Mónica Martins Tavares

LARGE-SCALE ANALYSES OF GENOTYPE-PHENOTYPE RELATIONSHIPS OF AMINOMETHYLTRANSFERASE MUTATIONS IN NONKETOTIC HYPERGLYCINEMIA DISEASE

Dissertação no âmbito do Mestrado em Neurociências Molecular e de Translação orientada pela Professora Doutora Irina Sousa Moreira e pela Professora Doutora Ana Cristina Rego apresentada à Faculdade de Medicina da Universidade de Coimbra.

Setembro de 2023

Large-scale analyses of genotype-phenotype relationships of aminomethyltransferase mutations in nonketotic hyperglycinemia disease

Mónica Martins Tavares

Master's Thesis in Molecular and Translational Neuroscience
September2023
Faculty of Medicine
University of Coimbra

ABSTRACT

Martins Tavares, Mónica. 2023. Large-scale analyses of genotype-phenotype relationships of aminomethyltransferase mutations in nonketotic hyperglycinemia disease. Master's Thesis in Molecular and Translational Neuroscience. University of Coimbra. Faculty of Medicine.

Nonketotic hyperglycinemia (NKH) is a rare inborn error of glycine metabolism characterized by the accumulation of glycine in all tissues, especially in the central nervous system (CNS). The disease can occur because of various loss-of-function mutations in the aminomethyltransferase (AMT) gene, which encodes the T-protein. NKH can be divided into two clinical forms: attenuated and severe NKH. Currently, therapy is based on reduced glycine levels (sodium benzoate) and the use of N-methyl-D-aspartate (NMDA) receptor site antagonists (dextromethorphan). Despite these two therapies, the clinical outcomes remain poor. To date, no effective treatment for NKH has been established.

In this study, a data-driven and computation-based approach is proposed for the large-scale analysis of T-protein mutations, including benign and pathologic mutations. Using *in silico* methodologies, including sequence-structure analysis and protein-cofactor interactions, we aimed to characterize the effects of AMT mutations on protein dynamics and function. Statistical analysis of the molecular features of the AMT mutations was performed to study the differences between benign and NKH-causing mutations.

In addition, a machine learning model that identifies key molecular features to predict pathological mutations in AMT has been developed. This could be an important tool for evaluating the efficacy of new NKH treatments.

Keywords: aminomethyltransferase, AMT, NKH, genotype-phenotype, large-scale

AGRADECIMENTOS

Agradeço à Professora Doutora Ana Cristina por todas as oportunidades concedidas a mim, tanto na realização do mestrado em Neurociências Molecular e de Translação, quanto pela experiência na rotação laboratorial no seu grupo *Mitochondria and Neurodegenerative Disorders* no CNC-UC.

Agradeço à minha orientadora, Professora Doutora Irina Sousa Moreira, pela oportunidade de realizar o meu projeto de mestrado e me ter proporcionado as condições necessárias para a realização desse mesmo projeto em Biologia Computacional no seu grupo *Data-Driven Molecular Design* no CNC-UC.

Aos meus colegas do grupo *Data-Driven Molecular Design*, pelo conhecimento partilhado e pelo seu sentido de colaboração. Em especial, agradeço à aluna de doutoramento Raquel Pina Gouveia pela sua paciência ao me ensinar todos os conhecimentos em *Molecular dynamics simulations* necessários para a realização deste projeto, e por ser uma boa companhia de trabalho.

Agradeço à Professora Doutora Kasturi Haldar da Universidade de Notre Dame por me lançar novos desafios na fase final deste projeto. Ao investigador Joseph Farris da Mayo Clinic, por me auxiliar no desenvolvimento do *Clinical Outcome Score survey* apresentado nesta tese.

Ao meu marido Nuno Vidal, agradeço pelo seu apoio e ajuda durante o desenvolvimento da tese, tanto em conhecimento computacional quanto em infraestrutura para executar programas específicos para a realização deste projeto.

Por último, mas não menos importante, agradeço aos meus sogros por estarem sempre disponíveis em auxiliar com o meu filho!

CONTENT

LIST OF ABREVIATIONS

AAMD all-atom or atomistic molecular dynamics
ACP acyl carrier protein
AI artificial intelligence
AMBER assisted model building with energy refinement
AMT aminomethytransferase
AUC-PR area under the precision-recall
AUC-ROC area under the receiver operating characteristic
BINANA BINding ANAlyzer
C1 one-carbon
CCA cross correlations analysis
CGMD coarse-grained molecular dynamics
CHARMM chemistry at harvard molecular mechanics
CNS central nervous system
CO2 carbon dioxide
COS clinical outcome score
DL deep learning
DLD dihydrolipoyl dehydrogenase
DNA deoxyribonucleic acid
DOPE discrete optimized protein energy
EBI european bioinformatics institute
EEG electroencephalogram
ESM-2 evolutionary scale modeling
FAIR facebook AI research
FOCM folate one-carbon metabolism
GB generalized born
GCS glycine cleavage system
GCSH glycine cleavage system H-protein
GLDC glycine decarboxylase
GROMOS GROningen MOlecular Simulation
H4folate tetrahydrofolate
H-bonds hydrogen bonds
indels insertions and deletions
LIAS lipoic acid synthetase
LIPT2 lipoyltransferase 2
MAE mean absolute Error
MD molecular dynamics
ML machine learning
MMPA methylmercaptopropionate
MMPBSA molecular mechanics poisson-bolzmann surface area

MOE molecular operating environment
MolPDF molecular probability density function
NAD+ nicotinamide adenine dinucleotide
NADH nicotinamide adenine dinucleotide reduced
NCBI national center for biotechnology information
NH3 ammonia
NIH national institutes of health
NKH Nonketotic hyperglycinemia
NLM National Library of Medicine
NMDA N-methyl-D-aspartate
NTD neuronal tube defects
OPLS-AA optimized potentials for liquid simulations - all atom
PB Poisson-Boltzmann
PDB protein data bank
PEG percutaneous endoscopic gastrostomy
RMSD root mean square deviation
RMSF root mean square fluctuations
SAV single amino acid variation
SD standard deviation
SHAP SHapley Additive exPlanations
SNP single nucleotide polymorphisms
SVN single nucleotide variation
THF tetrahydrofolate
THH 5-methyltetrahydrofolate
TM text-mining
tRNA transfer ribonucleic acid
UNRES United RESidue
VMD visual molecular dynamics
WT wild-type
ΔG binding free energy

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1 INTRODUCTION

NKH is a rare autosomal recessive metabolic disorder that constitutes a complex multisystemic disease. Although rare, it has an incidence of 1/55.000 (Finland) and 1/63.000 (British Columbia, Canada) at birth and a carrier rate of 1/125[1]. The clinical presentation is divided into three major categories: neonatal, infant, and atypical. Neonatal is characterized by attenuated to severe disease[2,3] manifestations starting within a few days of birth, including lethargy/coma, hypotonia, hiccups, myoclonic jerks, and breathing/swallowing disorders, with subsequent intellectual deficits, spasticity, and intractable seizures[4]. Most patients present with a neonatal clinical presentation. Infantiles are characterized by a smaller proportion of patients showing developmental delay and generally mild seizures in the infantile period, whereas others do not develop symptoms until late infancy or adulthood. Although patients usually have either an attenuated or a severe disease course, there is a continuous clinical spectrum[3]. Atypical glycine encephalopathy indicates hyperglycinemic patients whose clinical presentations are different (transient or late-onset hyperglycinemia and patients with spastic paraparesis).

Several NKH diagnostic methods that measure glycine levels (high plasma and CSF levels and high CSF: plasma ratio), assess the glycine cleavage system (GCS) activity of biopsied liver samples, or request genetic testing. Complementary tests can be requested such as Brain MRI (may reveal hypogenesis of corpus callosum, abnormal gyrus, and hypogenesis of cerebellum in the neonatal form) and EEG (Suppression burst and hypsarrhytmia)[5].

GCS is a complex of T-, L-, P- and H-protein. Protein P has glycine decarboxylase activity; and is encoded by glycine decarboxylase (GLDC) gene[6,7]. T-protein has aminomethyltransferase activity and it is encoded by aminomethyltransferase (AMT) gene. L-Protein has dihydrolipoyl dehydrogenase activity and it is encoded by dihydrolipoyl dehydrogenase (DLD) gene. Finally, H-protein, encoded by glycine cleavage system H-protein (GCSH) shuttles components among L-, P- and T-protein shown in (Figure 1; Han Zhang. *et al.*) [7] .

As its name suggests, GCS breaks down a molecule called glycine by cleaving it into smaller molecules. Glycine is an amino acid that acts as a building block for proteins. This amino acid also acts as a neurotransmitter and chemical messenger that transmits signals to the brain. Glycine is both an excitatory and inhibitory neurotransmitter and has an excitatory effect when bound to the NMDA receptor[8]. Elevated glycine levels in the CNS can cause seizures, hypotonia, and lethargy. The breakdown of excess glycine is necessary for normal development and function of nerve cells in the brain.

Currently, treatment is based on reduced glycine levels (sodium benzoate), the use of NMDA receptor site antagonist (dextromethorphan), and symptomatic treatment of complications (antiepileptic drugs, surgery, physical therapy, PEG).

NKH is a monogenic disorder, and its etiology is characterized by mutations in AMT, GLDC and GCSH of the complex enzymatic GCS respectively.

