Computational and Structural Biotechnology Journal 23 (2024) 446–451

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Mini-Review

Structural and functional prediction, evaluation, and validation in the post-sequencing era

Chang Li [a,b], Yixuan Luo [c], Yibo Xie [d], Zaifeng Zhang [b], Ye Liu [b], Lihui Zou [b,*], Fei Xiao [a,b,c,*]

[a] Clinical Biobank, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China
[b] The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China
[c] Beijing Normal University, Beijing, China
[d] Information Center, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

ARTICLE INFO

Keywords:
Missense variants
Artificial intelligence
Protein structure
Post-sequencing era
Clinical interpretation

ABSTRACT

The surge of genome sequencing data has underlined substantial genetic variants of uncertain significance (VUS). The decryption of VUS discovered by sequencing poses a major challenge in the post-sequencing era. Although experimental assays have progressed in classifying VUS, only a tiny fraction of the human genes have been explored experimentally. Thus, it is urgently needed to generate state-of-the-art functional predictors of VUS in silico. Artificial intelligence (AI) is an invaluable tool to assist in the identification of VUS with high efficiency and accuracy. An increasing number of studies indicate that AI has brought an exciting acceleration in the interpretation of VUS, and our group has already used AI to develop protein structure-based prediction models. In this review, we provide an overview of the previous research on AI-based prediction of missense variants, and elucidate the challenges and opportunities for protein structure-based variant prediction in the post-sequencing era.

1. Introduction

With the surge of high-throughput sequencing technologies, a number of large-scale genetic projects have been launched by several countries[1]. The exponentially increasing genome sequencing data have revealed that there are extensive genetic variations within human populations[2,3]. The vast majority of these variants are defined as variants of uncertain significance (VUS) because their phenotypic consequences and clinical significance are unknown. Therefore, in the post-sequencing era, the main task of genomic research has changed to converting the rich genetic data into useful information. This is a particularly acute problem for missense single-nucleotide variants (SNVs), which account for the majority of VUS. In Clinvar, one of the most widely used genetic databases, there are about 5% variants with category or clinical significance conflicts (e.g. benign versus pathogenic) among the variants with at least 2 submissions[4]. According to the commonly sequenced genes, the proportions of variants with uncertain significance or conflicting information are even higher (e.g. BRCA1 52% uncertain and 3% conflicting, as of September 2023)[5].

Various technologies that can assess the functional effects of mutations have emerged[6–8], but only a tiny fraction of the human disease-related genes have been explored. Experimental approaches can be quite expensive and time-consuming when applied to all the remaining VUS, which limits the clinical application of genetic information and the realization of a more efficient discovery tool. Therefore, generating state-of-the-art computational functional predictors for VUS remains crucial.

Evidence suggests that in silico analysis can accelerate the clinical interpretation of VUS[9–11]. State-of-the-art missense variant functional predictors have been developed by artificial intelligence (AI) [12,

Abbreviations: VUS, variants of uncertain significance; SNVs, single-nucleotide variants; AI, artificial intelligence; NMR, nuclear magnetic resonance; cryo-EM, cryo-electron microscopy; PDB, Protein Data Bank; HM, homology modeling; pLDDT, predicted local-distance difference test; MSA, multiple sequence alignment; ΔΔG, differences in free energy; WT, wild-type; AANs, amino acid networks; MD, molecular dynamics.

* Corresponding authors at: The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China.
E-mail addresses: zoulihui4371@bjhmoh.cn (L. Zou), xiaofei3965@bjhmoh.cn (F. Xiao).

https://doi.org/10.1016/j.csbj.2023.12.031
Received 20 October 2023; Received in revised form 20 December 2023; Accepted 22 December 2023
Available online 25 December 2023

13]. Notably, breakthroughs in protein structure prediction based on deep learning, such as AlphaFold2[14] and RoseTTAFold[15], have improved AI variant predictors by adding information of protein tertiary structures[16]. Here, we review the recent advances in AI prediction for missense variants, particularly protein structure-based methods, and elucidate the challenges and opportunities for this promising direction.

