REPUBLIC OF TURKEY
ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

PREDICTING a FUNCTIONAL SCORE for MUTATIONS in RARE DISEASES BASED ON THEIR STRUCTURAL IMPACT

UMUT GERLEVİK

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and BIOINFORMATICS

SUPERVISOR:

Prof. Dr. Osman Uğur SEZERMAN

ISTANBUL – 2021

REPUBLIC OF TURKEY
ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

PREDICTING a FUNCTIONAL SCORE for MUTATIONS in RARE DISEASES BASED ON THEIR STRUCTURAL IMPACT

UMUT GERLEVİK

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and BIOINFORMATICS

SUPERVISOR:

Prof. Dr. Osman Uğur SEZERMAN

ISTANBUL – 2021

DECLARATION

I declare that this thesis study is solely my original work. I had no unethical behavior at any stage, from planning to writing; I obtained all the information in this thesis by following academic and ethical rules, I listed in the bibliography all resources I have used. There is no violation of any patents or copyrights.

19/08/2021
Umut GERLEVİK

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Uğur Sezerman for his guidance that shapes my early research career and the whole of my scientific point of view, his endless patience, and his all-time continuous support when I feel confused, lost, or find a good idea. He has been truly an idol for me, as a great scientist, as a great mentor, and as a kind and caring person. It has been a great privilege and honor for me to be one of his students.

I would like to thank jury members and substitutive jury members for my thesis, who are Prof. Dr. Eda Tahir Turanlı, Prof Dr. Aslı Tolun, Assoc. Prof. Emel Timuçin, and Asst. Prof. Bülent Balta for their worthy contributions.

Sezerman Lab friends and alumni, I am grateful to have worked side-by-side with them all in both scientific and fun events; Rüçhan Ekren, Pınar Altıner, M.D. Ege Ülgen, Ph.D. Ceren Saygı, Asst. Prof. Dr. Aslı Yenenler Kutlu, Ph.D. Ozan Özışık, Hüseyin Okan Soykam, Nogayhan Seymen, Begüm Özemek Güner, Tayyip Karaman, Narod Kebabçı, Eray Şahin, Zehra Hazal Sezer, Tuğçe Bozkurt, Fatma İşleyen, Berk Gürdamar, Deniz Ece Kaya, Ph.D. Orhan Özcan, Baran Özbek, Gökhan Nalbant, Hüseyin Avni Taç, Milena Jakimovska, Ege Şeker, Barış Balaban, Berkay Ekren, İbrahim Sertdemir, Bengisu Karaköse, Oğulcan Cingiler and Sıla Karacan. And I beg forgiveness from the ones I could not write their name here by remembering all members of the huge Sezerman Lab network at once. On the other hand, I want to thank Tayyip Karaman once more for coming up with the name of the package developed in this study, namely “Rmut”. This is the perfect name that refers to several things at the same time, which are the name of the programming language “R” and the focus of this work “mutation” as “mut”. Also, it is similar to my first name “Umut”, and it sounds like the Turkish word “*armut*” meaning “pear”, which gave inspiration to design the logo of the package.

I would like to express my sincere gratitude to Prof. Dr. Koray Özduman for his guidance in both scientific and non-scientific subjects and his endless support during my Ph.D. applications. It has been a great privilege and honor for me to work with his mentorship.

I want to express the most intense thanks to the love of my life, Sıla Karacan (Gerlevik now), for always being with me, for her love, support, and patience, and especially for the discussions and questionings on everything.

I would like to mention my great appreciation to my family that has brought me up to this day and has done their best. I want to express my gratitude to my mother Gülseren, my father Şenol, and my elder brother Ufuk, who supported me in all my decisions and provided me the financial and moral support in all difficulties.

I am thankful to all my close friends -they know themselves- for their tolerance, support, sharing fun, and the deepest discussions that shaped my personality and mentality.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

ANM	Anisotropic Network Model
API	Application Programming Interface
BLOSUM	Blocks Substitution Matrix
Cα	Carbon Alpha
ddG	Gibbs Free Energy Difference
DefE	Deformation Energy
ENM	Elastic Network Model
GNM	Gaussian Network Model
HMM	Hidden Markov Model
ID	Identifier
MCC	Matthews Correlation Coefficient
MD	Molecular Dynamics
PAM	Point Accepted Mutation
PDB	Protein Data Bank
Pfam	Protein Family
PSI-BLAST	Position-Specific Iterative-Basic Local Alignment Search Tool
PSIC	Position-Specific Independent Counts
RSA	Relative Solvent Accessibility
RVIS	Residual Variation Intolerance Score
SASA	Solvent Accessible Surface Area
SMOTE	Synthetic Minority Over-sampling Technique
SNV	Single Nucleotide Variant
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
β-factor	Beta Factor

LIST OF FIGURES

LIST OF TABLES

SUMMARY

The main purpose of this study was to develop a tool that would improve the prediction performance for the functional impact of missense variants. Variant impact prediction is crucial in variant prioritization to filter out the irrelevant variants from analyzed genomes or exomes by finding the relationships of variants with disease and/or some damaging properties. However, variant effect prediction is a difficult problem for missense variants, especially, since their effect is usually small on the protein sequence, structure, and dynamics. Therefore, this prediction task is important to diagnose and treat the patients suffering from a genetic disease such as rare diseases and cancers, and to understand the molecular mechanisms of these diseases. In this study, two former approaches were integrated to build a predictor with improved performance: comparison of each type of feature between wild-type and mutant protein structures, and protein dynamics-based features extracted from realistic approaches called anisotropic network models. Furthermore, the elastic net classifier built in this study, Rmut, outperformed the two tools that inspired it, namely Rhapsody and Missense3D. Rmut is available on GitHub (https://github.com/ugerlevik/Rmut) to make this predictor available to the scientific community as a user-friendly and easily accessible R package.

Keywords: Elastic Net Classification, Functional Impact Prediction for Missense Variants, Rare Diseases and Cancers, Rmut R Package, Variant Prioritization.

ÖZET

Nadir Hastalık Mutasyonları İçin Yapısal Etkiye Göre İşlevsel Skor Tahmin Edilmesi

Bu çalışmanın temel amacı, yanlış anlamlı varyantların işlevsel etkisi için tahmin performansını iyileştirmekti. Varyant etki tahmini, varyantların hastalık ve/veya bazı zararlı özelliklerle ilişkilerini bularak, analiz edilen genomlardan veya ekzomlardan alakasız varyantları filtrelemek için varyant önceliklendirmesinde çok önemlidir. Bununla birlikte, varyant etkisinin tahmini, özellikle protein dizisi, yapısı ve dinamikleri üzerindeki etkileri genellikle az olduğundan, yanlış anlamlı varyantlar için zordur. Bu nedenle, böyle bir tahmin görevi, nadir görülen hastalıklar ve kanserler gibi genetik bir hastalığı olan hastaların tanı ve tedavisi ve bu hastalıkların moleküler mekanizmalarının anlaşılması açısından önemlidir. Bu çalışmada, geliştirilmiş performansa sahip bir tahmin edici oluşturmak için iki bilinen yaklaşım entegre edildi: referans ve mutant protein yapıları arasındaki her özellik türünün karşılaştırılması ve anizotropik ağ modelleri olarak adlandırılan gerçekçi yaklaşımlardan çıkarılan protein dinamiğine dayalı özellikler. Ayrıca, bu çalışmada oluşturulan elastik ağ sınıflandırıcı Rmut, onun geliştirilmesine ilham kaynağı olan iki araçtan, yani Rhapsody ve Missense3D'den daha iyi performans gösterdi. Bu tahmin edici araç Rmut bilim camiasına kullanıcı dostu ve kolay erişilebilir bir R paketi olarak GitHub'da (https://github.com/ugerlevik/Rmut) sunulmaktadır.

Anahtar Sözcükler: Elastik Ağ Sınıflandırması, Nadir Hastalık ve Kanser, Rmut R Paketi, Varyant Önceliklendirme, Yanlış Anlamlı Varyantların İşlevsel Etkisinin Tahmini.

1. BACKGROUND AND AIM OF THE STUDY

Precision medicine is the field that aims the specialized diagnosis and therapy for the subgroups of patients which are classified by their molecular profiles (1,2). Moreover, precision medicine approaches are particularly beneficial in cancers and rare diseases. In both type of diseases, finding the underlying molecular markers (e.g., single nucleotide variants, SNVs) is one of the most important steps for accurate diagnosis and effective treatment (1,3). Furthermore, the most effective and recent method to achieve this goal is sequencing the exome (whole exome sequencing, WES) or genome (whole genome sequencing, WGS) of patients (1,3). However, researchers mostly have hundreds of changes as a result of such sequencing studies when they mapped the exome/genome of the individuals to the reference genome (3,4). The process until this point is simply represented in Figure 1. Herein, a variant prioritization procedure should be applied to eliminate the variants irrelevant to the disease that can be called as neutral (3,4). Only after such filtering, the number of variants can be reduced to study their effect by literature search and functional experiments. Moreover, one of the most powerful approaches for the prioritization is predicting the functional impact of variants by using certain features that can be extracted from the relevant sequence and structure of proteins (3,4).

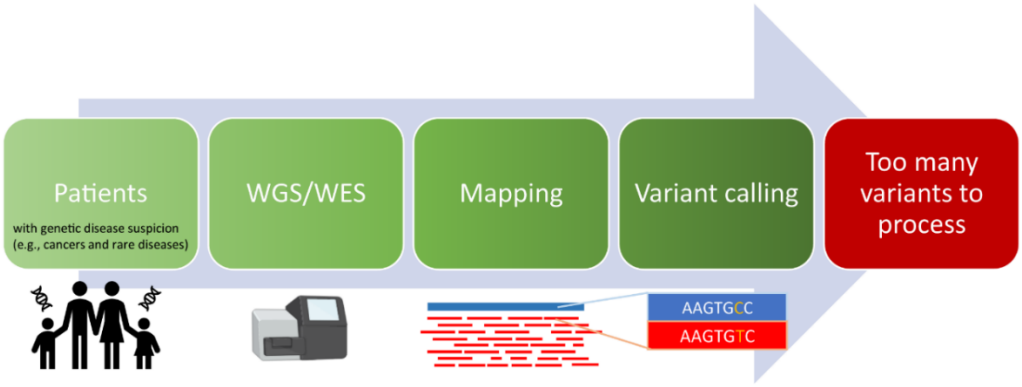

Figure 1. A simple representation of the process until obtaining the variants in genomes or exomes of the patients in precision medicine.

The first attempts of variant impact prediction date back to the late 1960s after the relationship between amino acid changes and diseases are proposed in 1962 (5), but it is still a major problem in the field even though the prediction accuracy is significantly increased in time (6–10). Since that time, many prediction tools have been developed, each with different advantages and disadvantages. In the literature, these approaches are commonly grouped based on the features they use such as evolutionary conservation-based, structure-based, combined methods, and meta-predictors (5,8). With the computational improvements over the past decade, recent methods utilize all types of features as combined by using state-of-the-art machine learning techniques (11–14). Examples for these features might be evolutionary conservations, allele frequencies in populations, being in functional sites of genome or protein, physicochemical properties of amino acids, secondary structure, solvent accessibility, difference in protein stability, residue flexibility and allosteric signals (3,12–15). Despite the diversity in features, even the performances of the most recent tools are limited at some point in accuracy (8,12). The reason for such limitations might be the known dataset-dependent biases (16,17), or some non-understood mechanisms that cannot be integrated into the predictors. For example, some variants might affect the structure and function of proteins but do not cause diseases (10), or vice versa (18).

In the near past, protein dynamics-based features have been started to be used for high-throughput functional impact prediction after they are extracted by using the realistic methods such as elastic network models, and they seem robust enough to obtain 80-90% accuracy (15,19). On the other hand, structure-based features are usually calculated for only wild-type instead of comparing it with the mutant. With the recent computational developments, making such comparisons efficiently is now possible. Bhattacharya *et al.* (2017) and Ittisoponpisan *et al.* (2019) showed the robustness of comparing structural features (10,18). Focus of this study was the prediction for missense variants rather than other types of variations. The first reason for this particular focus was that prediction for the missense variants, which are small changes, are often resulted with conflictions. Also they are frequently evaluated as variant of uncertain significance in the rare disease studies (20–22). This situation

makes the missense variant effect prediction a specific problem. The second reason was to take advantage of the forementioned improvements, which can be utilized for the missense variants well by examining the impact at protein level in detail. Therefore, in the present study, the hypothesis is that combining the successful features from the former studies with the comparison approach might improve the prediction performance for the effect of the missense variants. Furthermore, we also aimed to represent our method as a user-friendly R (23) package to the community, namely Rmut.

