Assessing predictions on fitness effects of missense variants in HMBS in CAGI6

Assessing predictions on fitness effects of missense variants in HMBS in CAGI6

Jing Zhang[1,2,3,6], Lisa Kinch[4,5], Panagiotis Katsonis[7], Olivier Lichtarge[7], Milind Jagota[8], Yun S. Song[8,9], Yuanfei Sun[10], Yang Shen[10], Nurdan Kuru[11], Onur Dereli[11], Ogun Adebali[11], Muttaqi Ahmad Alladin[12], Debnath Pal[12], Emidio Capriotti[13], Maria Paola Turina[13], Castrense Savojardo[13], Pier Luigi Martelli[13], Giulia Babbi[13], Rita Casadio[13] Zhiqiang Hu[22], Fabrizio Pucci[14], Marianne Rooman[14], Gabriel Cia[14], Matsvei Tsishyn[14], Alexey Strokach[18], Warren van Loggerenberg[15,16,17,18], Frederick P. Roth[15,16,17,18], Predrag Radivojac[19], Steven E. Brenner[20,21,22], Qian Cong[1,2,3,6,*], Nick V. Grishin[1,2,*].

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.
[2]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[3]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
[4]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [5]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[6]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
[7]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[8]Computer Science Division, University of California, Berkeley, CA 94720, USA
[9]Department of Statistics, University of California, Berkeley, CA 94720, USA
[10]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[11]Faculty of Engineering and Natural Sciences at Sabanci University, Turkey
[12]Department of Computational and Data Sciences, Indian Institute of Science, Bangaluru, 560012, India
[13]Department of Pharmacy and Biotechnology, University of Bologna, Via Selmi 3, 40126 Bologna, Italy
[14]Computational Biology and Bioinformatics, Université Libre de Bruxelles, 50 Roosevelt Ave, 1050, Brussels, Belgium
[15]Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[16]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada
[17]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON M5G 1X5, Canada
[18]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[19]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
[20]Biophysics Graduate Group, University of California, Berkeley, Berkeley, CA 94720, USA
[21]Center for Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[22]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA

*corresponding authors.

Abstract

This paper presents an evaluation of predictions submitted for the "HMBS" challenge, a component of the sixth round of the Critical Assessment of Genome Interpretation held in 2021. The challenge required participants to predict the effects of missense variants of the human HMBS gene on yeast

growth. The HMBS enzyme, critical for the biosynthesis of heme in eukaryotic cells, is highly conserved among eukaryotes. Despite the application of a variety of algorithms and methods, the performance of predictors was relatively similar, with Kendall tau correlation coefficients between predictions and experimental scores around 0.3 for a majority of submissions. Notably, the median correlation (>=0.34) observed among these predictors, especially the top predictions from different groups, was greater than the correlation observed between their predictions and the actual experimental results. Most predictors were moderately successful in distinguishing between deleterious and benign variants, as evidenced by an area under the receiver operating characteristic (ROC) curve (AUC) of approximately 0.7 respectively. Compared with the recent two rounds of CAGI competitions, we noticed more predictors outperformed the baseline predictor, which is solely based on the amino acid frequencies. Nevertheless, the overall accuracy of predictions is still far short of positive control, which is derived from experimental scores, indicating the necessity for considerable improvements in the field. The most inaccurately predicted variants in this round were associated with the insertion loop, which is absent in many orthologs, suggesting the predictors still heavily rely on the information from multiple sequence alignment.

Introduction

Understanding the relationship between genotype and phenotype is pivotal, as it underpins human trait diversity and plays a critical role in the onset and progression of diseases. Despite the advances in techniques to deduce the phenotypic effects of genomic variants (Adzhubei et al. 2013; Ancien et al. 2018; Calabrese et al. 2009; Capriotti et al. 2006; Choi and Chan 2015; Dehouck et al. 2009; Ioannidis et al. 2016; Katsonis and Lichtarge 2014; Ng and Henikoff 2001; Raimondi et al. 2017) and the formation of various global consortia (Genomes Project et al. 2015; International Cancer Genome et al. 2010; Lander et al. 2001; Turnbull et al. 2018) for the collection and analysis of genomic data, the exact link between genotype and phenotype remains elusive. This gap in knowledge persists despite advancements in comprehending diseases, including cancer, and their genetic bases. Echoing the objectives of The Critical Assessment of Protein Structure Prediction (CASP) (Kryshtafovych et al. 2021), which rigorously evaluates computational models for macromolecular structures and complexes, the Critical Assessment of Genome Interpretation (CAGI) (Critical Assessment of Genome Interpretation 2024) is established for a similar purpose in genomics. CAGI aims to rigorously assess computational methods for predicting the impacts of genomic variation and to gauge our proximity to the ultimate goal of *in silico* phenotype prediction from genotypes.

The CAGI, round six, includes 13 challenges, and here we present the assessment of a challenge called "HMBS". In this challenge, fitness scores were provided through a complementation assay developed in Frederick Roth's Lab (Warren van et al. 2023). The assay assessed the ability of human hydroxymethylbilane synthase (HMBS) missense variants to rescue a temperature-sensitive mutation of the yeast ortholog HEM3. The fitness score conceptually represents the relative growth rate of yeast expressing HMBS missense variants compared to yeast expressing wild-type HMBS. Deleterious missense variants have fitness scores closer to 0, while tolerated variants have fitness scores closer to 1. Participants were expected to predict fitness scores with experimental standard error for 6,589 variants of HMBS, including 310 synonymous, 317 nonsense, and 5,962 missense variants. Although the exact values of experimental fitness scores were not disclosed during the challenge, a distribution was provided to aid in normalizing predictions. In this assessment, we will only focus on the prediction performance of missense mutations.

HMBS is a protein involved in heme biosynthesis. It catalyzes the sequential polymerization of four molecules of porphobilinogen to form hydroxymethylbilane (Song et al. 2009). Dysfunction of the protein may lead to acute intermittent porphyria (AIP), a rare autosomal dominant disease with symptoms such as abdominal pain, nausea, vomiting, peripheral neuropathy, and seizures. HMBS served as a good target for evaluating predictors' ability to predict the effects of variants. First, HMBS is ubiquitous in most eukaryotic cells, providing numerous sequence homologs for sequence analysis. Secondly, there are numerous structures available for HMBS that aid in understanding the functional relevance of mutations (Bustad et al. 2021; Gill et al. 2009; Pluta et al. 2018; Sato et al. 2021; Song et al. 2009). Thirdly, various studies have explored how mutations affect the functions of the protein and the underlying mechanism by which they cause AIP (Kauppinen and von und zu Fraunberg 2002; Lenglet et al. 2018; Schneider-Yin et al. 2008; Ulbrichova et al. 2009). Overall, the wealth of existing knowledge regarding HMBS allows for the application of various methods, making it a suitable target for evaluating computational approaches.

In this round of CAGI, we received 50 predictions from 11 teams (Table 1 and detailed information in Supplementary material). Among them, Teams 1, 3, 5, 6, and 10 incorporated deep learning methods in some or all of their submissions. Team 1 applied two modules by combining a feature extractor using a long short-term memory network and a pathogenicity classifier composed of two fully connected layers; Team 3 combined pre-trained protein language models from bidirectional transformer encoder (BERT) (Devlin et al. 2018) with fine-tuning using HEM3_human multiple sequence alignment. Team 5 developed cross-protein transfer models (Jagota et al. 2023) that used deep mutational scanning data available in public databases along with predictions from REVEL (Ioannidis et al. 2016) , ESM-1v (Meier et al. 2021) , and DeepSequence (Riesselman et al. 2018). Notably, ESM-1v and DeepSequence are both deep learning methods. Team 6 also incorporated one predictor based on deep-learning method, Team 10 used ELASPIC2 (Strokach et al. 2021), ProteinSolver (Strokach et al. 2020), ProteinBert (Brandes et al. 2022), and ELASPIC2 with AlphaFold (Jumper et al. 2021) features for their submissions 1, 2, 3 and 5, respectively and these four methods are deep learning methods while submission 4 utilized Rosetta's cartesian_ddg protocol (Park et al. 2016). Team 7 applied Evolutionary Action scores (Katsonis and Lichtarge 2014), which accounts for phylogenetic divergence (Lichtarge et al. 1996) and amino acid substitution odds, calculated using protein evolution data and Katsonis and Lichtarge team also participated in the previous CAGI rounds using the similar methods (Katsonis and Lichtarge 2017, 2019). Team 9 used SNPMuSiC (Ancien et al., 2018) for submission 1, FiTMuSiC (Tsishyn et al., 2023) for submission 2, and PoPMuSiC (Dehouck et al. 2011) for submission 3. Team 2 developed a novel phylogeny-dependent probabilistic model that utilized phylogenetic tree information to measure the deleteriousness of a given variant. This draft version was an initial attempt that served as a foundation for PHACT (Kuru et al. 2022). PHACT differs from the approach submitted to CAGI in terms of considering position diversity through phylogenetically independent amino acid alterations as well as scaling the final score. On the GitHub page (https://github.com/CompGenomeLab/PHACT/tree/main/CAGI6_HMBS) of the tool, the authors demonstrated that PHACT outperformed both this draft version, PolyPhen-2, and the baseline predictor in various measures over the experimental results used in this challenge. Team 4 combined structural analysis with consensus of predictions from 3 stability predictors, namely FoldX (Schymkowitz et al. 2005), INPS3D (Savojardo et al. 2016), and PoPMuSiC 2.1 (Dehouck et al. 2011), while the Team 6 applied the random forest method to combine several published predictors. Team 11 applied PhyloP (Pollard et al. 2010), PhD-SNPg (Capriotti and Fariselli 2017, 2023), PhD-SNP (Capriotti et al. 2006), and SNPs-and-GO (Calabrese et al. 2009; Capriotti and Altman 2011; Capriotti et al. 2017) with/without structure and various linear transformations for different submissions. All the above methods involved

the multiple sequence alignment directly or indirectly, except for Team 8, which solely focused on structural information with molecular dynamics. Notably, Team 5 and 6 and submission2 from Team 9 also utilized published yeast complementation assay data for proteins such as UBE2I and CALM1 (Weile et al. 2017) to help train their models.