2. Recent developments in the determination of protein tertiary structures

Proteins are polymers of amino acids linked by peptide bonds that fold into specific spatial configurations. The functions of proteins are closely associated with their atoms and amino acid coordinates (tertiary structure), particularly the active domains. Protein structures have been recommended as a strong source of information in the Guidelines for Variant Classification[17]. Researchers have developed several experimental techniques to determine protein structures, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) [18–20]. While the number of structures stored in the Protein Data Bank (PDB)[21] has steadily increased to 210,000 (as of September 2023), it covers only about 17% of all human protein residues[22].

Since there are a large number of homologous sequences for the proteins that actually exist in nature, the relationship between the amino acid sequence and the structure of a protein follows a certain regularity. Therefore, various computational tools have been developed to predict 3D protein structures. Protein structure prediction methods commonly fall into two categories: theoretical analysis and statistical analysis. While the traditional prediction techniques, such as homology modeling (HM)[23], are relatively mature, recent advances in AI have initiated new concepts to enhance the quality and proteomic coverage of protein structure models.

In 2020, the AlphaFold2[14] project stood out in the challenging 14th Critical Assessment of Protein Structure Prediction (CASP14), followed by RoseTTAFold[15] and ESMFold[24]. On this basis, several multiple sequence alignment (MSA)-free protein structure prediction methods (e.g. OmegaFold[25] and HelixFold-Single[26]), filled the gap in structure prediction using evolutionary information. Compared with the widespread techniques (e.g. HM), these AI-based algorithms exhibit higher availability for several reasons: (i) there is no need for a close homolog solved experimentally; (ii) the whole-protein structure can be obtained; and (iii) the AI algorithms provide unlimited potential for refining the structures[16]. This breakthrough addressed the significant economic and temporal costs in determining protein structures, thus obtaining enough variant protein structures for structural analyses. Moreover, some human proteins possess intrinsically disordered regions that are not conducive to structural analysis. Therefore, the quality of models predicting protein structure varies across different structural domains. AlphaFold is considered to be particularly advantageous in such situations as it provides the predicted local-distance difference test (pLDDT) to reflect the confidence level of each atomic coordinate. It is preferable to use high-confidence regions to develop structure-based variant prediction models. Alternatively, visual inspection algorithms, such as phenix.process_predicted_model and ISOLDE may also be utilized[27].

3. Missense variant prediction using AI

Table 1 presents various state-of-the-art computational methods, many of which can be accessed through a web server. AI-based variant effect predictors based on primary amino acid sequences have emerged in the past few years. There are two main computational approaches. The first approach, supervised training, relies on clinical labels of pathogenic versus benign variants[28–31]. Given the influencing factors including label bias, label sparsity, label noise, and data leakage, this approach causes inflated prediction accuracy in the specific testing scenarios. The second approach, which involves unsupervised models of evolutionary sequences, has significantly contributed to the advancement of predicting the functional effects of variants[12,32,33]. These unsupervised generative models predict variant effects directly from MSA without relying on labels, which grants its theoretical generalizability. However, these sequence-based models have been limited in their ability to address clinical variant interpretation due to the lack of understanding of the protein structures that are more related to function [34].

Due to the non-negligible contribution of 3D structure to protein

Table 1
State-of-the-art computational methods for variant interpretation.

