Characterizing disease-associated human proteins without available protein structures or homologues

Neeladri Sen[1], Ivan Anishchenko[2,3], Nicola Bordin[1], Ian Sillitoe[1], Sameer Velankar[4,*], David Baker[2,3,5,*], Christine Orengo[1,*]

[1] Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK
[2] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
[3] Institute for Protein Design, University of Washington, Seattle, WA 98195, USA.
[4] Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[5] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

[*]Corresponding authors

Abstract

Mutations in human proteins lead to diseases. The structure of these proteins can help understand the mechanism of such diseases and develop therapeutics against them. With improved deep learning techniques such as RoseTTAFold and AlphaFold, we can predict the structure of these proteins even in the absence of structural homologues. We modeled and extracted the domains from 553 disease-associated human proteins. We noticed that the model quality was higher and the RMSD lower between AlphaFold and RoseTTAFold models for domains that could be assigned to CATH families as compared to those which could be assigned to Pfam families of unknown structure or could not be assigned to either. We predicted ligand-binding sites, protein-protein interfaces, conserved residues and destabilising effects caused by residue mutations in these predicted structures. We then explored whether the disease-associated mutations were in the proximity of these predicted functional sites or if they destabilized the protein structure based on ddG calculations. We could explain 80% of these disease-associated mutations based on proximity to functional sites or structural destabilization. Usage of models from the two state-of-the-art techniques provide better confidence in our predictions, and we explain 93 additional mutations based on RoseTTAFold models which could not be explained based solely on AlphaFold models.

Introduction

The sequence of amino acids dictates the structure of the protein (Anfinsen, 1973). A mutation in the amino acid sequence can lead to changes in/near catalytic site residues (Coleman et al., 1993; Joshi et al., 2018), ligand binding sites (Lee et al., 1997; Ricatti et al., 2019), protein-protein interface sites (Cheng et al., 2021; Jubb et al., 2017), allosteric sites (Tyukhtenko et al., 2018) among others; modifying the protein or rendering it inactive. Some mutations lead to changes in the structure of the protein (Baiardi et al., 2019; Soto, 2003), which make the protein non-functional.

Single nucleotide polymorphisms (SNPs) are common in genes, some of which lead to disease and are being collated in databases such as dbSNP (Smigielski et al., 2000) and the 1000 genomes project (Fairley et al., 2020). Databases such as ClinVar (Landrum et al., 2020) link mutations in humans with phenotypes. Specific databases such as COSMIC (Forbes et al., 2017), OncoVar (Wang et al., 2021), DoCM (Ainscough et al., 2016) contain information on cancer related mutations. Several other specialized databases containing subsets of these mutations exist, such as KinMutBase (for human disease-associated protein kinase mutations) (Stenberg et al., 1999), ActiveDriverDB (mutations at protein post-translational modification sites) (Krassowski et al., 2018) among others. UniProt-KB (The UniProt Consortium et al., 2021) and PDBe-KB (PDBe-KB consortium et al., 2020) also contain experimental and computational annotations of mutations and associated phenotypes. Humsavar (The UniProt Consortium et al., 2021) is a database of human variants from literature reports and therefore tends to have more substantial annotations from experimental evidence. DBSAV (Pei and Grishin, 2021) is a database that contains all SNPs in the human proteome and a predicted score of their deleteriousness.

Protein structures can help gain insights into the mechanism of these disease-associated mutations. However experimental determination of protein structures is time-consuming, expensive and difficult. With the increase in the size of proteins and complexes, it becomes increasingly challenging to experimentally determine their structures. Hence, computational models can help in such cases. Homology modelling techniques (such as MODELLER (Šali and Blundell, 1993), SWISS-MODEL (Waterhouse et al., 2018) etc) can be used to model sequences in presence of structural homologues having sequence identity greater than or equal to 30%, but a large number of protein sequences have no close homologues. *Ab initio* protein structure modelling techniques (ROSETTA (Rohl et al., 2004), I-TASSER (Roy et al., 2010) etc) are used in the absence of structural homologues. More recently deep neural networks trained on the massive body of available protein sequence and structure data have been shown to significantly outperform traditional homology modeling or *ab initio* approaches in terms of the accuracy of the resulting model (e.g. RaptorX (Xu, 2019), DMPFold (Greener et al., 2019), trRosetta (Anishchenko et al., 2017), AlphaFold.v1 (Senior et al., 2020) among others). Models like AlphaFold.v2 (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021) trained end-to-end to directly predict 3D model coordinates were shown to be able to reach near experimental accuracy. AlphaFold was shown to be the best performing technique for protein structure prediction in CASP14 (https://predictioncenter.org/casp14/). In July 2021, AlphaFold released the structures of the entire reference proteome of 21 organisms

(Tunyasuvunakool et al., 2021), which had about 50% of the residues (which could not be modeled using homology) predicted with confidence (Akdel et al., 2021). Recent studies have shown that these high-quality models can be used to predict binding sites and effects of mutations and these calculations/predictions are similar to those obtained from experiments/experimental structures (Akdel et al., 2021).

Along with the protein structure, experimental and computational determination of functional residues can help explain the effect of mutations on proteins. Mutations on/near functional sites are likely to affect the functioning of the proteins. The current strategies for the computational prediction of functional sites such as catalytic sites, ligand binding sites, allosteric sites, protein-protein interaction sites etc have been reviewed elsewhere (Ding and Kihara, 2018; Greener and Sternberg, 2018; He et al., 2019; Rauer et al., 2021). In addition to the predictions of functional sites, the energetic effect of mutations can also be calculated using various tools (Gapsys et al., 2016; Jespers et al., 2019; Schymkowitz et al., 2005; Steinbrecher et al., 2017), that determine the free energy change of the mutation. These tools can predict if the mutation will destabilize the protein structure, possibly affecting its functioning.

Homology models have been previously used to explain disease-associated mutations (Almqvist et al., 2004; Ittisoponpisan et al., 2019). In this manuscript, we model and exploit RoseTTAFold and AlphaFold models of human proteins without known protein structures or homologues in the PDB (sequence identity<30%) and explain the effect of deleterious mutations by checking if these mutations are on/near a predicted functional site, conserved site or lead to protein structure destabilization.

Results

1. *Protein domains for disease-associated proteins*

CATH (Orengo et al., 1997; Sillitoe et al., 2021) and Pfam (El-Gebali et al., 2019) superfamilies can be functionally diverse, with 62% of the protein space being covered by the largest 200 CATH superfamilies (Dessailly et al., 2009). Hence, only some of the members of these superfamilies will function similarly, requiring the need to subclassify them into functionally similar families called FunFams (Das et al., 2015). FunFams are coherent subsets of sequences predicted to have similar functions according to conserved specificity-determining positions in their multiple sequence alignment. Out of the 553 human disease-associated proteins, 198 proteins mapped to CATH, 341 mapped to Pfam leaving 14 proteins that could not be assigned to CATH or Pfam domains.