In this HMBS challenge, all top-performing predictors (Team 5 from the Yun Song group, Team 10 from the Alexey Strokach group, and Team 9 from the Fabrizio Pucci group) exhibit similarly moderate correlations with experimental scores, with a Kendall tau correlation coefficient around 0.3. When compared to prior CAGI rounds, a greater number of these predictors surpassed baseline performance, showcasing the progress in the field. Despite this progress, the application of deep learning methods did not achieve the groundbreaking performance seen in approaches like AlphaFold (Jumper et al. 2021) for CASP. Furthermore, regardless of whether deep learning methods were employed, these leading predictors showed stronger correlations with each other than with experimental scores. This pattern suggests a shared reliance on similar types of information, such as amino acid frequency and conservation, within the multiple sequence alignment. Additionally, top-performing predictions, submissions from Team 5 and submission 2 from Team 9, leveraged publicly available yeast complementation assay data for other proteins to transform the values of their raw predictions. This approach suggests that taking advantage of experimental data, such as deep mutational scanning, could improve predictions of mutation effects.

Results

The distribution of experimental scores and predicted scores

In the yeast complementation assay, three types of mutations were provided: nonsense, synonymous, and missense. Interestingly, we observed a wide distribution of relative growth scores for synonymous mutations, which overlapped with the distribution of nonsense mutations. Meanwhile, approximately 20% of the missense mutations in our dataset exhibited extreme deleterious effects with experimental scores of 0 (Fig 1A).

During the HMBS challenge, significant variations were observed in the distribution and value scaling of scores predicted by different teams (Fig 1B) although the distribution of experimental scores was provided to help participants rescale their raw predicted scores. Notably, submission 6 from Team 8 (submission 8_6) was the sole group with predictions not statistically different from the distribution of experimental scores, as confirmed by the Kolmogorov-Smirnov test ($P > 0.01$, detailed in Table S1). In contrast, submission 1 from Team 3 (submission 3_1) provided some predicted scores exceeding 89,700, with 134 missense mutations having predicted scores above 10. Furthermore, both Team 1 and Team 8 submitted predictions that included negative scores. To ensure a fair comparison across predictors and to comply with the guidelines of the challenge that submitted predictions should be numeric values on a log scale greater than or equal to 0, we implemented a quantile transformation to rescale their predictions and shift all negative scores to 0, adhering to the method employed in prior CAGI assessments (Zhang et al. 2019; Zhang et al. 2017). Additionally, we took into account the distributions of nonsense and synonymous mutations. We characterized mutations with growth scores below 0.3 as deleterious, a category that contained less than 5% of synonymous mutations. On the other hand, benign missense mutations were identified as those with growth scores ranging between 0.8 because there are less than 5% of nonsense mutations scored above this threshold. Additionally, we observed

several hyper-complementing mutations (Warren van et al. 2023; Weile et al. 2017). Some of these mutations could be deleterious to humans, while others might result from experimental errors. As such, we excluded mutations with experimental scores over 1.36, a threshold above which the top 5% of synonymous mutations reside. After eliminating these hyper-complementing mutations, our predictor evaluation dataset contained 5811 missense mutations, including 2043 deleterious and 1942 benign mutations, in line with our classification criteria.

Moderate performance achieved but falls significantly short of positive control

We have applied the same evaluation strategy (Table 2) as CAGI4 (Zhang et al. 2017) and CAGI5 (Zhang et al. 2019) to assess the predictions in terms of (1) classification of missense mutations, (2) ranking variants by fitness effects, and (3) numerical prediction of fitness scores with both positive control, from experimental scores, and a baseline predictor based on solely multiple sequence alignment from orthologs in orthoDB (Zdobnov et al. 2021) as references. We also included PolyPhen (Adzhubei et al. 2013) with the HumVar model in the comparison. Table 3 provides a detailed summary of the performance of the predictors against each of these criteria. With the exception of Team 8, which utilized molecular dynamics to predict the effects of variants, all participants demonstrated significantly better than random predictions, with the best-performing teams achieving Kendall's tau correlation coefficients of approximately 0.3. All predictions from Team 1 negatively correlate with experimental scores, which indicates a potential misinterpretation regarding the orientation of the scores.

For discriminating deleterious and non-deleterious mutations, the best-performing submissions for each team are displayed in Fig 2A. Although predictors still fall considerably behind the positive control, a number of them (Team 5, Team 7, Team 9, Team 10, and Team 11) show an improvement in performance compared with the baseline predictor. Interestingly, although submission 11_8 displays an overall better performance, its initial worse performance compared with the baseline predictor at a lower false positive rate suggests it is less specific for recognizing the most deleterious mutations compared with the baseline predictor. In contrast, submission 4 from Team 3 displayed a higher AUC at the initial of the ROC, while performance rapidly deteriorates when a false positive reaches around 0.08, suggesting it is able to predict extremely deleterious mutations but discrimination ability lowered for more benign cases. Team 8 is the only team showing nearly random predictions for deleterious mutations. Team 1 likely reversed the deleterious mutations and benign mutations in the submission. Upon inverting the predictions, Team 1's performance (AUC 0.73 for submission 1) aligns more closely with that of the other top-performing teams (0.75 for submission 10_5, 0.73 for submissions 9_2 (Matsvei et al. 2023) and 11_8), exhibiting comparable metrics. In addition, 7 teams (Team 1, Team 5, Team 6, Team 7, Team 9, Team 10 and Team 11) with predictors surpass the PolyPhen, signifying advancements in recent years.

In contrast to previous rounds of CAGI4 (Zhang et al. 2017) and CAGI5 (Zhang et al. 2019), the current challenge to predict the effects of missense mutations in HEM3 has witnessed the emergence of several predictors, including submissions 5_1, 10_5, 5_2, 9_2, 10_3, and 5_5, that surpass the performance of the baseline predictor, which relies solely on amino acid frequency in a multiple sequence alignment. Notably, submission 5_1 stands out as the overall top-performing predictor when encompassing rank-based scores, original value-based scores, and rescaled-value-based scores. However, a closer examination of the scores reveals that Team5's superior performance primarily stems from its proficiency in original-value-based scores. Conversely, when considering rank-based scores and rescaled

values, predictors from submissions 10_5 and 9_2 exhibit marginal superiority (with a difference around 0.01 to 0.03 in ranked-based measures) over Team5, indicating that Team5 excels over submission 10_5 and submission 9_2 in value assignment. Compared with Team 10 and Team 9, predictions from Team 5 align closer to the distribution of experimental scores (Table S1), and they used publicly available yeast complementation assay scores for other proteins to transform the values of their predictions, which may explain its superiority in original-score-based measures.

To ascertain the statistical significance of our evaluation, we undertook simulations involving 5,000 datasets, assuming a Gaussian distribution for the fitness scores of each variant. The mean and standard deviation for this distribution were derived from the experimental fitness scores and their corresponding standard errors, respectively. For every simulated dataset, we calculated the evaluation metrics, computed Z-scores for each set of predictions, and tallied the number of times one predictor outperformed another. The head-to-head comparisons, depicted in Fig 2B, reveal that submission 5-1 consistently outshone the other predictors across the majority of the simulated datasets, while submission 10_5, 9_2, and 5_2 appeared to be neck and neck. In alignment with the head-to-head analysis, the distribution of ranks for the predictors, as shown in Fig 2C, further supports the notion that submission 5_1 takes the lead in a significant number of simulated datasets. Concurrently, submissions 10_5, 9_2, and 5_2 demonstrate comparable rank distributions, indicating a virtual tie among them.

Inaccurate predictions on functional loops

To investigate the missense mutations where the predictions failed, we examined the absolute difference between the median of rescaled scores from top-performing predictors and experimental scores. Subsequently, we calculated the median absolute difference for each position and visualized it using a heat map (Fig 3A). We observed significant discrepancies between the predictions and experimental scores in/around two specific regions: the active site loop (56 to 76aa) of the diazomethane cofactor binding domain and “insertion regions” (296 to 324aa), a loop constraining the movement of domain 1 (residues 1-114, 219-236) and 2 (residues 120-212) relative to domain 3 (residues 241-361). Other small regions showing high disparities include the cofactor-binding loop (257 to 262aa) and 354-356aa. (Fig 3A and B). Remarkably, domain 3 stands out for its enrichment of missense mutations whose effects are challenging to predict, with 42 (35%) positions exhibiting absolute difference >= 0.4. This is in stark contrast to domain 1 and domain 2, which have only 17 positions (15%) and 19 positions (20%), respectively. Interestingly, domain 3 also has the lowest average alignment depth (Table S2) and conservation score (Table S3).

Upon detailed examination of the distributions of experimental scores and predicted scores from the top-performing predictor, submission 5_1, it is observed that the predictor tends to classify mutations on the active-site loop and cofactor-binding loop as deleterious, although many of them are actually benign. Conversely, around 200 mutations in the insertion regions are predicted to be benign, despite their deleterious effects (Fig 3C). All those regions are less conserved, and the insertion region is even missing in more than 50% of sequences in the HEM3 ortholog group we used to construct our baseline predictor (Table S2).

