A Review on the Computational Tools for Predicting the Functional Impact of Missense Mutation

Smile Olubukola Temilola[1*], Makolo Angela Uche[2], Chiaka Anumudu[3]

[1,2]*Department of Computer Science, Faculty of Science, University of Ibadan, Ibadan, Nigeria*

[3]*Department of Zoology, Faculty of Science, University of Ibadan, Ibadan, Nigeria*

Abstract—**Advances in genomic sequencing have left us with millions of genetic variants, the highest percentage of which constitutes missense mutation. Missense mutations are responsible for more than 50% of human-inherited diseases. The accurate characterization of which mutations lead to disease is an indispensable asset for the genomic era. Besides the endless potential for precision and personalized medicine, is the opportunity for timely and appropriate clinical interventions. To this end, the development of tools for predicting the functional impact of missense mutation remains an active area of research. In this review, we present a brief introduction and discussion on the state-of-the-art computational tools for predicting the functional impact of missense mutation. The focus is on the principles, features and methods employed by each tool. The methods employed were grouped into three; -Sequence-based methods, Machine learning methods and Graph-based methods. It was found that graph-based methods were able to capture the spatial structure of the protein in addition to features used by sequence-based and machine learning methods. Besides, graph-based models create opportunities for structure comparison of the wildtype and mutant protein by leveraging on the emerging field of graph representation learning.**

Index Terms—**missense mutation, pathogenicity prediction, graph, wildtype-mutant structure comparison.**

1. Introduction

The human genome project has helped to identify millions of genetic variants.[1] The most common of these perhaps is missense mutation. Although it occurs as a result of the substitution of a single nucleotide base, missense mutations are responsible for several human-inherited diseases. A few of these diseases include Sickle cell anemia, caused by the substitution of the GTG for GAG in the sixth codon of the β-globin gene [2]. Late-onset Alzheimer's disease associated with the substitution of cytosine for guanine, at nucleotide 2119 [3], and Cystic fibrosis caused by a mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene [4].

Despite the implications of missense mutations, only a few have been associated with functional significance. In general, the functional impact of missense mutation may be benign or pathogenic. Benign mutations have no effect on the function of the protein. Pathogenic mutations disrupt the function of protein and can lead to diseases. Other missense mutations which have not been characterized are referred to as variants of uncertain significance.

Characterization of missense mutations is the foundation on which precision medicine will be built. It creates opportunities for timely and appropriate clinical interventions. Therefore, the functional impact of missense mutation prediction remains an active area of research.

Several computational tools have been developed to predict the functional impact of missense mutation on proteins. This paper provides a comprehensive overview of standard tools with focus on the principles, features and methods employed by each of the tools.

2. Principles Behind Pathogenicity Prediction Tools

There are three main assumptions on which pathogenicity discrimination algorithms are based: Changes in the physicochemical properties, Evolutionary conservation features and Structural features

A. Changes in the Physicochemical Properties

The physicochemical properties of amino acids dictate their interaction with each other and consequently the shapes and function of protein [5]. The difference between two amino acids can be computed using 134 different physicochemical properties, some of which are size, charge, hydrophobicity, and polarity. The implication is that for a missense mutation, a change of physicochemical properties (with changed amino acids) in turn changes the behavior of the protein. Hydrophobic amino acids for example tend to cluster away in the core of a protein structure to stay away from water. Hydrophilic amino acids, on the contrary, remain at the surface in contact with water. Substituting a hydrophobic amino acid with a hydrophilic acid can destabilize protein structure. In the same vein, replacing a large amino acid with a smaller amino acid leaves a cavity in the protein structures. Physicochemical properties can be captured in missense mutation prediction algorithms using the following measures: the Grantham's distance, Sneath's index, Epstein's coefficient, Miyata's

*Corresponding author: et.olubukolasmile@gmail.com distance, Experimental Exchangeability and Local sequence context.

Grantham Deviation (GD): GD measures the biochemical difference between two amino acids in terms of the composition, polarity and molecular volume. A large GD indicates a significant difference in the behaviour of the amino acid [6]. GD is given by:

$$D_{ij} = \left[\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2\right]^{1/2} \qquad (1)$$

where c, p, v are composition, polarity, and molecular volume. α, β and γ and are constants.

