Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Protein structure-based evaluation of missense variants: Resources, challenges and future directions

Alessia David and Michael J. E. Sternberg

Abstract

We provide an overview of the methods that can be used for protein structure-based evaluation of missense variants. The algorithms can be broadly divided into those that calculate the difference in free energy (ΔΔG) between the wild type and variant structures and those that use structural features to predict the damaging effect of a variant without providing a ΔΔG. A wide range of machine learning approaches have been employed to develop those algorithms. We also discuss challenges and opportunities for variant interpretation in view of the recent breakthrough in three-dimensional structural modelling using deep learning.

Addresses

Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK

Corresponding author: David, Alessia (alessia.david09@imperial.ac.uk)

Current Opinion in Structural Biology 2023, **80**:102600

This review comes from a themed issue on **Sequences** and **Topology** (2023)

Edited by **Madan Babu** and **Rita Casadio**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2023.102600

Keywords

Missense variants, Prediction, Algorithms, Protein structure.

Introduction

Large scale genetic projects, such as the All of Us Research Program in the US, the 100 K Genomes Project in the UK, and similar studies in other countries, are generating large amount of genetic data the clinical significance and phenotypic impact for which is, in most cases, unknown. In ClinVar, one of the most used genetic databases, there is conflicting information regarding the variant categorization (e.g. benign versus damaging) for 3% of variants which increases to 7% when confidence in the annotation is also considered (e.g. benign versus likely benign). The proportion of variants with conflicting annotation is even greater in commonly sequenced genes, such as cancer-related ones (e.g. BRAC1 8% and 13%, respectively) and the LDLR gene, which is associated with familial hypercholesterolemia [1] (30% and 68%, respectively) (from https://clinvarminer.genetics.utah.edu/). In silico analysis is an invaluable tool to aid prioritization and characterization of genetic variants. Several bioinformatics resources are available (reviewed in the study by Zeng et al. [1–4]) and this review focusses on methods that implement data from protein three-dimensional (3D) structures.

Recent expansion in the 3D structural coverage of the proteome

Guidelines for the clinical interpretation of genetic variants recommend the use of several sources of information, including sequence conservation and population frequency [5]. Although structural interpretation of genetic variants using protein structures is one the methods recommended [5], the incomplete 3D structural coverage of the proteome remains a major limiting factor. The number of experimentally solved structures in PDB has massively grown in the last few years, however only about 17% of human residues have structural coverage [6]. Although this percentage can be greatly increased by the use of homology models obtained from algorithms, such as I-Tasser [7], Phyre [8], SwissModel [9] (reviewed in the study by Kuhlman et al. [10]), large protein complexes and proteins with an unusual topology are challenging to solve experimentally or through modelling.

In the last few years, one of the major developments in the field of structural biology has been the dramatic improvement in the cryoEM technique, especially its resolution, thus leading to a large expansion in the number of cryoEM structures released per year in PDB (from <1% in 2012 to 29% in 2022, from https://www.rcsb.org/stats/growth/growth-released-structures). CryoEM structures typically provide coordinates for large, challenging proteins, as well as large protein complexes [11], allowing the analysis of missense variants for which 3D coordinates were previously not available.