The high correlation between predictors and conservation plays a significant role in predictions

To evaluate the similarity among the predictors, the absolute Kendall tau correlation coefficients were computed to measure the association between their predictions. Interestingly, a notable degree of

correlation was observed among predictions from different teams, which exceeded the correlation between the experimental scores and the predictions themselves (Fig 4A). To discern what might be contributing to this high similarity among predictors, we analyzed the correlation between conservation scores and predictions, as well as between conservation scores and experimental scores (Fig 4B). This analysis was conducted given that a majority of predictors were based, either directly or indirectly, on multiple sequence alignment. Both the experimental and prediction scores demonstrated a correlation with the conservation index, with Kendall tau correlation coefficients of approximately 0.6 and 0.4, respectively. However, the range of prediction scores across different levels of the conservation index was considerably narrower compared to that of the experimental scores.

Furthermore, we calculated the proportion of deleterious missense mutations occurring at conserved positions versus benign mutations at non-conserved sites, as indicated by both experimental scores and predictions. The experimental scores indicated that about 60% of mutations at conserved sites were deleterious, whereas several predictors were inclined to predict a higher proportion of mutations at conserved sites as deleterious (Fig 4C above). On the flip side, experimental scores suggested that approximately 44% of mutations at non-conserved sites were benign, while many predictors, particularly those that performed well, tended to predict a higher proportion of benign mutations at non-conserved sites. For instance, submissions 5_1, 10_5, and 9_2 estimated that 72.5%, 76%, and 76%, respectively, of mutations at non-conserved positions were benign (Fig 4C below).

Discussions

Advantages and possible disadvantages of using yeast complementation assay for accessing effects of mutations.

The choice of datasets to evaluate mutation effects plays a vital role in shaping the conclusions drawn from the assessment. Many prediction models rely on publicly available datasets like OMIM (Hamosh et al. 2005), dbSNP (Sherry et al. 1999), and ClinVar (Landrum et al. 2014), extracting variant information from these sources. Thus, relying solely on public datasets for evaluation comes with inherent drawbacks: 1) the potential for biased assessments, 2) an overestimation of performance, 3) limitations in generalizing functional effects to new variants, and 4) the possibility of errors in public databases.

To overcome these limitations, the CAGI committee offers a unique dataset of experimentally determined variant fitness that is not publicly accessible. This dataset is distinct from the training data used by existing predictors and comprises a large number of missense variants. For the HMBS challenge in CAGI6, there are, on average, 17 missense mutations in each position, nearly harboring all the possible missense mutations for each position. By doing so, it fully challenges the predictive capabilities of existing models in determining the functional effects of new variants. Although employing such a dataset can avoid significant data overlapping with training dataset predictors used and thus avoid over-optimistic evaluation, the yeast system may also bring other disadvantages. Due to the disparity between humans and yeast, the protein properties may still be quite different between yeast and humans. For example, in the calmodulin challenge of CAGI5, the budding yeast, *Saccharomyces cerevisiae*, can survive with all EF-hands ablated although CALM1 is essential for yeast (Geiser et al. 1991) and predictors are most inconsistent with experiment scores around calcium binding sites, suggesting possible limitations of the map derived from this model system (Zhang et al. 2019) and yet the map is still useful as evidence for and against pathogenicity (Weile et al. 2017).

In addition, yeast only has around 6000 proteins, while the number of human proteins is more than 22000. Although yeast and human share a considerable number of orthologs and biological pathways, most human proteins lack yeast counterparts which suggests the lack of a yeast-based

complementation assay with which to assess human variants in these proteins. Even where a complementation assay exists, interactions that the complementing human protein might have in human cells may not exist in yeast. Thus missense mutations affecting those interactions may not show severe effects on yeast growth. However, they may severely affect protein functions in human.

Therefore, considering the predictor performance and the characteristics of the yeast complementation assay, it is recommended that future challenges involving the yeast assay as an evaluation dataset focus on protein targets that meet the following criteria: 1) Has a strong phenotype that is suitable for selection (e.g. growth or fluorescence reporter); 2) Demonstrate a high degree of similarity in protein function and properties between yeast and humans, with all functional regions in the human protein being also crucial for optimal yeast protein functioning; 3) Prioritize proteins with fewer interactions or those with interacting partners that have counterparts in yeast and share similar interacting interfaces. Notwithstanding, where many pathogenic and benign human variants are known, a yeast or any other functional assay may be considered empirically validated as accurate if it is able to accurately distinguish pathogenic from benign variation (Brnich et al. 2019; van Loggerenberg et al. 2023).

Participants applied deep learning methods for the first time in CAGI

With the remarkable success of Alphafold, "deep learning" has gained increasingly widespread recognition in the field. In this challenge, several teams, namely team 1, team 3, team 5, and team 10, directly or indirectly incorporated deep learning methods into several to all of their approaches. Additionally, team 9 also employed neural networks, albeit with a shallower architecture. However, no groundbreaking advancements were observed, akin to the Alphafold breakthrough in structure prediction. Team 3 also did not demonstrate a better performance compared to the other teams. Additionally, the high correlation between methods with and without the application of deep learning, as well as the strong correlations between conservation scores and predictions, suggest that amino acid conservation and frequency in each position may be the most important features captured by both types of methods. Consequently, the effective construction and analysis of multiple sequence alignments are crucial for accurate predictions. Studies have demonstrated that deep multiple sequence alignments can improve protein structure predictions by approximately 22%. This is also exemplified in the HMBS challenge, where predictors generally exhibited reduced performance in domain 3, which had a more shallow sequence alignment. In addition, one potential improvement lies in devising methods to derive overall statistics from alignments while taking into account the precise sequences and unique properties of the target proteins, especially when regions with a more shallow alignment depth.

The improvement of predictors compared with previous CAGI challenges

As assessors for CAGI4, CAGI5 and CAGI6 (this round), we noticed that the performance of predictors became comparable to CAIG4, higher than CAGI5 with median Kendall's tau correlation coefficient are 0.26 in CAIG4, 0.15 in CAGI5 and 0.25 in CAGI6 while top-performing predictors are 0.34, 0.17 and 0.31 for CAGI4, CAGI5 and CAGI6, respectively. One particularly exciting development is that several predictors have demonstrated superior performance compared to a baseline predictor based solely on

the frequency of amino acids in the sequence alignment. In CAGI4, only one group surpassed the performance of the baseline predictor, whereas in CAGI5, the baseline predictor itself performed the best. However, in the CAGI6, several teams (Team 5, Team 9, Team 10, and Team 1 if they do not reverse the score scale) have surpassed the performance of the dummy predictors, indicating substantial progress in the field. Furthermore, the top-performing predictors in CAGI6 have shown significantly improved performance compared to previous methods like PolyPhen, which was developed around a decade ago to predict the effects of missense mutations. This indicates advancements in predicting the impact of missense mutations and showcases the evolving capabilities of the top-performing predictors.

Overall, the performance of predictors in the CAGI challenges has shown promising advancements and highlights the ongoing progress in this field.

Methods

Positive control and the baseline predictor

As in CAGI4 (Zhang et al. 2017) and CAGI5 (Zhang et al. 2019), we defined a positive control and a dummy predictor serving as crucial reference points just as a marathon competition has a distinct start line and finish line. The positive control consists of fitness scores for each variant randomly drawn from an assumed Gaussian distribution with the given fitness score as the mean and the experimental standard error as the standard deviation. The baseline predictor was based on the frequency of amino acids at each position in an HMBS multiple sequence alignment (MSA). About 2360 ortholog/inparalog sequences of HMBS were extracted from orthoDB at the metazoa level and were aligned using Promals3D (Pei et al. 2008). The original predicted score for each variant was calculated using the following formula:

$$ln\frac{Q_m}{P_m} - ln\frac{Q_w}{P_w}$$

In this formula, Q_m denotes the estimated probability of the amino acid variant (mutated) occurring at the position where the mutation is located within the alignment, while Q_w is the estimated probability of the original (wild-type) amino acid at the same position. And P_m and P_w are the Robinson-Robinson background frequencies for the mutated amino acid and the wild-type amino acid, respectively. The original predicted scores were normalized according to the distribution of experimental fitness scores.

Quantile transformation of original predictions

Although the distribution of experimental fitness scores was provided, most participants did not calibrate their predictions using this information. Therefore, it was necessary to normalize the predictions in order to facilitate a fair and meaningful comparison among predictors, particularly for numerical assessment. To achieve this, we conducted quantile transformation on both the original predictions from participants and our baseline predictor. To accommodate the requirement that predictors cannot predict negative values, any negative competitive growth scores were adjusted to 0 prior to the transformation. The variants were then ranked based on their predicted values, and each variant was assigned the experimental score corresponding to its rank. In cases where multiple mutants were predicted to have the same rank, the assigned experimental scores were averaged to yield the final transformed predictions.

Measures for prediction assessment

Each predictor was evaluated by their ability (1) to classify variants into categories such as deleterious and non-deleterious variants (classification), (2) to rank variants by their impacts on yeast fitness (ordinal association), and (3) to predict experimental fitness scores (numeric comparison). For the assessment, variants were assigned to the following categories by their experimental fitness score: less than 0.3 for deleterious, between 0.3 and 0.8 for intermediate, and from 0.8 to 1.36 for wild type. Table 2 summarizes all scores used for the evaluation. One important aspect to note is that if the original root mean square deviation (RMSD) based on the predicted values from Team 3 exceeds a certain threshold, which is 1.05 times the maximum RMSD among all other predictors, due to the presence of very large numbers in their predictions, we replaced it with 1.05 times the maximum RMSD among all other predictors.

Evaluation of overall performance and its statistical significance

Four of the measures listed in Table 2 (i.e. the three ordinal associations and the AUC) are purely based on rank and are not sensitive to the distribution of numeric values. Five others depend on the distribution of numeric values and thus were calculated with both original and quantile-transformed predictions. For each measure, we transformed the original scores to Z scores, and positive control and baseline predictor were excluded from the calculation of mean and standard deviation of original scores to avoid their influence on the score distribution. The average Z scores of the rank-based, original-value-based, and transformed-value-based measures were computed and summed up to be the final score to assess the performance of each subset.