The 198 proteins assigned to CATH comprised 309 domains which were assigned to 297 FunFams and 117 superfamilies CATH. Leucine-rich repeat variants, Immunoglobulin superfamilies were the most represented in the dataset. The leucine-rich repeat variant fold, Rossman fold and Immunoglobulin-like fold were the most common folds (Figure S1).

The remaining proteins and regions which were not assigned to CATH domains were scanned against Pfam FunFams. 416 domains belonging to 341 proteins were matched to 376 unique Pfam FunFams.

A total of 469 regions of proteins (at least 50 residue long stretches) belonging to 332 proteins were designated as unassigned domains which did not map to CATH/Pfam domains.

Gene ontology analysis (molecular function) using PANTHER (Mi et al., 2013) for these 553 proteins showed enrichment of some gene functions, including phosphatidylinositol N-acetylglucosaminyltransferase activity (30 fold enrichment), melanocortin receptor binding (type 1,3 and 4) (28 fold enrichment), sodium-phosphate symporter activity (13 fold enrichment). The biological processes enriched were sodium-dependent phosphate transport (37-fold enrichment), protein localisation to the ciliary transition zone (29 fold enrichment), Endoplasmic reticulum tubular network organisation (28 fold enrichment), lysosomal lumen acidification (27 fold enrichment), attachment of GPI anchor to proteins (25 fold enrichment).

We wanted to check if certain superfamilies were modeled better than others. For this purpose, we used all the models available from the AlphaFold Protein Structure Database, we calculated the mean model quality for each of the CATH superfamilies. Only 3.5% of the superfamilies had a mean model quality score lesser than 70, indicating most superfamilies were modeled well irrespective of their type (Figure S2).

2. *Comparison of domain models generated using RoseTTAFold and AlphaFold*

The quality of the models was assessed using the average predicted lDDT score. The range of lDDT scores for RoseTTAFold models was between 0 and 1, while those of AlphaFold were between 0 and 100. RoseTTAFold models with a score of 0.7 and AlphaFold models with a score of 70 were considered as good models, as recommended by the developers of these techniques. 304/309 (98%), 328/416 (79%) and 198/469 (42%) CATH, Pfam and unassigned domains respectively were good AlphaFold models (Figure 1). Whereas for the RoseTTAFold models 274/309 (89%), 240/416 (58%) and 79/469 (17%) CATH, Pfam and unassigned domains respectively were good models. 72% AlphaFold models and 58% RoseTTAFold models were of high quality.

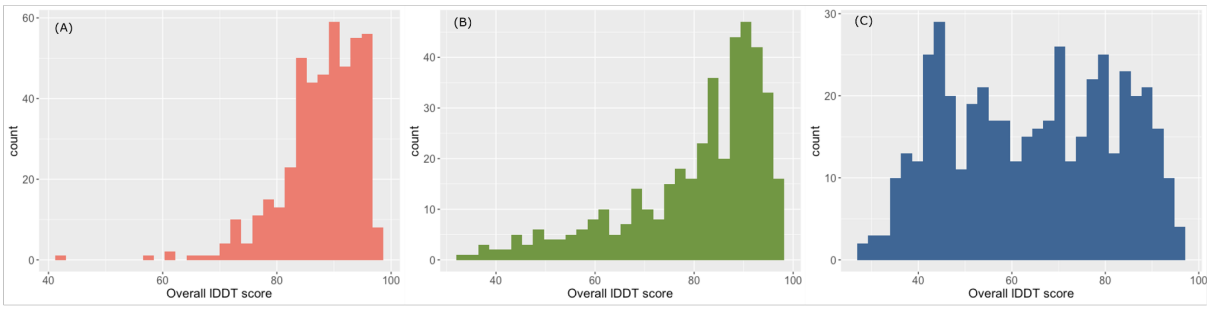

Figure 1- The model quality of AlphaFold models for A) CATH domains B) Pfam domains C) Unassigned domains.

We did not have access to the MSAs used to build the AlphaFold models, hence we analyzed the MSAs used to build the RoseTTAFold models. We calculated the Neff, DOPs, percent scorecons and number of taxons in the alignment to check the information content of the MSA

(see Methods). Neff and the number of taxons in the alignment are higher for CATH models compared to the other two groups providing more coevolutionary information for those models (Figure S3). However, no clear correlation was found between Neff, DOPs, percent scorecons, number of taxons in the alignment and model quality (Supplementary Figure S4).

Though the AlphaFold model quality was higher for 89% of the models compared to RoseTTAFold, 2, 27 and 45 CATH, Pfam and unassigned domains respectively had good RoseTTAFold models where the AlphaFold models were not of good quality (Figure 2).

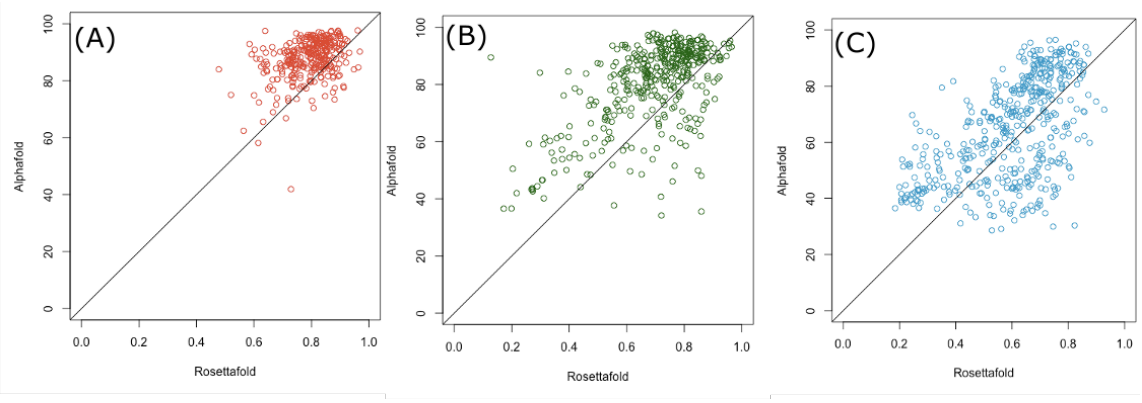

Figure 2- Model quality of AlphaFold models vs RoseTTAFold models for A) CATH domains B) Pfam domains C) Unassigned domains.