Mutations in the AMT gene account for approximately 20 percent of all NKH cases. More than 100 mutations have been identified in affected individuals. Most of these genetic changes alter amino acids in aminomethyltransferase (AMT; T-protein). Other mutations delete genetic material from the AMT gene or disrupt how genetic information from the gene is spliced together to create a blueprint for aminomethyltransferase production. AMT mutations alter the structure and function of aminomethyltransferase. For example, some AMT gene mutations reduce the activity of the glycine cleavage system, whereas others eliminate its activity, resulting in an excess of glycine.

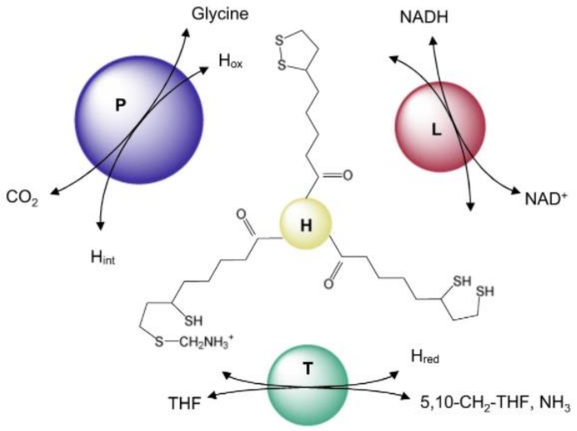


Figure 1. Glycine Cleavage System (GCS) showing the shuttle protein H and the molecules and reactions catalyzed by proteins P, T, and L, respectively.

1.1 Folate and T-protein Interactions

Since 1971, the protein data bank (PDB) archive has served as a single global repository for open access to atomic-level data on biological macromolecules. The archive currently holds more than 200.000 experimentally detected structures (more than 1 billion atoms). These structures are the molecules found in all organisms. Knowledge of the 3D structure of a biological macromolecule is essential for understanding its function, providing insights into health and disease, food and energy production, and other topics of concern for prosperity and sustainability. PDB data are freely publicly available, including T-protein, in (Figure 2; Okamura-Ikeda, K. *et al.*)[9], without restrictions on usage[10].

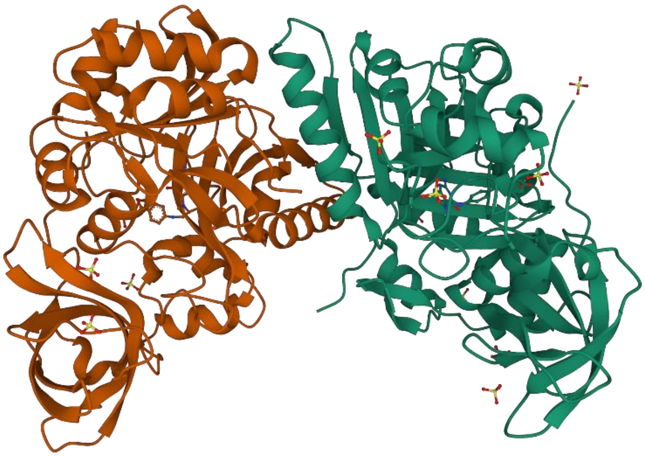


Figure 2. Crystal structure (1wsv) of aminomethyltransferase (T-protein; tetrahydrofolate-dependent) of glycine cleavage system from source: Okamura-Ikeda, K. et al., 2005.

T-protein, a component of the glycine cleavage system, catalyses the formation of ammonia and 5,10-methylenetetrahydrofolate from the aminomethyl moiety of glycine attached to the lipoate cofactor of H-protein. Several mutations in the human T-protein gene cause NKH. To gain insights into the effect of disease-causing mutations and the catalytic mechanism at the molecular level, crystal structures of human T-protein in free form and that bound to 5-methyltetrahydrofolate (5-CH3-H4folate) were determined at 2.0 Å and 2.6 Å resolution, respectively[9]. The overall structure consists of three domains arranged in a cloverleaf-like structure with a central cavity, where 5-CH3-H4folate is bound in a kinked shape with the pteridine group deeply buried in the hydrophobic pocket and the glutamyl group pointed to the C-terminal side surface, shown in (Figure 3; Okamura-Ikeda, K. *et al.*) [9]. Most of the disease-related residues cluster around the cavity, forming extensive hydrogen bonding networks. These hydrogen bonding networks are employed in holding not only the folate-binding space but also the positions and the orientations of α-helix G and the following loop in the middle region, which seems to play a pivotal role in the T-protein catalysis[9].

The human T-protein's crystal structure, as shown in (Figure 2; Okamura-Ikeda, K. *et al.*)[9] contains two monomers. Each monomer comprises three domains 1, 2, and 3, which are colored in blue, green, and red, respectively (Figure 3; Okamura-Ikeda, K. *et al.*) [9]. Domain 1 (residues 32–84 and 175-267) is made up of two and five-stranded antiparallel β-sheets, five α-helices and a β-strand (b13). Domain 2 (residues 85–174) consists of a five-stranded antiparallel β-sheet and two α-helices. Domain 3 (residues 268–402) consists of a loop-rich region and a distorted six-stranded jelly-roll that is oriented perpendicular to the β-sheets of domains 1 and 2, closing the ring-like structure of the protein[9].

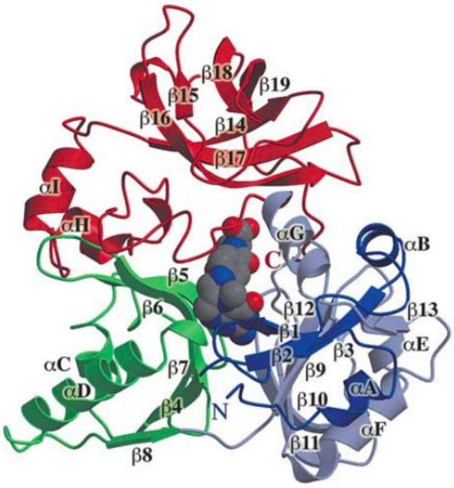


Figure 3. T-protein cloverleaf-like structure with the central cavity, where folates bind. Domains 1, 2 and 3 are colored in blue, green and red.

1.2 T- and H-protein Interactions

This highly conserved protein complex is in mitochondrial membrane in eukaryotes and is the major route of glycine catabolism[11]. GCS action involves oxidative cleavage of glycine with release of carbon dioxide (CO2) and ammonia (NH3) and transfer of a methylene group (–CH2–) to tetrahydrofolate, with concomitant reduction of NAD+ to NADH (Figure 3; Leung et al., 2021)[11].

The Figure 4 A illustrates the process of glycine breakdown. Glycine undergoes decarboxylation by GLDC, which then transfers the methylamine component to GCSH. Afterward, with the aid of AMT, ammonia is released, and the methylene part is passed on to tetrahydrofolate (THF). This THF enters the mitochondrial folate cycle, eventually leading to the formation of formate that moves to the cytoplasm.

The Figure 4 B shows that Lipoyl-GCSH is produced through an intermediate stage of lipoyltransferase 2 (LIPT2), which originates from octanoyl-ACP (acyl carrier protein). This is then transferred to GCSH. The sulfur incorporation process into this structure is mediated by the enzyme lipoic acid synthase (LIAS).

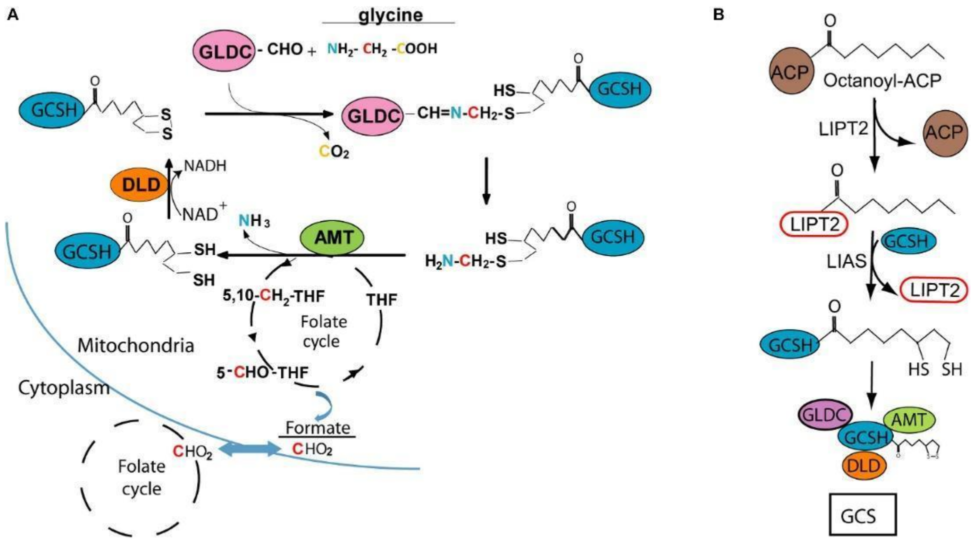


Figure 4. GCSH function and lipoylation (A) and Lipoyl-GCSH is generated via LIPT2 (lipoyltransferase 2) (B)

The GCSH plays a central role in the catalytic process, forming an amino-methyl intermediate with GLDC and acting as the acceptor of the methylamine group which is then transferred as the substrate for AMT[12].