To take experimental errors into consideration, we assumed that the fitness score for each variant can be randomly drawn from a Gaussian distribution defined by the reported fitness score and the standard error. We simulated 50 datasets using the above method. Then, we performed bootstrap resampling on each simulated dataset 100 times and thus generated 5000 mock datasets. We obtained the distribution of ranks for each group on 5000 mock datasets.

Identification of well-/poorly predicted mutations

We calculated median difference between top-performing predictions and experimental scores for mutations at each position. The conservation index was calculated by Al2CO (Pei and Grishin 2001) using multiple sequence alignment from orthologs of HEM3 with allowing gap ratio up to 0.8. We defined the positions with conservation index <= -0.95 as unconserved positions while conservation index >=1.42 as conserved positions.

Acknowledgment

The CAGI experiment is supported by NIH U24 HG007346. This work was supported by the following grants: NSF (DBI) 2224128 (to N.V.G.), NIH GM127390 (to N.V.G.), Welch Foundation I-2095-20220331 (to Q.C.), and I-1505 (to N.V.G.). J.Z. is supported by the Cancer Prevention and Research Institute of Texas training grant RP210041. Q.C. is a Southwestern Medical Foundation Scholar. M.J. and Y.S.S. are supported by NIH R35-GM134922. Y.F.S and Y.S. are supported by NIH/NIGMS R35GM124952. EC and MPT acknowledge funding from the Italian Ministry of Education, University, and Research (MIUR-PRIN-201744NR8S).

Legends

Table 1. A brief summary of methods employed by each team.

Table 2. Metrics for evaluating performances for predictors.

Table 3. Assessment of predictors.

Fig 1. **Distributions of experimental fitness scores and predicted scores.** (A) Histogram showing the distribution of experimental fitness scores for nonsense and synonymous mutations (left) and missense mutations (right); (B) histograms of predicted scores from a selected submission from each participating team. The Y-axis represents the proportion of mutations, while the X-axis represents experimental scores in panels (A) and (B)

Fig 2. **Performance assessment of predictors.** (A) Receiver Operating Characteristic (ROC) curves for predicting deleterious mutations; (B) Head-to-head comparison matrix of predictors, with colors indicating the number of datasets in which one predictor (row) outperforms another (column); (C) Boxplot of the distribution of ranks for predictors in simulated datasets. The box edges represent the first and third quartiles of the ranks, the line inside the box denotes the median rank, whiskers extend to 1.5 times the interquartile range from the box edges, and circles represent outliers beyond 1.5 times the interquartile range.

Fig 3. **Effects of mutations on functional loops were poorly predicted by top-performing predictors.** (A) Heatmap of the median differences between experimental scores and those of the top-performing predictors at each position, with blue indicating lower and red indicating higher differences; (B) Structural representation of HEM3 (PDB ID: 5m6r, chain A) highlighting the active-site loop, cofactor-binding loop, insertion region, and residues 354 to 356 in red. ES2 and the phosphate group are displayed as spheres; (C) Distributions of experimental scores (blue) and predicted scores from submission 5_1 (green) within the active-site loop, cofactor-binding loop, and insertion region.

Fig 4. **Correlation among predictors and the role of conservation in prediction.** (A) A heatmap displaying absolute Kendall's tau correlation coefficients between predictors. The absolute correlation coefficients are color-coded, with blue indicating lower and red indicating higher correlation; (B) Scatter plots depicting the correlation between the conservation index and the median of all predicted scores (left) or experimental scores (right) for mutations at each position. The Y-axis represents the median predicted/experimental score, while the X-axis represents the conservation index; (C) Bar graphs showing the ratio of deleterious mutations at conserved positions as indicated by experimental scores and predictors (upper graph) and the ratio of benign mutations at unconserved positions as indicated by experimental scores and predictors (lower graph).

453 **Table 1**

Teams	DL-based	Brief Summary	Used public data from yeast-based functional complementation assay to rescale predictions
Team 1	Yes	A feature extractor using a long short-term memory network and a pathogenicity classifier composed of two fully connected layer	No
Team 2	No	Phylogeny-Aware Amino Acid Substitution Scoring	No
Team 3	Yes	Protein language models (BERT)	No
Team 4	No	Combining of functional annotation analysis (e.g., active sites, post-modification sites and other biologically important sites) from sequences and structures with a consensus of stability predictions from consensus of INPS3D (Savojardo et al. 2016), PoPMuSiC 2.1 (Dehouck et al. 2011) and FoldX (Guerois et al. 2002)	No
Team 5	No, but they use the predictions of other deep learning methods	Ensemble of ordinary linear regression models combining sequence features and predictions from one or several of REVEL (Ioannidis et al. 2016), DeepSequence (Riesselman et al. 2018) and ESM-1v (Meier et al. 2021).	Yes
Team 6	No, but predictions from MetaRNN, a deep learning method, was used in their models.	Random forest models to combine several scores such as MetaSVM (Kim et al. 2017), MetaLR (Liu et al. 2020), MetaRNN (Li et al. 2022), REVEL, MPC (Kaitlin et al. 2017), PROVEAN (Choi and Chan 2015), GERP RS (Cooper et al. 2005), phyloP100way_vertebrate, GM12878_fitCons (Gulko et al. 2015) and H1.hESC_fitCons (Gulko et al. 2015).	Yes
Team 7	No	Evolutionary Action (Katsonis and Lichtarge 2014)	No
Team 8	No	Weighted changes of root mean square fluctuation between wild type and variants simulated by molecular dynamics	No
Team 9	No, but shallow artificial and probabilistic neural networks	SNPMuSiC (Ancien et al. 2018), PoPMuSiC (Dehouck et al. 2011), and FiTMuSiC (Tsishyn et al. 2024), a new linear regression model incorporating multiple predictions (Matsvei et al. 2023) including PoPMuSiC, Maestro (Laimer et al. 2015), EVCoupling (Hopf et al. 2019), PROVEAN (Choi and Chan 2015) and DEOGEN2 (Raimondi et al. 2017)	Yes, but only for submission 2
Team 10	Yes	ELASPIC2 (Strokach et al. 2021), ProteinSolver (Strokach et al. 2020), ProtBert (Elnaggar et al. 2022) and Rosetta's cartesian_ddg protocol (Park et al. 2016)	No
Team 11	No	PhD-SNPg (Capriotti and Fariselli 2023) , SNPs-and-GO (Capriotti et al. 2017) and PhyloP100	No

454

455

456

457

458

Table 2

Classification	
Area Under ROC	$\frac{1}{PN}\sum_{j=1}^{N}(R_j - j)\,P$, P: **number of true deleterious mutations based on experimental scores**; *N*: number of true non-deleterious mutations. All mutations are ranked by the predicted growth score. $(R_j - j)$ is the count of true deleterious mutations that are ranked no worse than the j^{th} true non-deleterious mutation. Each true deleterious mutation ranked the same as the j^{th} true non-deleterious mutation is counted as 0.5.
MCC	$(TP_i \times TN_i - FP_i \times FN_i)/\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}$, *i* ∈ (deleterious, intermediate, benig); TP: true positive; TN: true negative; FP: false positive; FN: false negative.
F1	$(2 \cdot precision \cdot recall)/(precision + recall)$, $precision = TP/(TP + FP); recall = TP/(TP + FN)$ TP: true positive; TN: true negative; FP: false positive; FN: false negative. Mutations were divided into three categories: deleterious, intermediate, and benign. We used f1_score from the sklearn.metrics package with the 'micro' for averaging
Ordinal association	
Kendall tau-b rank correlation	$(n_c - n_d)/\sqrt{(n_0 - n_1)(n_0 - n_2)}$, $n_0 = n(n-1)/2\,; n_1 = \sum_k t_k(t_k - 1)/2\,; n_2 = \sum_j u_j(u_j - 1)/2\,;$ n_c, the number of concordant pairs; n_d, the number of discordant pairs; n, the total number of pairs; t_k, number of values in the k^{th} group of ties by predictions; u_j, number of values in the j^{th} group of ties by experimental scores.
Spearman's rank correlation	$cov(R_{pred}, R_{exp})/\sigma_{R_{pred}}\sigma_{R_{exp}}$ $cov(R_{pred}, R_{exp})$, covariance between predicted and experimental ranks of mutants; $\sigma_{R_{pred}}$ and $\sigma_{R_{exp}}$, standard deviations of predicted and experimental ranks, respectively. Ties were randomly assigned distinct ranks first and then the average of these ranks were assigned to each of them.
Numeric comparison	
Pearson's correlation	$cov(pred, exp)/\sigma_{pred}\sigma_{exp}$, $cov(pred, exp)$, the covariance between predictions and experimental scores; σ_{pred}, the standard deviation of predictions; σ_{exp}, standard deviation of experimental scores
RMSD	$\sqrt{\frac{1}{N}\sum_{j=1}^{N}(pred_j - exp_j)^2}$ *N*, the size of a dataset; $pred_j$, j^{th} predictions; exp_j, j^{th} experimental scores
Value agreement test (value_diff)	$\sum Ci$ *C* is the percentage of mutants with the difference between the predicted and experimental growth scores below a certain cutoff *i*. The cutoffs are taken from 0 to 1 with an incremental of 0.01. The area under the curve was used as a measurement.

