Analysis of Protein Structural Features Associated with Pathogenic Missense Variants

Shuo Sha[1]

[1] *School of Physical Sciences, University of California, Irvine, 92617, United States. Tel: 1-949-668-3958. E-mail: shashuo0104@gmail.com*

Abstract

Accurate prediction of pathogenic missense variants is a key step in the advancement of genetic studies, clinical diagnosis, and drug design. One of the ways to improve the accuracy of pathogenicity prediction models is by feature engineering, extracting new features or modifying existing features to capture more relevant and biologically significant properties of missense variants and their potential impact on protein function. Here, we present various protein structural features correlated with pathogenic missense variants using new analytical methods with a larger, better-curated data set. By analyzing the protein structures of 20,000 missense variants from the Humsavar database, we find the structural significance of the protein's core for determining pathogenicity and numerous solvent accessibility features associated with pathogenic missense variants.

Keywords: Pathogenicity Prediction, Missense Variant, Protein Structure, Solvent Accessibility

1 Introduction

Missense variants, in which a single nucleotide change leads to a different amino acid being incorporated into the resulting protein, are the most common type of genetic variants and a major cause of various diseases. However, not all missense variants are disease-associated. Depending on the location of the variant, the specific amino acid substitution, and the functional role of the affected protein, missense variants can either be benign or pathogenic. Therefore, accurate identification of pathogenic missense variants is a critical step for early diagnosis, risk assessment, and the development of effective treatment strategies for genetic diseases.

Recent advancements in machine learning and natural language processing have significantly enhanced our ability to predict the pathogenicity of missense variants based on various protein features such as properties of the amino acid sequence and protein structure. One of the ways to continue improving the prediction power of missense variant pathogenicity prediction models is through feature engineering, by including additional features or modifying the existing features to capture more relevant characteristics of missense variants and their impact on protein function. To do so, we focus on deducing more complex and specific protein structural features that are statistically correlated with the pathogenicity of missense variants.

Previous studies [1, 2] found that disease-associated variants tend to occur in the conserved regions of the protein, and missense variants that cause a change in charge, especially those that introduce a charged residue into the protein core, are more likely to be pathogenic using less than 5000 missense variants in an uncurated database. However, a clear distinction between variants occurring in the core and those on the surface of a protein and their correlations with pathogenicity are absent, and not enough data have been studied. Moreover, other studies [2, 3] found that benign missense variants tend to have a higher average relative solvent accessibility (RSA) score than pathogenic variants. Yet, the average RSA score, accounting for all the residues in the wild type amino acid sequence, does not consider local patterns at different locations in the protein structure and the change of structure after the variant.

In this work, we examine various influences of the solvent accessibility of amino acid residues on the pathogenicity of the missense variants. Solvent accessibility, referring to the extent to which an amino acid residue in a protein is exposed to solvent or buried within the protein structure, provides important information about the protein's structure, function, and stability. We randomly sampled 10,000 benign and 10,000 pathogenic missense variants from the Humsavar database, an extraction of all human missense variant entries from the UniProtKB/Swiss-Prot database which are extensively curated with annotations generated through experimental methods. To obtain the corresponding structural information and solvent accessibility values, we input the amino acid sequence data into a deep-learning protein structure prediction model called NetSurfP-3.0. We present two main findings: (1) pathogenic missense variants tend to occur in the core of the protein whereas benign missense variants tend to occur at the surface of the protein, and (2) benign missense variants tend to have a higher average delta RSA value near by the variant site.

Biologically, the core of the protein includes regions that are responsible for maintaining the structural stability of the protein and protein-protein interactions. There-

fore, pathogenic variants occurring in these core regions can destabilize the protein's entire structure, leading to a loss of function or toxic effects on cellular processes. On the other hand, benign variants tend to occur on the surface of the proteins because these regions have less of an effect on protein function and stability and do not interfere with these crucial interactions. Furthermore, as benign variants occur at the surface of proteins, they are more likely to cause changes to the protein's interaction with its environment, thereby causing an increase in the average RSA value near the variant site.

2 Materials and Methods

Protein Folding Model. NetSurfP-3.0 [6] is a pre-trained protein language model that predicts the solvent accessibility, secondary structure, structural disorder, and backbone dihedral angles of each residue in a protein given its amino acid sequence. While attaining state-of-the-art prediction accuracy, NetSurfP-3.0 is also a time-efficient model.