Sneath's index (SI): SI is a function of the dissimilarity index D of the 134 physicochemical properties [7]. D is obtained by taking the sum of properties not shared between two amino acids Epstein's coefficient: This is a function of polarity and size of amino acids [8]. It is given by:

$$\Delta_{\mathrm{a \to b}} = (\delta^2_{\mathrm{polarity}} + \delta^2_{\mathrm{size}})^{1/2} \qquad (2)$$

for difference between small hydrophobic amino acid and a larger hydrophobic amino acid and

$$\Delta_{\mathrm{a \to b}} = (\delta^2_{\mathrm{polarity}} + [0.5\delta_{\mathrm{size}}]^2)^{1/2} \qquad (3)$$

for difference between polar residue and non-polar or large residue and a smaller amino acid

Miyata's distance: It is a function based on volume and polarity [10]. The distance between amino acids a_i and a_j is given by:

$$\mathrm{d_{ij}} = \sqrt{(\Delta \mathrm{p}_{\mathrm{ij}/\sigma_{\mathrm{p}}})^2 + (\Delta \mathrm{v}_{\mathrm{ij}/\sigma_{\mathrm{v}}})^2} \qquad (4)$$

where $\Delta \mathrm{p}_{\mathrm{i,j}}$ is value of polarity difference between replaced amino acids and $\Delta \mathrm{v}_{\mathrm{i,j}}$ and is difference for volume; σ_{p} and σ_{v} are standard deviations for $\Delta \mathrm{p}_{\mathrm{i,j}}$ and $\Delta \mathrm{p}_{\mathrm{i,j}}$.

Experimental exchangeability (EX): EX captures the difference between two amino acids using experimental data to the behavior of the protein when an amino acid is replaced with another [9].

Local sequence context: The local context of an amino acid refers to the immediate amino acid flanking its on both sides. This is usually measured in windows. A window size of 5 refers to five amino acids on each side. The Local context captures the local mutation effect.

B. Evolutionary Conservation Features

Some tools exploit the rate of conservation of residues in certain positions across different species. This starts with filtering out homologs of the protein sequence in question. A multiple sequence alignment is then carried out to identify regions that have been conserved over a period of evolutionary time. Conserved regions are said to be structurally and functionally important [11]. Evolutionary conservation features used by missense mutation prediction algorithms include Conservation score, Co-evolution strength, Evolutionary Preservation, Sequence profiles, Phylogenetic Tree Analysis, Evolutionary distance, Grantham Variation (GV), Position-specific entropy, Phylogenetic profiles and 2-gram features.

Conservation score: The conservation score of a position is a measure of how preserved that position has been in terms of the amino acid change [12]. Algorithms such as Amino Acid Substitution Matrices, and Shannon Entropy are used to compute conservation score. Mutations in highly conserved regions are classified as pathogenic.

Co-evolution: This measures the strength of occurrence of amino acids at two positions over an evolutionary period [13].

Evolutionary Preservation [14]: This is a measure of the evolutionary time that an amino acid has been in a particular position Sequence profiles: This shows the frequency of the amino acid at the position of mutation over a set of homologues.

Phylogenetic Tree Analysis: Phylogenetic tree is a visual representation of the evolutionary relationship among proteins of the same family [15]. Deviation from the unique pattern provides an understanding on the pathogenicity of the mutation.

Evolutionary distance: This is a measure in evolutionary space of the level of divergence of the mutant protein sequence from the wildtype sequence.

Grantham Variation (GV): GV is a function of the variability of composition, polarity and molecular volume of amino acid across protein sequences of the same family. A low GV shows conserved regions.

Position-specific entropy: This measures the level of variability of amino acids in the position of mutation [16]. A low entropy reflects a conserved position

Phylogenetic profiles (PP): PP is a vector representation of the observed variability of amino acids in a particular position across different species.

2-gram features: 2-gram features is the frequency at which a pair of amino acids occur together in specific positions.

C. Changes in Structural Features

Missense mutation sometimes disrupts the native structure of proteins. This can mean the breaking of important chemical bonds, such as the disulphide bond or hydrogen bond, or loss of interactions such as protein-protein interactions, protein-ligand interaction or change in protein conformation. Tools that exploit disruption in structure often access the effect of such structural changes on the function of protein. Missense 3D predicts pathogenicity by analyzing structural disruptions caused by the mutation. Measures of structural changes include solvent accessibility, B-factors, secondary structure, protein stability, conformational flexibility, Root Mean Square Deviation (RMSD), Template modelling (TM) scores, protein domains, proximity to splice sites and local context.