Table 3

group	Rank-based				original prediction						rescaled prediction					
	tau	spearman	dele_roc	wild_roc	rmsd	pearson	value_diff	mcc_dele	mcc_wild	f1	rmsd	pearson	value_diff	mcc_dele	mcc_wild	f1
1_1	-0.27	-0.38	0.29	0.31	0.70	-0.38	0.44	-0.20	-0.29	0.23	0.69	-0.38	0.44	-0.21	-0.30	0.24
1_2	-0.28	-0.39	0.29	0.31	0.70	-0.38	0.44	-0.20	-0.30	0.23	0.69	-0.38	0.44	-0.21	-0.30	0.23
1_3	-0.29	-0.42	0.27	0.30	0.71	-0.42	0.43	-0.23	-0.30	0.22	0.70	-0.41	0.43	-0.25	-0.30	0.22
1_4	-0.29	-0.41	0.28	0.31	0.70	-0.41	0.43	-0.24	-0.29	0.22	0.70	-0.41	0.43	-0.27	-0.28	0.22
2_1	0.16	0.22	0.63	0.60	0.51	0.21	0.61	0.20	0.08	0.40	0.52	0.22	0.61	0.20	0.12	0.40
3_1	0.12	0.17	0.57	0.60	0.93	0.02	0.46	0.02	0.07	0.36	0.54	0.14	0.57	0.09	0.08	0.37
3_2	0.11	0.15	0.56	0.59	0.93	0.01	0.45	0.03	0.07	0.36	0.55	0.12	0.57	0.07	0.07	0.36
3_3	0.13	0.19	0.58	0.61	0.88	0.00	0.46	0.03	0.07	0.36	0.53	0.17	0.58	0.15	0.08	0.39
3_4	0.15	0.22	0.60	0.62	0.87	0.01	0.47	0.05	0.08	0.37	0.52	0.20	0.59	0.17	0.10	0.40
3_5	0.15	0.21	0.60	0.62	0.86	0.02	0.48	0.09	0.10	0.38	0.53	0.19	0.59	0.14	0.11	0.39
3_6	0.12	0.18	0.58	0.60	0.89	0.01	0.48	0.07	0.08	0.37	0.54	0.16	0.58	0.10	0.11	0.38
4_1	0.21	0.27	0.64	0.63	0.53	0.27	0.61	0.23	0.20	0.44	0.49	0.27	0.63	0.23	0.20	0.44
5_1	0.30	0.42	0.72	0.71	0.43	0.43	0.68	0.37	0.25	0.50	0.44	0.42	0.67	0.38	0.26	0.50
5_2	0.29	0.41	0.71	0.70	0.44	0.41	0.68	0.37	0.24	0.49	0.45	0.41	0.67	0.37	0.24	0.49
5_3	0.28	0.38	0.70	0.69	0.44	0.39	0.67	0.35	0.21	0.48	0.46	0.39	0.66	0.35	0.22	0.48
5_4	0.25	0.35	0.68	0.68	0.46	0.35	0.66	0.28	0.22	0.45	0.47	0.35	0.64	0.27	0.22	0.45
5_5	0.29	0.40	0.71	0.70	0.44	0.40	0.67	0.36	0.23	0.48	0.45	0.40	0.66	0.37	0.23	0.49
6_1	0.24	0.35	0.68	0.68	0.48	0.35	0.64	0.27	0.22	0.45	0.47	0.35	0.64	0.27	0.23	0.45
6_2	0.24	0.34	0.68	0.67	0.47	0.35	0.64	0.25	0.23	0.45	0.47	0.34	0.64	0.25	0.23	0.45
6_3	0.22	0.33	0.67	0.67	0.48	0.34	0.63	0.25	0.25	0.45	0.48	0.34	0.63	0.24	0.25	0.45
6_4	0.25	0.36	0.69	0.68	0.47	0.36	0.64	0.27	0.24	0.46	0.47	0.36	0.64	0.27	0.24	0.46
6_5	0.23	0.33	0.67	0.67	0.48	0.34	0.63	0.23	0.22	0.44	0.48	0.34	0.64	0.23	0.22	0.44
7_1	0.28	0.40	0.72	0.70	0.47	0.40	0.65	0.34	0.23	0.47	0.45	0.40	0.66	0.34	0.24	0.48
7_2	0.27	0.39	0.71	0.69	0.47	0.38	0.65	0.33	0.22	0.48	0.46	0.39	0.66	0.34	0.23	0.48
7_3	0.28	0.40	0.71	0.69	0.46	0.39	0.66	0.34	0.23	0.48	0.46	0.39	0.66	0.34	0.24	0.48
7_4	0.28	0.41	0.72	0.70	0.46	0.40	0.66	0.35	0.23	0.48	0.45	0.40	0.66	0.35	0.23	0.48
7_5	0.26	0.37	0.70	0.68	0.48	0.36	0.64	0.33	0.20	0.47	0.47	0.37	0.65	0.33	0.20	0.47
7_6	0.28	0.40	0.72	0.70	0.46	0.40	0.66	0.34	0.23	0.47	0.45	0.40	0.66	0.34	0.24	0.48
8_1	0.04	0.06	0.52	0.53	0.58	0.05	0.53	-0.01	0.06	0.34	0.57	0.05	0.54	-0.01	0.06	0.34
8_2	0.04	0.05	0.52	0.53	0.62	0.06	0.51	0.01	0.03	0.34	0.57	0.05	0.55	0.01	0.04	0.34
8_3	0.04	0.06	0.52	0.53	0.58	0.05	0.53	-0.01	0.06	0.34	0.57	0.05	0.54	-0.01	0.06	0.34
8_4	0.04	0.05	0.52	0.53	0.62	0.06	0.51	0.01	0.03	0.34	0.57	0.05	0.55	0.01	0.04	0.34
8_5	0.04	0.06	0.52	0.53	0.58	0.05	0.54	-0.01	0.06	0.34	0.57	0.05	0.54	-0.01	0.06	0.33
8_6	0.04	0.05	0.52	0.53	0.59	0.05	0.54	0.01	0.04	0.34	0.57	0.05	0.55	0.01	0.04	0.34
9_1	0.27	0.39	0.71	0.69	0.43	0.38	0.66	0.32	0.15	0.44	0.46	0.39	0.66	0.32	0.26	0.48
9_2	0.30	0.43	0.73	0.71	0.39	0.42	0.67	0.37	0.10	0.43	0.44	0.43	0.67	0.37	0.26	0.49
9_3	0.15	0.22	0.62	0.62	0.44	0.24	0.65	0.21	0.10	0.39	0.52	0.21	0.60	0.19	0.14	0.41
10_1	0.28	0.41	0.72	0.70	0.43	0.37	0.66	0.33	0.21	0.45	0.45	0.40	0.66	0.35	0.25	0.48
10_2	0.15	0.22	0.63	0.61	0.48	0.22	0.62	0.16	0.16	0.40	0.52	0.22	0.60	0.17	0.16	0.42
10_3	0.20	0.28	0.64	0.64	0.51	0.26	0.61	0.20	0.18	0.43	0.50	0.27	0.62	0.21	0.18	0.43
10_4	0.21	0.30	0.67	0.65	0.47	0.32	0.64	0.28	0.21	0.43	0.49	0.29	0.63	0.27	0.16	0.44
10_5	0.31	0.45	0.75	0.72	0.51	0.36	0.63	0.32	0.26	0.48	0.44	0.45	0.68	0.41	0.26	0.51
11_1	0.11	0.16	0.58	0.58	0.57	0.15	0.56	0.14	0.09	0.39	0.54	0.17	0.59	0.14	0.10	0.39
11_2	0.11	0.16	0.58	0.58	0.58	0.15	0.56	0.14	0.10	0.40	0.54	0.17	0.59	0.14	0.10	0.39
11_3	0.19	0.25	0.62	0.60	0.43	0.20	0.63	0.00	0.15	0.33	0.47	0.25	0.61	0.21	0.16	0.43
11_4	0.19	0.25	0.62	0.60	0.45	0.20	0.62	0.00	0.16	0.34	0.47	0.25	0.61	0.21	0.16	0.43
11_5	0.26	0.36	0.69	0.68	0.51	0.35	0.62	0.28	0.22	0.46	0.47	0.36	0.65	0.29	0.23	0.46
11_6	0.26	0.38	0.72	0.68	0.51	0.36	0.62	0.30	0.23	0.46	0.46	0.38	0.65	0.31	0.23	0.46
11_7	0.27	0.38	0.70	0.68	0.52	0.36	0.62	0.31	0.24	0.47	0.46	0.38	0.65	0.31	0.25	0.47
11_8	0.29	0.41	0.73	0.70	0.49	0.40	0.63	0.34	0.25	0.48	0.45	0.42	0.66	0.35	0.25	0.48
polyphen	0.20	0.29	0.64	0.65	0.51	0.28	0.61	0.20	0.20	0.42	0.50	0.28	0.62	0.20	0.19	0.42
baseline	0.28	0.40	0.71	0.69	0.47	0.40	0.66	0.35	0.25	0.49	0.45	0.40	0.67	0.35	0.24	0.49
positive	0.82	0.95	0.98	0.98	0.13	0.95	0.92	0.87	0.85	0.87	0.12	0.96	0.92	0.87	0.85	0.87

