HMBS variants and better understand the pathogenic mechanisms leading to AIP [44]. This data was unpublished at the time of the CAGI6 experiment and was used as blind fitness values to assess predictors.

Among the 5963 HMBS single-site missense mutations with experimental fitness values from [44], the CAGI6 assessors discarded hyper-complementing mutations (with experimental scores above 1.36), leaving a final evaluation dataset of 5811 mutations [27]. Indeed, it has previously been reported by the authors of the experiments that such variants displaying increased fitness in yeast assays could be mostly disadvantageous in human [38, 44].

We applied our prediction models FiTMuSiC (Eq. (4)), SNPMuSiC (Eq. (5)) and PoPMuSiC (Eq. (6)) to the HMBS target. We also report the results of the two other top-performing methods among the 11 teams participating in the challenge, i.e. CalVEIR and ELAPSIC (called team 10_5 and 5_1 in [27]). Additionally, we provide the results of six widely used methods for deleteriousness prediction, i.e. FATHMM [13], PROVEAN [17], DEOGEN2 [15], PolyPhen–2.0 [19], EVE [14] and MutPred2 [23] as well as two recently developed deep-learning based predictors, Sequence UNET [22] and AlphaMissense [24]. To ensure consistency with the metrics provided by the CAGI6 HMBS challenge, all methods were benchmarked on the same dataset of 5811 mutations. The performance of the predictors was assessed by three types of correlations (i.e. Pearson correlation, and Spearman and Kendall rank correlations), and the root mean squared deviation (RMSD). The results are given in Table 1.

Note that the current version of FiTMuSiC (available on our webserver) slightly outperforms the version used for the CAGI6 HMBS challenge due to a small implementation modification. Namely, we now consider the SNP and PVS terms separately (as described in Methods), whereas they were aggregated into a single term in the previous version. The Kendall, Spearman and Pearson correlations improved from (0.30, 0.43, 0.42) to (0.31, 0.45, 0.45), respectively, between the first and second versions. However, to ensure the blind nature of the challenge, we presented in the table the performances of the older FiTMuSiC version.

Among CAGI6 participants, FiTMuSiC performs as well as the other two best performing predictors, ELAPSIC and CalVEIR with very similar performance metrics. CalVEIR shows the best results in rank-based metrics, ELAPSIC in Pearson correlation and FiTMuSiC in RMSD. These three predictors all perform significantly better than the other 8 teams participating in CAGI6 [27]. They also perform significantly better than the other methods tested (FATHMM, PROVEAN, DEOGEN2, PolyPhen-2, Sequence UNET and MutPred2), except for EVE and AlphaMissense. We observe that FiTMuSiC outperforms EVE in rank-based metrics but not in Pearson correlation and that, conversely, FiTMuSiC outperforms AlphaMissense in Pearson correlation but not in rank-based metrics. Overall, these five best performing methods display very comparable scores and their respective ranking depends on the metric considered.

Table 1 Fitness prediction results of the benchmarked methods on the 5811 variants used in the CAGI6 HMBS challenge [27]. The best score for each metric is indicated in bold

Method	CAGI6	Kendall	Spearman	Pearson	RMSD
FiTMuSiC	✓	0.30	0.43	0.42	**0.39**
$\mathcal{SNP}$	✓	0.27	0.39	0.38	0.43
$\mathcal{POP}$	✓	0.15	0.22	0.24	0.44
ELAPSIC team	✓	0.30	0.42	0.43	0.43
CalVEIR team	✓	0.31	0.45	0.36	0.51
FATHMM		0.16	0.23	0.17	–
PROVEAN		0.21	0.31	0.30	–
DEOGEN2		0.22	0.32	0.20	–
PolyPhen-2		0.21	0.28	0.22	–
EVE*		0.29	0.42	**0.43**	–
Sequence UNET		0.21	0.30	0.30	–
MutPred2		0.25	0.37	0.34	–
AlphaMissense		**0.32**	**0.46**	0.41	–

The performances were taken from the assessors' results for CAGI6 participants, while for the other methods we evaluated the performances ourselves. EVE's predictions are available for only 5152/5811 variants; missing values where set to the median

We also wish to underline the good performances of the SNPMuSiC deleterious variant predictor [16], which only slightly underperforms the best methods. In contrast, PoPMuSiC [29], which predicts stability changes upon mutations, does not work so well. This is not surprising given deleteriousness and fitness are very well correlated, while stability and fitness are less so. For example, all functional residues are highly important for fitness while very poorly optimized for stability [40].