Solvent accessibility: This is a measure of the surface area of the amino acid in contact with solvent. The hydrophobicity of each amino acid determines its Solvent accessibility.

B-factors: This refers to the vibrational amplitude of each atom within the amino acid.

Secondary structure: This shows if the mutated residue is in an alpha-helix, beta-sheet, or coil structure.

Proximity to active sites or binding interfaces: How close is the position of mutation to functionally important regions? Protein stability: Stability is a measure of the difference in free energy (ΔΔG) between the mutated and wild-type proteins.

conformational flexibility: Measures the ability of the protein to maintain the native 3D structure in different conformations

RMSD: This quantifies the structural similarity between two proteins by pairwise comparison of the atoms. It is given by:

$$RSMD = \sqrt{\left[\left(\frac{1}{N}\right) * \sum(d)^2\right]} \tag{5}$$

where N is the number of atoms being compared, and d is the distance between the corresponding atoms

Template modelling (TM) scores: TM-score quantifies the structural similarity between two protein structures, factoring in the arrangement of the secondary structures

Protein domains: The structural and functional importance of the domain where the mutation took place informs the pathogenicity of the mutation.

Proximity to splice sites: Mutations on sequences that defines the boundaries between introns and exons affect the mRNA splicing process

Local context: The structural Features and interactions around the position of mutation provides light on the role of the residue in the protein structure.

3. Classification of Pathogenicity Prediction Tools

Missense mutation pathogenicity prediction tools can be classified into three, namely, sequence-based tools, Machine learning tools and Graph-based tools. Each class of tool differs based on the approach, principles and features employed.

A. Sequence Based Tools

Sequence based tools work with protein sequences. They make use of features such as evolutionary conservation score, position specific scoring Matrices, pretrained language models, physicochemical properties of the amino acids and pairwise amino acid substitution likelihood.

The Sorting Intolerant from Tolerant (SIFT) [17] is the first model developed to determine the impact of missense mutation. Evolutionary information is used to predict the effect of a missense. A multiple sequence alignment of a homologous protein is carried out to identify conserved regions. The conservation score at a position c is as shown in equation (6). Mutations in highly conserved regions are predicted as pathogenic while those in less conserved regions are predicted as neutral. Sift score ranges from 0 to 1. Scores between 0 and 0.05 are classified as pathogenic.

$$R_c = log_2^{20} - \sum 20_{aa}\, p_{ca} log p_{ca} \tag{6}$$

where p_{ca} is the frequency of occurrence of amino acid a in position c.

Stone and Sidow [18] combined both physicochemical and evolutionary features in Multivariate Analysis of Protein Polymorphisms tool (MAPP). The evolutionary features of the protein sequence are determined with multiple sequence alignment and phylogenetic tree construction. Based on the phylogenetic correlation, weights are applied to each sequence. Each of the amino acids is represented by the alignment score. The Mean and variance scores of physicochemical properties such as polarity, volume and hydropathy of each amino acid are calculated. The deviation from each property is calculated and converted to a single score as a measure of the mutation.

Mathe et al. in Align-Grantham Variation and Grantham Deviation (Align-GVGD) [19] combined both evolutionary conservation and physiochemical properties together. For physicochemical properties, the Grantham difference is computed between the substituted amino acids. A multiple sequence alignment of orthologous sequences is used to determine evolutionary features. The mutation score known as the Grantham Difference score (GD) is a function of both evolutionary score and Grantham differences. GD assigns a class from C0, C15, C25, C35, C45, C55, or C65 to each mutation. C0 implies a lower impact and C65 an higher impact.

Reva et al [20] categorized functional impact of missense mutation based on evolutionary conservation features in Mutation Assessor. A multiple sequence alignment of homologous protein is carried out; A conservation score based on rate of substitution among homologous protein is computed. the functional specificity score is computed as a function of conservation pattern in protein subfamilies. The functional impact score is computed by combining the conservation score and the specificity score.