Figure 1

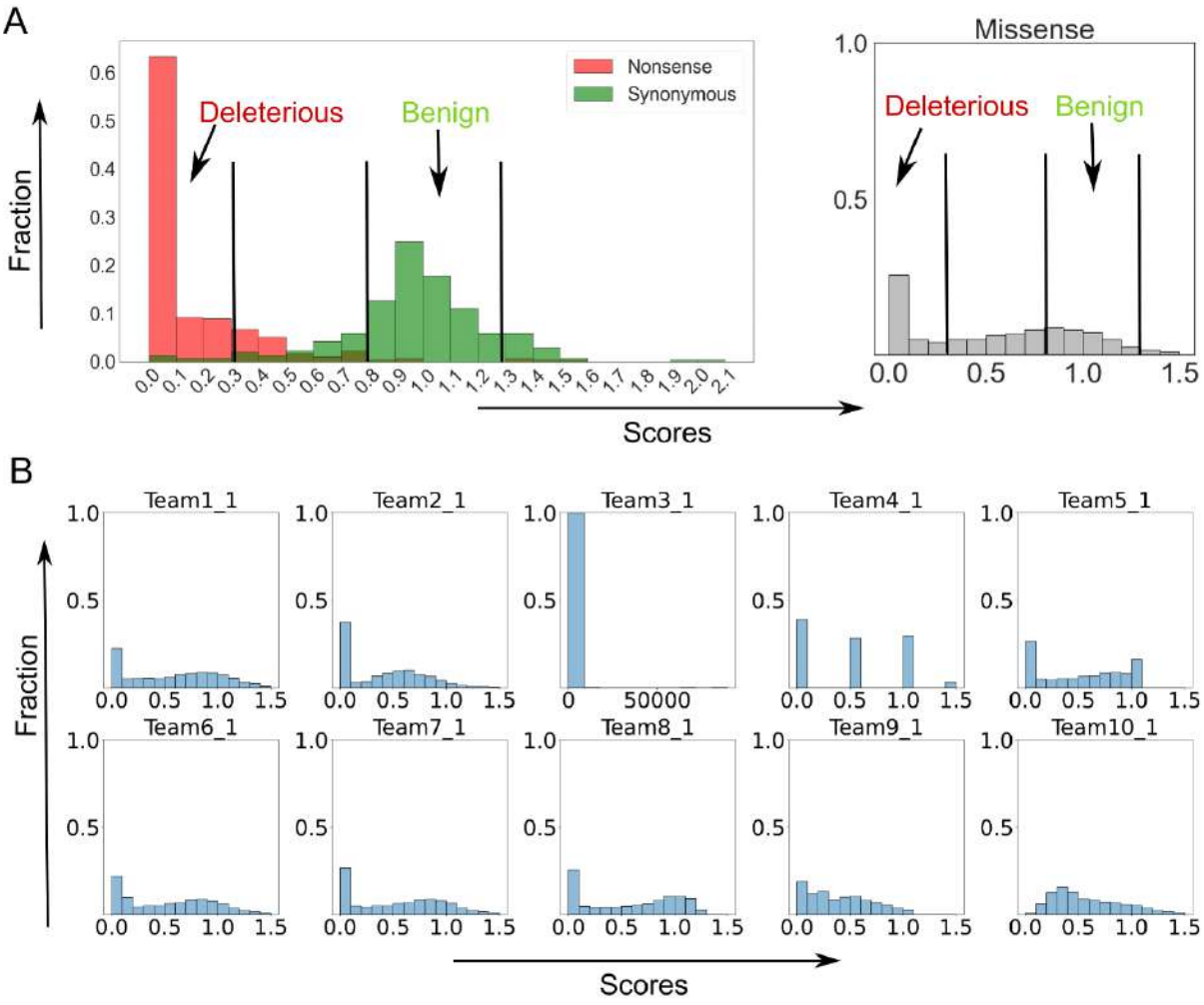

A
Nonsense
Synonymous
Deleterious
Benign
Fraction
Scores
Missense
Deleterious
Benign
B
Team1_1
Team2_1
Team3_1
Team4_1
Team5_1
Team6_1
Team7_1
Team8_1
Team9_1
Team10_1
Fraction
Scores

Figure 2

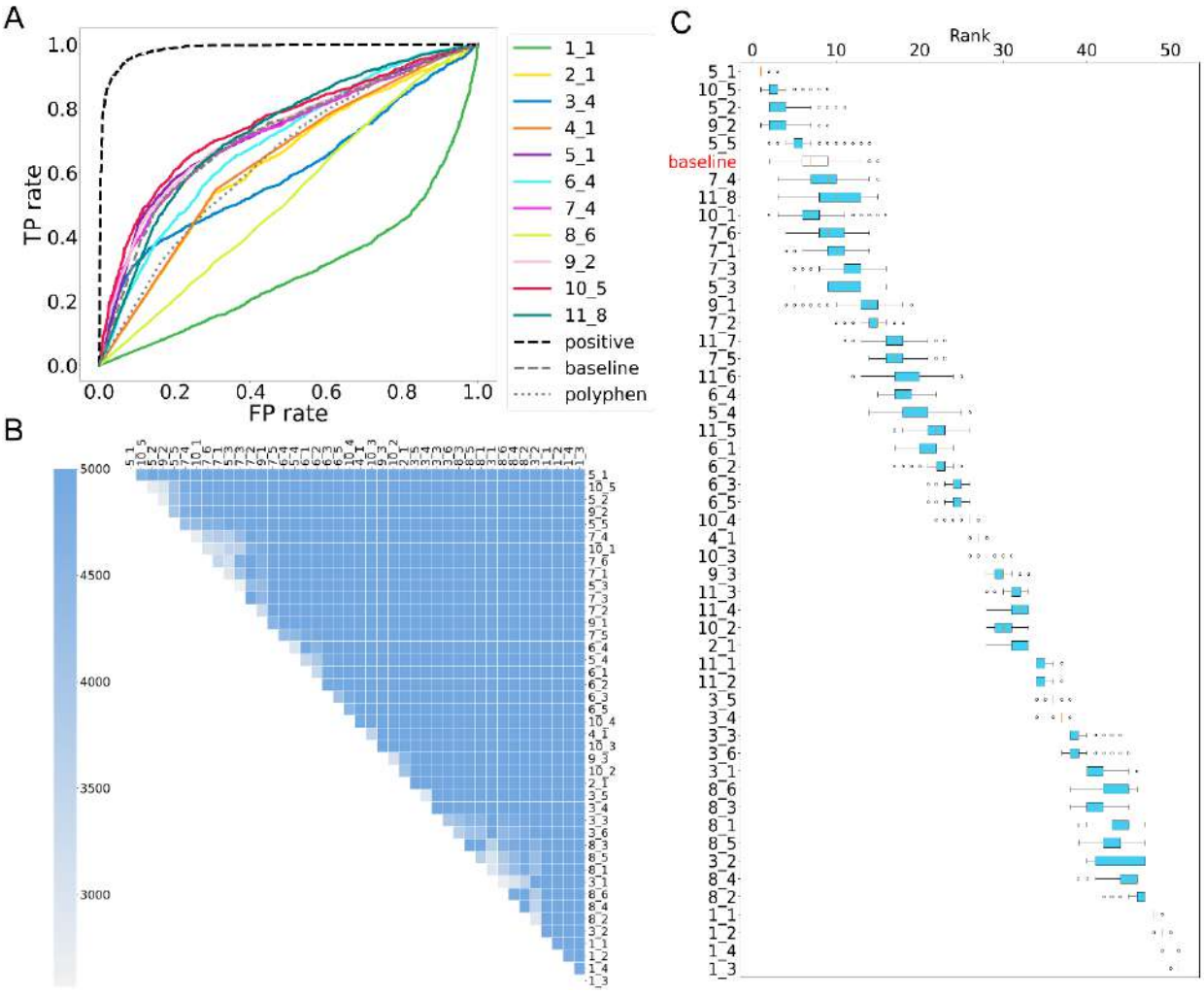

A
TP rate
FP rate
1_1
2_1
3_4
4_1
5_1
6_4
7_4
8_6
9_2
10_5
11_8
positive
baseline
polyphen
B
C
Rank
baseline