The performance of the tested methods can be considered as good considering that the HMBS data was not seen by any of the methods. However, there is still room for improvement as the Pearson correlation coefficient of all methods is below 0.5. Note, however, that the noisiness of deep-mutagenesis datasets (with both random and systematic errors) puts an upper bound to the precision of the predictors which cannot be surpassed without overfitting.

Feature analysis and model interpretation

It is well known that enzymes exhibit an activity-stability trade-off: residues in catalytic regions are optimized for functional reasons and less or not at all for stability, while other residues are very important for protein folding and stability and play little to no role in function [40]. FiTMuSiC can help in distinguishing these functional and structural contributions. Indeed, it outputs the z-scores $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$ (Eqs. 10–11) which inform us about the extent to which structural and/or evolutionary features contribute to protein fitness, and provides us with a molecular-level understanding of variant effects. It also gives us information about the RSA of the mutated residues, and thus about their location in the protein.

We focused here on three functionally or structurally important residue groups of HMBS, which are structurally represented in Fig. 1 and colored according to their average per-residue z-score values $\mathcal{Z}_{\text{evo}}$ and $\mathcal{Z}_{\text{str}}$. Paired $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$ values of all single-site mutations are

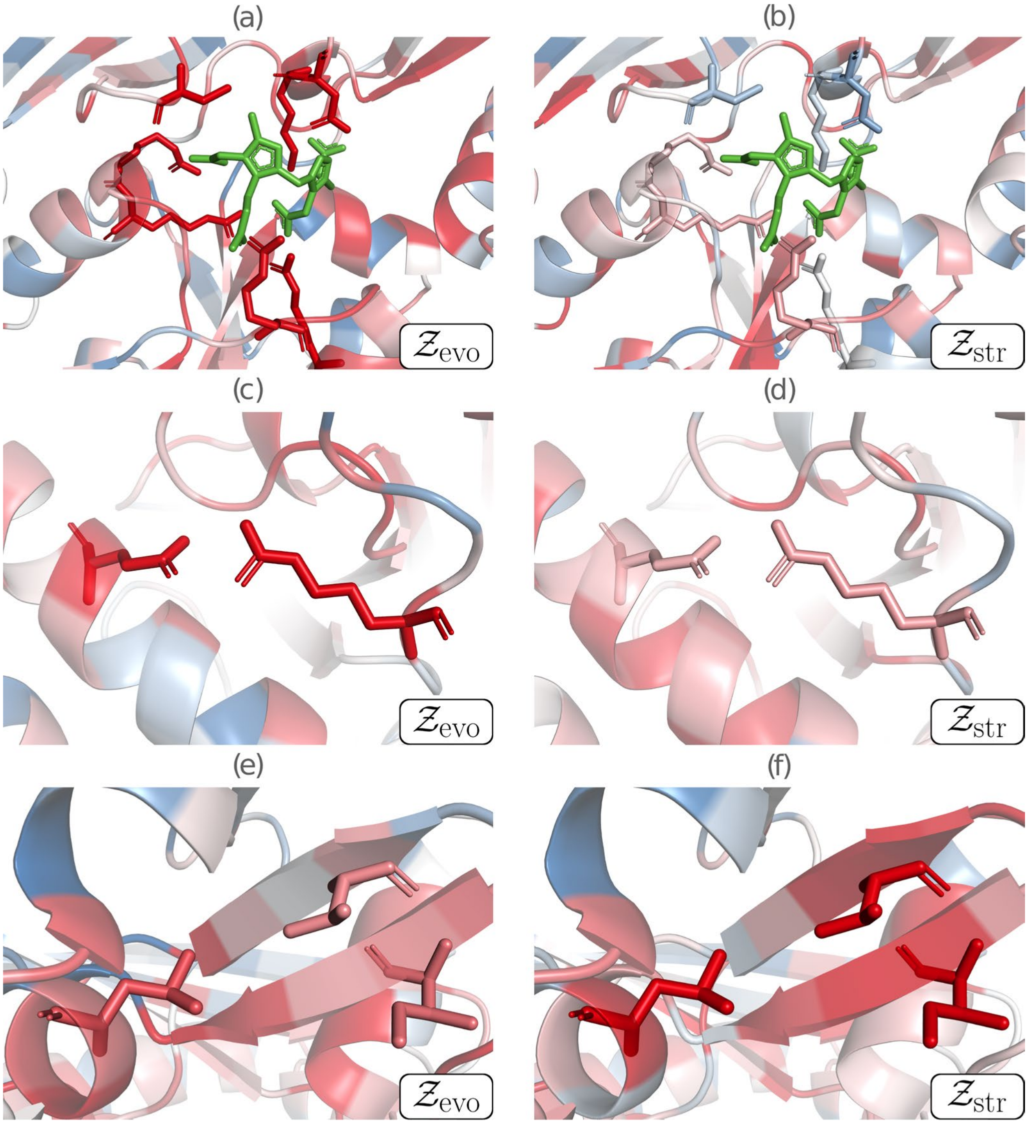
(a)
$\mathcal{Z}_{\text{evo}}$
(b)
$\mathcal{Z}_{\text{str}}$
(c)
$\mathcal{Z}_{\text{evo}}$
(d)
$\mathcal{Z}_{\text{str}}$
(e)
$\mathcal{Z}_{\text{evo}}$
(f)
$\mathcal{Z}_{\text{str}}$