Predictor	Structure accepted	Predicted structure accepted	△△G accepted	Website
DEOGEN2[28]	N	N	N	http://deogen2.mutaframe.com/
PERCH[29]	N	N	N	http://BJFengLab.org/
REVEL[30]	N	N	N	https://sites.google.com/site/revelgenomics/
CADD[31]	N	N	N	https://cadd.gs.washington.edu/
EVE[12]	N	N	N	https://evemodel.org/
EVmutation [32]	N	N	N	http://evmutation.org/
SIFT[33]	N	N	N	http://sift-dna.org/sift4g
AUTO-MUTE [41]	Y	Y	Y	http://binf.gmu.edu/automute/
CUPSAT[42]	Y	Y	Y	http://cupsat.tu-bs.de/
DDGun3D[43]	Y	Y	Y	https://folding.biofold.org/ddgun/
Dynamut2.0 [44]	Y	Y	Y	https://biosig.lab.uq.edu.au/dynamut2/
I-Mutant2.0 [45]	Y	Y	Y	https://folding.biofold.org/i-mutant/i-mutant2.0.html/
MutPred2[13]	Y	N	N	http://mutpred.mutdb.org/
PMut[46]	Y	N	N	http://mmb2.pcb.ub.es:8080/PMut/
VIPUR[47]	Y	N	N	https://osf.io/bd2h4/
gMVP[48]	Y	N	N	https://github.com/ShenLab/gMVP/
PolyPhen-2 [49]	Y	N	N	http://genetics.bwh.harvard.edu/pph2/
SNAP2[50]	Y	N	N	https://rostlab.org/services/snap2web/
AlphScore[36]	Y	Y	N	https://github.com/Ax-Sch/AlphScore/
Missense3D [75]	Y	Y	N	http://missense3d.bc.ic.ac.uk/
SNPMuSiC[76]	Y	Y	N	https://soft.dezyme.com/
SNPs&GO^{3d} [77]	Y	Y	N	https://snps.biofold.org/snps-and-go/snps-and-go-3d.html/
vERnet-B[16]	Y	Y	N	https://ai-lab.bjrz.org.cn/vERnet/
AlphaMissense [52]	Y	Y	N	https://console.cloud.google.com/storage/browser/dm_alphamissense/

function, the local structural properties have been widely used to be combined with sequence features. The combination of structural features showed significant improvement in the performance of variant prediction. Protein structure-based learning models can be classified into feature-based machine learning and graph-based deep learning. Classical machine learning methods rely on features that are manually designed by experts, such as transmembrane helices signal peptides or other motifs[35,36]. Deep learning methods usually consider structural data as graphs or images and extract task-specific features directly using convolutional neural networks (CNNs)[37,38] or graph convolutional networks (GCNs)[39,40]. Structure-based algorithms for predicting the functional effect of one amino acid substitution (Fig. 1) can be divided into two categories of methods based on their utilization of free energy. Energy-based methods utilize experimentally-measured differences in free energy (ΔΔG) between wild-type (WT) and variant structures to train prediction models[41–45], whereas non-energy-based methods directly use structural features such as hydrophobicity and surface accessibility[13,46–50].

Building upon the increased structural coverage achieved by incorporating AlphaFold models, our group successfully applied the predicted structure models to identify pathogenic missense SNVs[16]. This work demonstrated the potential of utilizing AlphaFold2-predicted protein tertiary structures for rich feature learning, although they had been considered to fall short in predicting the effect of point mutations[51]. Furthermore, AlphaMissense introduced an additional approach derived from AlphaFold2, enabling the development of a proteome-wide variant effect prediction method by leveraging the structural information[52]. AlphaMissense demonstrated an innovative approach to variant effect prediction by incorporating protein structural information which enabled the inclusion of the two key capabilities of AlphaFold2: the high-precision protein structure model and the ability to learn evolutionary constraints based on Evoformer block.

4. Construction of datasets for machine learning

For supervised AI prediction, benchmark datasets are used as information sources for training and testing models. The supervised variant prediction methods follow two broad strategies according to the type of training labels. The first class of methods directly leverage the variant interpretation on human-curated variation databases, which are generated by experimental assays or clinical evidence. As per the basis of classification, these variation benchmark databases can be further divided into effect-specific databases and disease-specific databases. Variants are classified in effect-specific databases according to the protein's specific function, such as stability[53–55], solubility[56], and binding free energy[54,57]. Disease-specific databases represented by Clinvar[5], collect disease-associated phenotypes of variants by clinical data sharing or functional assessments related to pathogenic mechanisms.

The second class of methods avoid using human classification and train with weak labels instead[58,59]. In this strategy, variants are defined based on their frequencies observed in human or other primate species. These approaches[52,59,60] mitigate the impact of biases introduced by human annotation and prevent data leaks between training and testing sets. Specifically, it is supported by the assertion that most of the common variants in primate species are clinically benign in human. Notably, as a result of the incorporation of numerous false labels, these methods still necessitate the use of known labels to assess their actual performance. Table 2 lists the databases widely used in protein effect prediction tasks.