Around 88%, 74% and 64% of the CATH, Pfam and unassigned domains had an RMSD < 2 Å between domains which had both good RoseTTAFold and AlphaFold models, indicating higher structural similarity for the CATH and Pfam domains (Figure S5). Some of the models, including regions of the C2 domain-containing protein 3 (Q4AC94) mapped onto CATH FunFams, showed RMSD < 1 Å between RoseTTAFold and AlphaFold models (Figure 3). The RMSD of the domains increased with a decrease in model quality (Figure 4). The AlphaFold model quality did not depend on the length of the domain (correlation coefficient between length and model quality of 0.16, -0.13, -0.03 for CATH, Pfam and unassigned domains) (Figure S6). However, the model quality was lower for longer Pfam and unassigned RoseTTAFold domain models (correlation coefficient between length and model quality of 0, -0.35, -0.52 for CATH, Pfam and unassigned domains respectively) (Figure S6).

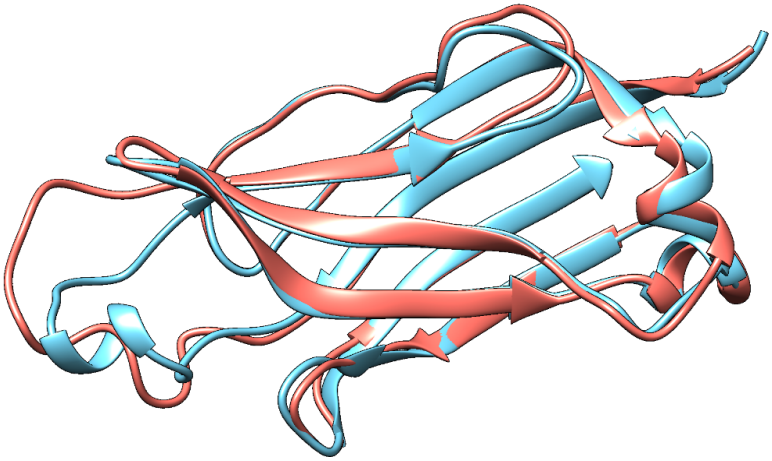

Figure 3 - Ribbon model of superimposition of AlphaFold model (salmon) and RoseTTAFold model (sky blue) A) CATH domain of C2 domain-containing protein 3, having an RMSD of 1 Å. Both the models are of good quality.

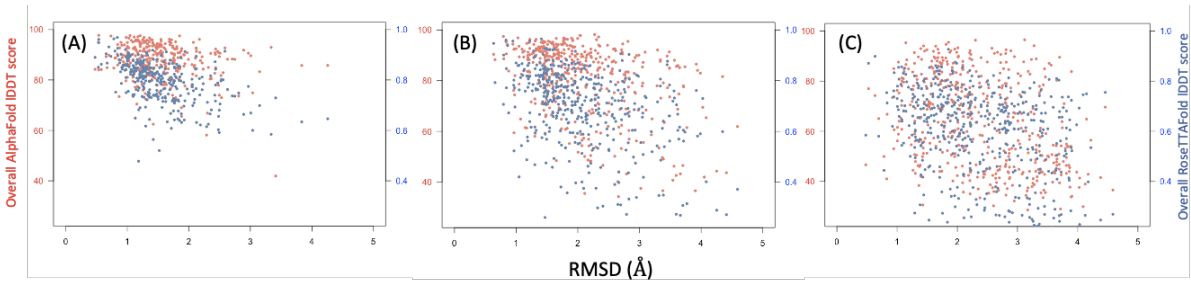

Figure 4 - Distribution of model quality (AlphaFold in red, RoseTTAFold in blue) against the RMSD between the models for a) CATH domains b) Pfam domains and c) Unassigned domains

2. Disorder in protein domains

The probability of a residue in the sequence to be disordered for the CATH, Pfam and unassigned domains was calculated using IUPred2A (Mészáros et al., 2018). We then calculated the percentage of the domain sequence that is disordered (IUPred2A predicted probability of a residue being greater than 0.5 is considered disordered). For the CATH, Pfam and unassigned domains around 3, 15 and 38% of the sequences respectively had greater than 40% of its sequence disordered (Figure 4). This indicates that the CATH domains had much lower disorder compared to Pfam which in turn was much lower compared to the unassigned domains. Given the fact that the disordered regions lack proper structure the model quality of those regions will be low, hence lowering the overall domain model quality. The percentage of the predicted disordered region in the sequences of the unassigned domains had an inverse correlation coefficient of 0.74 and 0.41 with model quality for the AlphaFold and RoseTTAFold models respectively (Figure S7), indicating that the model quality was lower because the sequences might be disordered.

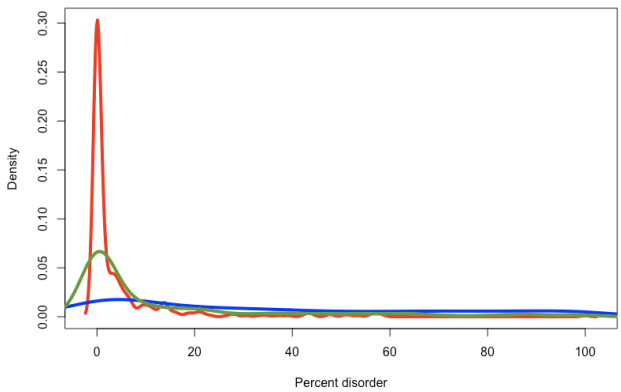

Figure 4 - Density plot showing the distribution of percent disorder of the domain sequence for CATH (in red), Pfam (in green) and unassigned (in blue) domains

3. Explaining disease-associated mutations

We extracted the disease-associated mutations from humsavar. We checked if these mutations were on or near a predicted ligand binding site (by P2Rank) or protein-protein interface (by meta-PPISP) or are conserved (by scorecons) in the FunFam or superfamily. Though there are a lot of servers and tools for predicting ligand binding sites and protein-protein interface we used meta-PPISP and P2Rank based on availability, ease of installation and ability to run the scripts on all the protein domains. We considered the residues within a 5 Å radius of a predicted functional site because mutations in those can lead to change in the environment of the functional site residue, potentially leading to modification or loss of activity. We also calculated the effect of the mutations on the stability of the protein using FoldX. Some of these proteins had more than 10 disease-associated mutations (Figure S8).

The total number of disease-associated mutations for our modeled proteins in humsavar is 1730, out of which 1317 mutations are in high confidence AlphaFold modeled domains and residues (lTTD>70). For these domains, 122 CATH (585 mutations), 143 Pfam (508 mutations) and 61 unassigned domains (224 mutations) had disease-associated mutations. 796 of these disease-associated mutations are in/near a predicted functional site (Figure 5). In addition, 227 mutations were predicted to be destabilizing using FoldX (Figure 5). Of all the disease-associated mutations, 488 were near the predicted protein-protein interface, 163 near predicted ligand binding sites, 473 near conserved sites and 689 mutations were predicted to be destabilizing by FoldX (ddG>1) (Figure S9). In total, this suggests that we can provide some rationale for 78% of the disease-associated mutations in the well modeled regions.