Fig. 1 Contributions of structural and evolutionary features to HMBS fitness, represented by $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$, respectively. Negative z-scores (indicating mutations less fit than average mutations) are in red, close to zero scores in white and positive scores (indicating mutations fitter than average mutations) in blue. **a**, **b** Catalytic region, with the catalytic residues K98, D99, R149, R150, R167, R173 and C261 shown in sticks, and the substrate in green; **c**, **d** Salt bridge partners E250 and R116 shown in sticks; **e**, **f** Cluster of the three buried hydrophobic residues V124, I186 and L193 shown in sticks

plotted in Fig. 2, with the mutations of the selected residue groups highlighted.

The region around the catalytic site of HMBS is represented in Figs. 1a,b and 2a. The catalytic residues (K98, D99, R149, R150, R167, R173 and C261) were identified by aligning the sequences of the considered human HMBS and of *Escherichia coli* HMBS, and by mapping the seven catalytic residues of the latter [45] annotated in the Catalytic Site Atlas [46]. These residues are thus functionally important, well conserved and very specific. As expected, mutating them results in very negative $\mathcal{Z}_{\text{evo}}$ values (between −2.41 and −0.59), which reflects drastic reduction or loss of function. In contrast, they contribute little to structural stability, as seen from the predicted $\mathcal{Z}_{\text{str}}$ values centered around zero (between −1.43 and +1.04).

The second region considered is the salt bridge between the negatively charged residue E250 and the positively charged residue R116 (Figs. 1c, d and 2b). It is a highly specific interaction that has been shown to play an essential role in the enzyme's fold by molecular dynamics simulations [44]. The $\mathcal{Z}_{\text{evo}}$ and $\mathcal{Z}_{\text{str}}$ values of these two residues are predicted to be negative on the average (−1.41 and −0.43 respectively), indicating fitness reduction upon mutations. $\mathcal{Z}_{\text{evo}}$ is negative for all mutations (≤ -0.72), whereas $\mathcal{Z}_{\text{str}}$ is only negative on the average (between −1.53 and +0.37). The high specificity of the interaction gives a particularly strong evolutionary signal, whereas the stabilizing effect of salt bridges is less marked compared to other interactions.

Finally, the hydrophobic cluster of the three residues V124, I186 and L193 (Figs. 1e, f and 2c) located in the core of the protein is very important for the stability of the protein fold. It thus shows strongly negative $\mathcal{Z}_{\text{str}}$ values, with some exceptions that correspond to mutations from one hydrophobic residue into another. In contrast, this cluster plays no direct role in the protein's enzymatic activity and, moreover, hydrophobic interactions have low specificity and are often substituted with other hydrophobic residues across evolution. This explains the large width of the $\mathcal{Z}_{\text{evo}}$ distribution (between −1.52 and +1.64), and its only weakly negative average value (−0.70). On the other hand, $\mathcal{Z}_{\text{str}}$ values are also sparse (between −3.05 and +0.38) but are more shifted towards negative values (average of −1.60).