5. Challenges in the development of structure-based variant prediction

Several studies[11], including ours, have observed significant variations in the accuracy of a particular method when being applied to different datasets. The training datasets used in this context are typically

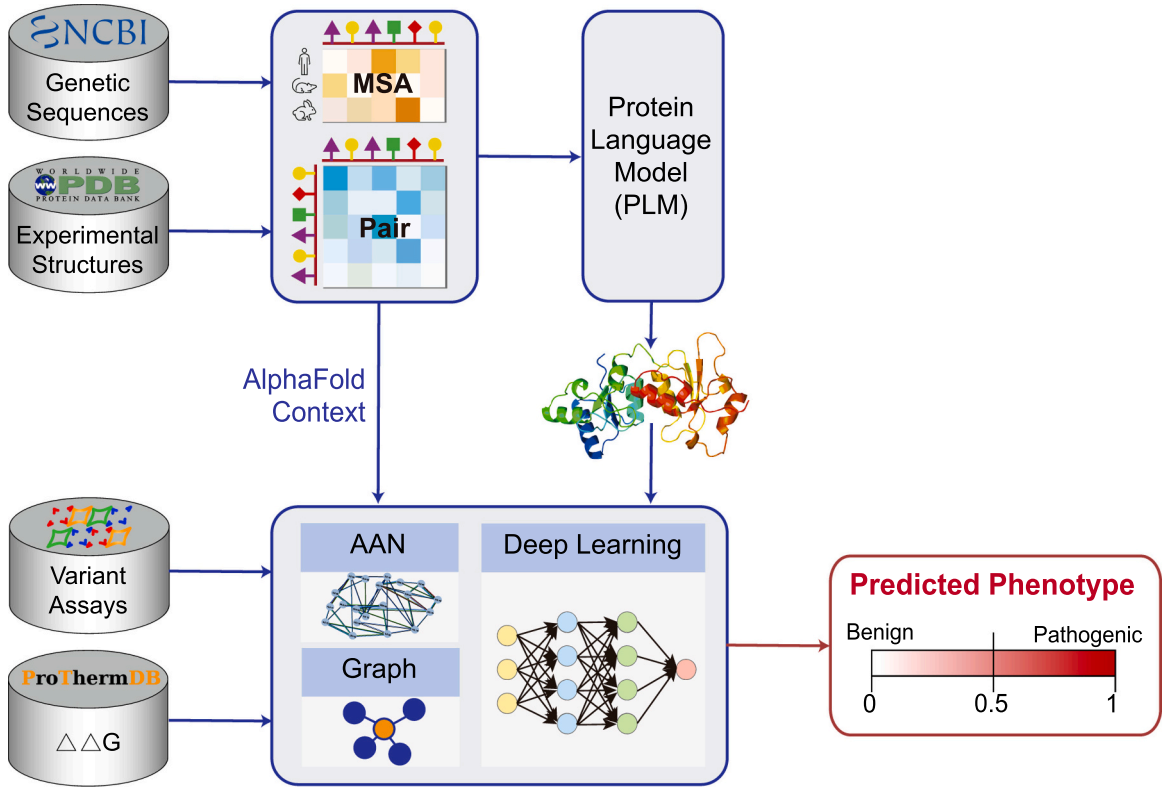

Fig. 1. Diversity of approaches adopted by structure-based variant prediction. The figure illustrates the main sources of information and multiple strategies that were used to develop variant effect predictors.

Table 2
Widely used variation benchmark databases for developing variant effect predictors.