2. INTRODUCTION

To be able to make improvements in the variant impact prediction field, the previous scores, predictors, and features should be understood well in terms of their strengths and weaknesses. Moreover, protein structure optimization and several machine learning approaches are also important to follow this study. Therefore, all these topics are detailed in the following sub-sections.

2.1 History of the Variant Impact Predictors

The following sub-sections explain the development of variant impact prediction procedures during the last ~55 years.

2.1.1 Early Variant Impact Prediction Approaches

The first variant impact scoring approaches focused on predicting the effect of amino acid changes in proteins by using substitution frequencies and physicochemical properties of amino acids (6,24–27). The first substitution scoring is represented by Eck and Dayhoff (1966), and the methodology was based on phylogenetic trees of several families (28). In the same time periods with the Dayhoff's score, there were also some other mutation distance approaches calculated by using structural differences and evolutionary similarities (29). In early 1970s, McLachlan merged these substitutions in homologous proteins-based distances with a search approach comparing segment pairs of two proteins and a statistical test for randomness of the correlations (29). In 1974, Grantham score, which uses the physicochemical changes such as polarity, volume and side chain composition of amino acids, was emerged as the first use of this kind of features (6). Grantham score is still used in some widely-

used methods such as PANTHER (27), CHASM (13), VEST (30) and CADD (9) whereas the formerly-developed scores are not. Almost all the methods explained in the next sub-sections are currently used in predictors, as predictors or in meta-predictors that uses the results of other predictors as features.

2.1.2 Variant Impact Scores

In early 1990s, BLOcks SUbstitution Matrix (BLOSUM) and Point Accepted Mutation (PAM) matrices are used to score replaceability of amino acids with each other in an evolutionary aspect, and they are used for sequence alignment scoring as well as for the tolerance of an amino acid change (31,32). The number following BLOSUM (e.g., BLOSUM62 is the most popular due to its high performance) refers to the identity level of the homologous proteins included in the substitution matrix calculation (31) whereas the number following PAM (e.g., PAM250 is the most popular because of its high performance) represents the distance from the identity level of the protein sequences included in the matrix calculation (32). Position-Specific Independent Counts (PSIC) scoring was published in 1999, but still included in the most popular and newest predictors (15,33). PSIC is the likelihood of an amino acid at a specific position in a protein by making a multiple sequence alignment of homologous sequences including the probabilities from BLOSUM62 substitution matrix (24). In early 2000s, Sorts Intolerant From Tolerant amino acid substitutions (SIFT) (26) is proposed to the literature, and it is still one of the most popular predictors and supported by computational performance improvements (34). SIFT estimates the probability for each amino acid at a given position of a protein based on the Position-Specific Iterative-Basic Local Alignment Search Tool (PSI-BLAST) (35) alignment of high identity protein sequences to score how much tolerant an amino acid at that position (26,34,36). LogR.E-value does a similar scoring but uses Protein family (Pfam) (37) sequences and Hidden Markov Model (HMM) profile search called HMMER (38) to predict protein motifs (25). After this point, predictors started to have more than one score/feature. For instance, MAPP utilizes the physicochemical

properties such as hydropathy, polarity, volume, and free energy by weighting them based on multiple sequence alignment to calculate the probability of being deleterious (39). PANTHER uses Grantham score, a position-specific and HMM-based evolutionary conservation score, and Ka/Ks ratio which corresponds to the ratio of the number of nonsynonymous changes per nonsynonymous site to the number of synonymous substitutions per synonymous site (27).

2.1.3 Variant Impact Prediction Approaches Involving Structural Biology

As the Human Genome Project is finished in 2003, many new variants were found within the human genome, and to understand the relation of those variants with diseases is so important for diagnosis and therapy (40). After several years, next-generation sequencing technologies began to emerge, and they became vital tools for diagnostics of the patients with genetic diseases such as cancers and rare diseases (3,41,42). However, the main problem here was finding hundreds of changes within the exome/genome of the patients when it is compared to the reference genome (3). Furthermore, since to find out the causative variants is so important for diagnosis and treatment by filtering out the irrelevant ones, predicting the effect of the variants has become more and more significant (3). Moreover, to develop successful predictors was possible by utilizing the enhanced computational power with the multicore usage as well as the structural information with the greatly-increased number of structures in the Protein Data Bank (PDB) (43) by 2000s (44,45).

Variant effect predictors have started to use features that can be calculated from protein structures, in other words, structural features. Frequently used features can be exemplified as follows: solvent accessibility of the mutation site, secondary structure containing the variant, being involved in an interaction (e.g., salt bridges, disulfide bonds or H-bonds), being involved in an interaction surface with other proteins or ligands, being at the functional sites (e.g., catalytic) of the proteins, and flexibility of

the amino acid at the variation site. Flexibility could be classified as dynamics feature instead of structural because it refers to the mobility capacity of a residue instead of a static situation (15,19). Moreover, this information might be based on the crystallographic β-factors or based on the predictors trained by using solvent accessibility and secondary structure as features to estimate the crystallographic β-factors (46) of the amino acid of change, and variant being at functional sites of the proteins. On the other hand, examples of the predictors include: SAAP (47), PolyPhen-2 (33), SuSPect (48), SNAP (49), HOPE (50), CHASM (13), VEST (30), ENTPRISE (51), and EvolutionaryAction (52). To sum up, these predictors utilize these features and/or conservation-based features, but almost each feature referring to the same concept is defined on different basis. For example, PolyPhen-2 uses PSIC as the conservation metric (33) whereas EvolutionaryAction has another substitution score based on a genotype-phenotype perturbation approach (52). Another example is that solvent accessibility is represented as numeric after normalization by dividing it to the theoretical maximum value in SuSPect (48) whilst CHASM integrated this feature by converting it into categorical variables as buried, intermediately-buried or exposed via applying cutoffs (13).

Despite all the new types of features used in this era, there are also new predictors entirely based on evolutionary conservation. Because of the enhanced sequencing technologies and analysis methods as well as the accumulated knowledge in the literature, it was possible to emerge the new conservation metrics with better performances. Most of these methods are still quite popular now. The following predictors can be given as examples: PROVEAN that uses the mutant protein's similarity to its functional homologous based on BLOSUM62 (53), MutationAssessor that uses entropy among family and subfamily (54), FATHMM that uses HMM within 90% identity proteins (55), GERP that utilizes phylogenetic trees to calculate evolutionary rates and substitution rejection score (56), LRT that makes a likelihood ratio test by using the rate of nonsynonymous to synonymous substitution rate (57), SIPHY that uses a novel probabilistic approach based on substitution patterns (58),

and EVmutation that utilizes an unsupervised model that takes epistasis and evolutionary statistical energy into account (59).

The previously-mentioned methods have quite good performances. However, with the improvements in the sequencing technologies and with their decreasing cost, sequencing studies and precision medicine have become more and more vital and popular for diagnostics (3,42). Also, with the growing popularity of whole genome sequencing and epigenetics studies, importance of the non-coding genomic regions has increased. Therefore, the need for more accurate predictors has greatly increased for the variant impact including for the non-coding variants. To meet this performance need, new features and approaches have emerged. Besides, integrating known methods as meta-predictors seem as improved the predictive power (8,60–64).

To predict the impact of all genomic changes including the non-coding ones, several scoring algorithms have emerged. They are important to support the variant prioritization process in whole genome analysis as well as to understand the non-coding variants underlying the diseases. Residual Variation Intolerance Score (RVIS) is one such scoring method that makes a gene-level assessment based on their frequencies in genomes for intolerance to variations (65). MutationTaster applies a naïve Bayes classifier for both exonic and intronic variants by using conservation, frequency and functional site-based features (66,67). CADD makes predictions for all type of genome-wide variations by utilizing more than 60 features related to conservations and functional sites as well as the structural features for protein-level variants (9). DANN uses the same datasets and features with CADD but utilizes a deep neural network which can catch the non-linear associations better instead of a support vector machine (68). GWAVA (69) and FATHMM-MKL (70) have also epigenetics-related features such as histone modifications and open chromatins besides with the conservations and functional sites. On the other hand, PrimateAI, which has also structural features for protein changes, is built by considering the bottleneck events during evolution (14). In detail, some variations in human might have low frequencies

in the population due to bottleneck events, and this situation could cause false positives as pathogenic. Moreover, integrating great ape genomes to the frequency analysis could improve the prediction performance (14).

Meta-predictors are the models that use the results of other predictors as their features. For instance, Condel (61) uses the LogR.E-value, MAPP, MutationAssessor, PolyPhen-2 and SIFT by applying a weighted averaging on their normalized scores. Similarly, KGGSeq (also called as Logit in some studies (16)) (71) utilizes the scores from SIFT, PolyPhen-2, LRT, MutationTaster and phyloP. Moreover, even meta-predictors could be used in other meta-predictors. For example, phyloP (72) used by Logit is also a meta-predictor that uses LRT, GERP and two other methods. On the other hand, MVP (11) and UNEECON (73) are deep-learning-based predictors, which utilize more than 30 features including the scores from other predictors, to identify intolerant genes and/or variants. Additionally, REVEL (62), CanDrA (60), M-CAP (63), MSC (64), ReVe (8), MetaSVM (74) and MetaLR (74) are other examples of meta-predictors with different basis and good performances.

Within the last two decades, novel, adapted and/or enhanced predictors that target the impact of missense variants have continued to be represented to the literature. For example, stability predictors for proteins (e.g., I-Mutant2.0 (75), FoldX (76), DynaMut (77), and PoPMuSiC (78)) were already available, but their focus was only on the folding/binding free energy change in the proteins, instead of the clinical effects of the variants. To predict causative relationships between stability changes and pathogenicity, the models should be trained on clinical datasets since there is no simple correlation between them, specifically referring that both stability decreases, and increases might cause diseases (79,80). Some predictors utilized this approach like SNPMuSiC that performs statistical potential calculations besides with the other structural features such as solvent accessibility (12). VIPUR uses Rosetta (81) to produce stability-based features (82). Of note, some other predictors mentioned in the previous section, such as CHASM (13) and VEST (30), include the stability scores

calculated by other known tools, which is FoldX in these cases. On the other hand, SNPs&GO uses gene ontology-based features and local environment information of the variant site besides the other common features (83). DEOGEN2 has new and improved features such as early folding predictions, interaction patches, Pfam log-odd score, a conservation index and pathway log-odd score (84). PhD-SNP uses an SVM-based classifier that utilizes protein sequence to evaluate the environment of the variant site and BLAST-based conservation profiles (85).

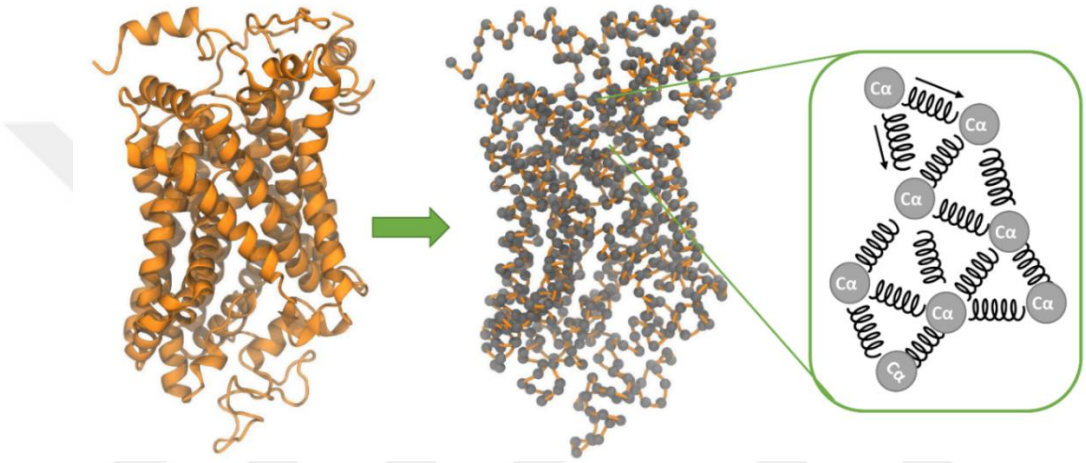

Figure 2. Elastic network model representation of a protein as Cα beads connected with springs.