This model is freely available as a web server and can as well be downloaded as a standalone local package. More information can be found on the web server: https://services.healthtech.dtu.dk/services/NetSurfP-3.0/.

Humsavar Database. To evaluate statistically significant protein structural features correlated with the pathogenicity of missense variants, we analyze the Humsavar database, a public database maintained by Universal Protein Resource (UniProt) that catalogs and curates all human missense variants in UniProtKB/Swiss-Prot human entries and their associated disease information.

Fig. 1 The first couple entries in the humsavar database

```
            LP/P:  32081
            LB/B:  39607
            US:    9684
            Total: 81372

Main      Swiss-Prot            AA             Variant
gene name AC         FTId       change         category dbSNP
Disease name
_________ _________ __________ _____________ ________
_____
A1BG      P04217    VAR_018369 p.His52Arg     LB/B     rs893184
-
A1BG      P04217    VAR_018370 p.His395Arg    LB/B     rs2241788
-
A1CF      Q9NQ94    VAR_052201 p.Val555Met    LB/B     rs9073
-
A1CF      Q9NQ94    VAR_059821 p.Ala558Ser    LB/B     rs11817448
-
A2M       P01023    VAR_000012 p.Arg704His    LB/B     rs1800434
-
A2M       P01023    VAR_000013 p.Cys972Tyr    LB/B     rs1800433
-
A2M       P01023    VAR_000014 p.Ile1000Val   LB/B     rs669
-
A2M       P01023    VAR_026820 p.Asn639Asp    LB/B     rs226405
```

For each missense variant, this database lists the gene name, Swiss-Prot AC (an access number labeled by Swiss-Prot for every protein), FTId (a set of variant identification numbers), AA Change (the amino acid change and its location), Variant Category (LB/B for likely benign or benign variants, LP/P for likely pathogenic or pathogenic variants, US for unclassified variants), and dbSNP (the variant identification number in the Single Nucleotide Polymorphism Database).

Processing Input Data. To process this data, we removed all the unclassified missense variants, randomly shuffled the data set, and obtained the full-length wild type amino acid sequence for 10,000 benign and 10,000 pathogenic missense variants, each with less than 1022 residues. The corresponding mutant amino acid sequences are computed using the "AA Change" column in Humsavar and the variant category is denoted a "0" for benign variants and a "1" for pathogenic variants. Both wild type and mutant sequences are input into NetSurfP-3.0 to obtain the solvent accessibility values for each amino acid residue and Fig. 2 is created for both the benign and pathogenic sequences:

Fig. 2 The first couple entries of the processed data

ID	N	J	seq_w	seq_m	rsa_w	rsa_m	delta_rsa
>sp_P04217_A1BG	1	51	M	M	0.68839943	0.68995637	-0.0015569
>sp_P04217_A1BG	2	50	S	S	0.62953651	0.6337024	-0.0041659
>sp_P04217_A1BG	3	49	M	M	0.54539949	0.54942077	-0.0040213
>sp_P04217_A1BG	4	48	L	L	0.51880127	0.52403903	-0.0052378
>sp_P04217_A1BG	5	47	V	V	0.4774645	0.4812237	-0.0037592
>sp_P04217_A1BG	6	46	V	V	0.45158371	0.44934663	0.00223708
>sp_P04217_A1BG	7	45	F	F	0.47816885	0.4720661	0.00610274
>sp_P04217_A1BG	8	44	L	L	0.46410516	0.46182725	0.00227791
>sp_P04217_A1BG	9	43	L	L	0.45797655	0.46324718	-0.0052706
>sp_P04217_A1BG	10	42	L	L	0.4655211	0.47001401	-0.0044929
>sp_P04217_A1BG	11	41	W	W	0.49386564	0.49343818	0.00042745
>sp_P04217_A1BG	12	40	G	G	0.4888474	0.49046928	-0.0016219
>sp_P04217_A1BG	13	39	V	V	0.48220345	0.48448703	-0.0022836
>sp_P04217_A1BG	14	38	T	T	0.56396109	0.56693679	-0.0029757
>sp_P04217_A1BG	15	37	W	W	0.495224	0.4986504	-0.0034264
>sp_P04217_A1BG	16	36	G	G	0.50519627	0.50364113	0.00155514

The columns are, respectively, ID: Swiss-Prot access number, N: an ascending order of residues, J: distance from the variant site (negative for residues after the variant), seq_w: the residue in the wild type sequence, seq_m: the residue in the mutant sequence, rsa_w: relative solvent accessibility for the corresponding residue in the wild type sequence, rsa_m: relative solvent accessibility for the corresponding residue in the mutant sequence, delta_rsa: rsa_m − rsa_w (a measure of structural change).