Loss of function of the GCS is predicted to cause accumulation of glycine and suppression of folate one-carbon metabolism (FOCM) and this is associated with post-natal neurometabolic disease and structural malformations of the developing brain including neural tube defects (NTDs) and ventriculomegaly. Accumulation of glycine in tissue and body fluids is a hallmark of NKH, a life-limiting inborn metabolic error that most of the cases occurs in the neonatal period with hypotonia, apnea, and seizures[9,10]. Subsequently, affected babies and children suffer complex epilepsy and profound developmental delays.

In GCS glycine is enzymatically cleaved into CO_2, NH_4+, and a methylene group (Figure 5; Zhang et al., 2019)[12]. The methylene group is accepted by tetrahydrofolate (THF), forming 5,10-methylene-THF as the one-carbon (C1)[11,13] source for purine synthesis and cell growth, and yielding one molecule of NADH as reducing power[14]. GCS also catalyzes the reversible reaction of glycine synthesis from CO_2, ammonium, 5,10-methylene-THF and NADH, especially in anaerobic bacteria such as Clostridium acidiurici[13,14].

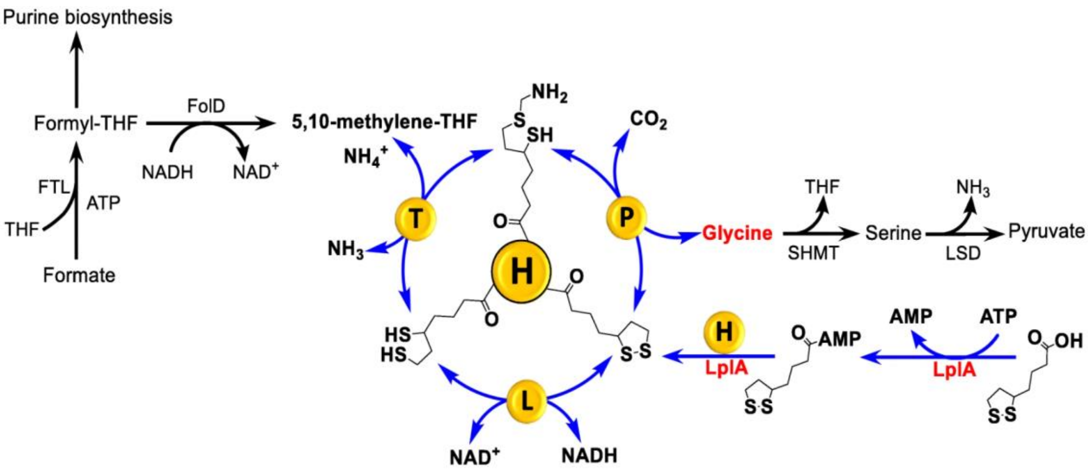


Figure 5. Glycine cleavage system (GCS) with H protein as a shuttle among its components, also shown are the lipoylation of H protein and the roles of GCS in formate and purine biosynthesis.

.

Recently, the reversed GCS reactions have been successfully employed to develop novel C1 assimilation pathways in Escherichia coli, facilitating the use of formate and CO2[15–20]. To achieve this, both endogenous GCS and exogenous formyl-methenyl-methylenetetrahydrofolate synthetase were overexpressed in engineered E. coli strains. This allowed the conversion of formate into amino acids like glycine and serine, which were subsequently integrated into the central metabolic pathway [12]. However, the rate or flux of glycine synthesis remains relatively low. Currently, only about 10% of the carbon necessary for cell growth can be sourced from this synthetic pathway. For a truly effective growth utilizing formate and efficient CO2 fixation, it's crucial to further understand and optimize the GCS mechanism.

1.3 Objective

In this thesis, we propose an integrated computational approach for large-scale analysis of genotype-phenotype relationships of AMT (T-protein) mutations in NKH disease. We proposed a collection of AMT mutations associated with benign, attenuated and severe phenotypes from various scientific gene mutation databases. The computational work consisted of three steps. First, I performed a survey of NKH-causing missense mutations on the T-protein, presenting the most frequently observed mutated positions in AMT, and distributing NKH-causing missense mutations across the length of the T-protein and the positions of the active site, active site tunnel, and dimerization domains. Second, I performed a careful analysis of binding (interactions) with the cofactor 5-methyltetrahydrofolate (5-CH3-H4folate) also known as THH with the T-protein, which can be used as a key molecular feature of the machine learning model to predict how mutations decrease T-protein activity. Finally, built a machine learning (ML) model that identifies key molecular features to predict whether mutations decrease T-protein function and classifies it as NKH-causing or benign mutations.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Data Sources

In this study, data were obtained from four sources: ClinVar, SimpleClinVar, PubTator, and a key journal article published in Genetics in Medicine[6], showed in Figure 6. ClinVar is a public archive of interpretations of the clinical significance of genetic variations. SimpleClinVar is a subset of ClinVar that focuses on providing simplified interpretations of the common genetic variants. PubTator is a text-mining (TM) platform that uses natural language processing to annotate genes, diseases, chemicals, and mutations in the scientific literature. The article (*"The genetic basis of classic nonketotic hyperglycinemia due to mutations in GLDC and AMT"*; Coughlin, C. R. et al., 2016)[6] published in Genetics in Medicine contains a comprehensive list of mutations in GLDC and AMT that are clinical tested.

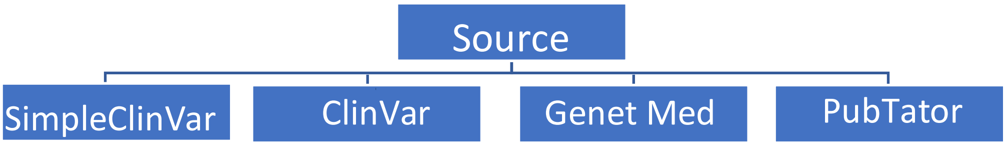


Figure 6. Data sources with AMT mutations used: SimpleClinVar, Clinvar, GenetMed and PubTator

Information gathered from these sources is crucial to understand the impact of these mutations on NKH. Combining the data from these four sources provides a comprehensive overview of the mutations and their impact on NKH.

3 Methods

3.1 Homology Modeling

Crystallography is the traditional method for determining protein structures; however, it is time consuming and challenging. In recent years, computational methods have been developed to predict protein structure. In this study, different computational methods such as homology modeling, AlphaFold, and Meta-IA were used to predict the 3D structures of T-proteins.

Homology modeling, also known as comparative modeling[21], is a computational method used to predict the three-dimensional structure of a protein based on its evolutionary relationship with proteins of known structures. This technique uses a known protein structure as a template to predict the structure of a target protein that shares a significant sequence similarity with the template protein.

The homology modeling process typically involves the following steps:

1. Template identification: Homologous protein structures are searched within protein structure databases using sequence alignment methods such as BLAST or PSI-BLAST.

2. Alignment: The target protein sequence should be aligned to the template protein sequence. Computer algorithms are utilized to identify regions of similarity between two proteins, which can be accomplished using various sequence alignment algorithms such as ClustalW, MUSCLE, and T-Coffee.

3. Model building: A three-dimensional model of the target protein should be built using the template structure as a guide. Several software tools including MODELLER, SWISS-MODEL, and I-TASSER are available for this purpose.

4. Model evaluation: The quality and reliability of the final model should be assessed using various validation methods, such as Ramachandran plot analysis, Procheck, MolProbity, Verify3D, MolPDF, DOPE score, and RMSD.

In this study, we applied a homology modeling protocol based on the MODELLER software[22] package to predict the missing structure corresponding to the 28 absent amino acids at the N-terminus of the protein. The protocol utilized homology modeling, as described in the tutorial from the paper Rosario-Ferreira et al.[23].

The crystal X-ray structures of, 1wsv[24], 1wsr[25], 1woo[26], 1wop [27], and 1wor [28] were downloaded from the Protein Data Bank. Details of the crystal X-ray structures are provided in Appendix 1. The properties of these structures, such as the origin organism, sequence length, number of chains, number of ligands, and resolution of the X-ray method, were compared. The 1wsv [24] structure was selected as a template[29] for homology modeling using the MODELLER[21] software (Version 10.1). 100 models for the T-Protein structure were created, and the top 5 models were selected using molecular probability density function (MolPDF)[30], discrete optimized protein energy (DOPE)[31] score and score the root mean square deviation (RMSD)[32,33]. These models were visually inspected using PYMOL (Version

2.5.2) software. Additionally, AlphaFold and Meta-IA[34] models for T-proteins were downloaded from their respective databases. AlphaFold, developed by DeepMind, can accurately predict 3D models of protein structures based on their amino acid sequences. Meta-IA, developed by the META team, can accurately predict 3D models from protein sequences[35]. The AlphaFold and Meta-IA models were added to the PYMOL[36] session along with the top five models generated by the MODELLER software.

3.1.1 MolPDF, DOPE and RMSD Scores

The MolPDF[30] is a scoring function used to evaluate the quality of protein homology models as it measures the agreement between the predicted protein structure and experimental data such as X-ray diffraction or NMR spectroscopy data. The MolPDF score is based on the calculation of a probability density function that describes the spatial distribution of the atoms in a protein structure. The probability density function was calculated from a set of experimental data and was used to compare the predicted protein structure with the experimental data. In homology modeling, the MolPDF score is used as a criterion to evaluate the accuracy of the predicted protein structure, rank, and select the best model. The higher the MolPDF score, the better the agreement between the predicted protein structure and experimental data, and the higher the quality of the prediction.