At present, there is an enormous number of variant impact predictors in the literature that could not be included in this comprehensive review. As the last two of the mentioned predictors, Rhapsody (15,19) and Missense3D (18) are especially important for this study. Rhapsody represented new protein dynamics features calculated by using realistic methods called Gaussian Network Models (GNM) (19) and then Anisotropic Network Models (ANM) in the improved version (15). GNM and ANM are together called elastic networks models (ENMs). In ENMs, proteins are represented as beads (mostly alpha-carbons (Cα)) connected with springs, including the spring forces as shown in Figure 2. And by using normal mode analysis, both major and allosteric movements of the protein can be successfully predicted as the results highly overlap with the molecular dynamics (MD) simulations and the experimental results (86–90). Moreover, these high-performance methods are generally able to give results in a few seconds with the R (23) package Bio3D (91,92), or python (93)

package ProDy (94). Therefore, using them have a great potential to improve the accuracy of dynamics-based features. The examples for those features are residue flexibility, mechanical stiffness, and sensitivity/effectiveness of amino acids in allosteric signal propagation (15,94). On the other hand, Missense3D makes predictions by comparing the structural features (e.g., solvent accessibility; polarity; charge; and interactions such as H-bonds, salt bridges and disulfide bonds) between the wild-type and mutant structures obtained by using homology modeling with the Phyre2 (95) method (18). Important point in Missense3D is the using a comparison between wild-type and mutant instead of calculating all structural/dynamics features only from the wild-type structure (18).

2.1.4 Approaches Beyond the State-of-the-art

All the explained predictors except Missense3D use only wild-type structure to calculate structural features whereas they also include the difference between wild-type and variant for in the sequence-based features. Based on the performances showed in both independent benchmarks and the comparisons involved in the original articles of variant impact predictors, it might be inferred that the predictive performances of these tools seem highly dataset-dependent since there is no such predictor that is consistently called the best in those benchmarks (16,17,74,96–104). Therefore, the seek for new approaches continues. On the other hand, Bhattacharya *et al*. (2017) showed the importance of the difference between wild-type and mutant structures to understand their relationship with diseases (10). Furthermore, Missense3D also supported the wild-type-mutant difference-based comparison by showing that 40% of the disease-associated missense variants disrupted the protein structures whilst only 11% of the neutral missense variants harm the structures (18). Therefore, we hypothesized that prediction performance could be enhanced for the impact of missense variants by integrating the structural difference methodology with the known successful features which can be grouped as sequence, structure, and dynamics-based features. Such procedures are applied in many functional studies including manually-

performed MD simulations, especially to investigate the mechanisms of the diseases associated with missense variants (105–111).

2.2 Databases

2.2.1 ClinVar

ClinVar (112) is a public database to deposit the associations between genetic variants found in human and disease/health status of the individuals. ClinVar also contains the information of assertion criteria for those variant-disease relationships. These criteria are called as review status (or stars) of ClinVar. Four review stars indicate that the association is discovered by performing the practice guideline provided by ClinVar. Three stars indicate that the assertion was performed by an expert panel. Two stars refer to similar interpretations and assessments provided by multiple submitters without any conflicts. One star means that the relationship is based on the assertion of a single submitter, or multiple submitters have conflicting interpretations (112). By using these review status, the ClinVar database might be filtered to keep only the variants interpreted with high confidence (17).

2.2.2 dbSNP

dbSNP is a public database of single nucleotide polymorphisms (SNP) within and across different species (113). dbSNP Reference SNP (rs or RefSNP) number is unique to each known variation and facilitates large-scale studies for database-to-database mappings and annotations (113). There are many tools and packages that use rs number.

2.2.3 UniProt

UniProt (114) is a public database that contains comprehensive and high-quality information about the sequence and function of the proteins within and across different species. Therefore, UniProt annotations could be used as features in terms of the function of the location that has the missense change such as catalytic residue, calcium-binding, or DNA-binding region.

2.2.4 Protein Data Bank

RCSB Protein Data Bank (PDB) (43) is the freely-accessible public archive for protein structures obtained by structural biology experiments such as X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy. PDB also contains the related information about the structures and experiments that could be useful to assess the reliability of the structure (e.g., atomic resolution) and annotations (e.g., ligand-binding residues).

2.2.5 PDBSWS

PDBSWS (115) includes three publicly-accessible maps that are useful for annotating the protein structures for functional information. First one maps the protein chains of PDB structures to the UniProt annotations. Second map is the residue-level map of PDB structures to the UniProt annotations. The last one is a map of all mutations found in the PDB structures when they are compared to the UniProt sequences.

2.3 Classification

Interpretability of classification models is important to understand the significance of the features. Therefore, it might be a good strategy to first build explainable models (e.g., logistic regression and decision tree) and then build more complex models (also called “black boxes”) with less interpretability (e.g., random forests and gradient boosting) to measure the trade-off between interpretability and prediction performance (116,117).

2.3.1 Regression-based

Logistic regression is a commonly-used classification methodology that makes estimations based on maximum likelihood approach (118). Depending on the imbalance between sample size and number of features, maximum likelihood might cause unstable estimation. Therefore, the regularization methods, namely lasso, ridge and elastic net, are widely-used to stabilize estimation by applying penalties (118). Lasso penalizes the cost function of the model by adding the absolute sum of the coefficients to lower the effect of coefficients with low performance on explaining the variance in data, and this penalty term makes lasso good in feature selection (119). Ridge adds the square of the magnitude of the coefficients to lower both bias and variance in the model (120). Furthermore, elastic net combines the lasso and ridge methodologies by taking the advantages of both techniques (120).

2.3.2 Decision Tree-based

Decision trees are one of the most popular supervised learning methods that can be used for both regression and classification tasks (121). A decision tree has root and

leaves, referring to first split and the last decision, respectively (121). Its branched structure makes features reusable at different levels, and its tree visualization significantly improves its interpretability. On the other hand, there are some metrics used to split data into nodes. Gini index is one of the most popular splitting criteria, and it measures the divergence between the probability distributions from the values of target attributes (121).

2.3.3 Ensemble Methods

In many studies with the purpose of building a classifier, numerous decision trees are used together as integrated by utilizing ensemble methods such as random forests and boosting (122). Random forest algorithm is basically the building many trees with randomly-selected subsets of the given features to reduce overfitting, and it makes a majority voting on the decisions of these decision trees to make its own prediction (122). By contrast with the random forest's multiple trees built in parallel, boosting builds a decision tree model by sequentially optimizing the weights based on splitting performance (122).

3. MATERIALS AND METHODS

All calculations were performed by using R v4.0.3 (23) via RStudio v1.4.1103 (123). The details about the called packages, programs, and application programming interfaces (APIs) were given in the following sub-sections.

3.1 Datasets

ClinVar (112) (release 10.06.2021) was used to build the dataset of missense variants. Number of all variants in this release was 950,237. As shown in Figure 3, the ClinVar database was filtered to obtain the variants that have both wild-type and mutant experimental structures.

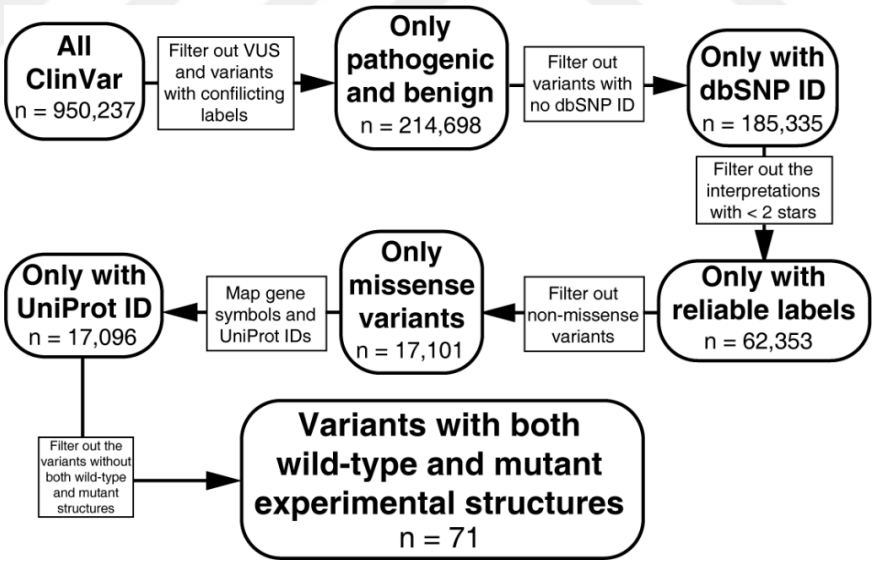

Figure 3. Summary of the dataset preparation process.

The ClinVar database was filtered for "Pathogenic", "Pathogenic/Likely_pathogenic", "Benign" and "Benign/Likely_benign" missense variants with two or more review stars to keep only the variants asserted with high confidence. The variants with no rs numbers were excluded to facilitate mappings and

annotations during this procedure. The high confidence missense variants selected totaled up to 17,101. Afterwards, the UniProt (114) accession identifiers (ID) were mapped by using biomaRt (124,125), and the total remaining variants was 17,096. Then, the corresponding amino acid changes for the variants were mapped from ClinVar by using NCBI's E-utilities (126,127). By using the UniProt IDs, variant positions and reference residues, the corresponding PDB IDs (43) of the wild-type structures were mapped from the residue-level database of PDBSWS (115). Similarly, the residue-level database and mutation map of PDBSWS were used to map the PDB structures including the variants of interest. To obtain comparable wild-type and mutant structures, the PDB structures with additional mutations besides the variant of interest were also excluded. After the whole filtering process, the sample size dropped to 71. Moreover, 52 of them were "Pathogenic" or "Pathogenic/Likely_pathogenic" (together will be called as "Pathogenic" from here onwards) whereas 19 of them were "Benign" or "Benign/Likely_benign" (together will be called as "Benign" from here onwards). This final dataset is presented in Appendix 1.

3.2 Structural Information Processing

The "clean.pdb" function of bio3d (91,92) was used to correct the modified amino acids observed in 4 of the 123 mapped wild-type and mutant PDBs. Bio3d was used to inspect the connectivity, and homology modeling was performed to fill missing residues within structures (if exists) and to model the mutant structures. To fill the missing residues, basic protocol of MODELLER v10.1 (128,129) was used.

3.3 Model Features

Features used to compare the structures of wild-type and mutant were grouped as sequence-based, structure-based, and dynamics-based. For each feature, a denotation

was used for further results, and they are given with the definitions in the next sub-sections. Of note, all difference-based features were calculated by extracting the value for wild-type from the mutant's value.

3.3.1 Sequence-based

Position-Specific Independent Counts (PSIC) (24) was calculated by using PolyPhen-2 webserver (http://genetics.bwh.harvard.edu/pph2/) (33). PSIC for wild-type (wtPSIC) and the difference between PSIC in the mutant and wild-type (dPSIC).

If the variant site is in a functional region defined in UniProt (114) (i.e., binding site for metal, DNA, calcium or any chemical group; and functional motifs, regions or domains such as zinc finger; and post-translational modification sites), the feature (hitUniProt) is TRUE, otherwise FALSE.

Amino acid alphabet used by the EMBL-EBI's PDBeXpress Service (https://www.ebi.ac.uk/pdbe-srv/pdbexpress/) (130) was used. Thus, the amino acid change groups (aa_groups) were aliphatic (Ala, Ile, Leu and Val), aromatic (Phe, Trp and Tyr), basic (Arg, Lys and His), acidic (Asp and Glu), polar neutral (Asn, Gln, Cys, Met, Ser and Thr), and unique (Pro and Gly). This feature (aa_groups) was used as categorical variable including each combination of the groups (e.g., aromatic_aliphatic, basic_basic, polar.neutral_unique and acidic_aliphatic).

Mass, volume and hydropathy index of amino acids were collected from IMGT's educational pages (131). These features were applied by extracting the values of the wild-type amino acids from the value of mutant amino acids, denoted as dMass, dVolume and dHydropathy.

The pre-calculated BLOSUM62 substitution scoring matrix was taken from NCBI (https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt). The feature was denoted as blosum62_score by matching the reference and altered amino acids in the matrix.