Data Availability. The Humsavar data set for all human missense variants with their corresponding disease information can be downloaded from: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/humsavar.txt.

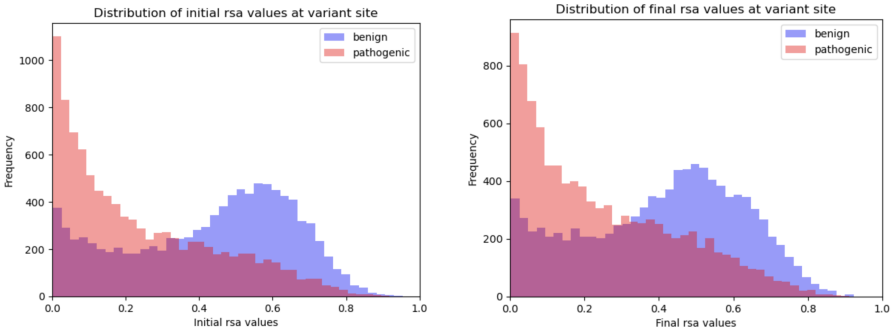
Code Availability. The full Python script for the processed data and figures is available on Github: https://github.com/shashuo0104/Protein-Features-Analysis.

3 Results

In paper [7], the authors analyzed the results from the Critical Assessment of Protein Structure Prediction (CASP) competition, where contestants designed algorithms to predict the 3D structure of an unknown amino acid sequence, and found models that accurately predicted the structure of the protein's core correctly tend to yield a significantly higher overall prediction accuracy than those that cannot. Yet, a protein's core only occupies a small portion of the entire area, thereby emphasizing the structural significance of its core.

The same structural significance is present for pathogenicity interpretations. Though proteins differ in size, we can generally assume that residues with relative solvent ac-

Fig. 3 Distribution of RSA values at variant site

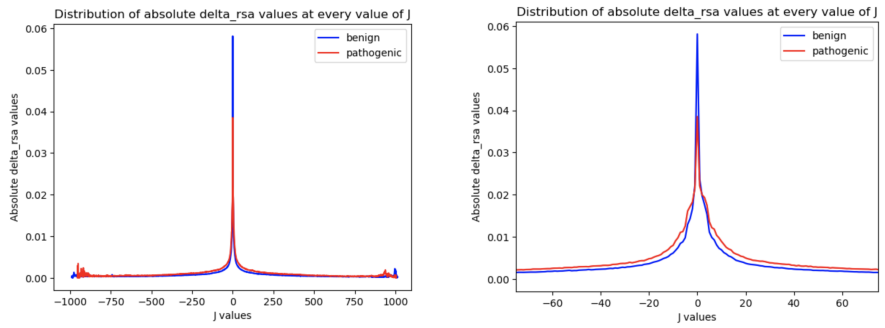


The graph on the left is the distribution of initial RSA values at the variant site before the variant has occurred, where the blue area represents the distribution of benign sequences, the red area represents the distribution of the pathogenic sequences, and the darker area represents the overlap between the two. The graph on the right is plotted the same way but for the distribution of final RSA values at the variant site after the variant has occurred. The x-axis is the initial/final RSA values (Å) evenly distributed in 40 bins and the y-axis is the frequency of each corresponding bin.

cessibility < 0.1 are buried in the core of the protein, enabling us to plot the frequency of the solvent accessibility before and after the variant for the benign and pathogenic sequences in Fig. 3. We found that pathogenic missense variants tend to occur in the core of the protein whereas benign missense variants tend to occur across the protein, typically at its surface, both before and after the variant.