Choi et al in the design of the protein variation effect analyzer (PROVEAN) [21] introduced an alignment-based score. The alignment-based score is a measure of the evolutionary features. It is obtained by comparing the change in sequence similarity of the wildtype and mutated protein sequences to aligned homolog protein sequence. A score of less than -2.5 implies pathogenic while scores >-2.5 are considered neutral.

Tang and Thomas [21] included evolutionary features using the Hidden Markov Models (HMMs) in the design of Protein Analysis Through Evolutionary Relationships (PANTHER). A multiple sequence alignment of homologous protein carried out. The phylogenetic tree obtained is traversed upward from the wild-type amino acid. This continues until a different amino acid is found. The distance between the wildtype amino acid and position where change is observed is the evolutionary path of the wildtype amino acid. Longer path indicates higher pathogenicity.

Malhis et al. in LIST-S2 [23] combined output from three modules, The Position Mutation Module (PMM) Position Module (PM) together. PMM and the mutation module, PMM assesses how conserved the mutant residue is in homologous sequences. The PM module evaluates the rate of change of amino acid in the mutated position. The mutation module (MM) finally assesses the likelihood of replacing the wildtype amino acid with the mutant amino acid.

B. Machine Learning Methods

Machine learning approaches combine sequence-based features and structural features to build a classification model.

Structural properties such as changes in protein stability, disruption of protein-protein interaction, disruption of the disulfide bonds are employed. Machine learning models such as Random Forest Algorithm, Support Vector machine and decision trees have been used.ML methods offer improved performance over the sequence-based approach. However, they rely on feature engineering. Also, integrating features from different sources of data is complex and laborious. The use of Deep Learning overcomes the challenge of feature extraction in traditional machine learning method [24].

Steinhaus et al. [25] in Mutation Taster combined evolutionary conservation features, amino acid properties, functional location of mutation, structural disruptions of splice sites and protein domain into a Random Forest classifier. The older version uses a Naive Bayes classifier

Bao et al. [26] in nsSNPAnalyzer combined evolutionary features, Physicochemical properties of amino acid and structural features into a Random Forest classifier. The polarity of the neighborhood of the substituted amino acid, the secondary structure, and the solvent accessibility defines the structural features of the protein.

Niroula et al., in Prediction of Pathogenicity of Missense Variants (PON-P2). [27] combined evolutionary sequence conservation features, physicochemical properties of the amino acid, Gene Ontology (GO) annotations and the functional feature of the location of the variants into a random Forest classifier.

Pejaver et al in the design of Mutation Prediction 2 (MutPred2) [28] extracts 1345 features from sequence-based features, substitution-based features, PSSM-based features, conservation-based features, homolog profiles and changes in predicted structural and functional properties into a feed forward Neural Network

Capriotti and Fariselli, in Predictor of Human Deleterious Single Nucleotide Polymorphisms (PhD-SNP) [29] combined sequence-based features. Structural features with physicochemical properties of the amino acid into a Support Vector classifier to make prediction. Given a list of SNVs, the tool analyses each variant by first generating a 25-element vector based on the five closest nucleotides to the mutated residue. The conservation indexes for the positions around the mutation site are extracted. The PhyloP7 and PhyloP100 scores are used to generate a 10-element vector, which represents the conservation features. This is added to the 25-element vector encoding for the sequence features.

Quinodoz, et al., in Combined Annotation Dependent Depletion (CADD) [30] combined more than 60 genomic features into one. The features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements, and functional predictions. Given a variant, CADD computes a CADD score from all the annotations. CADD score is used to rank human single nucleotide variants, short insertion and deletions.

Carter et al., [31] combined both functional enrichment analysis of mutations and statistical hypothesis in the design of Variant Effect Scoring Tool (VEST). The core classifier is the random forest algorithm. VEST score ranges between 0 and 1. The significance of each score is tested via the Statistical hypothesis testing framework.

Capriotti et al, 2013 in SNPs and Gene Ontology SNPs and Gene Ontology (SNPs&GO) [32] combined evolutionary and Gene Ontology features into a support Vector Machine. Given the amino acid in question, a 20-element vector is used to uniquely represent the substitution, with a value of -1 for the wild-type residue and 1 for the mutant residue. Other positions representing the remaining 18 amino acids are represented with zero. The second twenty element vector records the frequency of the frequency of the residue around the mutated residue. The sequence profile features are extracted from BLAST search and combined with GO annotation into an SVM classifier.