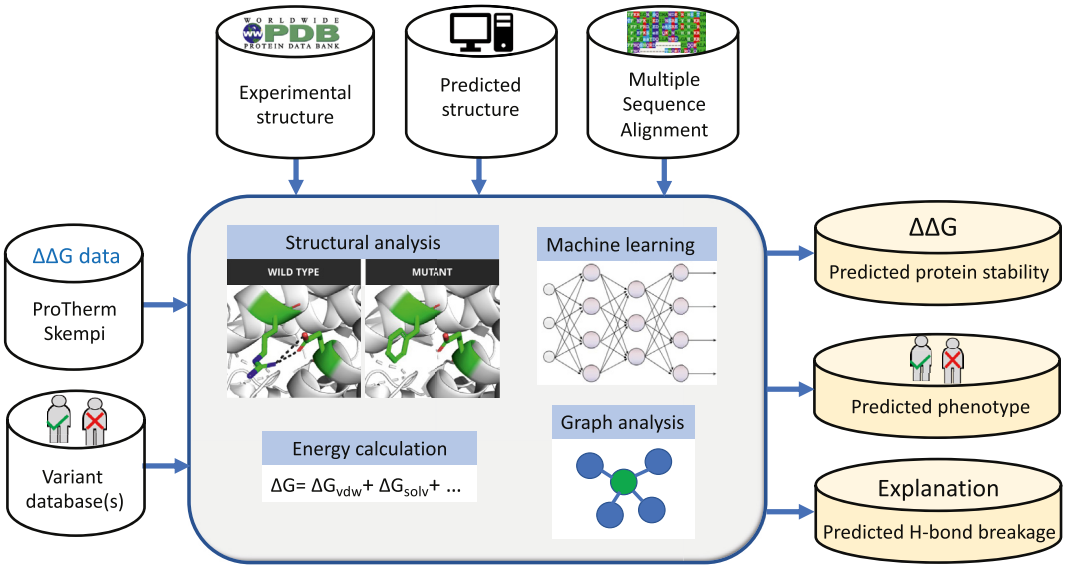
Another major recent breakthrough has been the development of algorithms based on deep learning, such as AlphaFold [12], RoseTTAFold [13], ESMFold [14] and OmegaFold [15]. These can build high quality models, even for difficult targets [16], thus increasing the structural coverage of the proteome and of the variants that can be structurally analysed. Experimental and homology predicted structures cover just over 50% of the residues in the human proteome and using AlphaFold models this can be extended to about 60% [17]. Thus, variants in many residues of the human proteome, including disordered regions, are not directly amenable to structure-based analysis. It is, however, important to remember that the quality of the variant structure prediction is dependent on the quality of the model or experimental structure used. Therefore, the accuracy of the models and/or the resolution of the experimental structures used for variant prediction must be considered when selecting the coordinates. This is particularly important in the case of AlphaFold models that include regions of high and low confidence in the same 3D coordinate file [17]. In particular, the relative positioning of confident substructures, typically domains, can be uncertain as indicated by the PAE (predicted aligned error) matrix provided with the model coordinates, for example, via the European Bioinformatics Institute. If one uses a structure, then it is best to divide it into confident substructures either by visual inspection of the PAE matrix or algorithmically using, for example, the approach reported in the study by Oeffner et al. [18].

Structure-based algorithms for variant prediction

Algorithms that use protein structure data to predict the effect of an amino acid substitution on protein stability and/or function can be broadly divided into those that calculate the difference in free energy (ΔΔG) between the wild type and variant structures and those that use structural features but without providing a ΔΔG (Figure 1). Table 1 presents some of the most popular methods, many of which are available from a web server. Molecular dynamics and free energy perturbation studies can yield valuable assessments of ΔΔG [19] but require expertise in running these approaches. Accordingly, we will now focus on resources readily accessible to the community, particularly via web servers.

Machine learning strategies, including decision trees, support vector machine and neural networks, have been widely used to develop variant predictors. Most methods have been trained using thermodynamic databases of experimentally measured changes in free energy in proteins and their engineered mutants, such as ProTherm [20] and SKEMPI [21]. Over the years, these databases have been greatly enhanced and the recently released ProthermDB [22] and SKEMPI 2.0 [23] contains 84% and 133% more data, respectively, compared to their previous versions, thus representing an invaluable resource for the training of future algorithms. Additional thermodynamic databases that can be used for the development of energy-based methods include MPTherm for membrane proteins [24],

Figure 1



Diversity of approaches adopted by structure-based variant effect predictors. The figure illustrates the main sources of information and strategies used to develop variant predictors.

Table 1

Resources for variant interpretation utilizing information from protein structure.