In Fig. 3, though most benign sequences occur at the outer surface of the protein, the rest of the sequences mainly occur in the core of the protein, with very few located at the inner surface. Evaluating the effect of the variant, pathogenic variants tend to slightly shift toward the surface and benign variants tend to shift toward the core.

Fig. 4 Distribution of absolute delta RSA values with distance

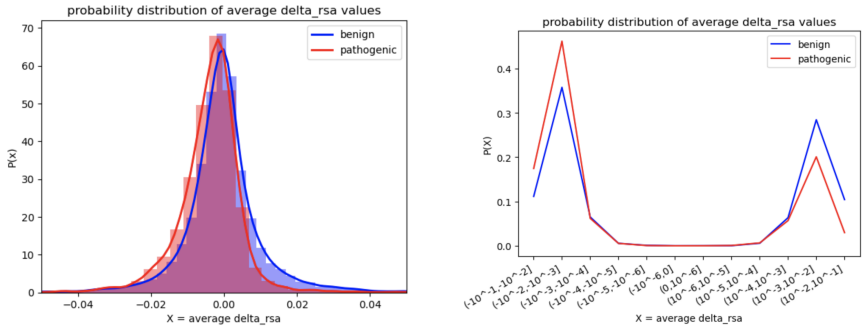


The graph on the left plots the distribution of the absolute value of the average delta RSA values at every value of J, where the blue curve represents the values for the benign sequences and the red curve represents the values for the pathogenic sequences. Since amino acid sequences differ in length, every sequence is first stretched to the same length by adding a "N/A" to the absent J values. The graph on the right shows a zoomed in view of the previous plot near the variant site, from -75 to +75 residues away from the variant site.

Mostly consistent with the study [5], we found that the absolute delta RSA decreases significantly with distance for the first 10-15 residues and decreases more gradually at greater distances in Fig. 4. However, using the absolute value of the average delta RSA value at every location is a unique approach and involves mathematical significance - it accounts for the absolute magnitude which avoids large values of opposite signs from canceling out when only taking the average. In effect, this operation smooths out the curve and produces a more apparent pattern. In Fig. 4, with a larger data set than study [5], we also found that benign variants tend to have a larger delta RSA value at the variant site than pathogenic variants, indicating a larger, overall structural change. At distances near ± 1000 residues, the absolute delta RSA values fluctuated owing to the lack of enough input sequence data close to 1022 residues.

The finding in Fig. 4 is a crucial step toward determining the statistically significant distance from the variant site when considering the effect of delta RSA on pathogenicity. As delta RSA approaches 0 after the first 20 residues, we plot Fig. 5 by computing the average delta RSA value of the 20 residues closest to the variant site.

Fig. 5 Probability density of average delta RSA values near the variant site



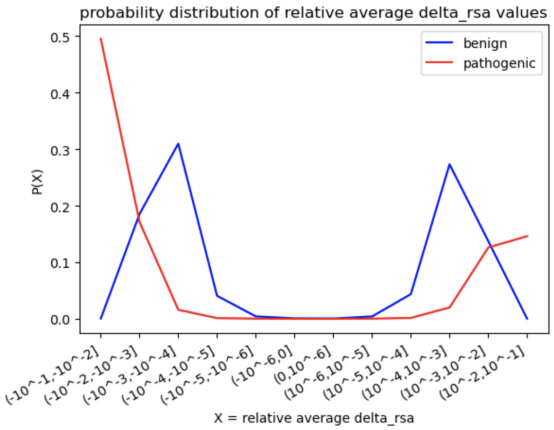
The graph on the left shows the probability distribution of the average delta RSA values of the 20 residues closest to the variant using 80 evenly distributed bins, where the blue area represents the probability distribution for the benign sequences and red area represents the probability distribution for the pathogenic sequences. The two curves are the Kernel smoothing of the corresponding histograms. The graph on the right is plotted the same way but using logarithmic bins, which shows more detailed distributions near X = 0 in an exponential manner.

Bisecting Fig. 5 at delta RSA = 0, we found that benign missense variants tend to have a higher average delta RSA value than pathogenic missense variants, consistent with the height differences in the peaks of Fig. 4. As the magnitude of average delta RSA measures the potential impact of the missense variant on the protein's structure and function, the difference in the heights of the peaks indicates that pathogenic missense variants are more disruptive to a protein's folding, stability, and interaction with other molecules. Since delta RSA is densely distributed near zero, the figure on the right describes a clearer and magnified pattern using logarithmic bins.