The DOPE[31] score is a computational method used in protein homology modeling to evaluate the quality of predicted 3D protein structures. The DOPE score is based on a statistical potential that incorporates information regarding the geometry and interaction patterns of amino acids in known protein structures. The DOPE score measures the energy of a predicted protein structure and compares it to a reference state, which is the average energy of amino acid interactions in known protein structures. The DOPE score provides a quantitative measure of the similarity of a predicted protein structure to known protein structures, and can be used to rank and select the best models from a set of predictions. The lower the DOPE score, the higher the quality of the predicted protein structure and its similarity to the known protein structures. In protein homology modeling, the DOPE score is used as a scoring function to evaluate the accuracy of the predicted protein structure. The DOPE score provides a measure of the energy of the predicted protein structure and its deviation from known protein structures, helping identify the most accurate and reliable models.

RMSD[32,33] is a measure of the difference between the two structures in protein homology modeling. RMSD was used to compare the differences in the three-dimensional structure between a predicted protein model and its corresponding experimental structure. RMSD was calculated by taking the root mean square of the deviations between the atomic positions of the equivalent residues in the two structures. It provides a measure of the average distance between the corresponding atoms in the two structures, expressed in Angstroms. The smaller the RMSD value, the more similar the two structures. RMSD is often used as a measure of the accuracy of the predicted protein structure, with lower RMSD values indicating better agreement between the predicted and experimental structures. The RMSD scores were used in combination with the DOPE score to assess the overall quality of the prediction and select the best models.

3.2 5-methyltetrahydrofolate and T-protein Interactions

BINding ANAlyzer (BINANA)[37] is a web server that analyzes ligand-binding. The program identifies key binding characteristics such as hydrogen bonds, salt bridges, and pi interactions.

Hydrogen bonds are formed between a hydrogen atom bonded to a highly electronegative atom (e.g., nitrogen, oxygen, or fluorine) and another electronegative atom. These bonds are weaker than covalent bonds but still play an important role in stabilizing protein structures such as α-helices and β-sheets.

Salt bridges, also known as ion pairs, are formed between oppositely charged amino acid side chains (such as a negatively charged aspartic acid side chain and a positively charged lysine side chain). These interactions are important for stabilizing the overall protein structure.

Pi interactions occur between two aromatic rings (such as those found in phenylalanine, tryptophan, and tyrosine). These interactions are important for stabilizing protein structures and are also involved in many other biological processes, such as DNA base-stacking.

As input, BINANA accepts receptor and ligand files in PDBQT (preferred) or PDB formats[38].

3.3 Domains, and functional sites of T-protein

One approach is to use a database of protein families to determine biological functional sites. In this study, we used InterPro, which is a database of protein families, domains, and functional sites. It provides a central resource for the analysis of protein sequences and is a key component of the biological research infrastructure[39]. InterPro was maintained at the European Bioinformatics Institute (EBI).

3.4 Pathogenicity Prediction

Several software tools can be used to predict the pathogenicity of protein mutations. This study included the following:

LYRUS: A machine learning model for predicting the pathogenicity of missense variants. LYRUS is trained using a large dataset of annotated protein mutations and uses machine learning algorithms to learn the characteristics of damaging and neutral mutations. It can then be used to predict the pathogenicity of new mutations, based on their impact on protein stability. LYRUS selected 15 features from these three categories for the prediction pipeline. Fifteen features belonging to the three categories were used. Each feature calculation requires either an amino acid sequence, PDB file, or both. SEQ, sequence-based feature; STR, structure-based feature; DYN, dynamics-based feature[40].

PROVEAN: This tool uses sequence-based information and evolutionary conservation to predict the impact of mutations on protein function. It can predict whether a mutation is likely to be tolerated or deleterious, based on its effects on protein stability and structure. PROVEAN scores suggested using a cut-off of -2.5 for the PROVEAN score when using the NCBI nr protein database released in August 2011. That is, consider a score higher than -2.5 to be neutral (tolerated) and that lower than or equal to -2.5 to be deleterious (damaging). The PROVEAN scores and optimal cutoff may vary slightly with different versions of the nr database because the scores are computed based on the homologs in the DB[41]. In this study, we used the NR protein database released in August of 2011.

SNAP2 (SNP Annotation and Proxy Search): is a software tool developed by the National Center for Biotechnology Information (NCBI) that can be used to predict the potential impact of single nucleotide polymorphisms (SNPs) on gene expression and function. SNAP2[41,42]uses a combination of sequence-based and functional annotation data to predict the effect of SNP on gene expression and protein function. It can be used to identify SNPs that are likely to have a significant impact on gene expression and protein function and to identify other SNPs that may be in linkage disequilibrium with these functional SNPs.

MUTAFRAME: This tool uses a combination of physical and chemical parameters to score the stability of a protein and can predict whether a mutation is likely to be damaging or neutral based on its effect on the stability score[43].

SIFT: This tool uses evolutionary conservation and structural information to predict the effect of a mutation on protein function. It can predict whether a mutation is likely to be tolerated or deleterious based on its effects on protein stability and structure[44,45]. It assigns a score to each mutation, where scores ≤0.05 are predicted to be deleterious. However, its accuracy in predicting the clinical significance of frameshift and stop-gain mutations is limited.

SuSPect: This tool is based on a statistical model that uses a set of training data to determine the characteristics of the damaging and neutral mutations. It then uses this knowledge to predict the pathogenicity of new mutations, based on their impact on protein stability. SuSPect can be used to predict the pathogenicity of both single-point mutations and large structural variations such as insertions, deletions, and inversions. The SuSPect scores ranged from 0 to 100, with a recommended cut-off of 50 for discriminating between neutral and disease-associated SAVs[46].

3.5 Evolutionary conservation profile of T-protein

Consurf: This tool predicts the functional and evolutionary importance of amino acid residues in proteins based on the analysis of their evolutionary conservation across multiple species[47]. It uses multiple sequence alignments to identify conserved residues, and then assigns a conservation score to each residue based on its degree of conservation across multiple alignments[48].

3.6 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations[49] are computational techniques used to investigate the behavior of molecular systems over time. These simulations are commonly classified into two primary types: explicit and implicit [50,51].

Explicit MD simulations facilitate the study of protein-solvent interactions, including those involving water or ions. In explicit MD simulations, the solvent is represented by a specific model that accounts for its atomic-level structure and properties. For example, water is often modeled using the TIP3P, TIP4P[52], TIP5P[53] or SPC, and SPC/E[54] water models, which represent water as a rigid molecule with fixed geometries and partial charges on oxygen and hydrogen atoms. The detailed information provided by explicit simulations allows for the study of solvation effects by including solvent molecules directly in the simulation. This approach offers a more comprehensive understanding of the molecular interactions, solvation dynamics, and role of the solvent in determining the properties and behavior of the solute. However, owing to their high computational expense, explicit simulations of large systems require significant computational resources. In addition, explicit representations of the system can be split into an all-atom or atomistic molecular dynamics approach (AAMD)[55], and coarse-grained molecular dynamics approach (CGMD)[56,57]. In AAMD simulations every atom within the molecular system is modeled, providing a detailed and accurate representation of the system's interactions and behavior. This method models all atoms in a system, including solvent molecules, protein residues, and other relevant molecules. Interactions among atoms are determined using physical laws such as Coulomb's law[58] and van der Waals interactions [59], determining the forces acting on each atom. In CGMD simulations simplify the system by grouping several atoms into a single particle or "bead." This reduction in the number of degrees of freedom leads to a less detailed representation of the system, but significantly reduces the computational cost of the simulations. CGMD is particularly useful for studying larger systems or longer timescales that may be computationally prohibitive for AAMD simulations.

On the other hand, implicit MD simulations simplify the system by not explicitly modeling all atoms. Instead, a continuum solvent model represents the solvent, assuming a uniform interaction with the protein. For example, the solvent can be simplified using the generalized Born (GB)[60–64] model. This approach is computationally less expensive and can be applied to larger systems or longer timescales. However, implicit simulations may not provide detailed information about protein-environment interactions compared to explicit simulations.

In summary, implicit, and explicit simulations differ in their representation of the solvent, with implicit simulations using a continuum model and explicit simulations modeling solvent molecules. AAMD and CGMD simulations differ in their representation of the molecular system, with AAMD modeling every atom, and CGMD simplifying the system by grouping atoms into single particles or beads.

MD simulations of proteins[65,66] typically involve three main phases: minimization, equilibration, and production[67] .Minimization: In the minimization phase, the geometry of the initial protein structure is optimized, and any steric clashes61 or bad contacts are removed[68] through a series of energy minimization calculations. This phase is essential to ensure that the protein structure is in a stable conformation before proceeding to equilibration and production phases. Energy minimization is typically achieved by applying an optimization algorithm, such as steepest descent or conjugate gradient[69], which iteratively adjusts the coordinates of the atoms to minimize the total energy of the system.

Equilibration: During the equilibration phase, the system was gradually brought to the desired temperature and pressure conditions, and the velocities of the atoms were randomized to a thermal distribution. This stage allows the system to relax and reach a state of thermal equilibrium where the temperature, pressure, and other thermodynamic properties are stable and consistent with the desired conditions. Equilibration typically involves a series of short simulations with gradually increasing time steps, during which temperature and pressure are controlled by thermostats and barostats, respectively. Equilibration simulations may also involve position restraints on certain parts of the protein, such as the backbone or the ligand, to prevent large conformational changes.