Pfam IDs were scanned by using the UniProt sequence of the proteins in the EMBL-EBI's HMMER web service (132). Pfam entropy was calculated by using "entropy" function of bio3d, after the Pfam alignment for the proteins were obtained by "pfam" function of bio3d. Pfam entropy features were Pfam_entropy_H (standard entropy score for a 22-letter amino acid alphabet), Pfam_entropy_H.10 (entropy score for a 10-letter alphabet), Pfam_entropy_H.norm (normalized entropy score for a 22-letter alphabet between 0 and 1 based on lowest and highest entropy), Pfam_entropy_H.10.norm (normalized entropy score for a 10-letter alphabet between 0 and 1 based on lowest and highest entropy), and Pfam_entropy_freq (residue frequency at the position of interest as percentage). Of note, there was no hit in one case (i.e., the protein encoded by *AGXT* gene with UniProt ID: P21549), mice package in R was used for imputation of these missing values by using predictive mean matching method with 100 maximum iterations (133).

3.3.2 Structure-based

The local environment of a residue was defined as the 7 Å radius around its alpha carbon atom. Residues within the local environment of the variant sites were detected and their hydropathy indices were summed to define how much hydrophobic or hydrophilic the position's environment was. The wild-type's local hydropathy (wtLocal_hydropathy) and the difference between these local environments (dLocal_hydropathy) were calculated.

Solvent accessible surface area (SASA) was calculated by using 1.4 Å probe radius via VMD. Relative solvent accessibility (RSA) refers to normalized SASA, which defined as the SASA over maximum theoretical SASA for the residue of interest (134). RSA of variant site in wild-type (wtRSA) and difference of RSA at the variant site (dRSA) were used as features.

Torsional angles were calculated by using bio3d package. Then, unfavorable torsions were detected based on the definition of which phi angle is equal or greater than 135 degrees, or whether psi angle equals or is between -150 and -160 degrees (135), where phi angle of residue *i* refers to the torsion C_{i-1}-N_i-$C\alpha_i$-C_i and psi angle to the torsion N_i-$C\alpha_i$-C_i-N_{i+1} (136). Moreover, the difference in the numbers of unfavorable torsions (dNum_unfavor_torsions) was used as feature.

Secondary (2D) structures were predicted by STRIDE (137) with by Visual Molecular Dynamics (VMD v1.9.3) (138) executables. 2D structures were represented in three features: 2D structure of the variant site in wild-type structures (wt2D), existence of a change in the 2D structure of the variant site (d2D) as True/False, and existence of any four consecutive 2D structure changes within the protein.

FoldX (v5.0) (139) was used to calculate the folding free energy indicating the stability of protein. This is used as the difference between wild-type and mutant by extracting the wild-type's stability from mutant's (ddG_FoldX).

Root-mean-square deviation (RMSD) shows the difference between the coordinates of the Cα atoms of wild-type and mutant protein structures, so basically indicates the differences in protein folding. It is calculated via bio3d by fitting the wild-type and mutant structures to each other.

After fitting the wild-type and mutant structures, center of mass was calculated by using bio3d package. Euclidean distance between their center of masses (distCoM) was used as feature.

Radius of gyration (Rg) is a parameter to measure the compactness of protein structure (140). Rg of wild-type structure (wtRg) and the difference between the Rg of wild-type and mutant structures (dRg) were used as features.

3.3.3 Dynamics-based

Flexibility of each residue in the protein structure was calculated based on the fluctuations derived from the first fourteen mode (as default) of anisotropic network modeling (ANM) at 310 K temperature with the default parameters of bio3d (v2.4.1). Spearman's rho correlation coefficient and root-mean-square inner product (RMSIP) were calculated to measure the similarity between the wild-type's and mutant's residue flexibility distributions (141), and these features were denoted as rhoFlexOverall and rmsipFlexOverall, respectively. Furthermore, flexibilities were also obtained for the variant site, and these features were flexibility of the mutation site in wild-type (wtFlexSite) and difference in the flexibilities of the variant site (dFlexSite).

Deformation energy (DefE), which is calculated via bio3d package, indicates the atomic motion relative to neighboring atoms by including local flexibility and force constants from normal mode analysis (91,142). The features related to this metric were contribution of the variant site in the wild-type to DefE at first three non-trivial modes separately (dDefE_site_contribution_1st_mode, dDefE_site_contribution_2nd_mode, dDefE_site_contribution_3rd_mode), and root-mean-square (rms) distance between the DefE of first three modes (together as a vector) of wild-type and mutant structures (rmsDefE_site_contributions). Similarly, dDefE_sum_1st_mode,

dDefE_sum_2nd_mode, dDefE_sum_3rd_mode and rms distance of DefE sums at first three non-trivial modes separately (rmsDefE_sums) were also used as separate features, where DefE sum indicates the total DefE at a mode. Of note, non-trivial modes refer to the modes after filtering out the modes related to the rotational and translational movements.

Stiffness is defined as the mechanical resistance of residues to external forces calculated by elastic network modeling such as ANM (143). MechStiff (144) of ProDy was used with default parameters for the calculations, and only extracted feature was rms distance of stiffness range between wild-type and mutant (rmsStiffnessRange).

Perturbation response scanning measures how sensitive and effective each residue in a protein for allosteric signals by sequentially distorting the residues up to 1.5 Å, and calculations were performed via ProDy (94,145,146), called from R via reticulate (147). Sensitivity of the variant site in wild-type (wtSensitivity), sensitivity difference (dSensitivity) and rms distance of range of sensitivity within protein (rmsSensitivityRange) were the features. Similarly, wtEffectiveness, dEffectiveness and rmsEffectivenessRange were the features for effectiveness.

3.4 Classification

Firstly, the training dataset was prepared to eliminate the class imbalance and small sample size bias, and then the different classification techniques were used and compared to each other to achieve high prediction performances in an unbiased way as explained in the following sub-sections.

3.4.1 Imbalanced Data Preparation

While training models on the small and imbalanced datasets, overfitting might be a significant problem. To overcome this problem, synthetic minority over-sampling technique (SMOTE) was developed by Chawla *et al.* (2002). SMOTE oversamples the minority class with synthetic examples based on the "k" nearest neighbors (148). When the datasets are balanced by SMOTE, classification performances are improved in comparison to the imbalanced datasets (149,150).

The final dataset included 71 variants as 19 benign and 52 pathogenic variants. Since this final dataset is small and imbalanced, SMOTE was applied on the training dataset (15 benign and 42 pathogenic) with k = 3, and amount of oversampling was 500% for the minority class (i.e., "benign" in this study) whilst 120% for the majority class (i.e., "pathogenic" in this study). The smoted training dataset included 90 benign and 90 pathogenic samples, and this dataset was used for training purposes. SMOTE was performed by using the DMwR (151) package in R. Of note, test dataset contained randomly-selected 4 benign and 10 pathogenic variants. Appendix 1 shows which variants were held in which of the training and test datasets.

3.4.2 Modeling Scheme

We initially preferred the most interpretable models, and then built less interpretable models to measure the trade-off between performance and interpretability. Therefore, the regularized logistic regression, namely elastic net model as well as decision tree, random forest and gradient boosting models were built. Random forest and gradient boosting models were also used to measure variable importance. Additionally, 5-times 5-fold cross validation, indicating separating the

training dataset into five parts and sequentially using one of them as an internal test data repeatedly for five times, was applied to avoid overfitting (152).

3.4.3 Regression-based

The caret (153) and glmnet (154) R packages were used for training of the binomial (binary classification) elastic net model. Penalties of the elastic net model were optimized as alpha = 0.8248367 and lambda = 0.007261618 by applying 5-times 5-fold cross validation and comparing their performance automatically by caret. Afterwards, these penalty values were used to build a single elastic net model by glmnet.

3.4.4 Decision Tree-based

The rpart (155) R package was used to build a rpart decision tree by using Gini index and 5-fold cross validation. The tree was pre-pruned by limiting the depth of tree as 4, by keeping at least 5 samples at each node of tree, and by using 0.01 as the step size. The rattle (156) R package was used for visualization of trees.

3.4.5 Random Forest-based

The randomforest (157) package was used to build random forest models. To optimize the selection of the number of random features, each value from 1 feature to all 44 features were tried to build random forest models with 10,000 trees, and out-of-bag error was calculated for each value. Out-of-bag error refers to the error rate in the test samples remained out of the training part of the dataset when a tree of the random

forest trained by randomly-chosen samples as a result of bagging process. The optimal values were 1, 2, 5, 6, 7, 8, 9, 24, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, and 44 with the equally minimum out-of-bag error. Therefore, each of these values were tried again, and based on the performance on the test dataset, 27 was chosen as the number of variables randomly sampled at each split.

3.4.6 Gradient Boosting-based

For gradient tree boosting modeling, the gbm (158) R package was used. The gradient tree boosting model was built by trying 10,000 trees with 0.01 shrinkage, 4 maximum tree depth, 5-fold cross-validation, at least 5 samples at each node, and Bernoulli distribution indicating binary classification. Based on the cross-validation error, the optimal number of trees was 844. Therefore, 844 trees were iterated for the gradient tree boosting model.

3.5 Performance Assessments

Accuracy, sensitivity referring to the accuracy level within the positive class (i.e., “Pathogenic”), specificity indicating the accuracy level within the negative class (i.e., “Benign”), and Matthews correlation coefficient (MCC), which is reliable metric that works well for imbalanced data as well, were used to assess the performance of predictors (159,160).

4. RESULTS

4.1 Models and Feature Importance

Using all features without weighting or filtering might cause redundancy and low prediction performance. However, all used modeling methods in this study (i.e., elastic net, decision tree, random forest and gradient tree boosting) prioritize the features with their classification powers. Moreover, observation for the importance levels of the features might be useful to understand which features are meaningful at also biological level. All coefficients calculated of the elastic net model are given in Table 1.

Table 1. Coefficients of the elastic net model for each feature, sorted by importance.

Features	Coefficients
intercept	10.42
aa_groups = polar.neutral_basic	5.25
aa_groups = aromatic_basic	-3.98
wtPSIC	3.55
hitUniProt	3.48
aa_groups = acidic_polar.neutral	3.13
aa_groups = unique_aliphatic	2.35
d2D	1.94
aa_groups = aliphatic_aromatic	-1.84
dDefE_site_contribution_1st_mode	1.83
wtRSA	-1.73
aa_groups = unique_polar.neutral	-1.56
dDefE_site_contribution_3rd_mode	-1.54
Pfam_entropy_H.norm	-1.38
rmsipFlexOverall	1.25
dPSIC	1.06
aa_groups = aliphatic_acidic	1.05

Only the features with non-zero coefficients influence the elastic net model since others are multiplied by zero, and importance of a feature is directly proportional to the absolute value of its coefficient. Thus, the features are ordered from the most important to less in Table 1, and if the value is below 1, it is assumed as unimportant and not shown in the Table 1. The ten best features were aa_groups =

polar.neutral_basic (indicates a polar neutral amino acid changed into a basic amino acid), aa_groups = aromatic_basic, wtPSIC, hitUniProt, aa_groups = acidic_polar.neutral, aa_groups = unique_aliphatic, d2D, aa_groups = aliphatic_aromatic, dDefE_site_contribution_1st_mode and wtRSA, respectively.