Alirezaie in ClinPred [33] combined the Allele frequencies (AFs) of each variant in different populations with prediction scores from 16 other tools into random forest and gradient boosting models. Variants not represented on the gnomAD database are assigned an AF of zero. The model was trained using data from ClinVar database with prediction score ranging from 0 to 1. The threshold is set to 0.5.

Jagadeesh et al., [34] combined 7 features from standard measures of base-pair, amino acid, genomic region, and gene conservation with 298 features from multiple sequence alignment of mammals and the 99 primates with pathogenicity scores from 9 other tools into a gradient boosting tree classifier in the design of Mendelian Clinically Applicable Pathogenicity (M-CAP).

Deleterious Annotation of genetic variants using Neural networks (DANN) [35] is an improvement over CADD. In order to capture linear relationships among features, DANN trained a deep neural network (DNN) over features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements, and functional prediction. Compared with DANN, it achieved a 19% relative reduction in the error rate and 14% increase in the area under the curve (AUC).

Hopf et al., [36] in Evolutionary Variance Mutation (EVmutation) compute evolutionary conservation features by accessing how well mutations from different site evolve together. The Markov model is used to compute evolutionary energy. Stochastic Gradient Boosting predicts the pathogenicity score

Jiaying et al., combined five sequence-based, six structure-based, four dynamics-based features and a new evolutionary-based feature that measures conservation score by MSA of different species into an XGBoost classifier in the design of LYRUS [37]

Alpha Missense [38] is a deep learning model. It predicts the pathogenicity of missense mutation by combining structural features with evolutionary features. By looking at the potential disruption a missense mutation may confer on protein structure, it classifies mutation impact.

C. *Graph-Based Methods*

In graph-based models, Protein structures are converted into graphs. The advantage of this method is that the models are able to capture the connections between the amino acids and the spatial structure of the protein. Besides, graph-based models create the opportunity for structure comparison of wildtype and mutant protein. This has the potential of providing insight into how structural disruptions affect protein function. The impact of mutation is measured using network centrality measures such as node degree, clustering coefficient, spectra properties. The difference in the network parameters of the wildtype is compared to the mutant to make predictions.

Sotomayor-Vivas et al., [39] worked with a set of proteins. Every amino acid on each of the proteins was mutated to the other amino acids. For each mutation, the corresponding mutant structure was used to construct the mutant network. A perturbation network was then obtained for each mutant by comparison to the corresponding wildtype. The size of the network (nodes), number of edges, sum of edge weights, and its diameter were used to make predictions.

Yamuna and Karthika [40] modelled each amino acid as a directed network using A, T, G, C codons as nodes. The sum of the in degree and out degree of each node is computed. The 4-digit output is a pattern that uniquely identifies each amino acid, a deviation from this is classified as a missense mutation.

The drawback of this approach is that the computational cost of computing centrality measures increases as the size of the graph increases [41] and cannot adapt to new mutation. Graph representation learning overcomes the challenge by learning meaningful representation from the graph and the node features.

gMVP (Zhang, [42] is a graph attention neural network mode. It uses a graph to represent a variant and its protein context. The node features include sequence conservation and local structural properties of the protein. It uses coevolution strength as edge features. All information is pulled towards the variant residue in a star -topology network approach. The model considers only 128 amino acids flanking the mutated residue. The drawback with this approach is that it does not cater for long interaction effect. Also, the graph generated does not reflect the structural connectivity of the protein residues and atoms.

4. Conclusion

The development of missense prediction algorithm remains an active area of research as the practice of medical science gradually shifts the focus from traditional curative medicine to preventive medicine. The need to develop tools that can improve prediction accuracy will be needed more than ever before. This work presents the features, techniques and principles employed by each missense mutation prediction tool. The review shows that sequence-based tools, while efficient, perform poorly when mutation disrupts structures. Machine learning models solve the problem by combining both sequence and structural features into a classifier. Machine learning methods, however, suffer from the problem of feature engineering and integration of multiple sources of data. An emerging approach is graph-based tools. Graph-based tools provide an avenue to represent the complete spatial structure of protein. This representation allows for comparison between the wildtype and mutant protein structure. Although graph-based missense mutation prediction tools compared wildtype-mutant structure using computationally intensive centrality measures, future graph-based predictors will leverage advances in graph representation learning to determine the pathogenicity of missense mutation.