Comparing the coefficients in Eqs. (8) and (9) when all features of the linear regression are normalized by their standard deviation, we found the contribution of the evolutionary features to the final score to be about 3 times greater than that of structural features, which indicates that evolutionary terms hold a relatively larger predictive power. However, it is the combination of both contributions that leads to the highest precision and structural terms thus improve the detection of deleterious variants. For instance, most mutations of residue L244 have very low experimental fitness a display a largely negative $\mathcal{Z}_{\text{str}}$ but a positive $\mathcal{Z}_{\text{evo}}$. We postulate that the deleterious nature of these mutations has not been detected by evolutionary features due to the relatively low frequency of leucine in the MSA at this position (about 0.02). Another advantage of the structural terms is that they are reliable on proteins or protein regions with low evolutionary information (resulting in low-depth MSAs regions), such as *de novo* designed proteins. Indeed, none of the structural terms rely on evolutionary information.

In summary, the combination of both structural and evolutionary terms makes it possible to interpret whenever the deleterious effect of a mutation is attributed to a loss of function or to a perturbation of the protein fold. Since evolution and structure are related, it is no surprise that we often observe correlated $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$ values. However, this correlation is limited (Pearson correlation

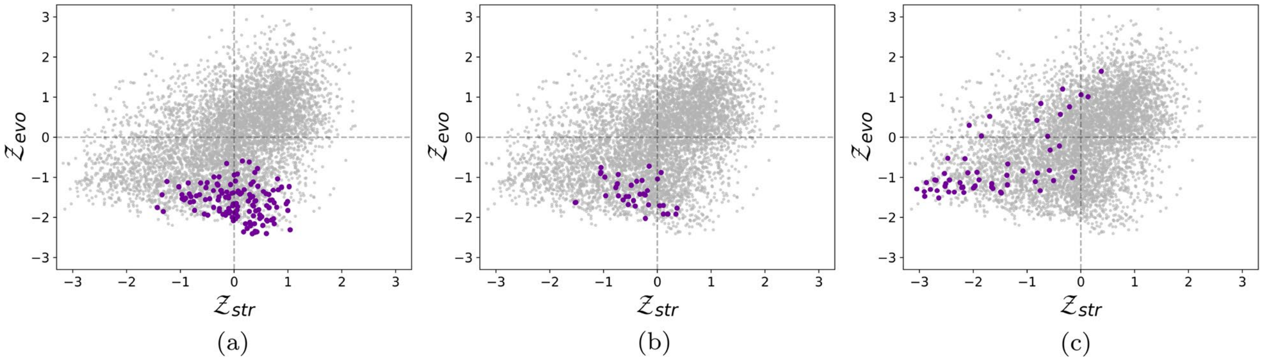

Fig. 2 Scatter plots of paired $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$ values for all single-site mutations in HMBS. Mutations of **a** the catalytic residues K98, D99, R149, R150, R167, R173 and C261, **b** the salt bridge residues E250 and R116 and **c** the hydrophobic cluster residues V124, I186 and L193 are highlighted in purple

of 0.40). As a matter of fact, there are a lot of counterexamples where $\mathcal{Z}_{\text{str}}$ and $\mathcal{Z}_{\text{evo}}$ have opposite signs, as seen in Fig. 2. This reflects the fact that the evolutionary and structural components of fitness are complementary, and that combining them into a single model increases both its accuracy and interpretability.

Prediction of HMBS gain-of-function variants

Variants displaying an increased fitness compared to the wild-type, sometimes referred as gain-of-function (GoF) variants are known to be difficult to predict and to interpret [47]. Furthermore, as pointed out above, very high fitness values in yeast assays tend to be deleterious in human, making their interpretation even more ambiguous. FiTMuSiC, as well as the other assessed predictors, cannot be used to accurately detect GoF variants. However, we still note that the set of variants with experimental fitness above 1.1 (about one tenth of all HMBS mutations) have both positive $\mathcal{Z}_{\text{evo}}$ and $\mathcal{Z}_{\text{str}}$ values (0.51 and 0.37, respectively). In addition, when comparing the average z-score of the GoF variant predictions, FiTMuSiC displays the highest value (0.54) among all tested methods.

FiTMuSiC application to HMBS variant pathogenicity and activity

Fitness predictors are expected to play a crucial role in the classification and interpretation of genetic variants by providing complementary information to the experimental characterizations [48]. It has however to be noted that the experimental HMBS fitness values of the CAGI6 challenge come from a deep mutagenesis experiment that uses functional complementation yeast assays, which cannot fully reflect the complex mechanisms underlying variants' pathogenicity and activity.