The Maxwell-Boltzmann[70] distribution equation describes the distribution of particle speeds in a system at a given temperature, is essential for determining the initial velocities of atoms in MD simulations and is applied during the equilibration phase. The equation is given by

$$f(v) = 4\pi\left(\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2 e^{\frac{-mv^2}{2kT}}\right)$$

Equation 1. Maxwell-Bolzmann

where m is the particle mass, v is the particle speed, k is Boltzmann's constant, and T is the temperature.

By initializing the velocities according to this distribution, the system reaches a state in which the temperature, pressure, and other thermodynamic properties are stable and consistent with the desired conditions. This ensures that the system is in a proper state for the subsequent production phase, during which the properties and behavior of the system are analyzed and studied over longer timescales.

Production: In the production phase, the equilibrated system undergoes simulation for an extended duration, recording and analyzing the trajectories of the protein atoms. Production simulations aim to study protein behavior and properties over longer timescales and generate statistically significant data for analysis. Production simulations can range from nanoseconds[71] to microseconds or longer, depending on the complexity of the protein and properties of interest. Trajectory data analysis can provide insights into protein structure, dynamics, and interactions with other molecules or environments.

Overall, the minimization, equilibration, and production stages of MD simulations are essential for ensuring the stability, accuracy, and reliability of simulation results, and they require careful consideration and optimization to achieve meaningful and informative results. Newton's second law[71,72], which forms the basis for the calculation of forces, accelerations, and subsequent updating of atom positions and velocities, is applied consistently across all these phases, driving the dynamics of molecular systems throughout the entire simulation process.

Newton's second law[73] states that the acceleration of an object is directly proportional to the net force acting on it, and inversely proportional to its mass. In MD simulations, this law is used to compute the trajectories of atoms over time by integrating the forces that act on them. Newton's second law is given by

$$F = ma$$

Equation 2. Newton's second law of motion

where F is the net force, m is the mass, and a is acceleration.

Force fields were employed at every stage of the MD simulation. They represent the potential energy function that describes the interactions between atoms in a system. This information is crucial for calculating the forces acting on each atom, which subsequently determine the atom positions and velocities throughout the simulation. These calculations are essential at every stage, from initializing and optimizing the geometry of the system to equilibrate and simulate its behavior over time during the production phase.

Different types of force fields were used in molecular dynamics simulations, with the choice depending on the level of detail required for the system under study. Examples of force-field approaches include AAMD, CGMD, and an implicit solvent.

In the AAMD simulations, every atom in the system is explicitly represented, providing high-resolution details suitable for studying atomic-level interactions and behavior. Accurate force fields based on experimental and/or quantum mechanical data are required for AAMD simulations. Some popular AAMD force fields include Chemistry at HARvard Molecular Mechanics (CHARMM)[74,75], Assisted Model Building with Energy Refinement (AMBER)[76–78], GROningen MOlecular Simulation (GROMOS)[79], and optimized potentials for liquid simulations - all atom (OPLS-AA)[80].

In CGMD simulations, groups of atoms are represented as single interaction sites or "beads," reducing the degree of freedom of the system. This enabled the study of larger time and length scales. CGMD simulations use specific force fields parameterized for the reduced representations. Some popular CGMD force fields are MARTINI[81] , which is a versatile CGMD force field for various biomolecular systems, GROMOS, which is applicable to both AAMD and CGMD approaches, and the United RESidue force field for proteins (UNRES)[82].

Implicit solvent models, on the other hand, simplify the representation of solvents by incorporating their effects as an averaged, continuous medium, instead of explicitly representing each solvent molecule. These models significantly reduce the computational costs, allowing the study of larger systems or longer timescales. Popular implicit solvent models include Generalized Born (GB)[83] models and Poisson-Boltzmann (PB)[84] models.

When choosing a force field for MD simulations, it is crucial to consider the required level of detail for the system and questions to be addressed. Atomistic simulations offer more detailed information but require more computational power, whereas coarse-grained and implicit solvent models provide more efficient exploration of larger systems and longer timescales, albeit with reduced detail. The choice of the force field also depends on the specific system being studied, as some force fields are optimized for types of molecules or interactions.

In summary, many force fields are available in the literature, each with varying degrees of complexity and tailored for different types of systems[85]. The following equation is one of the most widely used energy functions:

$$V = \sum_{bonds} \frac{1}{2} K_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{improper\ \ dihedrals} \frac{1}{2} k_\varepsilon (\varepsilon - \varepsilon_0)^2 + \\ \sum_{diherals} K_\theta [1 + cos\ cos\ (n\emptyset - \delta)\ ] + \sum_{atom\ pairs\ i.j.} (\frac{1}{4\pi\varepsilon_0} q_i q_j r_{ij}^{-1} + A_{ij} r_{ij}^{-12} - B_{ij} r_{ij}^{-6})$$

Equation 3. Force fields energy function

which typically consists of the first four terms relating to intramolecular or local contributions to the total energy (bond stretching, angle bending, dihedral, and improper dihedral), and the last terms describing the van der Waals interactions and electrostatic interactions contributing to non-bonded interactions. It is worth mentioning that van der Waals interaction terms, described by a Lennard-Jones potential function, include only dispersion or London interactions between transient dipoles, whereas Keesom interactions (between permanent dipoles) and Debye (induced dipole) interactions are included in the electrostatic (Coulomb) term, without explicit development into the charge distribution momenta[67].

MD simulations evaluate millions of interactions of particles for billions of time steps, which can require extraordinary amounts of computational hardware and time[86,87]. This huge application potential has led to the implementation of MD in many software packages, each with its own strengths and limitations, such as GROMACS[87], NAMD[88], AMBER[89], CHARMM[90], OpenMM[91] and LAMMPS[92].

In addition, several other tools are available that facilitate the generation of input files for MD simulations: visual molecular dynamics (VMD)[93], Chimera[94], molecular operating environment (MOE)[95], Maestro (Schrödinger Suite)[96] and CHARMM-GUI[97].

In this study, we used a powerful combination of CHARMM-GUI [98,99]and GROMACS[87]. CHARMM-GUI is a user-friendly web interface that streamlines the generation of input files for various simulations including MD simulations of proteins. Compatibility with several simulation packages, including GROMACS, CHARMM-GUI[99] simplifies the simulation setup process. GROMACS, on the other hand, is a powerful and versatile MD simulation package that is widely used to simulate proteins and other biomolecules.

To conduct an MD study of a protein using CHARMM-GUI, the following steps are typically performed.

1. Submit the protein structure to CHARMM-GUI and select the desired simulation setup options (such as force field [94], solvent type, and box size).

2. CHARMM-GUI generates input files for multiple simulation packages, including GROMACS, which can be downloaded and utilized to run the simulations.

3. CHARMM-GUI can offer guidance on specific aspects of simulation, such as setting up a solvated system or adding ions to the simulation.

To use GROMACS in an MD study of a protein, the following steps are typically performed.

1. Run energy minimization simulations to relax the system and remove any steric clashes[68]or bad contacts.

2. Run equilibration simulations were performed to bring the system to the desired temperature and pressure conditions, allowing it to reach a thermal equilibrium.

3. Run production simulations to generate trajectories of the protein over a longer time scale can be used to analyze its behavior and properties.

GROMACS offers advanced features and options for MD simulations, such as free-energy calculations, replica exchange simulations, and enhanced sampling techniques, which can be employed to study complex phenomena in proteins and other biomolecules.

To perform MD simulations of the T-Protein, we used CHARMM-GUI web site[95] and GROMACS software (Version 2019.4).

First, we submitted the protein model generated by homology modeling to the CHARMM-GUI website, as detailed in section 3.2. We included only residues 33-393 of the protein model, as the remaining residues belonged to the peptide transit on the N-terminal and/or did not have an associated secondary structure. Both chains A and B of the proteins and the THH ligand in both chains were also included.

In the second step of CHARMM-GUI, we employed CHARMM36 Force Field[69] to generate CHARMM top and par files and incorporate phosphorylation residue sites of the protein from literature sources such as phosphosite.org.

Next, we set a water box size of 20.0 and added NaCl ions at a concentration of 0.15M to calculate the solvent composition. We then built the system and generated input files for use in GROMACS MD simulations.

In GROMACS, we first ran a minimization stage to reduce the total energy of the system. Subsequently, equilibration of the system was performed, with a minimum distance of the periodic image of 1.2 nm. Finally, we ran three replicas of the MD simulations with a protein-run distance of 500 ns. Utilizing the strengths of both CHARMM-GUI and GROMACS, we effectively simulated and studied the behavior and properties of the T-protein.

3.6.1 Analysis of Molecular Dynamics Simulations

Molecular dynamics simulations are powerful tools for studying the behavior of proteins and their interactions with ligands. Several methods were used to analyze the molecular dynamics simulation data in this study, such as the density distribution of RMSD, MMPBSA decomposition method, hydrogen bond analysis, and root mean square fluctuation (RMSF).

Density Distribution of RMSD Analysis

The RMSD is a measure of the difference between the atomic positions of two structures. The density distribution of RMSD can be used to compare the structural similarity between the wild-type (WT) and mutant proteins[100]. The Gromacs tool *gmx rms* used to generate the density distribution. This tool calculates the RMSD between each frame of the simulation and a reference structure, and we then use the R-language to build RMSD density graphs. This tool can be used to calculate the density distribution of RMSD for both the WT and mutant T-protein systems.