Predictor	3D coordinates file accepted (either PDB Id or model structure)	ΔΔG	Website	Reference
AUTO-MUTE 2.0	Y	Y	http://binf.gmu.edu/automute/	[74]
BorodaTM	Y	N	https://www.iitm.ac.in/bioinfo/MutHTP/boroda.php	[40]
CUPSAT	Y	Y	http://cupsat.tu-bs.de/	[75]
DDGun3D	Y	Y	https://folding.biofold.org/ddgun/	[76]
Dynamut2.0	Y	Y	https://biosig.lab.uq.edu.au/dynamut2/	[77]
FlexddG	Y	Y	https://rosie.rosettacommons.org	[54]
FoldX	Y	Y	https://foldxsuite.crg.eu/	[78]
HOPE	N	N	https://www3.cmbi.umcn.nl/hope/	[33]
I-mutant2.0	Y	Y	https://folding.biofold.org/i-mutant/i-mutant2.0.html	[79]
INPS-3D	Y	Y	https://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D	[80]
Maestro	Y	Y	https://pbwww.services.came.sbg.ac.at/maestro/web	[81]
mCSM-PPI	Y	Y	https://biosig.lab.uq.edu.au/mcsm_ppi2/	[53]
mCSM-TM	Y	Y	https://biosig.lab.uq.edu.au/mcsm_membrane/	[45]
Missense3D	Y	N	http://missense3d.bc.ic.ac.uk/	[34]
Missense3D-PPI	Y	N	http://missense3d.bc.ic.ac.uk/	[55]
MutPred2	N	N	http://mutpred.mutdb.org/	[31]
PackPred	Y	N	http://cospi.iiserpune.ac.in/packpred/	[82]
PolyPhen2	N	N	http://genetics.bwh.harvard.edu/pph2/	[83]
PMut	Y	N	https://mmb.irbbarcelona.org/PMut/	[30]
PopMusic 2.0	Y	Y	https://soft.dezyme.com	[84]
PRECOGx	N	N	https://precogx.bioinfolab.sns.it/	[46]
Rosetta	Y	Y	https://rosie.rosettacommons.org	[85]
SAAPdap/SAAPpred	N	N	http://www.bioinf.org.uk/mutations/saapdap/	[35]
SDM	Y	Y	http://marid.bioc.cam.ac.uk/sdm2	[86]
SNAP2	N	N	https://rostlab.org/services/snap2web/	[87]
SNPMuSiC	Y	N	https://soft.dezyme.com	[88]
SNPs&GO^{3d}	Y	N	https://snps.biofold.org/snps-and-go/snps-and-go-3d.html	[89]
StructMan	N	N	https://structman.mpi-inf.mpg.de/upload.php	[90]
STRUM	Y	Y	https://zhanggroup.org/STRUM/	[91]

This list presents some of the most popular methods and is not exhaustive.

ProNAB for protein-nucleic acids complexes [25], Platinum for protein-ligand complexes [26] and PINT [27], dbMPIKT [28] and PROXiMATE [29] for protein−protein interactions.

3D protein structure data have also been widely used for non-energy based variant prediction algorithms. In most cases, the 3D structural features that characterize the residue harbouring the variant or its local environment, such as surface accessibility, hydrophobicity, etc., are combined with sequence based features, for example, PolyPhen2, PMut [30], MutPred2 [31] and VIPUR [32]. These predictors generally calculate the probability of a variant being damaging but do not return information on the mechanism by which the variant affects the phenotype. This information is instead provided by methods that use 3D structure coordinates to perform an in-depth atom-based study of the effect of a missense variant, e.g. HOPE [33], Missense3D [34] and SAAPdap/SAAPpred [35]. These predictors provide information on the structural damage, e.g breakage of a cysteine bond or a steric clash, thus providing the user with information on the mechanism by which a variant may disrupt protein folding and/or function. In the case of SAAPDab/SAAPpred and HOPE information on sequence conservation is also included in the variant analysis [35]. Some sequence-based methods have been shown to have a high number of false positives and, hence, a low precision score [36] and it has been suggested that using structural data can improve variant prediction performance. This has recently been demonstrated in VIPUR, which combines sequence and structure features in a logistic regression classifier and has an increased specificity and performance compared to sequence-only methods [32]. Nevertheless, it is worth noting that, at least in some cases, developers may favour a high sensitivity at the expenses of low specificity and the latter is therefore not necessarily a direct result of the sequence-based methodology used. Recent tools, such as E-SNP&GO, which is based on protein language models, shows similarly high sensitivity and precision [37].

Pre-calculated structural predictions for large sets of known, publicly available human variants have been made freely available from databases, such as Missense3D-DB [38] (also available via the DECIPHER website [39]), BorodaTM [40], PMut repository

[30] and PDBe-KB [41]. Moreover, visualization of variants within a 3D structure is also available from resources such as COSMIC3D [42], TopoSNP2 [43] and PhyreRisk [44] which allows to study the local environment surrounding the residue harbouring the variant of interest.