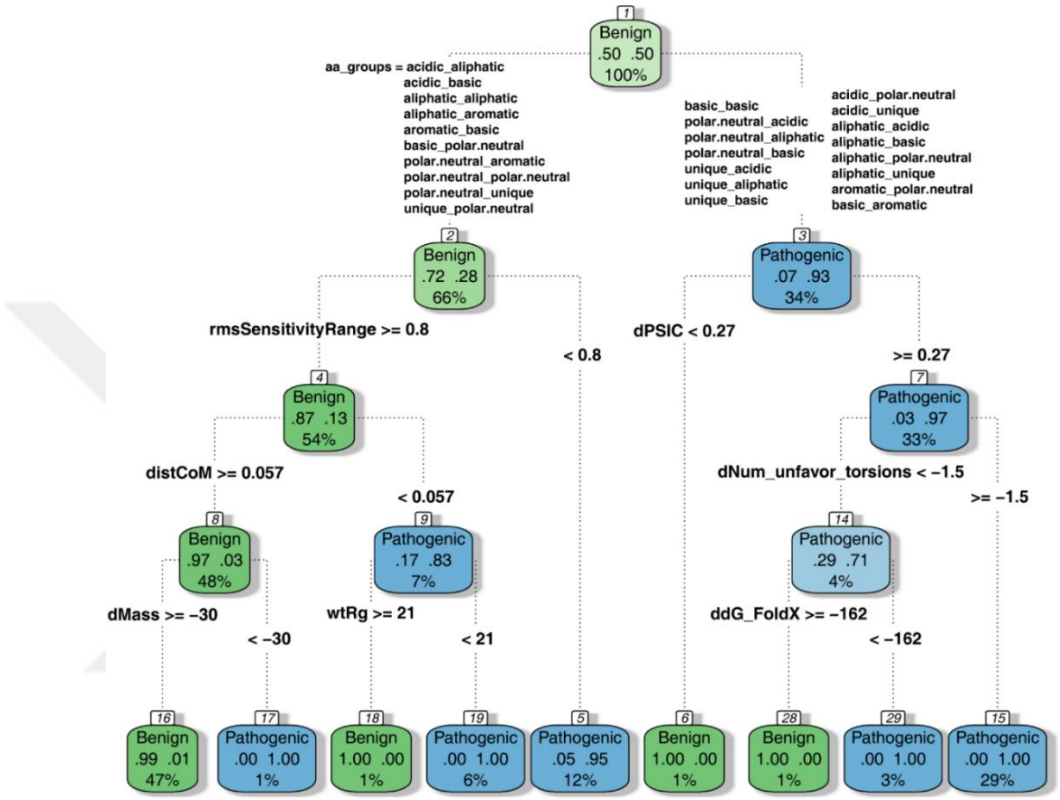

Figure 4. The decision tree model built on the smoted training dataset.

As shown in Figure 4, the most important feature (root) automatically selected by Gini index in the decision tree modeling was the aa_groups. The other features (i.e., rmsSensitivityRange, distCoM, dMass, wtRg, dPSIC, dNum_unfavor_torsions and ddG_FoldX) were also important to split the data subsets created by the former feature in the tree during classification.

Random forest model was built with 10,000 decision trees by randomly selecting 27 features to model each tree. Importance of each variable was calculated based on mean decrease in accuracy and mean decrease in Gini when the feature is removed, and the related plots for the best 15 features are given in Figure 5.

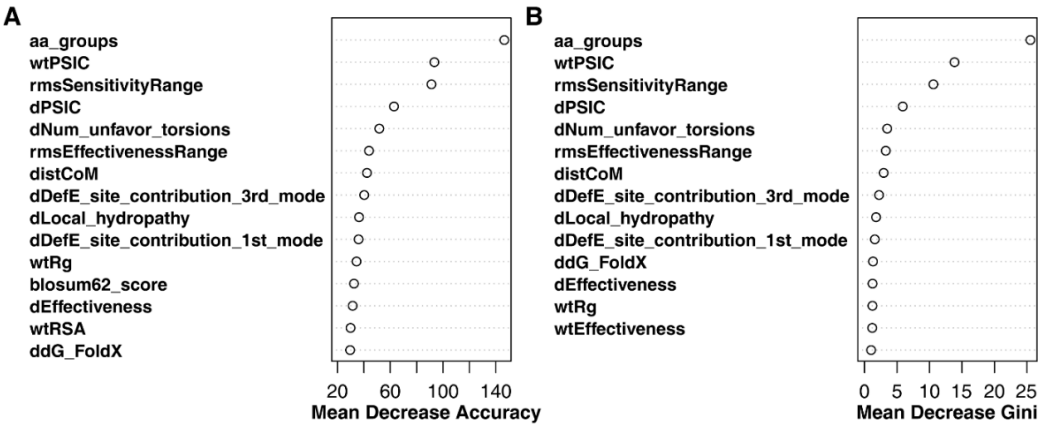

Figure 5. Variable importance plots extracted from the random forest model, (A) based on the mean decrease in accuracy and (B) based on the mean decrease in Gini.

Based on the random forest model, the best ten variables were aa_groups, wtPSIC, rmsSensitivityRange, dPSIC, dNum_unfavor_torsions, rmsEffectivenessRange, distCoM, dDefE_site_contribution_3rd_mode, dLocal_hydropathy and dDefE_site_contribution_1st_mode, respectively, according to both mean decrease in accuracy and Gini, as demonstrated in Figure 5.

Gradient decision tree boosting model was built with 844 iterations as the optimum. By using gradient tree boosting, variable importance for all features was calculated based on the relative influence of each feature on the model. The results are shown for the features with relative influence > 1 in Figure 6.

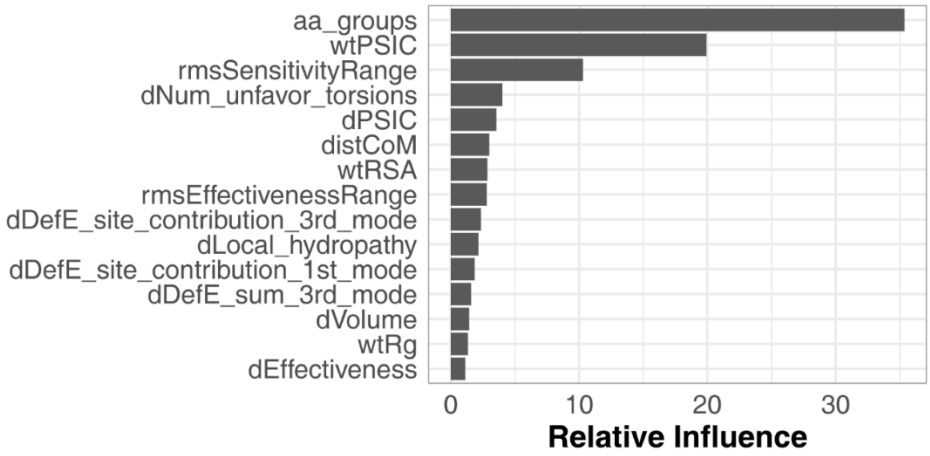

Figure 6. Variable importance plot extracted from the gradient tree boosting model, based on relative influence of each feature on the model.

Based on the gradient boosting model, the best ten features were aa_groups, wtPSIC, rmsSensitivityRange, dNum_unfavor_torsions, dPSIC, distCoM, wtRSA, rmsEffectivenessRange, dDefE_site_contribution_3rd_mode, dLocal_hydropathy and dDefE_site_contribution_1st_mode, respectively, as demonstrated in Figure 6.

4.2 Classification Performances

The elastic net, decision tree, random forest and gradient tree boosting models were trained by using the smoted training dataset. To compare their performances, the accuracy, sensitivity, specificity and MCC metrics were used. Comparisons were performed on the smoted training dataset (90 benign and 90 pathogenic), the training dataset (15 benign and 42 pathogenic) and the test dataset (4 benign and 10 pathogenic), and the results are shown in Table 2.

Table 2. Prediction performance metrics of the elastic net, decision tree, random forest and gradient tree boosting models on the smoted training dataset, the training dataset and the test dataset. i.e., accuracy, sensitivity, specificity and Matthews correlation efficient (MCC).

Model / Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Elastic net model				
Smoted training dataset	98.33	96.67	100.00	0.97
Training dataset	96.49	95.24	100.00	0.92
Test dataset	85.71	80.00	100.00	0.73
Decision tree model				
Smoted training dataset	98.89	98.89	98.89	0.98
Training dataset	94.74	95.24	93.33	0.87
Test dataset	78.57	80.00	75.00	0.52
Random forest model				
Smoted training dataset	100.00	100.00	100.00	1.00
Training dataset	98.25	97.62	100.00	0.96
Test dataset	85.71	100.00	50.00	0.65
Gradient tree boosting model				
Smoted training dataset	100.00	100.00	100.00	1.00
Training dataset	98.25	97.62	100.00	0.96
Test dataset	78.57	90.00	50.00	0.44

Based on the prediction metrics shown in Table 2, the best performing models on both smoted training dataset and training dataset were gradient tree boosting and random forest models with the same metrics. Moreover, the elastic net and decision tree models had only a few mistakes on these datasets as expected. The test dataset is the most important dataset for benchmarking models since it is a completely unseen dataset to the classifiers. On the test dataset, the elastic net model had the best performance with 0.73 MCC value. Even though the random forest was the second-best predictor with 0.65 MCC value and had the same accuracy level with the elastic net model, random forest and gradient boosting models seems overfitted to the positive class (i.e., pathogenic) and so cannot predictive the neutral variants well.

As a result of the benchmarking of the models, the elastic net model was chosen as the main predictor presented in this study to use for the benefit of scientific society. The elastic net model was named as Rmut, by integrating the name of "R" programming language with the "mut" as an abbreviation for "mutation". Also, its phonetics is similar to "*armut*" which is a Turkish word referring "pear". Thus, its logo is a pear with a mutation -bite-, as shown in Figure 7.

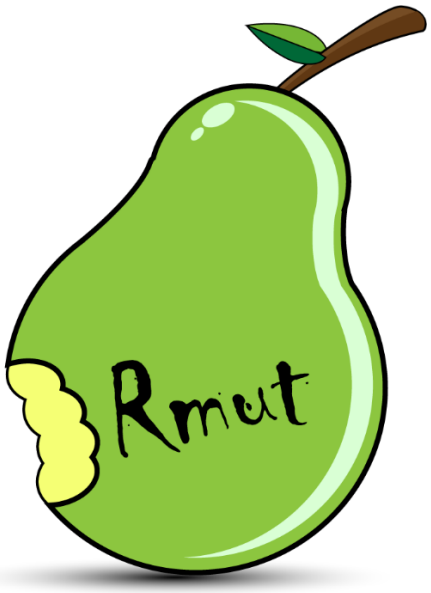

Figure 7. The logo of Rmut R package, which includes the functions to calculate the features used in this study as well the pre-trained elastic net model to predict the impact of variants.

Rmut R package has functions to replicate this study, and the functions required to individually calculate all features used in this study and related data and functions. Its dependencies are the packages and programs used in this study, e.g., bio3d, STRIDE, VMD and FoldX. The package also has the pre-trained elastic net model as the variant impact predictor called Rmut that will be available to predict the impact of new variants after a protein modeling approach is optimized as a future study. Furthermore, Rmut is available on GitHub (https://github.com/ugerlevik/Rmut).

4.3 Comparison with Missense3D and Rhapsody

Missense3D (18) and Rhapsody (15) were the comparators because the predictor developed in this study (namely Rmut) mainly integrates these two approaches. To obtain the predictions of Missense3D and Rhapsody, their webservers (i.e., Missense3D: http://missense3d.bc.ic.ac.uk/missense3d/ on "Position on 3D Structure" mode, and Rhapsody's batch query: http://rhapsody.csb.pitt.edu/batch_query.php) were used. Rhapsody has two "probably deleterious" and one "probably neutral" predictions within total, and these were assumed as "deleterious" and "neutral", respectively. Missense3D gives result as "No structural damage detected" or introduction of one of the structural changes of their interest (18). Thus, it was assumed that Missense3D's prediction for a variant with no structural damage is benign, otherwise pathogenic.

Rhapsody and Missense3D predictions for the training and test datasets were obtained from their webservers (http://rhapsody.csb.pitt.edu/batch_query.php and http://missense3d.bc.ic.ac.uk/missense3d/, respectively), and the corresponding results were given in Table 3.

Table 3. Prediction performance metrics of Rmut, Rhapsody and Missense3D on the training and test datasets. i.e., accuracy, sensitivity, specificity and Matthews correlation efficient (MCC).

Predictor	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Rmut				
Training dataset	96.49	95.24	100.00	0.92
Test dataset	85.71	80.00	100.00	0.73
Rhapsody				
Training dataset	82.46	80.95	86.67	0.62
Test dataset	85.71	80.00	100.00	0.73
Missense3D				
Training dataset	42.11	21.43	100.00	0.26
Test dataset	42.86	20.00	100.00	0.26

As shown in Table 3, performances of Rhapsody and Rmut were identical in the test dataset whereas Missense3D predictions were far worse. The performance of Rhapsody was also good in the training dataset with 0.62 MCC value. Since this is the training dataset of Rmut, it will not be fair to compare with Rhapsody at this point. Similarly, 4 of the 10 pathogenic variants and 1 of the 4 benign variants in the test dataset were in the training dataset of Rhapsody. Because this situation creates a bias towards overestimation of Rhapsody's performance, these variants were extracted from the test dataset and performances of Rhapsody and Rmut were compared again. Thus, this dataset contained 3 benign and 6 pathogenic variants (See "Rhapsody Training Info" column of Appendix 1), and the performances are given in Table 4.