Database	Description	Category	Website
ProTherm[53]	Thermodynamics of protein mutants	Effect-specific	https://web.iitm.ac.in/bioinfo2/prothermdb/
SKEMPI2.0[54]	Protein-protein binding energy, thermodynamics, and kinetics of protein mutations	Effect-specific	https://life.bsc.es/pid/skempi2/
MPTherm[55]	Thermodynamics of membrane protein mutants	Effect-specific	https://www.iitm.ac.in/bioinfo/mptherm/index.php/
Dataset used for PON-SOL[56]	Solubility of amino acid substitution	Effect-specific	http://structure.bmc.lu.se/VariBench/solubility.php/
PROXiMATE [57]	Thermodynamics of protein-protein complex mutations	Effect-specific	http://www.iitm.ac.in/bioinfo/PROXiMATE/
Dataset used for WALTZ-DB [78]	Amyloid forming for variants of hexapeptides	Effect-specific	http://structure.bmc.lu.se/VariBench/amyload1.php/
Nabe[79]	Binding energy of protein-nucleic acid complex	Effect-specific	http://nabe.dendglab.org/
Clinvar[5]	Clinical significance of disease-related gene variants	Disease-specific	https://www.ncbi.nlm.nih.gov/clinvar/
UMD[80]	Mutational landscape and significance across 12 major cancer types	Disease-specific	http://www.umd.be/VHL/
IARC TP53[81]	Phenotypes of TP53 mutants in different human tumors	Disease-specific	http://p53.iarc.fr/
DoCM[82]	Validated cancer-causing mutations	Disease-specific	http://www.docm.info/
OMIM[83]	Overviews of genetic phenotypes for disorders	Disease-specific	https://omim.org/
HGMD[84]	Published gene lesions responsible for human inherited disease	Disease-specific	https://www.hgmd.cf.ac.uk/ac/index.php/
BRCA Exchange [85]	A global resource for variants in BRCA1 and BRCA2	Disease-specific	https://brcaexchange.org/
gnomAD[58]	Allele frequencies of variants among human populations	Frequency	https://gnomad.broadinstitute.org/
Great ape genome sequencing project[59]	Genetic diversity among great ape populations	Frequency	http://biologiaevolutiva.org/greatape/

derived from experimental or clinical databases. Since these databases are regarded as benchmarks, the potential error labels in the benchmark databases arise the first challenge. This is evident through the inconsistent interpretations of several variants across different databases, indicating a need for caution during data interpretation. Besides the basic approach of deleting ambiguous data, our work has carried out a cross-training solution to eliminate the influence of the remaining unfaithful samples[16]. Another commonly used method is to train with weak labels instead of human-curated classification[52,59,60].

Another challenge is the class imbalance of variants in nature, especially between the pathogenic and benign ones. Due to their specific functional roles, some proteins are enriched in benign variants whereas others are enriched in pathogenic variants[61]. However, the skewed datasets can lead to a selection bias toward positive or negative samples. It is therefore essential to incorporate more efficient methods, such as EasyEnsemble[62], SMOTE[63], AdaCost[64], and RUSBoost[65], to address the imbalance by effectively using samples from the majority class. Class imbalance problems can be addressed at the data level, algorithm level, or through hybrid approaches[66]. These methods can be further grouped into the following techniques: re-sampling (under/over-sampling), cost-sensitive learning, and ensemble learning.

Inevitable conflict exists between precision and generalization ability, as highlighted by previous studies[16,52]. Compared to gene-specific techniques such as vERnet-B, generalizable methods like AlphaMissense do not necessitate the training of a separate recognition model for every protein. However, it may result in imprecision in identifying the effects of minor structural alterations caused by the substitution of a single amino acid. Furthermore, conformational changes caused by a mutation can potentially affect protein function via a variety of mechanisms[67], indicating that the same structural alteration may lead to completely opposite phenotypic consequences for different proteins. Therefore, further investigations on these methods are warranted to achieve a better balance between precision and generalization.

A challenge in implementing the structure-based deep learning is the representation of protein structures. These structures consist of polypeptide chains that can be hierarchically organized into primary, secondary, tertiary, and quaternary structures[68]. Due to such complexity, various protein representations can be used for deep learning, including molecular graph[69] and 3D projection[70] based on the protein's original 3D shape (Fig. 2). AlphaFold-derived methods can utilize the AlphaFold context[52], which is AlphaFold's intrinsic understanding of structure, to further learn its relationship with function. Constructing the amino acid networks (AANs)[16] is another mean to provide more detailed structural information, thus enabling the presentation of protein structures for deep learning purposes.