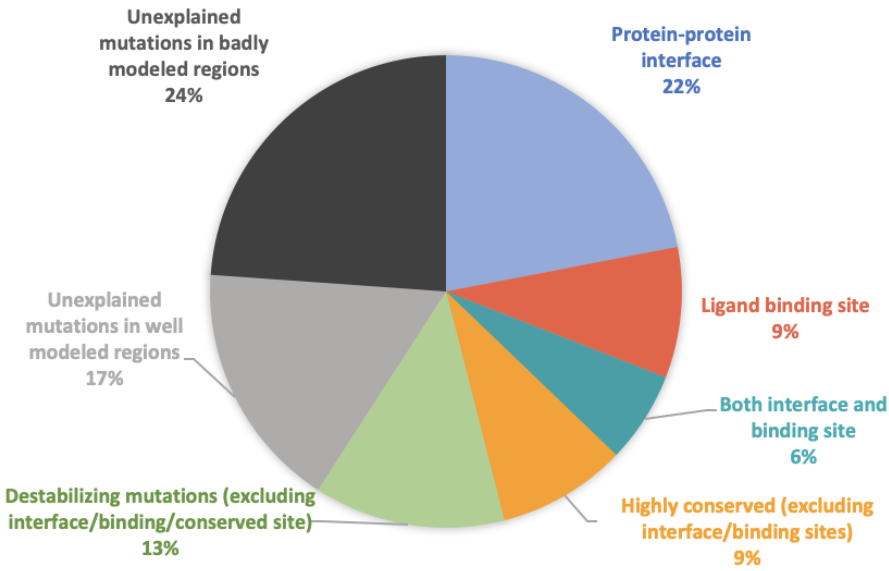

Figure 5 - Percentage of disease-associated mutations that are predicted to be in/near a ligand binding/protein-protein interface/conserved site in the AlphaFold domain models. Some of these mutations are predicted to be destabilizing.

When we consider whether mutations fall on overlapping functional sites, we observe that 108 mutations were predicted as both near a ligand binding site and a protein-protein interface, 101 mutations either near a conserved residue or ligand binding site, 221 mutations were either near an interface or conserved site (Figure 6). Around 44% of these mutations have multiple evidence for it lying on/near a functional site, which provides higher confidence in these sites having functional significance.

In addition, usage of the RoseTTAFold domain models also helped explain 93 disease-associated mutations which could not be explained using AlphaFold models. There was an overlap of 492 mutations predicted to be a functional site between the predictions derived from RoseTTAFold models and AlphaFold models, again providing higher confidence on these predictions.

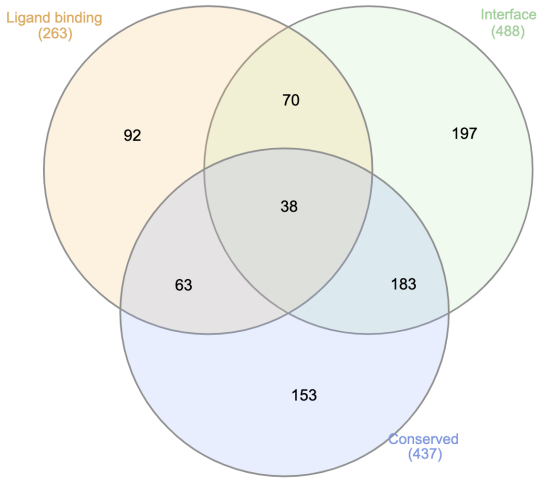

Figure 6 - The overlap between the number of disease-associated mutation sites that are either predicted to be on/near a ligand binding/protein-protein interface/conserved for the AlphaFold domain models.

The putative sodium-coupled neutral amino acid transporter 8 was mapped to the putative sodium-coupled neutral amino acid transporter 7 CATH FunFam. Mutations in residue numbers 32,33,233,236 and 412 lead to foveal hypoplasia 2. Other than residue 412, the others are spatially close to each other and are predicted either to be in a ligand binding site or near one. In addition, residue 233 is also predicted to be near a conserved site (Figure 7A). Mutations in Vitamin K Epoxide Reductase Complex submit 1 (Q9BQB6) leads to Coumarin resistance, which lie on/near the predicted ligand binding sites and interface residues (Figure 7B). We could also use RoseTTAFold models to explain mutations in cases where the AlphaFold model was of low quality, or the residue of interest was of low quality. Mutations in residue numbers, 21 and 78 in immediate early response-3 interacting protein-1 have been associated with microcephaly, epilepsy, and diabetes syndrome (MEDS). These residues are on the predicted protein-protein interface site (Figure 7C).

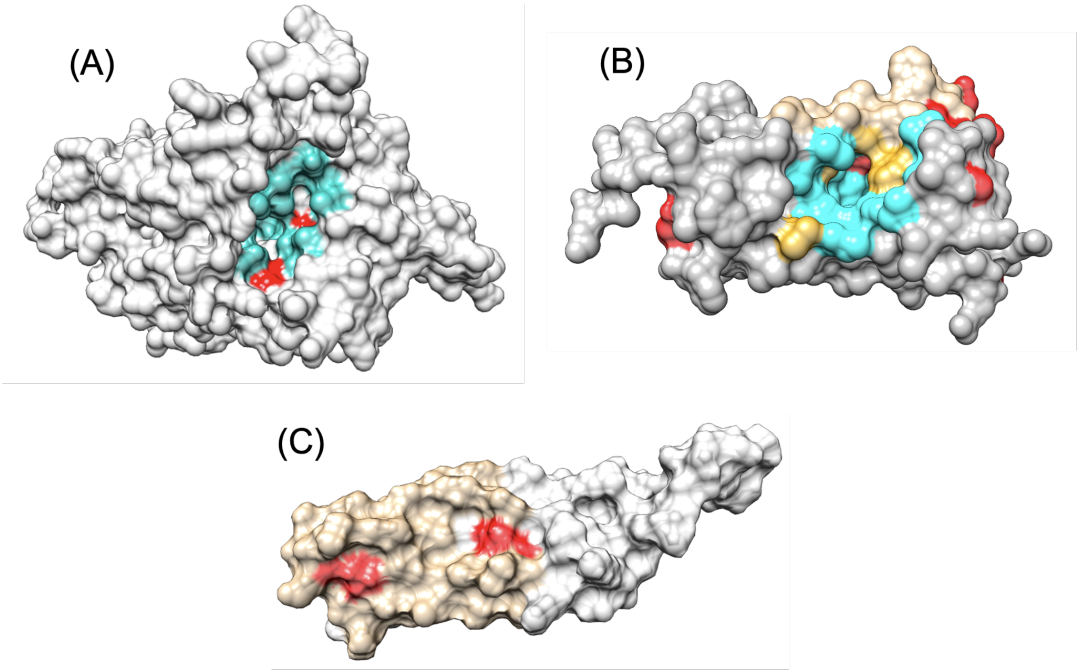

Figure 7 - A) AlphaFold models in grey surface representations for Putative sodium-coupled neutral amino acid transporter 8, the residues in cyan are the predicted ligand binding residues and in red are the disease-associated mutations. B) AlphaFold models in grey surface representations for Vitamin K Epoxide Reductase Complex submit 1, the residues in cyan are the predicted ligand binding residues, tan are the predicted interface residues, golden are the residues predicted both as ligand binding and interface, and in red are the disease-associated mutations. C) RoseTTAFold model of early response-3 interacting protein-1 in grey surface representation. The predicted interface is in tan and the disease-associated mutations are in red.