In this context, we assessed all the predictors considered as well as the experimental yeast assay data [44] on their ability to distinguish clinically annotated pathogenic and benign variants in humans. To that end, we collected the 53 pathogenic or likely pathogenic variants in HMBS that are related to AIP and the 13 benign or likely benign variants from ClinVar [43]. The metrics we used to assess the methods' performances are sensitivity, specificity and balanced accuracy (BACC), for which we used the default prediction thresholds provided by the methods (and 0.5 for FiTMuSiC), as well as a threshold-independent metric, the area under the receiver operating characteristic curve (AUC-ROC). We reported all performances in Table 2.

We observe that FiTMuSiC predicts with very high accuracy the pathogenicity of the variants with a BACC of 0.94 and an AUC-ROC of 0.98 only slightly outperformed by AlphaMissense with a BACC of 0.95 and an AUC-ROC of 0.99. It performs better than all other computational methods and also, notably, than the experimental high-throughput fitness data obtained by yeast complementation assays to evaluate variant pathogenicity. We found that some of the computational methods tested are heavily biased towards pathogenic variants, as for example PolyPhen-2 and FATHMM. This can be explained by the choice of the threshold values proposed by their authors. They have thus a very poor specificity and predict very few neutral variants. FiTMuSiC does not suffer from this bias and reaches almost perfect accuracy in identifying neutral variants. Note that EVE also shows good performances which are only slightly less accurate than FiTMuSiC.

Table 2 Performance on 66 HMBS variants with clear clinical annotations taken from ClinVar [44], using all predictors assessed as well as experimental fitness data obtained by yeast complementation assays [44]. The best score for each metric is indicated in bold

Method	Sensitivity	Specificity	BACC	AUC-ROC
Experimental	0.81	**0.92**	0.87	0.92
FiTMuSiC	0.96	**0.92**	0.94	0.98
FATHMM	**1.00**	0.00	0.50	0.79
PROVEAN	0.96	0.77	0.87	0.87
DEOGEN2	**1.00**	0.23	0.62	0.93
PolyPhen-2	0.98	0.31	0.64	0.91
EVE	0.94	**0.92**	0.93	0.98
Sequence UNET	0.70	0.77	0.73	0.82
MutPred2	**1.00**	0.54	0.77	0.96
AlphaMissense	0.98	**0.92**	**0.95**	**0.99**

Table 3 Correlation coefficients between experimental activity on 35 HMBS variants measured in [49] and the fitness values obtained by the assessed predictors and by experimental yeast complementation assays [44]. The best score for each metric is indicated in bold

Method	Spearman	Pearson
Experimental	**0.77**	0.72
FiTMuSiC	**0.53**	0.85
FATHMM	**0.53**	0.57
PROVEAN	0.19	0.50
DEOGEN2	0.41	0.57
PolyPhen-2	0.38	0.53
EVE*	0.40	0.71
Sequence UNET	0.08	0.35
MutPred2	0.23	0.49
AlphaMissense	**0.53**	**0.94**

*EVE's predictions are available for 34/35 variants; the missing value was set to the median

As an additional verification of FiTMuSiC robustness, we checked if it is able to predict the effect of variants on HMBS *in vitro* activity. We reported in Table 3 the correlations between the results of the predictors or high-throughput experiments and the experimentally measured activity of 35 variants described in [49]. These results show that FiTMuSiC performs very well. It even outperforms in Pearson correlation the experimental fitness data from [44] and is only outperformed by AlphaMissense.

FiTMuSiC application to human glucokinase

To further test the robustness of FiTMuSiC, we applied it to another blind test set containing experimental high-throughput fitness data of single-site variants in human glucokinase (GCK). This enzyme plays a key role in insulin secretion in pancreatic β-cells: it catalyzes the first step of the glycolysis by transforming glucose into glucose-6-phosphate [50]. Inactivating GCK variants were related to maturity-onset diabetes of the young as well as to permanent neonatal diabetes mellitus [50, 51]. Hyperactive GCK variants are also deleterious and lead to persistent hyperinsulinemic hypoglycemia of infancy.