Table 4. Prediction performance metrics (i.e., accuracy, sensitivity, specificity and Matthews correlation efficient (MCC)) of Rmut and Rhapsody on the test dataset of which the four pathogenic variants and one benign variant from the Rhapsody's training data were filtered out.

Predictor	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Rmut	88.89	83.33	100.00	0.79
Rhapsody	77.78	66.67	100.00	0.63

After filtering out the data found in the training of Rhapsody from out test dataset, Rmut's performance was higher with 0.79 MCC value as shown in Table 4.

5. DISCUSSION AND CONCLUSION

In this study, the purpose was improving the variant impact prediction performance and representing this predictor as an R package. Moreover, this first goal was achieved by combining the comparison methodology with the well-known sequence and structure-based features and protein dynamics-based features derived by realistic approaches such as elastic network models. Additionally, the comparison method refers to usage of the difference between wild-type and mutant to calculate features instead of only wild-type. The second goal was achieved by proposing the predictor to the community as an R package, called Rmut, which is available on GitHub (https://github.com/ugerlevik/Rmut).

Four different machine learning methods, i.e., elastic net, decision tree, random forest and gradient tree boosting, were used to find out which one can perform the best classification on this type of dataset with the most interpretability and with no overlearning. Moreover, all these modeling approaches perform an internal feature selection by prioritization them based on their separation power on classes. As a result of this prioritization, all models ended up with the same features as the most important features as shown in Table 1 and Figures 4-6. Therefore, it can be inferred that these features have probably the biological significance on the protein structure and dynamics. For example, aa_groups which indicates to the amino acid type change (e.g., aliphatic to aromatic or basic to unique) was the most important feature without exception. This could refer that change of a polar neutral amino acid to a basic amino acid might significantly disrupt the structure and/or function of the protein. It makes sense because the basic amino acid in the mutant protein may not adapt to the wild-type's neutral environment.

Besides the well-known separative features such as PSIC and RSA, the features related to the allosteric signals were found within the best ten features of the models.

These features were the allosteric signals (i.e., sensitivity and effectiveness), deformation energy contribution of the variant site and hydropathy change in the local environment (within 7 Å) of the variant site. Their importance in missense variant impact prediction might indicate that disease-driving changes related to missense variants, which are small with only one amino acid change, could be mostly related with the local changes rather than a global effect on the protein structure and dynamics.

Change in the number of unfavorable torsions within the structure (dNum_unfavor_torsion) and distance between center of mass of the mutant and wild-type proteins (distCoM) were uncommonly-used features within the variant impact predictors in the literature. However, they were found as important features here, and so they might be indicators for significant changes that could be investigated further.

Within the four different machine learning model, the best one was also the most interpretable one, namely elastic net model. Generally, it is expected that ensemble methods such as random forests and gradient boosting models are expected to have higher performances than the single models such as elastic net and decision tree models. However, small sample size might cause overlearning of one of the classes. Moreover, in the second-best model, which was the random forest model, this situation was observed for the positive class, as can be seen in Table 2. After this point, the elastic net model became the variant impact predictor, namely Rmut.

Rmut was benchmarked with the two predictors inherited their approaches to it, namely Rhapsody and Missense3D. As shown in Table 3, Missense3D had a lower performance than the others as well as than the randomly-selecting one of the classes which is theoretically expected 50% accuracy. However, this situation was not much surprising because Missense3D has only focus on the prediction of structural damage instead of a classification task for pathogenic and benign variants. Therefore, the assumption of pathogenicity in the presence of structural damage, which was made in

this study to make Missense3D comparable, was not correct. On the other hand, even though Rmut had better performance on the unseen data (Table 4), Rhapsody is also a good predictor with similar performances to Rmut (Table 3).

In this study, one of the limitations was the small sample size and imbalance of the dataset obtained by the process shown in Figure 3. However, SMOTE is a well-known and robust method specifically designed to overcome such problems, and it also worked well in this study, as can be seen with the prediction performances on the test dataset (Table 2). On the other hand, performances on the test dataset also demonstrate that another limitation, which was the subset bias created because of selecting only the variants with corresponding structures, was also overcome. The third limitation is the need of a reliable wild-type and mutant structure without missing residues, to make good predictions. The missing residue part can be surpassed by homology modeling as performed in this study. Moreover, there are many reliable and increasing number of structures for many proteins in the PDB database. Therefore, modeling the wild-type proteins will also not a major problem. However, modeling the mutant proteins might require some further optimizations because homology modeling algorithms are not expected to have one residue sensitivity although they can handle steric clashes very well. At this point, some advanced structure optimization techniques related to realistic approaches such as small optimization runs based on molecular dynamics simulations might be the solution to model the mutant structures by using the wild-type structures (161). This point will be investigated further.

In conclusion, representing the variant impact predictors as user-friendly packages is one the most crucial jobs for contribution to the scientific community and the progress in the field. Therefore, Rmut is presented as an easily-accessible R package. Rmut outperformed the two closest predictors that inspired its methodology, which was bringing the comparison approach mostly used for sequence-based features and structure-dynamics-based features together. However, it has also some limitations as expected, but they can be overcome in future studies.

6. APPENDICES

Appendix 1. The curated dataset from ClinVar.

It contains the variants with more than two ClinVar review stars that have protein structures in the PDB database for both wild-type and mutant structures.

rs Number	UniProt ID	Protein Change	Wild-type PDB ID	Mutant PDB ID	Label	Dataset	Rhapsody Training Info
rs76763715	P04062	Asn409Ser	2wkl_A	3ke0_A	Pathogenic	training	known deleterious
rs386134243	P02545	Arg335Trp	1x8y_A	3v4q_A	Pathogenic	training	new
rs57920071	P02545	Arg482Trp	1ifr_A	3gef_A	Pathogenic	test	new
rs2274064	P19878	Lys181Arg	1e96_B	1hh8_A	Benign	training	new
rs727502886	P35609	Ala119Thr	5a36_A	5a4b_A	Pathogenic	training	known deleterious
rs72470545	O43464	Gly399Ser	5m3n_A	5tny_A	Benign	training	new
rs13045	Q9NZJ5	Gln166Arg	5sv7_A	4yzs_A	Benign	test	new
rs1045485	Q14790	Asp285His	1qtn_A	1f9e_A	Benign	training	known neutral
rs121913500	O75874	Arg132His	1t09_A	3inm_A	Pathogenic	training	new
rs121913499	O75874	Arg132Cys	1t09_A	5k10_A	Pathogenic	training	new
rs121908529	P21549	Gly170Arg	1h0c_A	1j04_A	Pathogenic	training	known deleterious
rs587777331	P47897	Gly45Val	4ye6_A	4ye9_A	Pathogenic	training	known deleterious
rs10794537	P35475	His33Gln	3w81_A	4kgj_A	Benign	training	known neutral

Appendix 1. The curated dataset from ClinVar (cont.)

rs5030732	P09936	Ser18Tyr	2etl_A	2len_A	Benign	training	new
rs121913506	P10721	Asp816His	1t46_A	3g0f_A	Pathogenic	training	known deleterious
rs12514417	P49419	Lys439Gln	4x0t_A	2j6l_B	Benign	training	known neutral
rs116567033	O95363	Thr246Met	3cmq_A	5mgu_A	Benign	training	new
rs5987	P00488	Val651Ile	1evu_A	1fie_A	Benign	training	known neutral
rs1057090	Q8NEM0	Ala761Val	3sht_A	3szm_A	Benign	test	known neutral
rs121909329	P55072	Arg155His	3qc8_A	3hu3_A	Pathogenic	training	known deleterious
rs3208406	Q9NQT5	Tyr225His	2nn6_G	6d6q_G	Benign	training	known neutral
rs1800546	P05062	Ala150Pro	1qo5_A	1xdl_A	Pathogenic	training	known deleterious
rs74799832	P07949	Met918Thr	2ivs_A	4cki_A	Pathogenic	training	known deleterious
rs4935502	Q96QU1	Asp435Ala	5t4m_A	5t4n_A	Benign	training	new
rs1057519047	P21802	Lys641Arg	1gjo_A	2pzr_B	Pathogenic	training	known deleterious
rs77543610	P21802	Pro253Arg	1e0o_B	1iil_E	Pathogenic	test	new
rs79184941	P21802	Ser252Trp	1e0o_B	1ii4_E	Pathogenic	test	new
rs104894227	P01112	Lys117Arg	121p_A	2quz_A	Pathogenic	training	known deleterious
rs28933406	P01112	Gln61Lys	121p_A	2rgb_A	Pathogenic	training	known deleterious
rs104894230	P01112	Gly12Val	121p_A	1he8_B	Pathogenic	training	known deleterious
rs104894230	P01112	Gly12Asp	121p_A	1agp_A	Pathogenic	training	known deleterious

Appendix 1. The curated dataset from ClinVar (cont.)

rs104894229	P01112	Gly12Cys	121p_A	4l9s_A	Pathogenic	training	known deleterious
rs33950507	P68871	Glu27Lys	1a3n_B	1nqp_B	Pathogenic	training	new
rs33930165	P68871	Glu7Lys	1a3n_B	1k1k_B	Pathogenic	training	new
rs3729989	Q14896	Ser236Gly	2v6h_A	2avg_A	Benign	training	new
rs118204095	P08397	Arg167Gln	3ecr_A	3eq1_A	Pathogenic	training	known deleterious
rs1800973	P61626	Thr88Asn	1c46_A	1w08_A	Benign	test	new
rs121918461	Q06124	Asp61Gly	3tkz_A	4h1o_A	Pathogenic	training	known deleterious
rs397507520	Q06124	Glu139Asp	4dgp_A	4nwg_A	Pathogenic	test	known deleterious
rs121918456	Q06124	Tyr279Cys	3b7o_A	4dgx_A	Pathogenic	test	known deleterious
rs121918463	Q06124	Phe285Ser	3b7o_A	5i6v_A	Pathogenic	training	known deleterious
rs28933386	Q06124	Asn308Asp	3b7o_A	4nwf_A	Pathogenic	training	known deleterious
rs121918470	Q06124	Gln510Pro	3b7o_A	4h34_A	Pathogenic	training	known deleterious
rs17560	P03950	Lys84Glu	1a4y_B	5m9j_A	Benign	test	new
rs121913628	P12883	Glu924Lys	2fxm_B	2fxo_B	Pathogenic	training	known deleterious
rs28933990	P12271	Arg234Trp	3hy5_A	3hx3_A	Pathogenic	test	known deleterious
rs28940279	P45381	Glu285Ala	2i3c_A	4nfr_A	Pathogenic	training	known deleterious
rs41293463	P38398	Met1775Arg	1jnx_X	1n5o_X	Pathogenic	training	known deleterious
rs41293463	P38398	Met1775Lys	1jnx_X	2ing_X	Pathogenic	training	known deleterious

Appendix 1. The curated dataset from ClinVar (cont.)

rs41293459	P38398	Arg1699Gln	1jnx_X	3pxc_X	Pathogenic	training	known deleterious
rs1800458	P02766	Gly26Ser	1bz8_A	1bzd_A	Benign	training	new
rs11541795	P02766	Ala39Asp	1bm7_A	5dej_A	Pathogenic	training	new
rs28933979	P02766	Val50Met	1bm7_A	1eta_1	Pathogenic	test	new
rs121918069	P02766	Leu78His	1bm7_A	3djr_A	Pathogenic	training	new
rs121918070	P02766	Thr80Ala	1bm7_A	1tsh_A	Pathogenic	training	new
rs121918071	P02766	Ser97Tyr	1bm7_A	2try_A	Pathogenic	training	new
rs76992529	P02766	Val142Ile	1bm7_A	1ttr_A	Pathogenic	training	new
rs17570	P12955	Leu435Phe	5m4g_A	2iw2_A	Benign	training	new
rs74315403	P04156	Asp178Asn	1hjm_A	2iv6_A	Pathogenic	test	new
rs28933385	P04156	Glu200Lys	1hjm_A	1fkc_A	Pathogenic	test	new
rs1800014	P04156	Glu219Lys	1hjm_A	2lft_A	Benign	training	new
rs74315431	O95292	Pro56Ser	3ikk_A	2mdk_A	Pathogenic	training	new
rs121912443	P00441	His47Arg	1hl4_A	1oez_W	Pathogenic	training	known deleterious
rs121912441	P00441	Ile114Thr	1hl4_A	1uxl_A	Pathogenic	training	known deleterious
rs28934891	P35520	Asp444Asn	4coo_A	4l27_A	Pathogenic	training	known deleterious
rs148865119	P35520	Pro49Leu	1jbq_A	5mms_A	Pathogenic	training	new
rs864309504	Q9Y6X9	Ser87Leu	5of9_A	5ofb_A	Pathogenic	training	known deleterious