We have also tabulated all the disease-associated positions, their predicted functional site impacts, ddG of mutations and other characteristics here. The table also lists the known or predicted functions of the proteins given by the experimental GO terms for the FunFam in which the proteins are classified. All the data and the models can be accessed here.

Discussion

With the current improved state of art computational structural predictors like RoseTTAFold and AlphaFold, we can model regions of proteins without any known homologues for about 25% of the residues in humans. In this study, we modeled the disease-associated human proteins without known homologues using RoseTTAFold and AlphaFold. We used these models to check if the disease-associated mutations were near predicted functional residues (i.e. ligand binding site/protein-protein interface/conserved residues). Some of these disease-associated mutations were also predicted to be structurally destabilizing.

The AlphaFold models were largely of better quality when compared to those of RoseTTAFold models. 70% of the modeled AlphaFold domains were of good quality whereas 50% of the modeled RoseTTAFold domains were of good quality. However, there were a total of 74 models where RoseTTAFold produced a good quality model whereas AlphaFold had a low quality model. We should keep in mind that model quality calculations for the two techniques were independent of each other, and no consensus model quality assessment tool was used. To date, there has been no evaluation of the model quality estimates of AlphaFold and RoeTTAFold. However, optimizing model quality cut-offs for such consensus estimators will involve benchmarking and is out of the scope of the current study.

Around 56% of the good quality models from both methods, used for the analyses, had an RMSD <2 Å. This indicates that predicted structures from the state of art techniques are similar, which gives greater confidence in the models. In some others, the predicted structure of the local regions are similar, however, changes in loop orientations increase the RMSD. These high confidence models could be used to predict/design inhibitors/drugs against them (Sen et al., 2019), predict off-target effects of putative drugs (Nguyen et al., 2019), explain the functioning of proteins, the impact of variations (Waman et al., 2021) or model protein complexes (Kanitkar et al., 2021). RMSDs between the models built using AlphaFold and RoseTTAFold had an inverse correlation with model quality. The model quality of the RoseTTAFold models was reduced with an increase in the length of the sequence which was not the case with the AlphaFold models.

It should also be noted that the AlphaFold domains were extracted of the full proteins, however, RoseTTAFold models were built directly from the domain sequences. The inherent differences in the model building techniques might have added differences in the models. Residues in AlphaFold domains were modeled in the presence of the residues from other interacting domains, which was not the case for RoseTTAFold. Hence added constraints during modeling might have led to the better model quality of the AlphaFold models as compared to RoseTTAFold models. Simple orientation changes of helices or sheets with respect to one another in RoseTTAFold domain models would help in better superimposition to those of their AlphaFold counterparts.

The domains that matched to CATH FunFams belonged to diverse CATH folds. The commonly occurring folds predicted from diverse sequences in the dataset, such as the Rossmann fold and Immunoglobulin folds are some of the most commonly occurring folds in the CATH database.

The CATH domains were modeled better compared to Pfam or unassigned domains largely because of the sequence diversity of the multiple sequence alignments used to build the models. These diverse sequences provided better coevolutionary information which is used explicitly by RoseTTAFold and implicitly by AlphaFold to build computational models. With the ease of sequencing and large-scale sequencing projects, we expect to further increase the repertoire and quality of protein sequences in the near future. With the increased number of sequences, the model quality of the Pfam and unassigned domains will hopefully improve.

In addition, the CATH domains had lower disorder compared to Pfam which was lower than that of unassigned domains. Around 40% of the unassigned domains had greater than 40% of their sequence as disordered, hence leading to lower model quality of those regions by both the modeling techniques and also the high RMSD of the models built by the two techniques.

Functional sites such as ligand binding sites, interface residues, allosteric sites, post-translational modification sites etc are more conserved compared to the rest of the protein, hence mutation of these conserved sites and their neighbours might lead to a loss/modification in protein function. We considered residues close to the predicted functional site because these residues can have an effect on the structure and functioning of the functional site residues and hence changes to them might also alter the functioning of the protein. Also, previous studies have shown that in a large number of cases, the disease-associated mutations are near functional sites (Ashford et al., 2019; Gao et al., 2015). For the assessment, we used a model quality and residue quality lDDT cut off of 70 for AlphaFold and 0.7 for RoseTTAFold models. We assume that the neighbourhood of the functional sites will be well predicted at this level of model quality even if the functional residues are not always correct. Interface residues and ligand binding sites have been shown to be highly conserved (Das et al., 2021). In our dataset, 38% of the disease-associated ligand binding sites and 45% of the protein-protein interfaces are conserved, hence providing higher confidence that these mutations could impact structure or function. We were able to explain 78% of the disease-associated mutations in the good quality regions. Also, around 50% of these mutations were predicted to be structurally destabilizing, which might also impact structure and function. We also had 37% of the mutations being predicted near functional sites using both AlphaFold and RoseTTAFold models, which provide additional confidence in those predictions. Given the fact that the models built using the state of art tools AlphaFold and RoseTTAFold might still have inaccuracies, cases, where the structure predicted using multiple techniques, are similar provide higher confidence in the prediction accuracies.

In summary, the strategy of using good quality protein structures to predict functional sites can help explain the mechanism of disease-associated mutations. Similar strategies can be used on good quality models of the entire human proteome in the future to explain these mutations.

Methods

1. *Assignment of domain boundaries to disease-associated human proteins without structural homologues*

The VarSite database (Laskowski et al., 2020) contains 4444 human proteins which are disease-associated as annotated in Uniprot or ClinVar. Out of these, 553 human disease proteins are without any structural homologues (match determined by running the sequence with BLAST against all PDB sequences with evalue cutoff of 1e-06). All the 553 proteins were scanned against CATH-FunFams v4.3 (212,872 HMMs) using HMMER3 (Mistry et al., 2013) with an e-value cut off of 1e-03 to assign domains. The domain boundaries were resolved using cath-resolve-hits (Lewis et al., 2019) with a bitscore cut-off of 25 and coverage of 80%. The proteins/regions of proteins that were not assigned to CATH domain boundaries were scanned against the Pfam32-FunFams (102,712 HMMs) using the same protocol. The regions not belonging to domains identified by scans against CATH/Pfam FunFams, were modeled separately. These regions have been referred to as unassigned domains.