MMPBSA Decomposition Analysis

The molecular mechanics poisson-boltzmann surface area (MMPBSA) decomposition method can be used to calculate the binding free energy (ΔG) of a protein-ligand complex [101]. This method decomposes the ΔG into individual contributions from different components, such as van der Waals interactions, electrostatic interactions, and solvation effects. A study of botulinum neurotoxin A complexed with synaptic vesicle protein 2C used the MMPBSA method to calculate the interactions at the binding interface and propose a relationship between conformational changes and interfacial interactions[102]. To use MMPBSA, we first extracted the protein-ligand complex from the simulation trajectory using the Gromacs tool *gmx trjconv*. We then need to prepare the input files for the MMPBSA, which involves running a series of scripts to set up the calculations. Once the calculation is complete, the results can be analyzed to determine the contribution of each energy component to the ΔG. We used the decomposed MMPBSA results to compare the ΔΔG of the ΔG WT - ΔG mutant T-protein systems.

H-bonds Analysis

Hydrogen bonds (H-bonds) are important for protein-ligand interactions [103]. A group boxplot of the H-bonds between chains A and B interactions with the ligand was used to compare the WT and mutant proteins. A study of SAPAP in complex with Shank3 used energy decomposition and hydrogen bond analysis to show that mutations disrupt interactions with the canonical PDZ domain[104]. To generate the group boxplot, we first calculated the number of H-bond interactions between chains A and B and the ligand for each system using the Gromacs tool *gmx hbond*. We then used the R-language to generate a group boxplot.

Root Mean Square Fluctuations Analysis

RMSF analysis serves as a quantitative tool, evaluating the mobility of individual atoms across a MD trajectory. This technique elucidates the dynamic behavior of proteins, highlighting regions that are either notably flexible or remarkably stable. Specifically, the RMSF values for the atoms were determined by the standard deviation of their deviations from a selected reference position over the course of the simulation[105]. One study used RMSF to investigate the dynamics of zfP2X4 receptors[106]. This research used RMSF analysis to find the flexibility across protein residues and to identify areas of the protein experiencing considerable conformational alterations.

RMSF plots drawn for WT and mutant proteins can shed light on how mutations can alter flexibility across distinct protein regions. Such shifts in flexibility can alter the function and stability of proteins.

Before starting the RMSF analysis, it is essential to combine individual trajectory files to form a continuous, unified trajectory, especially when simulations involve multiple replicas, as seen in this study. To facilitate this, GROMACS offers the *gmx trjcat* command[107].

Post-concatenation, the unified trajectory file in xtc format must be converted into the. dcd format to ensure compatibility with the bio3d package in R, a package renowned for its comprehensive tools for analyzing biological structures and MD simulation data.

For the RMSF analysis, dcd trajectory file and its corresponding chain structures pdb format was loaded into R. Using the rmsf function of bio3d, RMSF values were calculated for each residue throughout the trajectory. Subsequently, the RMSF values for the WT and mutants were plotted against the residue numbers. Such a visual representation streamlines the comparison of mobility across different protein regions in both the mutant and wild type. The peaks in the RMSF plot often indicate regions exhibiting increased flexibility.

Dynamic Contact Extraction from MD Simulations

Understanding the frequency of H-bond interactions in molecular dynamics simulations provides valuable insights into the nuanced behavior of proteins and their interactions with ligands. The get_dynamic_contacts.py script was used to obtain the H-bond residue contact frequency[108]. Using this tool, we used the trajectory of our protein systems, uncovering H-bonds that were pivotal to understanding protein dynamics between chains and between chain and ligand binding affinities.

Using the system topology and MD trajectory, we focused on interactions that exhibited hydrogen bonds greater than zero. The ligand THH, identified by its unique residue name, and the protein discerned by its molecular signature, became the subject of our analysis. This methodology allowed the H-bond frequency interactions to be logged throughout the simulation period.

To discern the differences in hydrogen bond patterns between the WT and mutant proteins (especially Chains A and B) and (between Chain A and ligand THH A and Chain B and ligand THH B), the get_contact_frequencies.py script was used for each protein system[109]. This tool provides the frequency of interactions, revealing the distinct dynamics of different WT and mutant proteins.

Finally, using the R-language, we visualized these intricate gradient patterns using heatmaps. These heatmap plots offered a comprehensive view of the H-bond interactions across both the WT and mutant protein configurations.

3.7 Machine learning for T-protein pathogenicity prediction

3.7.1 Software and Tools

The development and execution of the ML model in this project were made possible by a combination of various software tools and libraries. The specific versions used for each library are as follows:

Python (version 3.10.12): the programming language used in the project. Python is popular in the data science community owing to its simplicity and scientific computing libraries.

Pandas (version 1.5.3): A powerful open-source data manipulation and analysis library.

NumPy (version 1.22.4): A fundamental package for scientific computing that provides support for arrays, matrices, and numerous mathematical functions.

ESM (version 2.0.0) was used for the protein embedding.

SHAP (version 0.42.1): A library for explaining machine-learning model predictions.

XGBoost (version 1.7.6): A highly efficient, flexible, and portable library for gradient boosting.

SciPy (version 1.10.1): a library used for entropy calculation, softmax for probability calculation, and evaluation statistics.

Scikit-learn (version 1.2.2): A user-friendly library for machine learning and evaluation metrics.

re (version 2.2.1): Module for working with regular expressions.

CatBoost (version 1.2): A machine learning library is designed to handle categorical data automatically.

Matplotlib (version 3.7.1): A plotting library for creating static, animated, and interactive visualizations.

Seaborn (version 0.12.2): A data-visualization library based on Matplotlib provides a high-level interface for drawing attractive statistical graphics.

All code to implement the ML in this study was written and executed on Google Colab[110], a cloud-based Python development environment that allows running Jupyter notebooks. Google Colab provides a flexible platform for writing and executing code and is equipped with robust computational resources, including GPUs and TPUs, making it an ideal choice for data-intensive tasks. The Jupyter notebook, with the implementation of machine learning, can be accessed here.

3.7.2 Datasets

ML models have emerged as powerful tools in the field of protein pathogenicity prediction to identify disease-causing mutations. An example is LYRUS[40], which uses an XGBClassifier to predict the pathogenicity of missense variants. It was trained on a dataset comprising 4363 protein structures corresponding to 22.639 SAVs from the ClinVar database[40]. Other studies leveraged datasets containing both pathogenic and benign variants to enhance the performance of the ML model in predicting the deleteriousness of single amino acid variations[111]. Some prediction methods have been developed based on limited data, such as disease-causing single amino acid variation (SAV) from the OMIM database or databases that categorize variants based on their effects on protein function or structure[111].

MVP is another ML model that predicts the pathogenicity of missense variants using deep learning (DL)[112]. Transformer-based DL models have also been used to predict protein properties including the effects of mutations [113]. Yamaguchi and Saito used pre-training and fine-tuning of a transformer network for mutation prediction [113].

Beyond these sources, understanding the structural dynamics is very important, as it significantly influences the functional impact of missense variants. It has been suggested that the predictive power of computational models can be improved by analyzing factors such as sequence conservation and structural features[114].

Building and evaluating such ML models require well-selected and correctly labeled datasets for training and testing. In this section, we describe the construction of two missense variant datasets that were used for training and testing our ML models for predicting protein mutation pathogenicity.

Construction of the First Dataset

The first dataset constructed in this study was derived from the article "The genetic basis of classic nonketotic hyperglycinemia due to mutations in GLDC and AMT" by Coughlin et al. (2016)[6]. This article provided a comprehensive list of AMT mutations, of which 74 mutations resulted in missense changes in T-Protein. Among these 74 mutations, 35 were unique missense mutations that were identified as pathogenic in causing NKH[6].

From this list, a subset of 25 missense variants was selected for inclusion in the first dataset. All 25 variants were confirmed pathogenic based on the study's findings[6]. To augment the dataset, an additional 10 missense variants were included, consisting of 4 benign variants and 6 pathogenic variants. These additional variants were selected by filtering the ClinVar and Simple ClinVar databases, specifically targeting missense variants labeled as benign/likely benign or pathogenic/likely pathogenic[115,116]. The details of this dataset and these variants can be found in Appendix 21.

Construction of the Second Dataset

The second dataset comprised the remaining missense variants from the AMT mutations list provided by Coughlin et al. (2016)[6]. After selecting 25 variants for the first dataset, 10 missense variants remained. To ensure diversity in the dataset, one additional pathogenic variant and one benign variant were included. These variants were also selected by filtering the ClinVar and Simple ClinVar databases following the same criteria as mentioned before[115,116]. The details of these variants are provided in Appendix 22.

Splitting into Training, Validation, and Test Sets

To ensure proper evaluation of ML models, it is important to have a good validation strategy and solid evaluation metrics[117]. One way to do this is to divide the data into training, validation, and test sets, which allows for the evaluation of the generalization capability of the models[118]. The training set enables the models to learn patterns and relationships from known pathogenic and benign variants, whereas the validation set facilitates the tuning of hyperparameters and model selection. Finally, the test set, which is completely independent, provides a reliable assessment of the performance of the model on unseen data[119].