In order to shed light on the molecular effects that lead to these disorders, the GCK activity of 8570 single-site variants have been experimental assessed using functional complementation yeast assays [52]. We used this set of variants as independent test set to assess the fitness predictors. To ensure homogeneity between this dataset and data provided for HMBS, we floored all negative fitness values to zero and excluded all values with standard error exceeding 0.3, as was done in the experimental data from [44]. This gives a final number of 6862 missense mutations; note that the experimental data appears to be noisier on GCK than on HMBS, as experiments on the latter were repeated twice. We show the performances of FiTMuSiC and other computational tools on GCK in Table 4. FiTMuSiC is among the top ranked predictors on this additional test set, with performance metric values in line with those of the HMBS benchmark.

We also evaluated the ability of the methods to classify deleterious and benign GCK variants that are defined based on clinical annotations. For that purpose, we curated a set of variants in GCK from ClinVar [43] with clear clinical interpretation. This led us to a collection 69 pathogenic or likely pathogenic variants, and 3 benign or likely benign variants. The very low number of benign variants and the bias of predictors towards deleterious variants make this test case relatively easy, and most methods thus reach very high scores: five methods have an AUC-ROC of at least 0.97 (Table 5). FiTMuSiC also shows good performance with a BACC of 0.89 and an AUC-ROC of 0.99. Due to the strong imbalance of this test set, we suggest to consider these results with caution.

It has to be noted that the use of experimental fitness data from complementation yeast assays to predict deleteriousness does not perform very well for GCK variants (Table 5). The BACC and AUC-ROC values are even lower than in the case of HMBS. Some reported pathogenic variants such as V62M, T65I and H137R, seem to be benign in the experimental fitness map. Their deleteriousness has been suggested to be related to effects such as modest structural instability which are not captured by the assay [52]. This observation underlines the importance of reliable and robust prediction methods to complement experimental data for annotation and interpretation of variants.

Table 4 Correlations between fitness values obtained by high-throughput experiments using functional complementation yeast assays [52] on 6862 GCK variants and those predicted by all the methods assessed. The best score for each metric is indicated in bold

Method	Spearman	Pearson
FiTMuSiC	0.49	**0.40**
FATHMM	0.39	0.30
PROVEAN	0.41	0.33
DEOGEN2	0.51	0.35
PolyPhen-2	0.36	0.23
EVE*	0.47	**0.40**
Sequence UNET	0.30	0.24
MutPred2	0.51	0.34
AlphaMissense	**0.53**	**0.40**

*EVE's predictions are available for only 6414/6862 GCK variants; missing values where set to the median

Table 5 Performance on 72 GCK variants with clear clinical annotations taken from ClinVar [43], using all predictors assessed as well as experimental fitness data obtained by yeast complementation assays [52]. The best score for each metric is indicated in bold

Method	Sensitivity	Specificity	BACC	AUC-ROC
Experimental	0.59	**1.00**	0.80	0.732
FiTMuSiC	0.78	**1.00**	0.89	0.990
FATHMM	**1.00**	0.00	0.50	0.766
PROVEAN	0.84	**1.00**	0.92	0.976
DEOGEN2	0.97	**1.00**	**0.99**	**0.995**
PolyPhen-2	0.93	**1.00**	0.96	0.978
EVE	0.58	**1.00**	0.79	0.807
Sequence UNET	0.52	0.67	0.59	0.638
MutPred2	0.91	0.67	0.79	0.865
AlphaMissense	0.87	**1.00**	0.93	**0.995**

Webserver

In order to make FiTMuSiC readily available to the scientific community, we have developed an easy-to-use webserver at http://babylone.ulb.ac.be/FiTMuSiC/. Users need to input a 3D structure of the target protein in one of three ways:

1. Provide its PDB ID if it is available in the Protein Data Bank (PDB) [53]; the structure is automatically retrieved.
2. Provide its UniProt ID; the corresponding AlphaFold DB structure [39] is then retrieved.
3. Provide a personal structure in PDB format (`.pdb`).

Since FiTMuSiC provides results on a per-chain basis, users need to select which chain they want the results for. Note that FiTMuSiC only outputs the results of a single chain, but the structural components of the model take into account all the chains contained in the structure file when computing the fitness score. Therefore, we recommend that users provide protein structures that correspond to biological units, especially when dealing with multimers.