In addition to considering the tertiary structure of globular proteins, groups are now developing structure-based algorithms tailored to variants occurring in transmembrane (TM) regions and protein−protein interfaces (PPIs). Several genetic variants occur within the transmembrane regions of single and multi-pass proteins, as well as in drug targets, such as GPCRs, and ion channels. The variant prediction methods so far described have been trained without distinguishing between globular and transmembrane proteins. The membrane lipid bilayer environment is fundamentally different to the aqueous one, thus requiring dedicated algorithms. However, to date, methods that use structures and have been specifically designed to predict the effect of variants occurring in transmembrane proteins are scarce, possibly because of the paucity of experimentally determined structures of transmembrane regions/proteins. mCSM-TM [45], BorodaTM [40], and PRECOGx [46] and our recently developed Missense3D-TM (soon to be available from the Missense3D homepage at http://missense3d.bc.ic.ac.uk/) are dedicated resources for TM proteins and use structural information to predict the effect of TM variants. mCSM-TM calculates the effect of a variant on TM protein stability by calculating the energy change using coordinates selected by the user, whereas BorodaTM offers precalculated predictions only for variants occurring in a TM region with an experimentally solved structure. PRECOGx is specifically designed for variants in GPCRs and uses sequence and structure features. The best coordinates are automatically selected by the machine learning algorithm among all PDB entries and AlphaFold models. Missense3D-TM predicts the effect of variants within transmembrane regions and allows the user to select experimental structures and/or 3D models.

The algorithms described above assess the effect of a variant using the 3D coordinates of a protein single chain and are therefore not tailored to analyse variants at PPI. PPIs are enriched in disease-causing variants [47,48] and prediction and characterisation of PPI variants can help to assess their effect on biological pathways. Experimental and modelled 3D coordinates of complexes are publicly available from databases, such as PDB [49], GWYRE [50], PrePPI [51] and Interactome3D [52]. However, currently, there is a limited number of algorithms that use 3D structures of complexes for the analysis of variants at PPI. Freely available resources include mCSM-PPI [53], FlexddG [54] and our recently released Missense3D-PPI [55]. At the time of writing this review, the structural coverage of the protein interactome, especially human, is limited, possibly due to the challenges in solving the structure of large macromolecule complexes. However, a large increase in the number of modelled 3D coordinates of protein complexes is expected to occur in the near future with the use of deep learning approaches [56,57] and this will likely encourage the development of more resources for the analysis of variants at PPI.

In this review, we focus on software that analyse the impact of variants on a protein single chain and on protein−protein interactions. It is worth noting that additional algorithms are available for assessing the effect of variants on protein-RNA, -DNA, -ligand and carbohydrates binding, such as mCSM-NA [58], SAMPDI [59], PremPDI [60], DeepCLIP [61] and PCA-MutPred [62]. However, a description of these methods is beyond the scope of this review.

Challenges in the development and benchmarking of bioinformatics resources

Several groups, for example, the study by Livesey et al. [4] have highlighted the challenges in developing and benchmarking variant predictors and noted the widely different accuracies obtained when a particular algorithm is applied to different data sets. Typically, there will be training and testing data sets derived from a particular larger data collection. The first problem is the errors in the assignment of either the ΔΔG for a variant or whether it is benign or pathogenic. If ΔΔG data are used and the aim is to predict pathogenicity, then a cut off value for the change in ΔΔG must be identified. Next is the balance of the data set, typically between benign and pathogenic. Biases in training and evaluation can result from heavily skewed data sets. Some proteins, possibly due to their functional role, are enriched in pathogenic variants whereas other proteins in benign variants [63]. It is essential therefore not to have the same protein in the training and testing datasets. In addition, ideally one should not have proteins from the same homologous superfamily in the training and testing datasets and this is particularly important in a structure-based approach since structure is more conserved than sequence. However, strict adhesion to the requirement to remove homologues can cause problems due to the reduction in the number of training/testing data and sometimes a compromise removing close homologues is required. As algorithms become more sophisticated, for example using deep learning, there is an increased opportunity for memorisation between training and testing and these confounding factors are likely to become increasingly important to address to obtain the best evaluation of accuracy. Given these difficulties, the CAGI (Critical Assessment of Genome Interpretation) provides a route to identify the most accurate approaches [64−66]. The expansion in available protein

structures, both experimental and predicted, should facilitate a large-scale comparative evaluation of structure-based methods for variant prediction.

Another limitation of structure-based approaches is that nearly all the methods, particularly those readily accessible to the community via web servers, employ the fixed backbone approximation and only model the side-chain conformation of the variant. However, it has been shown that observing an RMSD between equivalenced Cα atoms in pairs of crystal structures varying by one amino acid is rarely greater than 1 Å [67]. This suggests that although a limitation, the fixed backbone approximation is often effective in modelling the impact of a missense variant as major structural rearrangements to accommodate the substitution rarely occur.