References

[1] S. Aganezov, S.M. Yan, D.C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, D.J. Taylor, K.Shafin, A. Shumate, C. Xiao, J. Wagner," A complete reference genome improves analysis of human genetic variation", Science, 1;376(6588):eabl3533, Apr. 2022.

[2] L. Li, P.K. Mandal, "Recent advancements in gene therapy for sickle cell disease and β-thalassemia", Frontiers in Hematology, 27;3:1468952, Sep. 2024.

[3] V. Kumar, K. Roy, "Recent progress in the treatment strategies for Alzheimer's disease", Computational Modeling of Drugs Against Alzheimer's Disease. 1:3-47, Jul. 2023.

[4] A. Zaher, J. ElSaygh, D. Elsori, H. ElSaygh, A. Sanni, H. Elsaygh, A. Sanni, "A review of Trikafta: triple cystic fibrosis transmembrane conductance regulator (CFTR) modulator therapy", Cureus, 3;13(7), Jul. 2021.

[5] R.D. Ludescher. Physical and Chemical Properties of Amino Acids," Food proteins: Properties and characterization",17;65(65):23. Dec.1996.

[6] R. Grantham, "Amino acid difference formula to help explain protein evolution". Science, 6;185(4154):862-4. Sep. 1974.

[7] P.H. Sneath," Relations between chemical structure and biological activity in peptides", Journal of theoretical biology, 1;12(2):157-95, Nov. 1966.

[8] C.J. "Epstein, Non-randomness of ammo-acid changes in the evolution of homologous proteins", Nature, 22;215(5099):355-9, Jul. 1967.

[9] L.Y. Yampolsky, A. Stoltzfus, "The exchangeability of amino acids in proteins", Genetics, 1;170(4):1459-72, Aug. 2005.

[10] T. Miyata, S. Miyazawa, T. Yasunaga, "Two types of amino acid substitutions in protein evolution", Journal of molecular evolution, 12(3):219-36, Mar. 1979.

[11] G. Nimrod, M. Schushan, D.M. Steinberg, N. Ben-Tal, "Detection of functionally important regions in "hypothetical proteins" of known structure", Structure. 12;16(12):1755-63, Dec. 2008.

[12] A, C. Ben, G. Masrati, A. Kessel, A. Narunsky, J. Sprinzak, S. Lahav, H. Ashkenazy, N.Ben-Tal, "ConSurf-DB: an accessible repository for the evolutionary conservation patterns of the majority of PDB proteins", Protein Science,29(1):258-6, . Jan. 2020.

[13] S. de Oliveira, C. Deane, "Co-evolution techniques are reshaping the way we do structural bioinformatics". F1000Research, 25;6:1224, Jul. 2017

[14] H. Tang, P. D. Thomas, "PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation", Bioinformatics. 15;32(14):2230-2, Jul. 2016.

[15] T.H. Nguyen, C.C. Dang, V. S Le, "Rooting phylogenetic trees from protein alignments". I5th International Conference on Knowledge and Systems Engineering (KSE) (pp. 1-5). IEEE, Oct. 2023.

[16] K. Kouser, B.S. Rashmi, L. Rangarajan L. "Entropy Based Feature Selection for Lacunarity Analysis of Position Specific Motif Matrices of Promoter Sequences", In Proceedings of the International Conference on Informatics and Analytics (pp. 1-6), Aug. 2016.

[17] N.L. Sim, P.Kumar , J. Hu, S. Henikoff, G. Schneider, P.C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins", Nucleic acids research. 1;40(W1): W452-7, Jul. 2012.

[18] E.A. Stone, A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity". Genome research. 1;15(7):978-86 Jul. 2005.

[19] E. Mathe, M. Olivier, S. Kato, C. Ishioka, P. Hainaut, S.V. Tavtigian, "Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis-based methods", Nucleic acids research, 1;34(5):1317-25, Jan. 2006.

[20] B. Reva, Y. Antipin, C. Sander, "Predicting the functional impact of protein mutations, application to cancer genomics," Nucleic acids research, 1;39(17): e118, Sep 2011.

[21] Y. Choi, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels.

[22] H. Tang, P.D. Thomas, "PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation", Bioinformatics,15;32(14), 2230-2, Jul. 2016.