Once the chain has been selected and submitted, the computation starts. Depending on the length of the query protein and the depth of its MSA, users should expect the computation to be completed in a few minutes for short proteins to a few hours for very long proteins. Once the computation is done, a CSV file with the results is sent to the email address provided during the submission. This file contains the RSA of all residues in the protein and the predicted fitness scores for all possible single-site variants. The last four columns contain fitness score information, i.e. the raw FiTMuSiC score and the z-scores $\mathcal{Z}$, $\mathcal{Z}_{\text{evo}}$ and $\mathcal{Z}_{\text{str}}$ (Eqs. 4, 7, 10, 11). More information about the webserver and its usage is available on the help page (http://babylone.ulb.ac.be/FiTMuSiC/help.php).

Conclusion

We presented here FiTMuSiC, our new computational model based on a combination of structural and (co) evolutionary information, which predicts the impact of single-site amino acid substitutions on protein fitness. We applied it to predict variants in HMBS, one of the targets of the CAGI6 challenge. It was rated as one of the top three predictors by the CAGI6 assessors [27]. The strengths of FiTMuSiC can be summarized as follows:

- It is based on a simple model, which is less prone to overfitting and biases towards the training set than machine learning models with thousands of parameters. This allows for very good performances on blind, independent test sets as we have shown here for variants in HMBS and GCK.
- It retains interpretability by providing the $\mathcal{Z}_{\text{evo}}$ and $\mathcal{Z}_{\text{str}}$ scores which allow distinguishing between variants that impact more on function or on stability.
- It is available through an easy-to-use webserver, which allows users to get FiTMuSiC results in a simple way even without bioinformatics background.

For all these reasons, FiTMuSiC is of interest to the large community of scientists interested in the prioritization, classification and interpretation of genetic variants. Moreover, it represents a reliable, complementary and cheaper approach compared to experimental methods.

Acknowledgements

We would like to thank the assessors of the CAGI6 HMBS challenge Jing Zhang, Qian Cong and Nick V. Grishin for their assessment work and for providing us detailed information about it. We thank Steven Brenner, Predrag Radivojac and Constantina Bakolitsa for organizing the CAGI competition as well as the experimental group led by Frederick P. Roth for providing the data on HMBS variant fitness used as blind test in the HMBS challenge.

Author contributions

MR and FP conceived and supervised this study. MT, GC, MR and FP performed the investigation. MT, GC, PH and JK curated the data and developed the webserver associated to the method. MT, GC, MR and FP wrote the original draft. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare that they have no Conflict of interest.

Received: 1 August 2023 Accepted: 2 April 2024
Published online: 16 April 2024

References

1. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. Am J Hum Genet. 2019;105(3):448–55.
2. Eichler EE. Genetic variation, comparative genomics, and the diagnosis of disease. N Engl J Med. 2019;381(1):64–74.
3. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. J Hum Genet. 2021;66(1):11–23.
4. Morrison KL, Weiss GA. Combinatorial alanine-scanning. Curr Opin Chem Biol. 2001;5(3):302–7.
5. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. Variant interpretation: functional assays to the rescue. Am J Hum Genet. 2017;101(3):315–25.
6. Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. Hum Genet. 2018;137(9):665–78.
7. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20(1):1–11.
8. Karczewski KJ, Solomonson M, Chao KR, Goodrich JK, Tiao G, Lu W, Riley-Gillis BM, Tsai EA, Kim HI, Zheng X, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genomics. 2022;2(9):100168.
9. Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, Karczewski KJ, O'Connor LJ. Polygenic architecture of rare coding variation across 394,783 exomes. Nature. 2023;614(7948):492–9.