Future direction and concluding remarks

The remarkable advance in protein structure prediction obtained by AlphaFold has highlighted the value of advanced machine learning approaches in protein bioinformatics [68]. However, a small-scale study suggests that one cannot assess the impact of a variant by running AlphaFold on the wild-type and the variant and assessing if there is a conformational change [69]. This is in keeping with the previous status that structure prediction programs, particularly those that are template based, should not be used to assess the impact of a single missense variant [8].

Accordingly, groups are now developing algorithms specifically designed as variant predictors employing sequence information using methods, such as deep generative models [70], graph attention neural networks [71] and deep residual networks [72]. Recently, a preprint from the Baker group reported enhanced prediction of the mutational effects as the result of mutational scanning incorporating both sequence and structural information using a version of RoseTTAFold [73]. The challenge will be to develop novel deep learning approach based on modelling protein structure to yield enhanced variant predictions together with human comprehensible explanations for the predicted phenotype in terms of protein structure, function and interactions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the Wellcome Trust under grant 218242/Z/19/Z to MJES and AD and BBSRC grant BB/P023959/1, BB/T010487/1, BB/V018558/1 to MJES. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Zeng Z, Bromberg Y: **Predicting functional effects of synonymous variants: a systematic review and perspectives**. *Front Genet* 2019, **10**:914.

2. Ganakammal SR, Alexov E: **Evaluation of performance of leading algorithms for variant pathogenicity predictions and designing a combinatory predictor method: application to Rett syndrome variants**. *PeerJ* 2019, **7**, e8106.

3. Marabotti A, Scafuri B, Facchiano A: **Predicting the stability of mutant proteins by computational approaches: an overview**. *Briefings Bioinf* 2021, **22**:bbaa074.

4. * Livesey BJ, Marsh JA: **Interpreting protein variant effects with computational predictors and deep mutational scanning**. *Dis Model Mech* 2022, **15**, dmm049510.
This is a recent review emphasizing the challenges in the effective evaluation of variant prediction

5. Ellard S, Baple EL, Callaway A, Berry I, Forrester N, Turnbull C, Owens M, Eccles DM, Abbs S: *ACGS best practice guidelines for variant classification in rare disease 2020*. 2021.

6. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A: **The structural coverage of the human proteome before and after AlphaFold**. *PLoS Comput Biol* 2022, **18**, e1009818.

7. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: **The I-TASSER Suite: protein structure and function prediction**. *Nat Methods* 2015, **12**:7–8.

8. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE: **The Phyre2 web portal for protein modeling, prediction and analysis**. *Nat Protoc* 2015, **10**:845–858.

9. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, *et al.*: **SWISS-MODEL: homology modelling of protein structures and complexes**. *Nucleic Acids Res* 2018, **46**:W296–W303.

10. Kuhlman B, Bradley P: **Advances in protein structure prediction and design**. *Nat Rev Mol Cell Biol* 2019, **20**:681–697.

11. Murata K, Wolf M: **Cryo-electron microscopy for structural analysis of dynamic biological macromolecules**. *Biochim Biophys Acta Gen Subj* 2018, **1862**:324–334.

12. ** Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, https://doi.org/10.1038/s41586-021-03819-2.
This paper describes AlphaFold, a major breakthrough in protein structure prediction yielding accurate model

13. * Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**:871–876.
This paper reports a new powerful deep learning method for protein structure prediction

14. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, *et al.*: **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science* 2023, **379**:1123–1130.

15. Ruidong W, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B, *et al.*: *High-resolution de novo structure prediction from primary sequence*. 2022, https://doi.org/10.1101/2022.07.21.500999.

16. Kinch LN, Pei J, Kryshtafovych A, Schaeffer RD, Grishin NV: **Topology evaluation of models for difficult targets in the 14th**

round of the critical assessment of protein structure prediction (CASP14). *Proteins* 2021, **89**:1673–1686.

17. * David A, Islam S, Tankhilevich E, Sternberg MJE: **The AlphaFold database of protein structures: a biologist's guide**. *J Mol Biol* 2022, **434**, 167336.
This paper highlights opportunities and limitations of using AlphaFold models

18. Oeffner RD, Croll TI, Millán C, Poon BK, Schlicksup CJ, Read RJ, Terwilliger TC: **Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE**. *Acta Crystallogr D Struct Biol* 2022, **78**:1303–1314.