[23] N. Malhis, M. Jacobson, S.J. Jones, J. Gsponer, “LIST-S2: taxonomy-based sorting of deleterious missense mutations across species”, Nucleic acids research. 2;48(W1),W154-61. Jul. 2020.

[24] Y. Liu, H. Pu, D.W. Sun, “Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices”, Trends in Food Science & Technology. 1; 113:193-204, Jul. 2021.

[25] R. Steinhaus, S. Proft, M. Schuelke, D.N. Cooper, J.M. Schwarz, D. Seelow, “Mutation Taster 2021”, Nucleic Acids Research, 2;49(W1): W446-5, Jul. 2021.

[26] L. Bao, M. Zhou, Y.Cui, ” nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms”, Nucleic acids research. 1;33(suppl_2):W480-2, Jul. 2005.

[27] A. Niroula, S. Urolagin, M. Vihinen, “PON-P2: prediction method for fast and reliable identification of harmful variants”, PloS one. 3;10(2), e0117380, Feb 2015.

[28] V. Pejaver, J. Urresti, J. Lugo-Martinez, K.A. Pagel, G.N. Lin, H.J. Nam, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature communications.;11(1):5918, Nov. 2020.

[29] E. Capriotti, P. Fariselli, “PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants”, Nucleic acids research, 3;45(W1): W247-52. Jul. 2017.

[30] P. Rentzsch, D. Witten, G.M. “Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome”, Nucleic acids research. 8;47(D1): D886-94, Jan. 2019.

[31] H. Carter, C. Douville, P.D. Stenson, D.N. Cooper DN, R. Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC genomics. 28;14(Suppl 3):S3, May. 2013.

[32] E. Capriotti, R. Calabrese, P Fariselli, P.L. Martelli, R.B. Altman, R. Casadio, “WS-SNPs & GO: a web server for predicting the deleterious effect of human protein variants using functional annotation”. BMC genomics,14(Suppl 3): S6 May. 2013.

[33] N. Alirezaie, K.D. Kernohan, T. Hartley, J. Majewski, T.D. Hocking,” ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants”. The American Journal of Human Genetics,103(4):474-83, Oct. 2018.

[34] K.A. Jagadeesh, A.M. Wenger, M.J. Berger, H. Guturu, P.D. Stenson, D.N. Cooper, J.A. Bernstein, G. Bejerano, “M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity” Nature genetics. 48(12):1581-6. Dec.2016.

[35] D. Quang, Y. Chen Y, X. Xie,” DANN: a deep learning approach for annotating the pathogenicity of genetic variants”, Bioinformatics. 31(5):761-3, Oct 2014.

[36] T.A. Hopf, J. B. Ingraham, F.J Poelwijk, C.P. Schärfe, Springer M, Sander C, Marks DS. “Mutation effects predicted from sequence co-variation”. Nature biotechnology35(2):128-35, Feb. 2017.

[37] J. Lai, J. Yang J, E.D. Gamsiz, “Rubenstein BM, Sarkar IN. LYRUS: A machine learning model for predicting the pathogenicity of missense variants”. Bioinformatics Advances. 2(1): vbab045, Jan. 2022.

[38] H. Tordai, O. Torres, M. Csepi, R. Padanyi, G. L. Lukács, T. Hegedűs, “Analysis of AlphaMissense data in different protein groups and structural context”, Scientific data. 14;11(1):495, May. 2024.

[39] C. Sotomayor-Vivas, E. Hernández-Lemus, R. Dorantes-Gilardi, “Linking protein structural and functional change to mutation using amino acid networks”, Plos one;17(1):0261829 Jan. 2022.

[40] M. Yamuna, K. Karthika, “Point mutation determination using graph theory”, Der Pharmacia Lettre.;7(7):58-66. 2015.

[41] B. Meng, A. Rezaeipanah, “Development of a multidimensional centrality metric for ranking nodes in complex networks.”, Chaos, Solitons & Fractals. 1;191:115843, Feb. 2025.

[42] H. Zhang, M.S. Xu, X. Fan, W.K. Chung, Y. Shen, “Predicting functional effect of missense variants using graph attention neural networks.” Nature Machine Intelligence.; 4(11):1017-28, Nov. 2022.