However, the absolute delta RSA method does not account for the magnitude of the original RSA value - a large absolute difference might be, after all, less significant owing to a relatively large final or initial value. Therefore, we propose to use rela-

tive RSA value ($\frac{rsa_f - rsa_i}{rsa_i}$) for a more accurate and localized determination of general patterns. In fact, this complements the calculation of RSA, which accounts for the maximum solvent accessibility relative to that protein type, by now considering the relative change in solvent accessibility given both the original composition and protein type.

Fig. 6 Probability density of relative RSA values near the variant site



This graph shows the probability density of the relative RSA values, calculated by $\frac{rsa_f - rsa_i}{rsa_i}$, where rsa_f is the average final RSA values of the 20 residues near the variant site and rsa_i is the average initial RSA values of the 20 residues near the variant site. In the plot, the blue curve represents the distribution of the benign sequences and the red curve represents the distribution of the pathogenic sequences. The x-axis is plotted using logarithmic bins to show a clearer pattern near X = 0.

In Fig. 6, we found that pathogenic missense variants either have extremely large or small delta RSA values relative to their initial RSA values, indicating their deleterious nature. On the other hand, benign missense variants have delta RSA values that distribute evenly relative to their initial RSA values, which are less disruptive in general.

4 Discussion

We developed various statistical correlations between specific protein features associated with missense variants and their pathogenicity which contribute to the feature engineering of pathogenicity prediction models using 20,000 missense variants from a curated clinical database. Using the predictions results of NetSurfP-3.0, we showed that pathogenic missense variants tend to occur in the core of the protein, benign missense variants tend to occur at the surface of the protein, and these patterns are consistent after the impact of the variant.

Since studies [3, 12, 13, 14] have found that pathogenic variants are more disruptive to protein's structure and function than benign variants, they tend to occur in the core because it is a region that can destabilize the overall structure, interfere

with protein-protein interactions, or disrupt enzymatic activity. On the other hand, benign variants tend to occur on the surface which has higher solvent accessibility and can tolerate more variation without disrupting protein function. This conclusion not only enhances the accuracy of previous studies by using 4 times more input data and better-curated database but also contributes to the development of machine learning pathogenicity prediction models.

This finding emphasizes that future pathogenicity prediction models should explicitly account for the specific location of the variant rather than simply considering the conserved areas of a protein to further improve prediction accuracy. Currently, state-of-the-art models like MVP [8] and mCSM-membrane [9] have directly accounted for the location of the missense variant as a feature, whereas PROVEAN [10], SIFT [11], and Mutation Assessor [12] indirectly consider the location information through the level of protein conservation at the variant site, and PolyPhen-2 [3] analyzes location through the proximity of functional sites. Though this information provides a functional significance of a particular amino acid residue, it does not directly indicate the location of the residue within the protein structure, which is a crucial feature associated with pathogenicity.

Consistent with previous findings [5], we also validated that absolute delta RSA values decrease significantly with distance for the first 10-15 residues and decrease more gradually at further distances with a larger, curated data set. Armed with the pattern, we found a more accurate and efficient method to analyze the overall RSA score for a missense variant. Previous studies [5] simply computed an average initial RSA value for all residues, which does not capture the effect of the variant. However, we accounted for the impact of the variant by plotting the probability density of the change in RSA value. Then, we considered the localized pattern along with the influences of the variant through the relative RSA value which evaluates the solvent accessibility difference centering at the initial RSA value. Instead of considering the entire sequence, we only used a couple of residues near the variant site to attain a more in-depth conclusions about both the entire variant's structural alterations and local patterns.

Though NetSurfP-3.0 achieves state-of-the-art prediction efficiency and accuracy, it cannot formulate greater than 40 protein sequences with more than 1022 residues. This also caused fluctuations in Fig. 4 as the number of sequences of length close to 1022 is lacking. Future studies should continue to validate and extend the current conclusions using protein sequences larger than 1022 residues. Furthermore, the input data from the Humsavar database can also be categorized based on their disease information. Studies can also further filter the input data and select sequences with greater clinical significance to better suit the more practical applications of pathogenicity prediction models. Lastly, studies can actually generate the 3D protein structures of the wild type and mutant amino acid sequences using AlphaFold [15] to conduct further analysis with rSASA, a more absolute measure of solvent accessibility that takes into account of the size and shape of the residue.