Figure 3

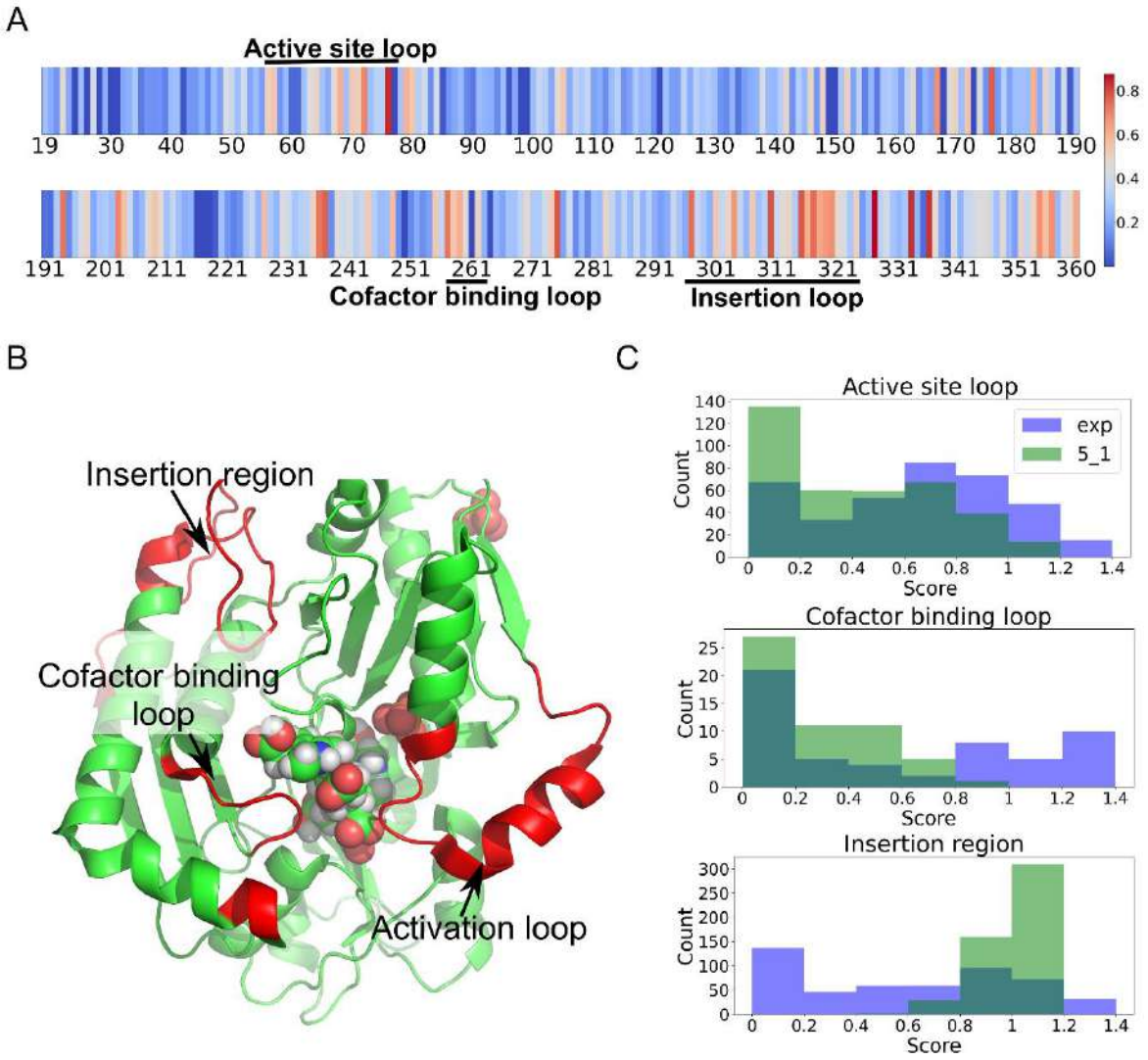
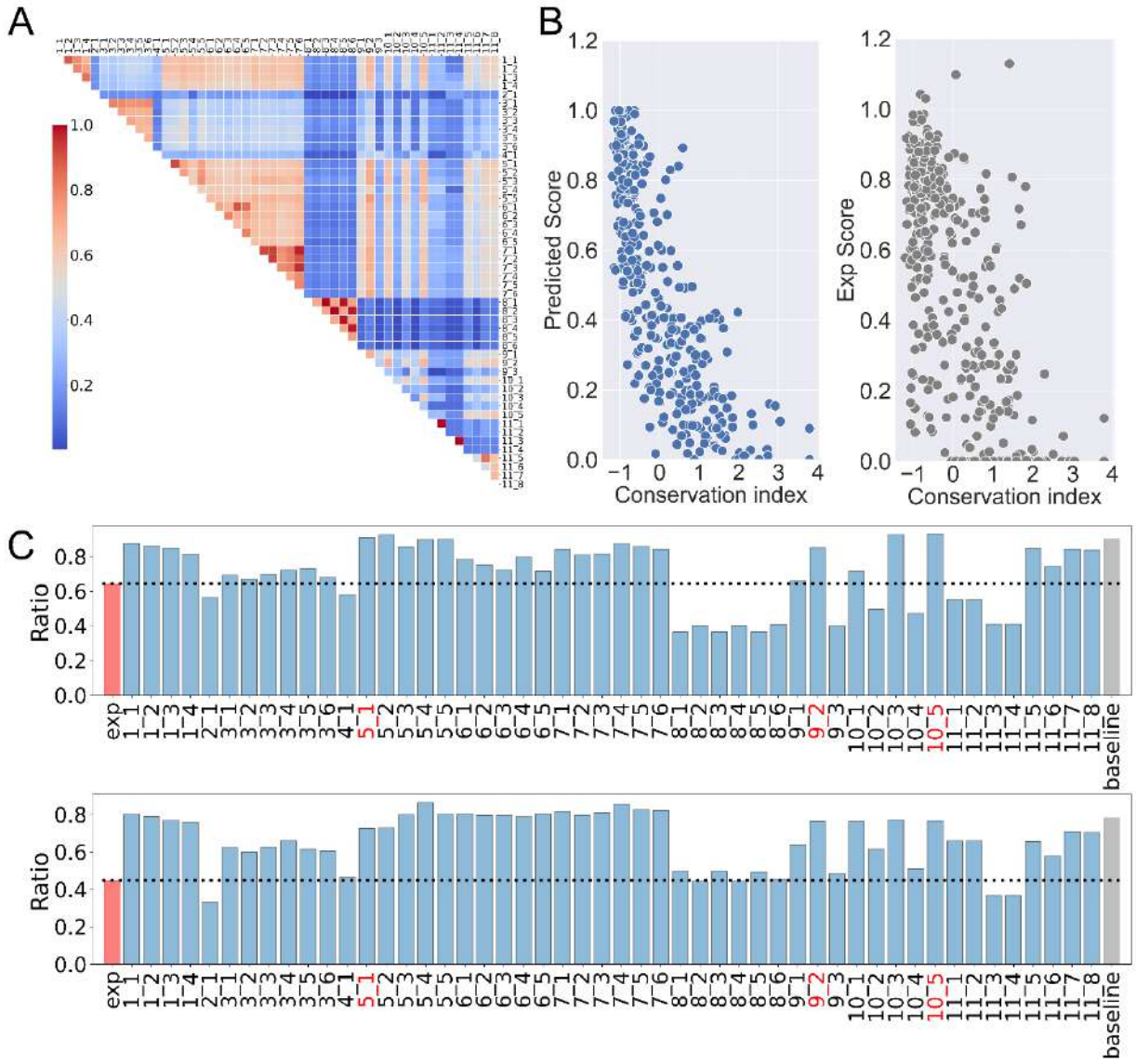

A
Active site loop
19 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
0.8
0.6
0.4
0.2
191 201 211 221 231 241 251 261 271 281 291 301 311 321 331 341 351 360
Cofactor binding loop
Insertion loop
B
Insertion region
Cofactor binding loop
Activation loop
C
Active site loop
exp
5_1
Count
Score
Cofactor binding loop
Count
Score
Insertion region
Count
Score

Figure 4


A
B
Predicted Score
Exp Score
Conservation index
Conservation index
C
Ratio
Ratio
exp
baseline

References

Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7: Unit7 20. doi: 10.1002/0471142905.hg0720s76

Ancien F, Pucci F, Godfroid M, Rooman M (2018) Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep 8: 4480. doi: 10.1038/s41598-018-22531-2

Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38: 2102-2110. doi: 10.1093/bioinformatics/btac020

Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, Kanavy DM, Luo X, McNulty SM, Starita LM, Tavtigian SV, Wright MW, Harrison SM, Biesecker LG, Berg JS, Clinical Genome Resource Sequence Variant Interpretation Working G (2019) Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med 12: 3. doi: 10.1186/s13073-019-0690-2

Bustad HJ, Kallio JP, Laitaoja M, Toska K, Kursula I, Martinez A, Janis J (2021) Characterization of porphobilinogen deaminase mutants reveals that arginine-173 is crucial for polypyrrole elongation mechanism. iScience 24: 102152. doi: 10.1016/j.isci.2021.102152

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30: 1237-44. doi: 10.1002/humu.21047

Capriotti E, Altman RB (2011) Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics 12 Suppl 4: S3. doi: 10.1186/1471-2105-12-S4-S3

Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22: 2729-34. doi: 10.1093/bioinformatics/btl423

Capriotti E, Fariselli P (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res 45: W247-W252. doi: 10.1093/nar/gkx369

Capriotti E, Fariselli P (2023) PhD-SNPg: updating a webserver and lightweight tool for scoring nucleotide variants. Nucleic Acids Res. doi: 10.1093/nar/gkad455

Capriotti E, Martelli PL, Fariselli P, Casadio R (2017) Blind prediction of deleterious amino acid variations with SNPs&GO. Hum Mutat 38: 1064-1071. doi: 10.1002/humu.23179

Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31: 2745-7. doi: 10.1093/bioinformatics/btv195

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-13. doi: 10.1101/gr.3577405

Critical Assessment of Genome Interpretation C (2024) CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. Genome Biol 25: 53. doi: 10.1186/s13059-023-03113-6

Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25: 2537-43. doi: 10.1093/bioinformatics/btp445

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 12: 151. doi: 10.1186/1471-2105-12-151

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell 44: 7112-7127. doi: 10.1109/TPAMI.2021.3095381

Geiser JR, van Tuinen D, Brockerhoff SE, Neff MM, Davis TN (1991) Can calmodulin function without binding calcium? Cell 65: 949-59. doi: 10.1016/0092-8674(91)90547-c

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526: 68-74. doi: 10.1038/nature15393

Gill R, Kolstoe SE, Mohammed F, Al DBA, Mosely JE, Sarwar M, Cooper JB, Wood SP, Shoolingin-Jordan PM (2009) Structure of human porphobilinogen deaminase at 2.8 A: the molecular basis of acute intermittent porphyria. Biochem J 420: 17-25. doi: 10.1042/BJ20082077

Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369-87. doi: 10.1016/S0022-2836(02)00442-4

Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 47: 276-83. doi: 10.1038/ng.3196

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514-7. doi: 10.1093/nar/gki033

Hopf TA, Green AG, Schubert B, Mersmann S, Scharfe CPI, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, Kang C, Sheridan R, Draizen EJ, Dallago C, Sander C, Marks DS (2019) The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 35: 1582-1584. doi: 10.1093/bioinformatics/bty862

International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolas P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clement B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayes M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, Lopez-Otin C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, et al. (2010) International network of cancer genome projects. Nature 464: 993-8. doi: 10.1038/nature08987

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 99: 877-885. doi: 10.1016/j.ajhg.2016.08.016