The first dataset was divided into training and validation sets, with 80% of the missense variants randomly selected as the training set and the remaining 20% as the validation set. The validation set was not used during training but served as an independent dataset for model evaluation[117]. The second dataset was reserved as the final test set to ensure independence of the datasets. There was no overlap between the variants included in the first and second dataset [120].

3.7.3 Feature Engineering

Feature engineering is a crucial step in the development of effective ML models for predicting mutation protein pathogenicity. The dataset used in this study presented limitations in terms of size, which directly affected the options available for feature engineering. The limited amount of data required careful consideration and utilization of relevant features to enhance the predictive capabilities of the ML models.

To augment the predictive capabilities of our models, we utilized the available output features provided by the evolutionary scale modeling (ESM-2) model, a cutting-edge protein model[121,122] trained on a masked language modeling objective[123]. This model has demonstrated promising results in mutation protein pathogenicity prediction[123], making it a valuable resource for feature engineering. The ESM-2 model is available as part of the ESM package "fair-esm-2.0.0"[124], which provides code and pre-trained weights for Transformer protein language models from Meta AI's Fundamental AI Research Team [125]. The ESM-2 model is trained with ~65 million unique sequences and is an order of magnitude faster (60x) for high-resolution structure prediction[123,126]. The ESM-2 model was used to create an open atlas of 617 million predicted metagenomic protein structures[123,127], which can be used to predict how a protein folds based on the primary amino acids [128]. One of the key features provided by the ESM-2 model is the "attention contact " output. This output predicts the distance between amino acids in the protein sequence and provides insights into protein folding, structural stability, and potential interactions with other molecules[129]. By incorporating this feature, we can capture the spatial relationships between amino acids and gain a deeper understanding of how mutations may affect AMT functionality. Another important feature derived from the ESM-2 model is the "amino acid predictions" probabilities[130]. The ESM-2 model provides "amino acid predictions" probabilities, which allows to estimate the probabilities of different amino acids at each position[123] in the T-protein sequence. This feature is useful for analyzing the potential impact of specific mutations on the enzymatic activity and stability of T-proteins. By considering these probabilities as features, we can quantify the likelihood of different amino acids occurring at specific positions, aiding the identification of critical residues affected by mutations[123]. In addition, the ESM-2 model enables the calculation of entropy based on the predicted probabilities of each position in the AMT protein sequence. Entropy values provide insight into the variability or conservation of amino acids at different positions. By incorporating entropy as a feature, we can assess the diversity or conservation of specific positions in the protein sequence, further enhancing our understanding of mutational effects [131]. To calculate the entropy of the predicted probabilities at each position in the AMT protein sequence, the formula for the entropy of a probability distribution can be used as follows:

$$Entropy = -\Sigma\ (p\_i\ *\ log(p\_i))$$

Equation 4. Entropy function

where p_i represents the predicted probability of each amino acid at a given position in the sequence, and the sum of all possible amino acids. The log is usually taken as base 2 if entropy is measured in bits. In this context, entropy measures the uncertainty or randomness of the distribution of amino acids at a given position. A high entropy would indicate a high degree of uncertainty about which amino acid would appear at that position, suggesting that many different amino acids could be functionally substituted at that position. On the other hand, low entropy indicates a high degree of certainty about which amino acid would appear at that position, suggesting that the specific amino acid is functionally important. While ESM models do not explicitly calculate entropy, the predicted probabilities provided for each position can be used to calculate the entropy. This provides an entropy profile along the protein sequence, potentially highlighting regions of functional importance or variability. To calculate entropy[132] the SciPy 1.10.1 entropy package[133] was used.

By leveraging the outputs of the ESM-2 model[122], such as attention contacts, amino acid prediction probabilities, and entropy, we designed specific features tailored to predict the effects of mutations in the T-Protein. These features integrate structural, sequence-based, and variability-based information, providing a comprehensive representation for accurate prediction of protein pathogenicity due to mutations.

The key features chosen for predicting the consequences of mutations in the T-Protein are as follows:

Attention Contacts Sum: Attention contacts are metrics that indicate which parts of the protein are predicted to be close together in 3D space based on the attention mechanism of the ESM transformer model which provides insights into the spatial relationships between amino acids[124], protein folding, structural stability, and interactions[129]. To obtain an understanding of these structural details, we called the ESM-2 model to return the attention contacts for each residue[134] which retrieved a predicted contact map for the WT T-Protein. This contact map provides a comprehensive view of the proximity details between each pair of amino acid residues in the protein's 3D structure. To facilitate the visualizing, the attention contact map[135] are represented as a 2D image, where each pixel corresponds to a pair of residues in the protein sequence. The color of each pixel signifies the attention contact score between the corresponding residue pair.

The 'Attention Contacts sum' feature is an indication of the number of contacts a residue makes with other residues within the protein structure[121]. As the value of 'Attention Contacts sum' increases, it suggests that the residue is likely part of the protein core.

Contacts sum inverse: Contact sum inverse feature, which is computed as the reciprocal of the attention contact sum. The contact sum value indicates the number of contacts a residue makes with other residues in the protein structure. By considering its reciprocal, higher values of the contact sum inverse correspond to residues that are less likely to be part of the protein's core structure, providing a more interpretable measure for subsequent analyses[122],[136].

Amino acids Prediction probabilities: This feature estimates the likelihood of specific amino acids at each position in a protein sequence[129]. It represents the probability associated with the amino acid that substitutes the original residue at the mutated position in the protein sequence. This probability, in essence, provides an estimate of how likely a specific amino acid substitution is to occur at the given position.

These amino acid predictions can be accessed through the *'results['logits']'* when calling the ESM model[113] to obtain the logits for each position in the T-protein sequence, which is one of the 20 common amino acids[123]. To convert the logits into probabilities, the 'softmax'[137] from the 'scipy.special' module was applied, which normalizes the output and makes it easier to interpret[138,139].

The resulting predictions ranged from indices 4 to 24, each corresponding to one of the 20 common amino acids[123]. To visualize these probabilities, a 2D image is generated where each position in the sequence is represented by a different column and each amino acid by a different row[123]. The color of each cell in this matrix represents the probability that the corresponding position in the sequence is the respective amino acid[123] serving as a valuable metric in our ML model[140,141, 142]

Normalization of Amino Acid Predictions

To further refine the features, we performed normalization of the raw values for each amino acid prediction logits. Normalization of amino acid predictions is a technique used to refine the features and highlight the relative differences among the amino acid probabilities[143].

In this process, the highest probability value at each position is subtracted from all 20 amino acid prediction probabilities for the same position, effectively setting the maximum value to zero[144].This technique, known as "max normalization," helps emphasize the model's preferences for certain amino acids at each position[144, 145].

The normalization of amino acid predictions not only makes the data more interpretable but also potentially improves the ability of the ML model to detect crucial patterns. By normalizing the logits, the model can better distinguish the relative likelihoods of different amino acids at each position in the protein sequence[146].

To visualize the normalized amino acid prediction predictions, a heatmap 2D image plot was created using function from the Seaborn 0.12.2[147] library in Python used for creating heatmaps[148].

Entropy

Entropy values offer insights into the variability or conservation of amino acids in protein sequences. This helps in understanding the level of diversity or conservation at each position in the protein[131]. The entropy[133] function from the scipy[149] module accepts the probabilities matrix as an input and for each position, the function computes the entropy for the probabilities at that position.

Amino acids count

A protein sequence is an arrangement of amino acids held together by peptide bonds[150].Proteins can be made from 20 different kinds of amino acids, and the prevalence of certain amino acids might be related to the function and structure of the protein[150]. The most common amino acid in a protein sequence can be determined by analyzing the count of each amino acid in the sequence[151].

3.7.4 ML Models

In this study, we followed the best ML practice by training and validating different models: XGBRegressor, XGBClassifier, CatBoostRegressor, CatBoostClassifier, RandomForestRegressor, and RandomForestClassifier[152,153]. These models were chosen because of their unique strengths and versatility in handling diverse datasets and prediction tasks[154–157]. During the validation stage, the performance of each model was evaluated using the validation set. The performance metrics of the models were compared to identify the one that best predicted the pathogenicity of the protein mutations [154–157] The models were evaluated and compared using Spearman correlations and mean absolute error (MAE)[158] in the validation phase[159,160]. The calculations of these metrics allow us to assess the predictive accuracy of the model and its alignment with computational predictions. Overall, the use of multiple models and comparison of their performance is a common and recommended practice in ML for selecting the most suitable model for a specific task.

The **XGBRegressor:** is an ML model used for regression tasks. It is based on the gradient boosting framework and uses decision trees as base learners. The XGBRegressor is known for its high accuracy and speed, making it a popular choice for predicting the pathogenicity of protein mutations[161–163].

XGBClassifier: This is another ML model based on the gradient boosting framework, but it is used for classification tasks. It is similar to XGB Regressor, but instead of predicting a continuous value, it predicts the probability of a sample belonging to a certain class. XGBClassifier has been used to predict the pathogenicity of single amino acid variants (SAVs) in proteins, achieving high accuracy levels[40,161]

CatBoostRegressor: This ML can predict the continuous impacts of protein mutations by formulating rules based on decision trees [164].

CatBoostClassifier: This ML algorithm can be used to predict the impact of mutations on protein structure and interactions [164,165]. It is a type of gradient boosting algorithm, such as CatBoostRegressor.

RandomForestRegressor: An ML method that uses multiple decision trees to predict continuous outcomes[166,167]. It is suitable for predicting the quantitative effects of protein mutations by factoring in the randomness of tree construction to optimize predictions.