2. *Modeling the proteins using RoseTTAFold*

   a) *Multiple sequence alignment (MSA) generation for protein domains*

The domains that were assigned to the CATH/Pfam FunFams were aligned to the FunFams using MAFFT (version 7.471) (Katoh and Standley, 2013) using the default options. The resulting seed alignments were enriched with additional homologs by performing iterative sequence *hhblits* (Steinegger et al., 2019a) searches against uniclust30 (version UniRef30_2020_01) and BFD (Steinegger and Söding, 2018; Steinegger et al., 2019b) databases as outlined by Anischenko and colleagues (Anishchenko et al., 2021).

   b) *Assessment of MSA quality*

The quality of the alignment was accessed using 4 parameters - Neff, DOPs (Diversity of Positions), percent scorecons and number of taxon IDs in the alignment. The number of effective sequences (Neff) is defined as the number of sequences in the multiple sequence alignment after removing the rows and columns with greater than 25% gaps and clustering the remaining sequences at 80% sequence identity cutoff. Scorecons (Valdar, 2002) is an entropy-based method that calculates the conservation of each position in the alignment with 0 being completely unconserved and 1 being completely conserved. Percent scorecons is defined as the percentage of positions (calculated with respect to mean length of the sequences in the alignment) that have a scorecons value >=0.8. This measure provides an estimate of the percentage of conserved positions. DOPs scores calculate the diversity based on the conservation scores and their frequencies, with 0 indicating no diversity and 100 indicating no positions having the same scores i.e highly diverse.

   c) *Modeling of the protein domains*

Using multiple sequence alignments from Section 2a above, structure templates were identified by *hhsearch* searches against the PDB100 database, and 10 best scoring templates were selected. The identified templates along with the multiple sequence alignment were then used as inputs to RoseTTAFold (Baek et al., 2021) to predict residue-residue distances and orientations followed by the PyRosetta-based (Chaudhury et al., 2010) 3D structure reconstruction protocol from trRosetta (Anishchenko et al., 2017). 15 structure models were generated and subsequently rescored according to lDDT scores (Mariani et al., 2013) predicted by DeepAccNet (Hiranuma et al., 2021). The model with the highest predicted global lDDT score was selected for further analysis.

3. *Creating domains from AlphaFold models*

The domains were extracted from the AlphaFold models downloaded from the AlphaFold DataBase portal. The domain boundaries were the same as obtained by the CATH/Pfam scans or unassigned domains as mentioned in Section 1. The AlphaFold domains were superimposed on the domains built using RoseTTAFold using GESAMT (Krissinel, 2012).

4. Scoring the local and global quality of the models

The model quality for both RoseTTAFold and AlphaFold models were calculated using the predicted lDDT scores (Mariani et al., 2013). The global quality of the model is the average of the local lDDT scores of the constituent residues.

5. Disorder prediction on the domain sequences

The disorder percentage was calculated for the unassigned regions using IUPred2A (Mészáros et al., 2018) using default parameters. A residue was considered disordered if the prediction probability was greater than 0.5.

6. Prediction of functional sites on protein domains and known mutational sites

The ligand binding sites were predicted using P2Rank (Krivák and Hoksza, 2018), with default parameters. Positions were annotated as ligand binding if the probability of the ligand binding was >=0.5 (intuitively chosen). The protein-protein interaction sites were predicted using the meta-PPISP webserver (Qin and Zhou, 2007).

The conserved positions were identified as positions in the MSA with a high scorecons value (>=0.8). FunFams have been shown to produce MSAs that can be used to identify conserved residues highly enriched in functional sites (Das et al., 2015) Sometimes these are not large enough to be informative enough to detect conserved residues. Wherever possible we created MSAs (such that Neff>30, DOPs>70 and percent scorecons<20%) by either combining the closely related functional families or aligning the query sequence with members within its superfamily. The FunFam of the queried domain was merged to other FunFams in the superfamily (provided that the average length is within 80% of the queried FunFam), provided the e-value cut off of 1e-03 was met and the Neff of the alignment increased. This process was iterated until Neff was maximised. The domains whose MSA could not be expanded by merging FunFams were searched using HMMER3 within the specific superfamily with an e-value cut off of 1e-03. These alignments were used if the DOPs>70, percent scorecons<20% and Neff>30. For the remaining domains, the conserved positions were identified using the MSA used to build the models, provided they followed the DOPs and percent scorecons cut-off. These MSAs were built using metagenomic sequences and therefore tended to be much larger which sometimes might affect the detection of conserved sites. The neighbourhood of the predicted functional residues were identified using the implementation of the cell list algorithm (https://github.com/neeleshsoni21/Cell_list).

The disease-associated mutations were identified from humsavar (The UniProt Consortium et al., 2021). For each of these sites, the effect of mutation on the stability of the protein was identified using FoldX (Schymkowitz et al., 2005). Mutations were labelled as destabilizing if the predicted ddG of mutation was greater than 1 (intuitively chosen). The mutation was predicted to affect the functioning of the protein if the site was either on or near (within 5 Å) of a predicted functional site by P2Rank/meta-PPISP/highly conserved position. We only considered predictions for those sites where the entire domain had a good model quality score

and the local quality score of the mutated residue was good (lDDT score of 70 for AlphaFold models and 0.7 for RoseTTAFold models).