10. Federici G, Soddu S. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. J Exp Clin Cancer Res. 2020;39:1–12.
11. Frederiksen SD, Avramović V, Maroilley T, Lehman A, Arbour L, Tarailo-Graovac M. Rare disorders have many faces: in silico characterization of rare disorder spectrum. Orphanet J Rare Dis. 2022;17(1):1–18.
12. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med. 2023;29(7):554–66.
13. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57–65.
14. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–5.
15. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):201–6.
16. Ancien F, Pucci F, Godfroid M, Rooman M. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep. 2018;8(1):4480.
17. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745–7.
18. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361–2.
19. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;76(1):7–20.
20. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
21. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):886–94.
22. Dunham AS, Beltrao P, AlQuraishi M. High-throughput deep learning variant effect prediction with sequence UNET. Genome Biol. 2023;24(1):1–19.
23. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11(1):5918.
24. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):7492.
25. Andreoletti G, Pal LR, Moult J, Brenner SE. Reports from CAGI: CAGI the critical assessment of genome interpretation. Hum Mutat. 2019;40(9):1197–201.
26. Consortium GI, et al. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. arXiv e-prints, 2022;2205
27. Zhang J. et al. Assessing predictions on fitness effects of missense variants in HMBS in CAGI6. submitted
28. Dalkas GA, Teheux F, Kwasigroch JM, Rooman M. Cation-*π*, amino-*π*, *π*-*π*, and H-bond interactions stabilize antigen-antibody interfaces. Proteins Struct Funct Bioinform. 2014;82(9):1734–46.
29. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics. 2009;25(19):2537–43.
30. Dehouck Y, Gilis D, Rooman M. A new generation of statistical potentials for proteins. Biophys J. 2006;90(11):4010–7.
31. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO-multi agent stability prediction upon point mutations. BMC Bioinform. 2015;16(1):1–13.
32. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(s2):29–37.
33. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., Consortium, U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–32.
34. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009;30(8):1237–44.
35. Raimondi D, Gazzo AM, Rooman M, Lenaerts T, Vranken WF. Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. Bioinformatics. 2016;32(12):1797–804.
36. Pucci F, Zerihun M, Rooman M, Schug A. pycofitness-Evaluating the fitness landscape of RNA and protein sequences. Bioinformatics 2024;btae074.
37. Zerihun MB, Pucci F, Peter EK, Schug A. pydca v.10: a comprehensive software for direct coupling analysis of RNA and protein sequences. Bioinformatics. 2020;36(7):2264–5.
38. 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. Mol Syst Biol. 2017;13(12):957.
39. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):439–44.
40. Hou Q, Rooman M, Pucci F. Enzyme stability-activity trade-off: New insights from protein stability weaknesses and evolutionary conservation. J Chem Theory Comput. 2023;19(12):3664–71.
41. Bustad HJ, Kallio JP, Laitaoja M, Toska K, Kursula I, Martinez A, Jänis J. Characterization of porphobilinogen deaminase mutants reveals that arginine-173 is crucial for polypyrrole elongation mechanism. Iscience. 2021;24(3):102152.
42. Simon A, Pompilus F, Querbes W, Wei A, Strzok S, Penz C, Howe DL, Hungate JR, Kim JB, Agarwal S, et al. Patient perspective on acute intermittent porphyria with frequent attacks: a disease with intermittent and chronic manifestations. Patient-Patient-Center Outcomes Res. 2018;11:527–37.
43. 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 Res. 2018;46(D1):1062–7.
44. van Loggerenberg W, Sowlati-Hashjin S, Weile J, Hamilton R, Chawla A, Sheykhkarimli D, Gebbia M, Kishore N, Frésard L, Mustajoki S, et al. Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. Am J Hum Genet. 2023;110(10):1769–86.
45. Woodcock SC, Jordan PM. Evidence for participation of aspartate-84 as a catalytic group at the active site of porphobilinogen deaminase obtained by site-directed mutagenesis of the hemC gene from Escherichia coli. Biochemistry. 1994;33(9):2688–95.
46. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 2004;32(s1):129–33.
47. Li MM, Awasthi S, Ghosh S, Bisht D, Coban Akdemir ZH, Sheynkman GM, Sahni N, Yi SS. Gain-of-function variomics and multi-omics network biology for precision medicine, pp. 357–372. Springer, New York 2023.
48. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–23.
49. Bung N, Roy A, Chen B, Das D, Pradhan M, Yasuda M, New MI, Desnick RJ, Bulusu G. Human hydroxymethylbilane synthase: Molecular dynamics of the pyrrole chain elongation identifies step-specific residues that cause AIP. Proc Natl Acad Sci. 2018;115(17):4071–80.
50. Gloyn AL. Glucokinase (GCK) mutations in hyper-and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy. Hum Mutat. 2003;22(5):353–62.
51. Osbak KK, Colclough K, Saint-Martin C, Beer NL, Bellanné-Chantelot C, Ellard S, Gloyn AL. Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum Mutat. 2009;30(11):1512–26.
52. Gersing S, Cagiada M, Gebbia M, Gjesing AP, Coté AG, Seesankar G, Li R, Tabet D, Weile J, Stein A, et al. A comprehensive map of human glucokinase variant activity. Genome Biol. 2023;24(1):1–23.
53. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.