References

[1] Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016 Jun 20;44(11):e107. doi: 10.1093/nar/gkw226. Epub 2016 Apr 15. PMID: 27084946; PMCID: PMC4914104.

[2] 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 May 17;431(11):2197-2212. doi: 10.1016/j.jmb.2019.04.009. Epub 2019 Apr 14. PMID: 30995449; PMCID: PMC6544567.

[3] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9. doi: 10.1038/nmeth0410-248. PMID: 20354512; PMCID: PMC2855889.

[4] Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013 Jan;34(1):42-9. doi: 10.1002/humu.22204. Epub 2012 Oct 11. PMID: 22903802.

[5] Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013 Sep;34(9):E2393-402. doi: 10.1002/humu.22376. Epub 2013 Jul 10. PMID: 23843252; PMCID: PMC4109890.

[6] Magnus Haraldson Høie, Erik Nicolas Kiehl, Bent Petersen, Morten Nielsen, Ole Winther, Henrik Nielsen, Jeppe Hallgren, Paolo Marcatili, *NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning*, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W510–W515, https://doi.org/10.1093/nar/gkac439

[7] Grigas, AT, Mei, Z, Treado, JD, Levine, ZA, Regan, L, O'Hern, CS. *Using physical features of protein core packing to distinguish real proteins from decoys*. Protein Science. 2020; 29: 1931– 1944. https://doi.org/10.1002/pro.3914

[8] Qi, H., Zhang, H., Zhao, Y. et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 12, 510 (2021). https://doi.org/10.1038/s41467-020-20847-0

[9] Douglas E V Pires, Carlos H M Rodrigues, David B Ascher, mCSM-membrane: predicting the effects of mutations on transmembrane proteins, Nucleic Acids Research, Volume 48, Issue W1, 02 July 2020, Pages W147–W153, https://doi.org/10.1093/nar/gkaa416

[10] Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015 Aug 15;31(16):2745-7. doi: 10.1093/bioinformatics/btv195. Epub 2015 Apr 6. PMID: 25851949; PMCID: PMC4528627.

[11] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4. doi: 10.1093/nar/gkg509. PMID: 12824425; PMCID: PMC168916.

[12] Boris Reva, Yevgeniy Antipin, Chris Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Research, Volume 39, Issue 17, 1 September 2011, Page e118, https://doi.org/10.1093/nar/gkr407

[13] Kumar, P., Henikoff, S. Ng, P. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4, 1073–1081 (2009). https://doi.org/10.1038/nprot.2009.86

[14] Li J, Shi L, Zhang K, Zhang Y, Hu S, Zhao T, Teng H, Li X, Jiang Y, Ji L, Sun Z. VarCards: an integrated genetic and clinical database for coding variants in the human genome. Nucleic Acids Res. 2018 Jan 4;46(D1):D1039-D1048. doi: 10.1093/nar/gkx1039. PMID: 29112736; PMCID: PMC5753295.

[15] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[16] Front. Mol. Biosci., 10 March 2021 Sec. Biological Modeling and Simulation. Volume 7 - 2020. https://doi.org/10.3389/fmolb.2020.620554

[17] Gao M, Zhou H, Skolnick J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. Structure. 2015 Jul 7;23(7):1362-9. doi: 10.1016/j.str.2015.03.028. Epub 2015 May 28. PMID: 26027735; PMCID: PMC4497952.

[18] Livesey BJ, Marsh JA (2022) The properties of human disease mutations at protein interfaces. PLOS Computational Biology 18(2): e1009858. https://doi.org/10.1371/journal.pcbi.1009858

[19] Woodard J, Zhang C, Zhang Y. ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities. J Mol Biol. 2021 May 28;433(11):166840. doi: 10.1016/j.jmb.2021.166840. Epub 2021 Feb 2. PMID: 33539887; PMCID: PMC8119349.

[20] Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JPI, Marsh JA. The role of protein complexes in human genetic disease. Protein Sci. 2019 Aug;28(8):1400-1411. doi: 10.1002/pro.3667. Epub 2019 Jul 1. PMID: 31219644; PMCID: PMC6635777.