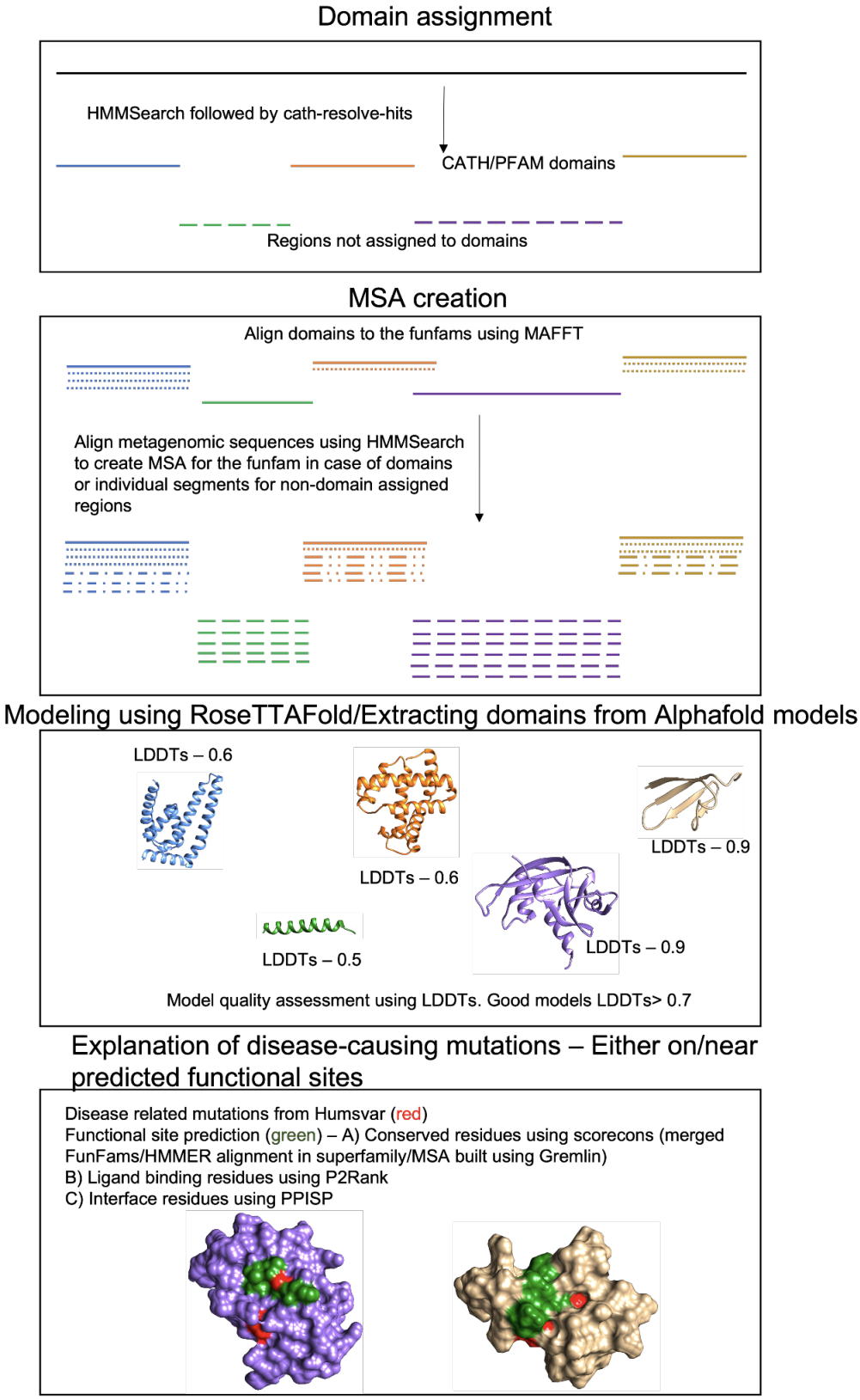

Figure 8- Schematic explaining the methodology to model protein domains and explain disease-associated mutations.

Conflict of Interest

None.

Acknowledgements

The following people acknowledge the Biotechnology and Biological Sciences Research Council for their funding: NS [BB/S020144/1], NB [BB/R009597/1], IS [BB/R014892/1] and SV[BB/S017135/1]. IA is supported by National Science Foundation Award # DBI 1937533 and DB is supported by Audacious Project at the Institute for Protein Design.

References

Ainscough, B.J., Griffith, M., Coffman, A.C., Wagner, A.H., Kunisaki, J., Choudhary, M.N., McMichael, J.F., Fulton, R.S., Wilson, R.K., Griffith, O.L., et al. (2016). DoCM: a database of curated mutations in cancer. Nat. Methods *13*, 806–807.

Akdel, M., Pires, D.E.V., Porta Pardo, E., Jänes, J., Zalevsky, A.O., Mészáros, B., Bryant, P., Good, L.L., Laskowski, R.A., Pozzati, G., et al. (2021). A structural biology community assessment of AlphaFold 2 applications (Biophysics).

Almqvist, J., Huang, Y., Hovmöller, S., and Wang, D.-N. (2004). Homology Modeling of the Human Microsomal Glucose 6-Phosphate Transporter Explains the Mutations That Cause the Glycogen Storage Disease Type Ib. Biochemistry *43*, 9289–9297.

Anfinsen, C.B. (1973). Principles that Govern the Folding of Protein Chains. Science *181*, 223–230.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. Proc. Natl. Acad. Sci. *114*, 9122–9127.

Anishchenko, I., Baek, M., Park, H., Hiranuma, N., Kim, D.E., Dauparas, J., Mansoor, S., Humphreys, I.R., and Baker, D. (2021). Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. Proteins Struct. Funct. Bioinforma. prot.26194.

Ashford, P., Pang, C.S.M., Moya-García, A.A., Adeyelu, T., and Orengo, C.A. (2019). A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. Sci. Rep. *9*, 263.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876.

Baiardi, S., Rossi, M., Capellari, S., and Parchi, P. (2019). Recent advances in the histo-molecular pathology of human prion disease: Histo-molecular pathology of human prion disease. Brain Pathol. *29*, 278–300.

Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691.

Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W.R., Wang, R., Huang, J., Hao, T., et al. (2021). Comprehensive characterization of protein–protein interactions perturbed by disease mutations. Nat. Genet. *53*, 342–353.

Coleman, C.S., Stanley, B.A., and Pegg, A.E. (1993). Effect of mutations at active site residues on the activity of ornithine decarboxylase and its inhibition by active site-directed irreversible inhibitors. J. Biol. Chem. *268*, 24572–24579.

Das, S., Lee, D., Sillitoe, I., Dawson, N.L., Lees, J.G., and Orengo, C.A. (2015). Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics *31*, 3460–3467.

Das, S., Scholes, H.M., Sen, N., and Orengo, C. (2021). CATH functional families predict functional sites in proteins. Bioinformatics *37*, 1099–1106.

Dessailly, B.H., Nair, R., Jaroszewski, L., Fajardo, J.E., Kouranov, A., Lee, D., Fiser, A.,