Another limitation of the structure-based approaches is that the existing structure prediction algorithms only provide an optimal static structure, which may not account for conformational changes in proteins resulting from compound binding, complexes, or condensates with various quaternary structures. Although the molecular dynamics (MD) simulation is theoretically appealing, this method remains a huge challenge even for moderately sized proteins[14]. Traditional MD simulations are limited by the computational intractability of algorithms as well as the context dependence of protein stability. Recent studies have demonstrated that more sampling can enhance the prediction of multiple conformational states[71–73]. However, further evidence is needed to establish the reliability and usability of these methods in revealing the conformational landscape. Nevertheless, the combination of AlphaFold2 with molecular dynamics simulation[74] holds great promise for optimizing structure-based variant predictors.

6. Future direction and concluding remarks

Deep learning for structural and functional prediction is becoming a popular direction, as evidenced by the increasing number of related publications over the last few years[75–77]. Despite posing challenges [34] and presenting a range of perspectives, previous studies have shown the feasibility of this approach and provided insights for generating even more precise variant effect learning models. The remarkable advances in protein structure prediction not only highlight the potential value of AI in structural biology but also pave the way for future genetic information research with AI. Structural AI prediction promises to tackle the crucial problems in the post-sequencing era and to enrich the toolbox of precision treatment.

In addition, using well-trained models to predict a batch of random mutations can specify the evolutionary tendency of a protein's function, thereby creating novel variants with the expected function. This approach can expedite drug discovery by reducing the number of wet experiments that validate the function of the selected variants, implying a favorable application of AI-based structure prediction for protein function. Therefore, the ongoing studies are developing novel algorithms specifically for modeling protein structures to enhance the precision of functional prediction, coupled with reliable validation for the

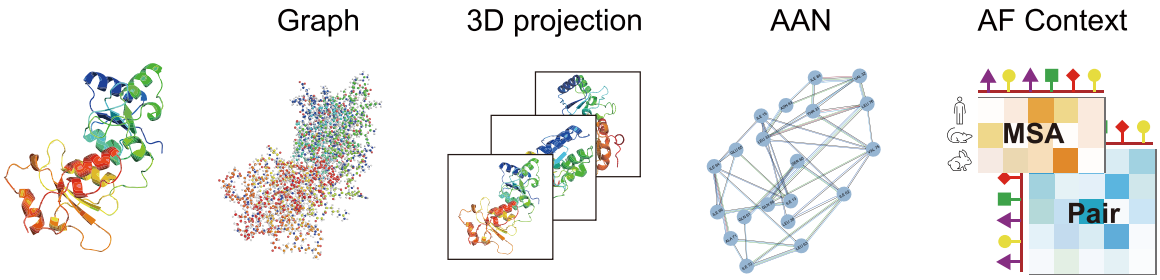

Fig. 2. Representations of protein structure for deep learning. Molecular graph (a graph whose nodes are atoms and edges are bonds), 3D projection (a 3D array reflecting the shapes measured in three directions), AAN (an undirected weighted network whose nodes are amino acids and edges are their non-covalent interactions), and AlphaFold Context (a structural context generated by AlphaFold, with MSA and Pair representations).

newly predicted functional proteins.

Funding

This work was supported by the National High Level Hospital Clinical Research Fund (Grant BJ-2023-077), the National Natural Science Foundation of China (Grant 82372314).

CRediT authorship contribution statement

Fei Xiao: Conceptualization, Writing – review & editing. **Lihui Zou**: Conceptualization, Writing – review & editing. **Chang Li**: Writing – original draft, Writing – review & editing. **Yixuan Luo**: Writing – original draft, Writing – review & editing. **Yibo Xie**: Writing – original draft, Writing – review & editing. **Zaifeng Zhang**: Writing – original draft, Methodology, Visualization. **Ye Liu**: Writing – original draft, Methodology, Visualization.

Declaration of Competing Interest

The authors declare that there are no known competing financial interests.