Appendix 1. The curated dataset from ClinVar (cont.)

rs11479	P19971	Ser471Leu	2wk5_A	1uou_A	Benign	training	new
rs61752159	O15537	Arg141His	3jd6_O	5n6w_A	Pathogenic	test	known deleterious
rs397515417	Q9BY41	Thr311Met	1t64_A	4qa3_A	Pathogenic	training	known deleterious
rs1603069440	Q9BY41	Ala188Thr	1t64_A	4qa1_A	Pathogenic	training	new

7. REFERENCES

1. Friedenreich CM, Neilson HK, Farris MS, Courneya KS. Physical activity and cancer outcomes: A precision medicine approach. Clinical Cancer Research. 2016.
2. Juengst E, McGowan ML, Fishman JR, Settersten RA. From "Personalized" to "Precision" Medicine: The Ethical and Social Implications of Rhetorical Reform in Genomic Medicine. Hastings Cent Rep. 2016;
3. Sezerman U, Bozkurt T, Sadife Isleyen F. Integrating Evolutionary Genetics to Medical Genomics: Evolutionary Approaches to Investigate Disease-Causing Variants. In: Methods in Molecular Medicine. 2021.
4. Eilbeck K, Quinlan A, Yandell M. Settling the score: Variant prioritization and Mendelian disease. Nature Reviews Genetics. 2017.
5. Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics. 2016.
6. Grantham R. Amino acid difference formula to help explain protein evolution. Science (80-). 1974;185(4154):862–4.
7. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61–80.
8. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;46(15):7793–804.
9. 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(3):310–5.
10. Bhattacharya R, Rose PW, Burley SK, Prlić A. Impact of genetic variation on three dimensional structure and function of proteins. PLoS One. 2017;
11. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun [Internet]. 2021 Dec 1 [cited 2021 Jun 10];12(1):1–9. Available from: https://doi.org/10.1038/s41467-020-20847-0
12. 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(1):1–11.
13. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res [Internet]. 2009 Aug 15 [cited 2018 Dec 10];69(16):6660–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19654296
14. 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(8):1161–70.
15. Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. Rhapsody: Predicting the pathogenicity of human missense variants. Bioinformatics. 2020;36(10):3084–92.

16. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513–23.

17. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. Hum Genomics. 2017;11(1):1–8.

18. 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 [Internet]. 2019;431(11):2197–212. Available from: https://doi.org/10.1016/j.jmb.2019.04.009

19. Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A. 2018;115(16):4164–9.

20. Wyrwoll MJ, Temel ŞG, Nagirnaja L, Oud MS, Lopes AM, van der Heijden GW, et al. Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. Am J Hum Genet. 2020;107(2):342–51.

21. Sessa G, Ehlén Å, Nicolai C von, Carreira A. Missense Variants of Uncertain Significance: A Powerful Genetic Tool for Function Discovery with Clinical Implications. Cancers 2021, Vol 13, Page 3719 [Internet]. 2021 Jul 23 [cited 2021 Aug 6];13(15):3719. Available from: https://www.mdpi.com/2072-6694/13/15/3719/htm

22. Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, et al. MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One [Internet]. 2020 Jul 1 [cited 2021 Aug 6];15(7):e0236962. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236962

23. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.r-project.org

24. Sunyaev SR, Eisenhaber F, Rodchenkov I V., Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 1999;12(5):387–94.

25. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics. 2004;20(7):1006–14.

26. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863–74.

27. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A. 2004;101(43):15398–403.

28. Eck R V., Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science (80-) [Internet]. 1966 [cited 2021 Jun 7];152(3720):363–6. Available from: https://pubmed.ncbi.nlm.nih.gov/17775169/

29. McLachlan AD. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. J Mol Biol. 1971 Oct 28;61(2):409–24.

30. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics [Internet]. 2013 [cited 2018 Dec 10];14 Suppl 3(Suppl 3):S3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23819870

31. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol [Internet]. 1991 Jun 5 [cited 2021 Jun 7];219(3):555–65. Available from: /pmc/articles/PMC7130686/

32. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915–9.

33. 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 [Internet]. 2010 Apr [cited 2018 Dec 10];7(4):248–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20354512

34. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1–9.

35. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Vol. 25, Nucleic Acids Research. 1997. p. 3389–402.

36. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–82.

37. Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 1998 Jan 1;26(1):320–2.

38. Eddy SR. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol [Internet]. 1995 [cited 2021 Jun 7];3:114–20. Available from: www.aaai.org

39. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978–86.

40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. 2001 Feb 15 [cited 2021 Jun 8];409(6822):860–921. Available from: www.nature.com

41. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed [Internet]. 2013 Dec [cited 2021 Jun 8];98(6):236–8. Available from: /pmc/articles/PMC3841808/

42. Hartman P, Beckman K, Silverstein K, Yohe S, Schomaker M, Henzler C, et al. Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. Mol Genet Metab Reports. 2019 Jun 1;19:100464.

43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2019 Jan 21];28(1):235–42. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235

44. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, et al. RCSB Protein

Data Bank: Enabling biomedical research and drug discovery [Internet]. Vol. 29, Protein Science. Blackwell Publishing Ltd; 2020 [cited 2021 Jun 8]. p. 52–65. Available from: http://www.whocc.no/atc_

45. Leiserson CE, Thompson NC, Emer JS, Kuszmaul BC, Lampson BW, Sanchez D, et al. There's plenty of room at the top: What will drive computer performance after Moore's law? Science (80-) [Internet]. 2020 Jun 5 [cited 2021 Jun 8];368(6495). Available from: http://science.sciencemag.org/

46. Schlessinger A, Yachdav G, Rost B. PROFbval: Predict flexible and rigid residues in proteins. Bioinformatics [Internet]. 2006 Apr 1 [cited 2021 Jun 8];22(7):891–3. Available from: http://www.rostlab.org/services/profbval

47. Al-Numair NS, Martin AC. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. BMC Genomics 2013 143 [Internet]. 2013 May 28 [cited 2019 Jan 30];14(3):S4. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-S3-S4#Abs1

48. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol [Internet]. 2014;426(14):2692–701. Available from: http://dx.doi.org/10.1016/j.jmb.2014.04.026

49. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease. BMC Genomics. 2016;16(Suppl 8):1–12.

50. Venselaar H, te Beek TAH, Kuipers RKP, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010;11.

51. Zhou H, Gao M, Skolnick J. Entprise: An algorithm for predicting human disease-associated amino acid substitutions from sequence entropy and predicted protein structures. PLoS One. 2016;11(3):1–24.

52. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. Genome Res. 2014;24(12):2050–8.

53. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One. 2012;7(10).

54. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res [Internet]. 2011 Sep 1 [cited 2018 Dec 10];39(17):e118–e118. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr407

55. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013;

56. Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high

fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12).

57. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61.

58. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. In: Bioinformatics [Internet]. Oxford Academic; 2009 [cited 2021 Jun 8]. p. 54–62. Available from: http://www.broadinstitute.org/science/software/.

59. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128–35.

60. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. Adamovic T, editor. PLoS One [Internet]. 2013 Oct 30 [cited 2018 Dec 10];8(10):e77945. Available from: https://dx.plos.org/10.1371/journal.pone.0077945

61. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet [Internet]. 2011 Apr 8 [cited 2018 Dec 10];88(4):440–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21457909

62. 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 [Internet]. 2016;99(4):877–85. Available from: http://dx.doi.org/10.1016/j.ajhg.2016.08.016

63. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;

64. Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-Barricarte R, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. Nat Methods. 2016;13(2):109–10.

65. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet [Internet]. 2013 Aug [cited 2021 Jun 10];9(8):1003709. Available from: www.plosgenetics.org

66. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods [Internet]. 2010;7(8):575–6. Available from: http://dx.doi.org/10.1038/nmeth0810-575

67. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods [Internet]. 2014 Apr 1 [cited 2018 Dec 10];11(4):361–2. Available from: http://www.nature.com/articles/nmeth.2890

68. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics [Internet]. 2015 Mar 1 [cited 2021 Jun 10];31(5):761–3. Available

from: https://cbcl.ics.uci.edu/

69. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6.

70. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536–43.

71. Li MX, Kwan JSH, Bao SY, Yang W, Ho SL, Song YQ, et al. Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet. 2013;9(1):1–11.

72. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110–21.

73. Huang YF. Unified inference of missense variant effects and gene constraints in the human genome. PLoS Genet [Internet]. 2020 Jul 1 [cited 2021 Jun 10];16(7):e1008922. Available from: https://doi.org/10.1371/journal.pgen.1008922

74. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet [Internet]. 2015 Apr 15 [cited 2021 Jun 10];24(8):2125–37. Available from: http://hgdownload.soe.

75. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res [Internet]. 2005 Jul [cited 2021 Jun 12];33(SUPPL. 2):W306. Available from: /pmc/articles/PMC1160136/

76. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res [Internet]. 2005 Jul 1 [cited 2019 Jan 16];33(Web Server):W382–8. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki387

77. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res [Internet]. 2018 Jul 2 [cited 2021 Jun 12];46(W1):W350–5. Available from: https://academic.oup.com/nar/article/46/W1/W350/4990022

78. Gilis D, Rooman M. PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. Protein Eng [Internet]. 2000 Dec 1 [cited 2021 Jun 12];13(12):849–56. Available from: https://academic.oup.com/peds/article/13/12/849/1565539

79. Casadio R, Vassura M, Tiwari S, Fariselli P, Luigi Martelli P. Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. Hum Mutat [Internet]. 2011 Oct 1 [cited 2021 Jun 10];32(10):1161–70. Available from: www.wiley.com/humanmutation

80. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat [Internet]. 2010 Jun 1 [cited 2021 Jun 10];31(6):675–84. Available from: http://www.enzim.hu/scpred/pred.html

81. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins Struct Funct Bioinforma [Internet]. 2011 Mar [cited 2021 Jun 10];79(3):830–8. Available from: https://pubmed.ncbi.nlm.nih.gov/21287615/

82. Baugh EH, Simmons-Edler R, Müller 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(6):2501–13.

83. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009 Aug;30(8):1237–44.

84. Raimondi D, Tanyalcin I, FertCrossed JSD, 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(W1):W201–6.

85. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics [Internet]. 2006 Nov 15 [cited 2021 Jun 12];22(22):2729–34. Available from: http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi

86. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. Phys Rev Lett. 1997;79(16):3090–3.

87. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular simulations and analytical approaches: Application to α-amylase inhibitor. Proteins Struct Funct Genet. 2000;40(3):512–24.

88. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997 Jun 1;2(3):173–81.

89. Haliloglu T, Bahar I. Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. Proteins Struct Funct Genet. 1999;37(4):654–67.

90. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001 Jan 1;80(1):505–15.

91. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics [Internet]. 2006 Nov 1 [cited 2021 Jun 11];22(21):2695–6. Available from: http://mccammon.ucsd.edu/bgrant/bio3d/

92. Skjaerven L, Yao XQ, Scarabelli G, Grant BJ. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics [Internet]. 2014 Dec 10 [cited 2021 Jun 11];15(1):1–11. Available from: http://thegrantlab.org/bio3d/.

93. van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA. 2009.

94. Bakan A, Meireles LM, Bahar I. ProDy: Protein dynamics inferred from theory and experiments. Bioinformatics. 2011;

95. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun 30;10(6):845–58.

96. Tang N, Sandahl TD, Ott P, Kepp KP. Computing the Pathogenicity of Wilson's Disease ATP7B Mutations: Implications for Disease Prevalence. J Chem Inf Model. 2019;59(12):5230–43.

97. Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. Sci Rep [Internet]. 2020;10(1):1–10. Available from: https://doi.org/10.1038/s41598-020-72404-w

98. Hart SN, Polley EC, Shimelis H, Yadav S, Couch FJ. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. npj Breast Cancer [Internet]. 2020;6(1):1–4. Available from: http://dx.doi.org/10.1038/s41523-020-0159-x

99. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol [Internet]. 2014 Oct 28 [cited 2018 Dec 10];15(10):484. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25348012

100. Ferreira KC do V, Fialho LF, Franco OL, de Alencar SA, Porto WF. Benchmarking analysis of deleterious SNP prediction tools on CYP2D6 enzyme. Chem Biol Drug Des. 2020;96(3):984–94.

101. Seifi M, Footz T, Taylor SAM, Walter MA. Comparison of Bioinformatics Prediction, Molecular Modeling, and Functional Analyses of FOXC1 Mutations in Patients with Axenfeld-Rieger Syndrome. Hum Mutat. 2017;38(2):169–79.

102. Seifi M, Walter MA. Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms. PLoS One. 2018;13(4):1–23.

103. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A. 2015;112(37):E5189–98.

104. Anderson D, Lassmann T. A phenotype centric benchmark of variant prioritisation tools. npj Genomic Med [Internet]. 2018;3(1). Available from: http://dx.doi.org/10.1038/s41525-018-0044-9

105. Singh G, Jayadev Magani SK, Sharma R, Bhat B, Shrivastava A, Chinthakindi M, et al. Structural, functional and molecular dynamics analysis of cathepsin B gene SNPs associated with tropical calcific pancreatitis, a rare disease of tropics. PeerJ [Internet]. 2019 Oct 3 [cited 2021 Jun 10];2019(10):e7425. Available from: http://doi.org/10.7717/peerj.7425

106. Aslan T, Yenenler-Kutlu A, Gerlevik U, Aktuğlu Zeybek AÇ, Kıykım E, Sezerman OU, et al. Identifying and elucidating the roles of Y198N and Y204F mutations in the PAH enzyme through molecular dynamic simulations. J Biomol Struct Dyn [Internet]. 2021;0(0):1–12. Available from: https://doi.org/10.1080/07391102.2021.1921619

107. Saygı C, Alanay Y, Sezerman U, Yenenler A, Özören N. A possible founder mutation in FZD6 gene in a Turkish family with autosomal recessive nail dysplasia. BMC Med Genet [Internet]. 2019 Dec 14 [cited 2019 Jan 29];20(1):15. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/30642273

108. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. Vol. 425, Journal of Molecular Biology. Academic Press; 2013. p. 3919–36.

109. Padhi AK, Gomes J. A molecular dynamics based investigation reveals the role of rare Ribonuclease 4 variants in amyotrophic lateral sclerosis susceptibility. Mutat Res - Fundam Mol Mech Mutagen. 2019 Jan 1;813:1–12.

110. Parveen A, Mirza MU, Vanmeert M, Akhtar J, Bashir H, Khan S, et al. A novel pathogenic missense variant in CNNM4 underlying Jalili syndrome: Insights from molecular dynamics simulations. Mol Genet Genomic Med [Internet]. 2019 Sep 1 [cited 2021 Jun 10];7(9):902. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/mgg3.902

111. Urreizti R, Mayer K, Evrony GD, Said E, Castilla-Vallmanya L, Cody NAL, et al. DPH1 syndrome: two novel variants and structural and functional analyses of seven missense variants identified in syndromic patients. Eur J Hum Genet [Internet]. 2020 Jan 1 [cited 2021 Jun 10];28(1):64–75. Available from: https://doi.org/10.1038/s41431-019-0374-9

112. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: Improvements to accessing data. Nucleic Acids Res [Internet]. 2020 Jan 1 [cited 2021 Jun 16];48(D1):D835–44. Available from: www.ncbi.nlm.nih.gov/clinvar/

113. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. DbSNP: The NCBI database of genetic variation. Nucleic Acids Res. 2001;

114. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res [Internet]. 2021 Jan 8 [cited 2021 Jun 15];49(D1):D480–9. Available from: www.earthbiogenome.org

115. Martin ACR. Mapping PDB chains to UniProtKB entries. Bioinformatics [Internet]. 2005 Dec 1 [cited 2021 Jun 17];21(23):4297–301. Available from: http://www.bioinf.org.uk/pdbsws/

116. Azodi CB, Tang J, Shiu SH. Opening the Black Box: Interpretable Machine Learning for Geneticists. Vol. 36, Trends in Genetics. Elsevier Ltd; 2020. p. 442–55.

117. Baryannis G, Dani S, Antoniou G. Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Futur Gener Comput Syst. 2019 Dec 1;101:993–1004.

118. Park H, Konishi S. Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. J Stat Comput Simul [Internet]. 2016 May 2 [cited 2021 Jun 20];86(7):1450–61. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=gscs20http://dx.doi.org/10.1080/00949655.2015.1073290

119. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B [Internet]. 1996 Jan 1 [cited 2021 Jun 20];58(1):267–88. Available from: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1996.tb02080.x

120. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat

Methodol. 2005;67(2):301–20.

121. Rokach L, Maimon O. Top-down induction of decision trees classifiers - A survey. IEEE Trans Syst Man Cybern Part C Appl Rev. 2005;35(4):476–87.

122. Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. In: Advances in Experimental Medicine and Biology. Springer Science+Business Media; 2011. p. 191–9.

123. RStudio Team. RStudio: Integrated Development for R [Internet]. Boston, MA: RStudio, Inc.; 2015. Available from: http://www.rstudio.com/.

124. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics [Internet]. 2005 Aug 15 [cited 2021 Jun 16];21(16):3439–40. Available from: https://academic.oup.com/bioinformatics/article/21/16/3439/215235

125. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat Protoc [Internet]. 2009 [cited 2021 Jun 16];4(8):1184–91. Available from: https://pubmed.ncbi.nlm.nih.gov/19617889/

126. Sayers E. A General Introduction to the E-utilities. 2010 [cited 2021 Jun 16]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/

127. Sayers E. E-utilities Quick Start. 2018 Oct 24 [cited 2021 Jun 17]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK25500/

128. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol [Internet]. 1993 Dec 5 [cited 2021 Jun 18];234(3):779–815. Available from: https://pubmed.ncbi.nlm.nih.gov/8254673/

129. Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinforma [Internet]. 2016 [cited 2021 Jun 18];2016:5.6.1-5.6.37. Available from: /pmc/articles/PMC5031415/

130. Velankar S, Alhroub Y, Best C, Caboche S, Conroy MJ, Dana JM, et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Res [Internet]. 2012 Jan 1 [cited 2021 Jun 18];40(D1):D445–52. Available from: https://academic.oup.com/nar/article/40/D1/D445/2903750

131. Elodie Foulquier CG. IMGT Aide-memoire: Amino acids [Internet]. 2020 [cited 2021 Jun 18]. Available from: http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html

132. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res [Internet]. 2018 Jul 2 [cited 2021 Jun 18];46(W1):W200–4. Available from: https://lucene.apache.org/core/]

133. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw [Internet]. 2011 Dec 12 [cited 2021 Jun 20];45(3):1–67. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v045i03/v45i03.pdf

134. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent

accessibilites of residues in proteins. PLoS One [Internet]. 2013 Nov 21 [cited 2021 Jun 19];8(11). Available from: /pmc/articles/PMC3836772/

135. Maxwell PI, Popelier PLA. Unfavorable regions in the ramachandran plot: Is it really steric hindrance? The interacting quantum atoms perspective. J Comput Chem [Internet]. 2017 Nov 5 [cited 2021 Jun 19];38(29):2459–74. Available from: /pmc/articles/PMC5659141/

136. Kleywegt GJ, Jones TA. Phi/Psi-chology: Ramachandran revisited. Structure. 1996;4(12):1395–400.

137. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins Struct Funct Bioinforma. 1995;

138. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph [Internet]. 1996 Feb [cited 2019 Jan 30];14(1):33–8, 27–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8744570

139. Delgado J, Radusky LG, Cianferoni D, Serrano L, Valencia A. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;

140. Lobanov MY, Bogatyreva NS, Galzitskaya O V. Radius of gyration as an indicator of protein structure compactness. Mol Biol [Internet]. 2008 Aug 10 [cited 2019 Feb 28];42(4):623–8. Available from: http://link.springer.com/10.1134/S0026893308040195

141. Fuglebakk E, Echave J, Reuter N. Measuring and comparing structural fluctuation patterns in large protein datasets. Bioinformatics [Internet]. 2012 Oct 1 [cited 2021 Jun 19];28(19):2431–40. Available from: https://academic.oup.com/bioinformatics/article/28/19/2431/288157

142. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins Struct Funct Genet. 1998;33(3):417–29.

143. Eyal E, Bahar I. Toward a molecular understanding of the anisotropic response of proteins to external forces: Insights from elastic network models. Biophys J [Internet]. 2008 May 1 [cited 2021 Jun 19];94(9):3424–35. Available from: /pmc/articles/PMC2292382/

144. Mikulska-Ruminska K, Kulik AJ, Benadiba C, Bahar I, Dietler G, Nowak W. Nanomechanics of multidomain neuronal cell adhesion protein contactin revealed by single molecule AFM and SMD. Sci Rep [Internet]. 2017 Dec 1 [cited 2021 Jun 19];7(1):1–11. Available from: www.nature.com/scientificreports

145. General IJ, Liu Y, Blackburn ME, Mao W, Gierasch LM, Bahar I. ATPase Subdomain IA Is a Mediator of Interdomain Allostery in Hsp70 Molecular Chaperones. PLoS Comput Biol [Internet]. 2014 [cited 2021 Jun 19];10(5):1003624. Available from: www.ploscompbiol.org

146. Atilgan C, Atilgan AR. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS Comput Biol [Internet]. 2009 Oct [cited 2021 Jun 19];5(10):1000544. Available from: www.ploscompbiol.org

147. Allaire JJ, Ushey K, Tang Y, Eddelbuettel D. reticulate: R Interface to Python [Internet]. 2017. Available from: https://github.com/rstudio/reticulate

148. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling

technique. J Artif Intell Res [Internet]. 2002 Jun 1 [cited 2021 Jun 20];16:321–57. Available from: https://www.jair.org/index.php/jair/article/view/10302

149. Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Inf Sci (Ny). 2019 Dec 1;505:32–64.

150. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. Med Decis Mak [Internet]. 2016 Jan 1 [cited 2021 Jun 20];36(1):137–44. Available from: https://journals.sagepub.com/doi/abs/10.1177/0272989X14560647

151. Torgo L. Data mining with R: Learning with case studies, second edition [Internet]. Data Mining with R: Learning with Case Studies, Second Edition. 2016 [cited 2021 Jun 21]. 1–405 p. Available from: https://www.routledge.com/Data-Mining-with-R-Learning-with-Case-Studies-Second-Edition/Torgo/p/book/9780367573980

152. Burman P. A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. Biometrika. 1989 Sep;76(3):503.

153. Kuhn M. caret: Classification and Regression Training. 2017.

154. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw [Internet]. 2010 Feb 2 [cited 2021 Jun 21];33(1):1–22. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v033i01/v33i01.pdf

155. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. 2017.

156. Williams G. Data Mining with Rattle and R. Data Mining with Rattle and R. Springer New York; 2011.

157. Liaw A, Wiener M. Classification and Regression by randomForest. R News [Internet]. 2002 [cited 2021 Jun 21];2(3):18–22. Available from: http://www.stat.berkeley.edu/

158. Greenwell B, Boehmke B, Cunningham J, Developers GBM. gbm: Generalized Boosted Regression Models [Internet]. 2020. Available from: https://cran.r-project.org/package=gbm

159. Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emerg (Tehran, Iran) [Internet]. 2015 [cited 2019 Jan 21];3(2):48–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26495380

160. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One [Internet]. 2017 [cited 2019 Jan 22];12(6):e0177678. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28574989

161. Hacisuleyman A, Erman B. ModiBodies: A computational method for modifying nanobodies in nanobody-antigen complexes to improve binding affinity and specificity. J Biol Phys. 2020;46(2):189–208.

8. CURRICULUM VITAE