Godzik, A., Rost, B., and Orengo, C. (2009). PSI-2: Structural Genomics to Cover Protein Domain Family Space. Structure *17*, 869–881.
Ding, Z., and Kihara, D. (2018). Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features. Curr. Protoc. Protein Sci. *93*.
El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.
Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. *48*, D941–D947.
Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.
Gao, M., Zhou, H., and Skolnick, J. (2015). Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. Structure *23*, 1362–1369.
Gapsys, V., Michielssens, S., Seeliger, D., and de Groot, B.L. (2016). Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. Angew. Chem. Int. Ed Engl. *55*, 7364–7368.
Greener, J.G., and Sternberg, M.J. (2018). Structure-based prediction of protein allostery. Curr. Opin. Struct. Biol. *50*, 1–8.
Greener, J.G., Kandathil, S.M., and Jones, D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. *10*, 3977.
He, W., Wei, L., and Zou, Q. (2019). Research progress in protein posttranslational modification site prediction. Brief. Funct. Genomics *18*, 220–229.
Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., and Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. Nat. Commun. *12*, 1340.
Ittisoponpisan, S., Islam, S.A., Khanna, T., Alhuzimi, E., David, A., and Sternberg, M.J.E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J. Mol. Biol. *431*, 2197–2212.
Jespers, W., Isaksen, G.V., Andberg, T.A.H., Vasile, S., van Veen, A., Åqvist, J., Brandsdal, B.O., and Gutiérrez-de-Terán, H. (2019). QresFEP: An Automated Protocol for Free Energy Calculations of Protein Mutations in Q. J. Chem. Theory Comput. *15*, 5461–5473.
Joshi, S., Virdi, S., Etard, C., Geisler, R., and Strähle, U. (2018). Mutation of a serine near the catalytic site of the choline acetyltransferase a gene almost completely abolishes motility of the zebrafish embryo. PLOS ONE *13*, e0207747.
Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montaño, B., Blundell, T.L., and Ascher, D.B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog. Biophys. Mol. Biol. *128*, 3–13.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
Kanitkar, T.R., Sen, N., Nair, S., Soni, N., Amritkar, K., Ramtirtha, Y., and Madhusudhan, M.S. (2021). Methods for Molecular Modelling of Protein Complexes. In Structural Proteomics, R.J. Owens, ed. (New York, NY: Springer US), pp. 53–80.
Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.
Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzneladze, I., Ouellette, B.F.F., Yamada, J.T., Fradet-Turcotte, A., and Reimand, J. (2018). ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. Nucleic Acids Res. *46*, D901–D910.

Krissinel, E. (2012). Enhanced fold recognition using efficient short fragment clustering. J. Mol. Biochem. *1*, 76–85.
Krivák, R., and Hoksza, D. (2018). P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J. Cheminformatics *10*, 39.
Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844.
Laskowski, R.A., Stephenson, J.D., Sillitoe, I., Orengo, C.A., and Thornton, J.M. (2020). VarSite: Disease variants and protein structure. Protein Sci. Publ. Protein Soc. *29*, 111–119.
Lee, B., Thirunavukkarasu, K., Zhou, L., Pastore, L., Baldini, A., Hecht, J., Geoffrey, V., Ducy, P., and Karsenty, G. (1997). Missense mutations abolishing DNA binding of the osteoblast-specific transcription factor OSF2/CBFA1 in cleidocranial dysplasia. Nat. Genet. *16*, 307–310.
Lewis, T.E., Sillitoe, I., and Lees, J.G. (2019). cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. Bioinformatics *35*, 1766–1767.
Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722–2728.
Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. *46*, W329–W337.
Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. *8*, 1551–1566.
Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121–e121.
Nguyen, M.N., Sen, N., Lin, M., Joseph, T.L., Vaz, C., Tanavde, V., Way, L., Hupp, T., Verma, C.S., and Madhusudhan, M.S. (2019). Discovering Putative Protein Targets of Small Molecules: A Study of the p53 Activator Nutlin. J. Chem. Inf. Model. *59*, 1529–1546.
Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. (1997). CATH – a hierarchic classification of protein domain structures. Structure *5*, 1093–1109.
PDBe-KB consortium, Varadi, M., Berrisford, J., Deshpande, M., Nair, S.S., Gutmanas, A., Armstrong, D., Pravda, L., Al-Lazikani, B., Anyango, S., et al. (2020). PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. *48*, D344–D353.
Pei, J., and Grishin, N.V. (2021). The DBSAV Database: Predicting Deleteriousness of Single Amino Acid Variations in the Human Proteome. J. Mol. Biol. *433*, 166915.
Qin, S., and Zhou, H.-X. (2007). meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics *23*, 3386–3387.
Rauer, C., Sen, N., Waman, V.P., Abbasian, M., and Orengo, C.A. (2021). Computational approaches to predict protein functional families and functional sites. Curr. Opin. Struct. Biol. *70*, 108–122.
Ricatti, J., Acquasaliente, L., Ribaudo, G., De Filippis, V., Bellini, M., Llovera, R.E., Barollo, S., Pezzani, R., Zagotto, G., Persaud, K.C., et al. (2019). Effects of point mutations in the binding pocket of the mouse major urinary protein MUP20 on ligand affinity and specificity. Sci. Rep. *9*, 300.
Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein Structure Prediction Using Rosetta. In Methods in Enzymology, (Elsevier), pp. 66–93.
Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738.
Šali, A., and Blundell, T.L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. *234*, 779–815.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382–W388.
Sen, N., Kanitkar, T.R., Roy, A.A., Soni, N., Amritkar, K., Supekar, S., Nair, S., Singh, G., and Madhusudhan, M.S. (2019). Predicting and designing therapeutics against the Nipah virus. PLoS Negl. Trop. Dis. *13*, e0007419.
Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710.
Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273.
Smigielski, E.M., Sirotkin, K., Ward, M., and Sherry, S.T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. *28*, 352–355.
Soto, C. (2003). Unfolding the role of protein misfolding in neurodegenerative diseases. Nat. Rev. Neurosci. *4*, 49–60.
Steinbrecher, T., Zhu, C., Wang, L., Abel, R., Negron, C., Pearlman, D., Feyfant, E., Duan, J., and Sherman, W. (2017). Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability-Large-Scale Validation of MD-Based Relative Free Energy Calculations. J. Mol. Biol. *429*, 948–963.
Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. *9*, 2542.
Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019a). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473.
Steinegger, M., Mirdita, M., and Söding, J. (2019b). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods *16*, 603–606.
Stenberg, K.A.E., Riikonen, P.T., and Vihinen, M. (1999). KinMutBase, a database of human disease-causing protein kinase mutations. Nucleic Acids Res. *27*, 362–364.
The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., et al. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596.
Tyukhtenko, S., Rajarshi, G., Karageorgos, I., Zvonok, N., Gallagher, E.S., Huang, H., Vemuri, K., Hudgens, J.W., Ma, X., Nasr, M.L., et al. (2018). Effects of Distal Mutations on the Structure, Dynamics and Catalysis of Human Monoacylglycerol Lipase. Sci. Rep. *8*, 1719.
Valdar, W.S.J. (2002). Scoring residue conservation. Proteins Struct. Funct. Genet. *48*, 227–241.
Waman, V.P., Sen, N., Varadi, M., Daina, A., Wodak, S.J., Zoete, V., Velankar, S., and Orengo, C. (2021). The impact of structural bioinformatics tools and resources on SARS-CoV-2 research and therapeutic strategies. Brief. Bioinform. *22*, 742–768.
Wang, T., Ruan, S., Zhao, X., Shi, X., Teng, H., Zhong, J., You, M., Xia, K., Sun, Z., and Mao, F. (2021). OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. Nucleic Acids Res. *49*, D1289–D1301.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. *46*, W296–W303.
Xu, J. (2019). Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. *116*, 16856–16865.