19. Steinbrecher T, Zhu C, Wang L, Abel R, Negron C, Pearlman D, Feyfant E, Duan J, Sherman W: **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* 2017, **429**:948–963.

20. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions**. *Nucleic Acids Res* 2006, **34**:D204–D206.

21. Moal IH, Fernández-Recio J: **SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models**. *Bioinformatics* 2012, **28**: 2600–2607.

22. ** Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM: **ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years**. *Nucleic Acids Res* 2021, **49**: D420–D424.
This paper reports a major update to the ProTherm database, a widely used resource for the development of variant predictors

23. Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH: **SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation**. *Bioinformatics* 2019, **35**: 462–469.

24. Kulandaisamy A, Sakthivel R, Gromiha MM: **MPTherm: database for membrane protein thermodynamics for understanding folding and stability**. *Briefings Bioinf* 2021, **22**: 2119–2125.

25. Harini K, Srivastava A, Kulandaisamy A, Gromiha MM: **ProNAB: database for binding affinities of protein-nucleic acid complexes and their mutants**. *Nucleic Acids Res* 2022, **50**: D1528–D1534.

26. Pires DEV, Blundell TL, Ascher DB: **Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes**. *Nucleic Acids Res* 2015, **43**: D387–D391.

27. Kumar MDS, Gromiha MM: **PINT: protein-protein interactions thermodynamic database**. *Nucleic Acids Res* 2006, **34**: D195–D198.

28. Liu Q, Chen P, Wang B, Zhang J, Li J: **dbMPIKT: a database of kinetic and thermodynamic mutant protein interactions**. *BMC Bioinf* 2018, **19**:455.

29. Jemimah S, Yugandhar K, Michael Gromiha M: **PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics**. *Bioinformatics* 2017, **33**:2787–2788.

30. López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL: **PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update**. *Nucleic Acids Res* 2017, **45**: W222–W228.

31. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM, *et al.*: **Inferring the molecular and phenotypic impact of amino acid variants with MutPred2**. *Nat Commun* 2020, **11**:5918.

32. Baugh EH, Simmons-Edler R, Müller CL, Alford RF, Volfovsky N, Lash AE, Bonneau R: **Robust classification of protein variation using structural modelling and large-scale data integration**. *Nucleic Acids Res* 2016, **44**:2501–2513.

33. Venselaar H, Te Beek TAH, Kuipers RKP, Hekkelman ML, Vriend G: **Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces**. *BMC Bioinf* 2010, **11**:548.

34. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE: **Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?** *J Mol Biol* 2019, **431**:2197–2212.

35. Al-Numair NS, Martin ACR: **The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations**. *BMC Genom* 2013, **14**(Suppl 3):S4.

36. Niroula A, Vihinen M: **How good are pathogenicity predictors in detecting benign variants?** *PLoS Comput Biol* 2019, **15**, e1006481.

37. Manfredi M, Savojardo C, Martelli PL, Casadio R: **E-SNPs&GO: embedding of protein sequence and function improves the annotation of human pathogenic variants**. *Bioinformatics* 2022, **38**:5168–5174.

38. Khanna T, Hanna G, Sternberg MJE, David A: **Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants**. *Hum Genet* 2021, **140**: 805–812.

39. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP: **DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources**. *Am J Hum Genet* 2009, **84**: 524–533.

40. Popov P, Bizin I, Gromiha M, K A, Frishman D: **Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure**. *PLoS One* 2019, **14**, e0219452.

41. * **PDBe-KB consortium: PDBe-KB: collaboratively defining the biological context of structural data**. *Nucleic Acids Res* 2022, **50**:D534–D542.
PDBe-KB is a major collaborative resource which presents extensive structural data including variant predictions

42. Jubb HC, Saini HK, Verdonk ML, Forbes SA: **COSMIC-3D provides structural perspectives on cancer genetics for drug discovery**. *Nat Genet* 2018, **50**:1200–1202.

43. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J: **topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association**. *Nucleic Acids Res* 2004, **32**:D520–D522.

44. Ofoegbu TC, David A, Kelley LA, Mezulis S, Islam SA, Mersmann SF, Strömich L, Vakser IA, Houlston RS, Sternberg MJE: **PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants**. *J Mol Biol* 2019, **431**: 2460–2466.