References

[1] Green ED, Watson JD, Collins FS. Human genome project: twenty-five years of big biology. Nature 2015;526:29–31.
[2] Gudmundsson S, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, et al. Addendum: the mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2021;597:E3–4.
[3] Karczewski KJ, Solomonson M, Chao KR, Goodrich JK, Tiao G, Lu W, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom 2022;2:100168.
[4] Henrie A, Hemphill SE, Ruiz-Schultz N, Cushman B, DiStefano MT, Azzariti D, et al. ClinVar miner: demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. Hum Mutat 2018;39:1051–60.
[5] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48. D835-D44.
[6] Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature 2018;562:217–22.
[7] Fowler DM, Adams DJ, Gloyn AL, Hahn WC, Marks DS, Muffley LA, et al. An atlas of variant effects to understand the genome at nucleotide resolution. Genome Biol 2023;24:147.
[8] Ransburgh DJ, Chiba N, Ishioka C, Toland AE, Parvin JD. Identification of breast tumor mutations in BRCA1 that abolish its function in homologous DNA recombination. Cancer Res 2010;70:988–95.
[9] Zeng Z, Bromberg Y. Predicting functional effects of synonymous variants: a systematic review and perspectives. Front Genet 2019;10:914.
[10] Marabotti A, Scafuri B, Facchiano A. Predicting the stability of mutant proteins by computational approaches: an overview. Brief Bioinform 2021;22.
[11] Livesey BJ, Marsh JA. Interpreting protein variant effects with computational predictors and deep mutational scanning. Dis Model Mech 2022;15.
[12] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021;599:91–5.
[13] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11:5918.
[14] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[15] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373:871–6.
[16] Li C, Zhang L, Zhuo Z, Su F, Li H, Xu S, et al. Artificial intelligence-based recognition for variant pathogenicity of BRCA1 using AlphaFold2-predicted structures. Theranostics 2023;13:391–402.
[17] Beygo J, Buiting K, Ramsden SC, Ellis R, Clayton-Smith J, Kanber D. Update of the EMQN/ACGS best practice guidelines for molecular analysis of Prader-Willi and Angelman syndromes. Eur J Hum Genet 2019;27:1326–40.
[18] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. 3-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 1958;181:662–6.
[19] Nitta R, Imasaki T, Nitta E. Recent progress in structural biology: lessons from our research history. Microsc (Oxf) 2018.
[20] Murata K, Wolf M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochim Biophys Acta Gen Subj 2018;1862:324–34.
[21] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[22] 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.
[23] Hameduh T, Haddad Y, Adam V, Heger Z. Homology modeling in the time of collective and artificial intelligence. Comput Struct Biotechnol J 2020;18:3494–506.
[24] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[25] Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv 2022. 2022.07.21.500999.
[26] Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. Nat Mach Intell 2023;5:1087–96.
[27] Oeffner RD, Croll TI, Millan C, Poon BK, Schlicksup CJ, Read RJ, et al. Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE. Acta Crystallogr D Struct Biol 2022;78:1303–14.
[28] Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res 2017;45. W201-W6.
[29] Feng BJ. PERCH: a unified framework for disease gene prioritization. Hum Mutat 2017;38:243–51.
[30] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 2016;99:877–85.
[31] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47. D886-D94.
[32] Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol 2017;35:128–35.
[33] Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc 2016;11:1–9.
[34] Marsh JA, Teichmann SA. Predicting pathogenic protein variants. Science 2023;381:1284–5.
[35] Cozzetto D, Minneci F, Currant H, Jones DT. FFPred 3: feature-based function prediction for all gene ontology domains. Sci Rep 2016;6:31865.
[36] Schmidt A, Roner S, Mai K, Klinkhammer H, Kircher M, Ludwig KU. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. Bioinformatics 2023;39.
[37] Jimenez J, Doerr S, Martinez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 2017;33:3036–42.
[38] Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ 2018;6:e4750.