606 Jagota M, Ye C, Albors C, Rastogi R, Koehl A, Ioannidis N, Song YS (2023) Cross-protein transfer learning
607 substantially improves disease variant prediction. Genome Biol 24: 182. doi: 10.1186/s13059-
608 023-03024-6
609 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A,
610 Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S,
611 Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M,
612 Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D
613 (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596: 583-589. doi:
614 10.1038/s41586-021-03819-2
615 Kaitlin ES, Jack AK, Konrad JK, Anne HOD-L, Emma P-H, Daniel GM, Benjamin MN, Mark JD (2017)
616 Regional missense constraint improves variant deleteriousness prediction. bioRxiv: 148353. doi:
617 10.1101/148353
618 Katsonis P, Lichtarge O (2014) A formal perturbation equation between genotype and phenotype
619 determines the Evolutionary Action of protein-coding variations on fitness. Genome Res 24:
620 2050-8. doi: 10.1101/gr.176214.114
621 Katsonis P, Lichtarge O (2017) Objective assessment of the evolutionary action equation for the fitness
622 effect of missense mutations across CAGI-blinded contests. Hum Mutat 38: 1072-1084. doi:
623 10.1002/humu.23266
624 Katsonis P, Lichtarge O (2019) CAGI5: Objective performance assessments of predictions based on the
625 Evolutionary Action equation. Hum Mutat 40: 1436-1454. doi: 10.1002/humu.23873
626 Kauppinen R, von und zu Fraunberg M (2002) Molecular and biochemical studies of acute intermittent
627 porphyria in 196 patients and their families. Clin Chem 48: 1891-900.
628 Kim S, Jhong JH, Lee J, Koo JY (2017) Meta-analytic support vector machine for integrating multiple
629 omics data. BioData Min 10: 2. doi: 10.1186/s13040-017-0126-8
630 Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2021) Critical assessment of methods of protein
631 structure prediction (CASP)-Round XIV. Proteins 89: 1607-1617. doi: 10.1002/prot.26237
632 Kuru N, Dereli O, Akkoyun E, Bircan A, Tastan O, Adebali O (2022) PHACT: Phylogeny-Aware Computing
633 of Tolerance for Missense Mutations. Mol Biol Evol 39. doi: 10.1093/molbev/msac114
634 Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P (2015) MAESTRO--multi agent stability prediction
635 upon point mutations. BMC Bioinformatics 16: 116. doi: 10.1186/s12859-015-0548-6
636 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh
637 W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P,
638 McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M,
639 Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D,
640 Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A,
641 Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S,
642 Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S,
643 Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier
644 LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe
645 SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson
646 DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T,
647 Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial
648 sequencing and analysis of the human genome. Nature 409: 860-921. doi: 10.1038/35057062
649 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public
650 archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:
651 D980-5. doi: 10.1093/nar/gkt1113
652 Lenglet H, Schmitt C, Grange T, Manceau H, Karboul N, Bouchet-Crivat F, Robreau AM, Nicolas G, Lamoril
653 J, Simonin S, Mirmiran A, Karim Z, Casalino E, Deybach JC, Puy H, Peoc'h K, Gouya L (2018) From

654 a dominant to an oligogenic model of inheritance with environmental modifiers in acute
655 intermittent porphyria. Hum Mol Genet 27: 1164-1173. doi: 10.1093/hmg/ddy030
656 Li C, Zhi D, Wang K, Liu X (2022) MetaRNN: differentiating rare pathogenic and rare benign missense
657 SNVs and InDels using deep learning. Genome Med 14: 115. doi: 10.1186/s13073-022-01120-z
658 Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces
659 common to protein families. J Mol Biol 257: 342-58. doi: 10.1006/jmbi.1996.0167
660 Liu X, Li C, Mou C, Dong Y, Tu Y (2020) dbNSFP v4: a comprehensive database of transcript-specific
661 functional predictions and annotations for human nonsynonymous and splice-site SNVs.
662 Genome Med 12: 103. doi: 10.1186/s13073-020-00803-9
663 Matsvei T, Gabriel C, Pauline H, Jean K, Marianne R, Fabrizio P (2023) FiTMuSiC: Leveraging structural
664 and (co)evolutionary data for protein fitness prediction. bioRxiv: 2023.08.01.551497. doi:
665 10.1101/2023.08.01.551497
666 Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A (2021) Language models enable zero-shot prediction of
667 the effects of mutations on protein function. Advances in neural information processing systems
668 34: 29287-29303.
669 Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863-74. doi:
670 10.1101/gr.176601
671 Park H, Bradley P, Greisen P, Jr., Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F (2016) Simultaneous
672 Optimization of Biomolecular Energy Functions on Features from Small Molecules and
673 Macromolecules. J Chem Theory Comput 12: 6201-6212. doi: 10.1021/acs.jctc.6b00819
674 Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment.
675 Bioinformatics 17: 700-12. doi: 10.1093/bioinformatics/17.8.700
676 Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure
677 alignments. Nucleic Acids Res 36: 2295-300. doi: 10.1093/nar/gkn072
678 Pluta P, Roversi P, Bernardo-Seisdedos G, Rojas AL, Cooper JB, Gu S, Pickersgill RW, Millet O (2018)
679 Structural basis of pyrrole polymerization in human porphobilinogen deaminase. Biochim
680 Biophys Acta Gen Subj 1862: 1948-1955. doi: 10.1016/j.bbagen.2018.06.013
681 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on
682 mammalian phylogenies. Genome Res 20: 110-21. doi: 10.1101/gr.097857.109
683 Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W (2017)
684 DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in
685 human proteins. Nucleic Acids Res 45: W201-W206. doi: 10.1093/nar/gkx390
686 Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture the
687 effects of mutations. Nat Methods 15: 816-822. doi: 10.1038/s41592-018-0138-4
688 Sato H, Sugishima M, Tsukaguchi M, Masuko T, Iijima M, Takano M, Omata Y, Hirabayashi K, Wada K,
689 Hisaeda Y, Yamamoto K (2021) Crystal structures of hydroxymethylbilane synthase complexed
690 with a substrate analog: a single substrate-binding site for four consecutive condensation steps.
691 Biochem J 478: 1023-1042. doi: 10.1042/BCJ20200996
692 Savojardo C, Fariselli P, Martelli PL, Casadio R (2016) INPS-MD: a web server to predict stability of
693 protein variants from sequence and structure. Bioinformatics 32: 2542-4. doi:
694 10.1093/bioinformatics/btw192
695 Schneider-Yin X, Ulbrichova D, Mamet R, Martasek P, Marohnic CC, Goren A, Minder EI, Schoenfeld N
696 (2008) Characterization of two missense variants in the hydroxymethylbilane synthase gene in
697 the Israeli population, which differ in their associations with acute intermittent porphyria. Mol
698 Genet Metab 94: 343-6. doi: 10.1016/j.ymgme.2008.03.001
699 Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online
700 force field. Nucleic Acids Res 33: W382-8. doi: 10.1093/nar/gki387

Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677-9.

Song G, Li Y, Cheng C, Zhao Y, Gao A, Zhang R, Joachimiak A, Shaw N, Liu ZJ (2009) Structural insight into acute intermittent porphyria. FASEB J 23: 396-404. doi: 10.1096/fj.08-115469

Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM (2020) Fast and Flexible Protein Design Using Deep Graph Neural Networks. Cell Syst 11: 402-411 e4. doi: 10.1016/j.cels.2020.08.016

Strokach A, Lu TY, Kim PM (2021) ELASPIC2 (EL2): Combining Contextualized Language Models and Graph Neural Networks to Predict Effects of Mutations. J Mol Biol 433: 166810. doi: 10.1016/j.jmb.2021.166810

Tsishyn M, Cia G, Hermans P, Kwasigroch J, Rooman M, Pucci F (2024) FiTMuSiC: leveraging structural and (co)evolutionary data for protein fitness prediction. Human Genomics 18: 36. doi: 10.1186/s40246-024-00605-9

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereau A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, Project G (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. Bmj-British Medical Journal 361. doi: ARTN k1687

10.1136/bmj.k1687

Ulbrichova D, Schneider-Yin X, Mamet R, Saudek V, Martasek P, Minder EI, Schoenfeld N (2009) Correlation between biochemical findings, structural and enzymatic abnormalities in mutated HMBS identified in six Israeli families with acute intermittent porphyria. Blood Cells Mol Dis 42: 167-73. doi: 10.1016/j.bcmd.2008.11.001

van Loggerenberg W, Sowlati-Hashjin S, Weile J, Hamilton R, Chawla A, Sheykhkarimli D, Gebbia M, Kishore N, Fresard L, Mustajoki S, Pischik E, Di Pierro E, Barbaro M, Floderus Y, Schmitt C, Gouya L, Colavin A, Nussbaum R, Friesema ECH, Kauppinen R, To-Figueras J, Aarsand AK, Desnick RJ, Garton M, Roth FP (2023) Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. Am J Hum Genet 110: 1769-1786. doi: 10.1016/j.ajhg.2023.08.012

Warren van L, Shahin S-H, Jochen W, Rayna H, Aditya C, Marinella G, Nishka K, Laure F, Sami M, Elena P, Elena Di P, Michela B, Ylva F, Caroline S, Laurent G, Alexandre C, Robert N, Edith CHF, Raili K, Jordi T-F, Aasne KA, Robert JD, Michael G, Frederick PR (2023) Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. bioRxiv: 2023.02.06.527353. doi: 10.1101/2023.02.06.527353

Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, Wu Y, Pons C, Wong C, van Lieshout N, Yang F, Tasan M, Tan G, Yang S, Fowler DM, Nussbaum R, Bloom JD, Vidal M, Hill DE, Aloy P, Roth FP (2017) A framework for exhaustively mapping functional missense variants. Mol Syst Biol 13: 957. doi: 10.15252/msb.20177908

Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV (2021) OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Res 49: D389-D393. doi: 10.1093/nar/gkaa1009

Zhang J, Kinch LN, Cong Q, Katsonis P, Lichtarge O, Savojardo C, Babbi G, Martelli PL, Capriotti E, Casadio R, Garg A, Pal D, Weile J, Sun S, Verby M, Roth FP, Grishin NV (2019) Assessing predictions on fitness effects of missense variants in calmodulin. Hum Mutat 40: 1463-1473. doi: 10.1002/humu.23857

Zhang J, Kinch LN, Cong Q, Weile J, Sun S, Cote AG, Roth FP, Grishin NV (2017) Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. Hum Mutat 38: 1051-1063. doi: 10.1002/humu.23293

748