45. Pires DEV, Rodrigues CHM, Ascher DB: **mCSM-membrane: predicting the effects of mutations on transmembrane proteins**. *Nucleic Acids Res* 2020, **48**:W147–W153.

46. Matic M, Singh G, Carli F, Oliveira Rosa ND, Miglionico P, Magni L, Gutkind JS, Russell RB, Inoue A, Raimondi F: **PRECOGx: exploring GPCR signaling mechanisms with deep protein representations**. *Nucleic Acids Res* 2022, **50**: W598–W610.

47. David A, Razali R, Wass MN, Sternberg MJE: **Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs**. *Hum Mutat* 2012, **33**:359–363.

48. David A, Sternberg MJE: **The contribution of missense mutations in core and rim residues of protein-protein interfaces to human disease**. *J Mol Biol* 2015, https://doi.org/10.1016/j.jmb.2015.07.004.

49. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, *et al.*: **RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy**. *Nucleic Acids Res* 2019, **47**:D464–D474.

50. Malladi S, Powell HR, David A, Islam SA, Copeland MM, Kundrotas PJ, Sternberg MJE, Vakser IA: **GWYRE: A resource**

for mapping variants onto experimental and modeled structures of human protein complexes. *J Mol Biol* 2022, **434**, 167608.

51. Garzón JI, Deng L, Murray D, Shapira S, Petrey D, Honig B: **A computational interactome and functional annotation for the human proteome**. *Elife* 2016, **5**, e18715.

52. Mosca R, Céol A, Aloy P: **Interactome3D: adding structural details to protein networks**. *Nat Methods* 2013, **10**:47–53.

53. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB: **mCSM-PPI2: predicting the effects of mutations on protein-protein interactions**. *Nucleic Acids Res* 2019, **47**:W338–W344.

54. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, Kortemme T: **Flex ddG: rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation**. *J Phys Chem B* 2018, **122**:5389–5399.

55. Pennica C, Hanna G, Islam SA, Sternberg MJE, David A: **Missense3D-PPI: a web resource to predict the impact of missense variants at protein interfaces using 3D structural data**. *JMB (J Mol Biol)* 2023, https://doi.org/10.1016/j.jmb.2023.168060.

56. Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A, *et al.*: **Towards a structurally resolved human protein interaction network**. *Nat Struct Mol Biol* 2023, https://doi.org/10.1038/s41594-022-00910-8.

57. Bryant P, Pozzati G, Elofsson A: **Improved prediction of protein-protein interactions using AlphaFold2**. *Nat Commun* 2022, **13**:1265.

58. Pires DEV, Ascher DB: **mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions**. *Nucleic Acids Res* 2017, **45**:W241–W246.

59. Peng Y, Sun L, Jia Z, Li L, Alexov E: **Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver**. *Bioinformatics* 2018, **34**:779–786.

60. Zhang N, Chen Y, Zhao F, Yang Q, Simonetti FL, Li M: **PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions**. *PLoS Comput Biol* 2018, **14**, e1006615.

61. Grønning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, Hansen MB, Hartung A-M, Baumbach J, Andresen BS: **DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning**. *Nucleic Acids Res* 2020, **48**:7099–7118.

62. Siva Shanmugam NR, Veluraja K, Michael Gromiha M: **PCA-MutPred: prediction of binding free energy change upon missense mutation in protein-carbohydrate complexes**. *J Mol Biol* 2022, **434**, 167526.

63. Alhuzimi E, Leal LG, Sternberg MJE, David A: **Properties of human genes guided by their enrichment in rare and common variants**. *Hum Mutat* 2018, **39**:365–370.

64. Andreoletti G, Pal LR, Moult J, Brenner SE: **Reports from the fifth edition of CAGI: the critical assessment of Genome interpretation**. *Hum Mutat* 2019, **40**:1197–1201.

65. * The Critical Assessment of Genome Interpretation Consortium TCA of GIC: *CAGI, the critical assessment of Genome interpretation, establishes progress and prospects for computational genetic variant interpretation methods*. 2022, https://doi.org/10.48550/arXiv.2205.05897.
This paper presents the report from the last comparative assessment of variant predictors

66. Savojardo C, Petrosino M, Babbi G, Bovo S, Corbi-Verge C, Casadio R, Fariselli P, Folkman L, Garg A, Karimi M, *et al.*: **Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge**. *Hum Mutat* 2019, **40**:1392–1399.