[39] Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model 2017;57:1757–72.
[40] Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12:3168.
[41] Masso M, Vaisman II. AUTO-MUTE 2.0: a portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. Adv Bioinforma 2014;2014:278385.
[42] Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006;34:W239–42.
[43] Montanucci L, Capriotti E, Birolo G, Benevenuta S, Pancotti C, Lal D, et al. DDGun: an untrained predictor of protein stability changes upon amino acid variants. Nucleic Acids Res 2022;50. W222-W7.
[44] Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci 2021;30:60–9.
[45] 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. W306-10.
[46] Lopez-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpi JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Res 2017;45. W222-W8.
[47] Baugh EH, Simmons-Edler R, Muller CL, Alford RF, Volfovsky N, Lash AE, et al. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic Acids Res 2016;44:2501–13.
[48] 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;4:1017–28.
[49] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[50] Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genom 2015;16(Suppl 8):S1.
[51] Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol 2022;29:1–2.
[52] Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 2023;381:eadg7492.
[53] Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 2006;34. D204-6.
[54] Moal IH, Fernandez-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 2012;28:2600–7.
[55] Kulandaisamy A, Sakthivel R, Gromiha MM. MPTherm: database for membrane protein thermodynamics for understanding folding and stability. Brief Bioinform 2021;22:2119–25.
[56] Yang Y, Niroula A, Shen B, Vihinen M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics 2016;32:2032–4.
[57] Jemimah S, Yugandhar K, Michael Gromiha M. PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics. Bioinformatics 2017;33:2787–8.
[58] Chen S., Francioli L.C., Goodrich J.K., Collins R.L., Kanai M., Wang Q., et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv. 2022: 2022.03.20.485034.
[59] Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018;50:1161–70.
[60] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5.
[61] Alhuzimi E, Leal LG, Sternberg MJE, David A. Properties of human genes guided by their enrichment in rare and common variants. Hum Mutat 2018;39:365–70.
[62] Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cyber B Cyber 2009;39:539–50.
[63] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[64] Fan W, Stolfo SJ, Zhang JX, Chan PK. AdaCost: misclassification cost-sensitive boosting. Mach Learn Proc 1999:97–105.
[65] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE T Syst Man Cy A 2010;40:185–97.
[66] Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and class imbalance in oncologic data-towards inclusive and transferrable AI in large scale oncology data sets. Cancers 2022;14.
[67] Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet 2018;50:874–82.
[68] Nekrasov AN, Kozmin YP, Kozyrev SV, Ziganshin RH, de Brevern AG, Anashkina AA. Hierarchical structure of protein sequence. Int J Mol Sci 2021:22.
[69] Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for molecular property prediction. ACS Cent Sci 2018;4:1520–30.
[70] Tavanaei A., Maida A.S., Kaniymattam A., Loganantharaj R. Towards Recognition of Protein Function based on its Structure using Deep Convolutional Networks. Ieee Int C Bioinform. 2016: 145–9.
[71] Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Homberger M, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 2023.
[72] Wallner B. AFsample: improving multimer prediction with AlphaFold using massive sampling. Bioinformatics 2023;39.
[73] Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods 2022;19:679–82.
[74] Zheng S, He J, Liu C, Shi Y, Lu Z, Feng W, et al. Towards predicting equilibrium distributions for molecular systems with deep learning. arXiv Prepr arXiv 2023:230605445.
[75] 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;431:2197–212.
[76] 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.
[77] 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.
[78] Beerten J, Van Durme J, Gallardo R, Capriotti E, Serpell L, Rousseau F, et al. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformatics 2015;31:1698–700.
[79] Liu J, Liu S, Liu C, Zhang Y, Pan Y, Wang Z, et al. Nabe: an energetic database of amino acid mutations in protein-nucleic acid binding interfaces. Database 2021;2021.
[80] Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 2000;15:86–94.
[81] Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, et al. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 2007;28:622–9.
[82] Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods 2016;13:806–7.
[83] McKusick VA. Mendelian inheritance in man and its online version, OMIM. Am J Hum Genet 2007;80:588–604.
[84] Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014;133:1–9.
[85] Cline M.S., Liao R.G., Parsons M.T., Paten B., Alquaddoomi F., Antoniou A., et al. BRCA Challenge: BRCA Exchange as a global resource for variants in and. Plos Genet. 2018; 14.