67. Arodź T, Płonka PM: **Effects of point mutations on protein structure are nonexponentially distributed**. *Proteins* 2012, **80**:1780–1790.

68. ** Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, *et al.*: **A structural biology community assessment of AlphaFold2 applications**. *Nat Struct Mol Biol* 2022, **29**:1056–1067.
This paper provides an extensive analysis of the use of AlphaFold models in biology including variant prediction

69. * Buel GR, Walters KJ: **Can AlphaFold2 predict the impact of missense mutations on structure?** *Nat Struct Mol Biol* 2022, **29**:1–2.
This paper highlights that running AlphaFold on the wild type and variant sequence is not a viable strategy to assess whether a variant is damaging

70. Riesselman AJ, Ingraham JB, Marks DS: **Deep generative models of genetic variation capture the effects of mutations**. *Nat Methods* 2018, **15**:816–822.

71. Zhang H, Xu MS, Fan X, Chung WK, Shen Y: **Predicting functional effect of missense variants using graph attention neural networks**. *Nat Mach Intell* 2022, **4**:1017–1028.

72. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y: **MVP predicts the pathogenicity of missense variants by deep learning**. *Nat Commun* 2021, **12**:510.

73. Mansoor S, Baek M, Juergens D, Watson JL, Baker D: *Accurate mutation effect prediction using RoseTTAFold*. 2022, https://doi.org/10.1101/2022.11.04.515218.

74. Masso M: **Vaisman II: auto-mute 2.0: a portable framework with enhanced capabilities for predicting protein functional consequences upon mutation**. *Adv Bioinformatics* 2014, **2014**:7, https://doi.org/10.1155/2014/278385. Article ID 278385.

75. Parthiban V, Gromiha MM, Schomburg D: **CUPSAT: prediction of protein stability upon point mutations**. *Nucleic Acids Res* 2006, **34**:W239–W242.

76. Montanucci L, Capriotti E, Birolo G, Benevenuta S, Pancotti C, Lal D, Fariselli P: **DDGun: an untrained predictor of protein stability changes upon amino acid variants**. *Nucleic Acids Res* 2022, https://doi.org/10.1093/nar/gkac325.

77. Rodrigues CHM, Pires DEV, Ascher DB: **DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations**. *Protein Sci* 2021, **30**:60–69.

78. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field**. *Nucleic Acids Res* 2005, **33**:W382–W388.

79. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure**. *Nucleic Acids Res* 2005, **33**:W306–W310.

80. Savojardo C, Fariselli P, Martelli PL, Casadio R: **INPS-MD: a web server to predict stability of protein variants from sequence and structure**. *Bioinformatics* 2016, **32**:2542–2544.

81. Laimer J, Hiebl-Flach J, Lengauer D, Lackner P: **MAESTROweb: a web server for structure-based protein stability prediction**. *Bioinformatics* 2016, **32**:1414–1416.

82. Tan KP, Kanitkar TR, Kwoh CK, Madhusudhan MS, Packpred: **Predicting the functional effect of missense mutations**. *Front Mol Biosci* 2021, **8**, 646288.

83. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2**. *Curr Protoc Hum Genet* 2013. Chapter 7:Unit7.20.

84. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M: **PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality**. *BMC Bioinf* 2011, **12**:151.

85. Kellogg EH, Leaver-Fay A, Baker D: **Role of conformational sampling in computing mutation-induced changes in protein structure and stability**. *Proteins* 2011, **79**:830–838.

86. Pandurangan AP, Blundell TL: **Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning**. *Protein Sci* 2020, **29**:247–257.

87. Hecht M, Bromberg Y, Rost B: **Better prediction of functional effects for sequence variants**. *BMC Genom* 2015, **16**(Suppl 8): S1.

88. Ancien F, Pucci F, Godfroid M, Rooman M: **Prediction and interpretation of deleterious coding variants in terms of protein structural stability**. *Sci Rep* 2018, **8**:4480.

89. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R: **WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation**. *BMC Genom* 2013, **14**(Suppl 3):S6.

90. Gress A, Ramensky V, Büch J, Keller A, Kalinina OV: **StructMAn: annotation of single-nucleotide polymorphisms in the structural context**. *Nucleic Acids Res* 2016, **44**:W463–W468.

91. Quan L, Lv Q, Zhang Y: **STRUM: structure-based prediction of protein stability changes upon single-point mutation**. *Bioinformatics* 2016, **32**:2936–2946.