RandomForestClassifier: This is an ML algorithm that can be used to predicting protein mutations[163]. It has been used to predict the effects of mutations on protein stability with high accuracy[162,168–170] and is based on multiple decision trees, similar to RandomForestRegressor, but it is used for classification tasks.

3.7.5 Test Evaluation Metrics

The evaluation methods used in the test stage include several evaluation metrics to assess the performance of the model on the test dataset. The evaluation metrics are as follows:

Accuracy: This measures the proportion of correctly classified instances among the total number of instances in the test dataset [171,172].

Area Under the Receiver Operating Characteristic curve (AUC-ROC) measures the model's ability to distinguish between positive and negative classes. The true positive rate is plotted against the false positive rate at different classification thresholds[173,174].

Precision: This measures the proportion of true positive predictions to the total number of positive predictions made by the model[174].

Recall: This measures the proportion of true positive predictions out of the total number of actual positive instances in the test dataset [175].

Area Under the Precision-Recall curve (AUC-PR): This measures the trade-off between precision and recall for different classification thresholds [176].

F1-score: This measure is commonly used for classification models, particularly when dealing with imbalanced datasets. The F1-score is a weighted average of precision and recall, where the relative contributions of precision and recall to the F1-score are equal[117]. The formula for the F1-score is:

$$F1 = \frac{2 * Precision * Recall}{Precison + Recall}$$

Equation 5. F1-Score

Precision and recall are two metrics that consider class imbalance. The F1 score is a measure of a test's accuracy, combining both precision and recall into a single value. Precision is the fraction of true positives among all predicted positives, whereas recall is the fraction of true positives among all actual positives. The F1-score reaches its best value at 1 and the worst score at 0. All of these evaluation metrics were used to assess the performance of the model during the test phase.

3.8 Clinical outcome score to assess patient disease severity status

In this study, we aimed to apply a clinical severity scale for NKH, a multisystemic neurometabolic disorder caused by mutations in the T protein. Previous classifications of the disease were based on the presence or absence of brain malformations or developmental outcomes, but a quantitative, dynamic progression of different NKH symptoms across severe and attenuated disease is lacking [2]. The authors of the article titled "Large scale analyses of genotype-phenotype relationships of glycine decarboxylase mutations and neurological disease severity" developed a clinical outcome score (COS) for P-protein[177]. COS is structured into four major domains: cognitive disorders, seizures, muscle and movement dysfunction, and brain malformations. Utilizing a Likert-like scale with scores ranging of 0–3 based on severity in each domain[177]. The seizure domain was assigned a non-linear step increase of 1 to 3 corresponding to the transition from controlled seizure activity to uncontrolled seizure activity. The brain malformation domain was assigned a binary choice of 0 or 3, because any brain malformation is expected to seriously impact neurological disease. Summation of all four domains of COS, with a maximal score of 12. In this assessment we reviewed case-reports and publications that listed patients with NKH due to T-protein mutations. Our objective was to create a COS that assessed disease severity status in patients with T-protein mutations.

4 RESULTS AND DISCUSSION

4.1 Data Sources

The article ("The genetic basis of classic nonketotic hyperglycinemia due to mutations in GLDC and AMT"; Coughlin, C. R. et al., 2016)[6] contains a list of AMT mutations (74 mutations, of which 35 are unique proteins change missense mutations) presented per mutation consequence in Figure 7 A and per Mutation's type in Figure 7 B. Tables in Appendix 2 and Appendix 3.

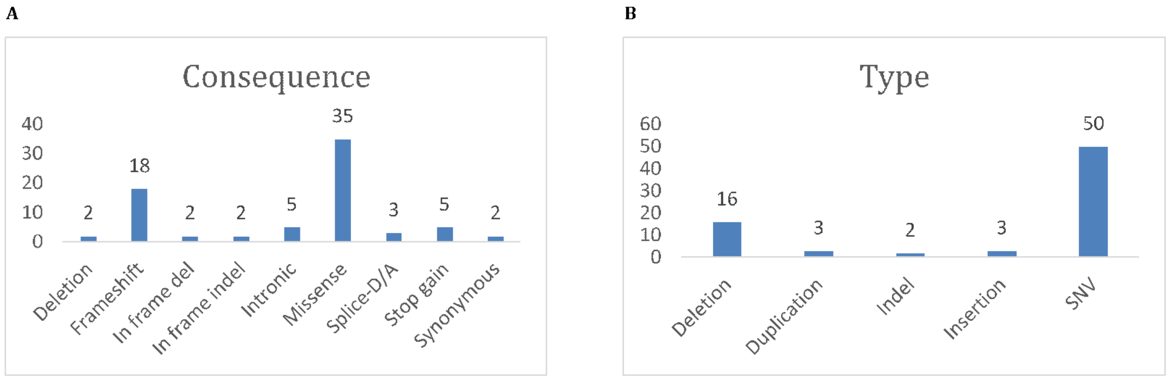


Figure 7. Mutation consequence (A) and Mutation's type (B) listed in Coughlin, C. R. et al., 2016.

Simple ClinVar (https://simple-clinvar.broadinstitute.org/)[115] is a website that is based on the ClinVar database version of July 14, 2021, but the data is better structure[178]. Simple ClinVar contains a list of AMT mutations (235 unique mutations, of which 99 are missense) presented per mutation consequence, per mutation's type and per clinical significance in Figures 8 A, B and C. Table can be found in Appendix 8.

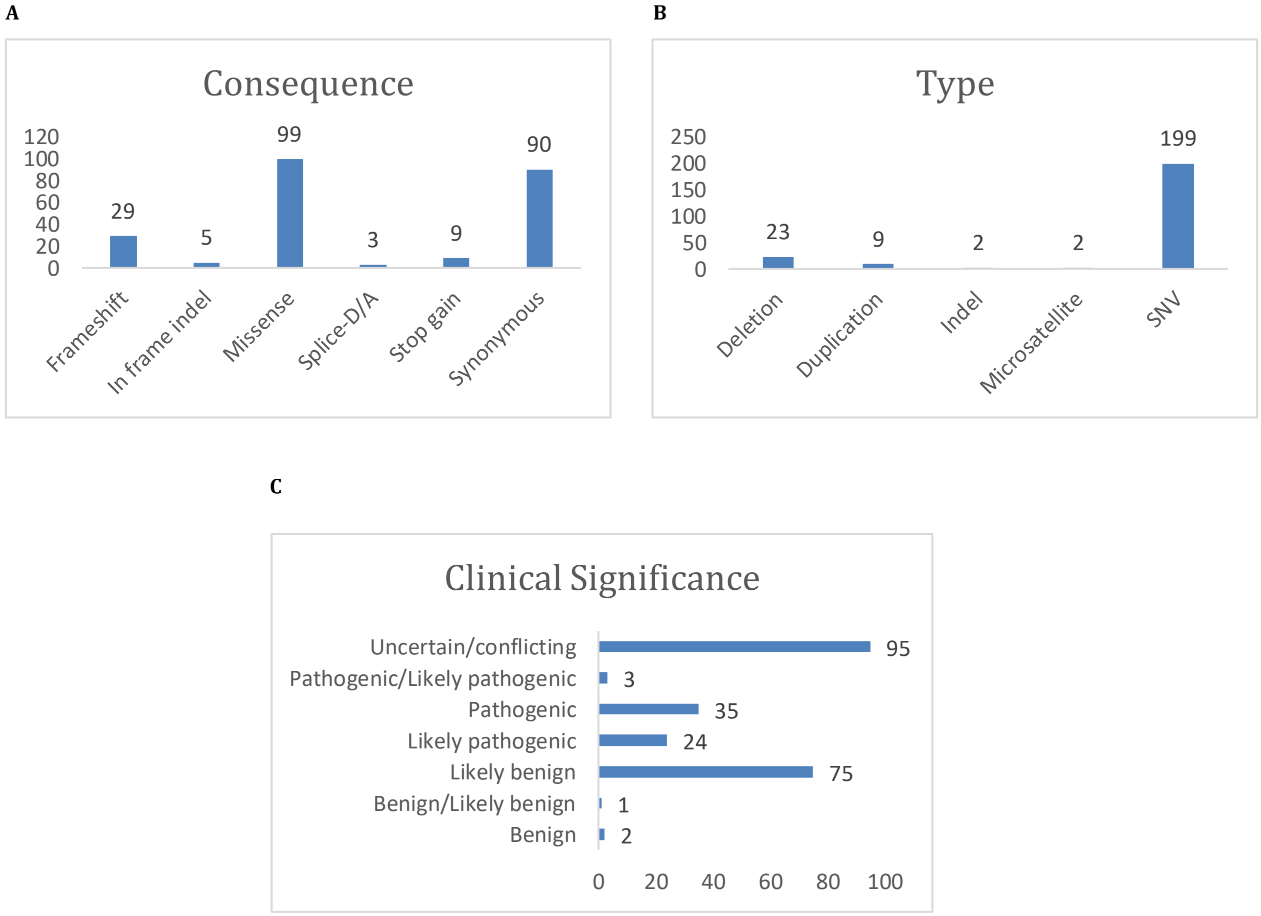


Figure 8. Mutation's consequence (A), type (B) and clinical cignificance (C) listed in Simple ClinVar.

ClinVar is a publicly available database of genetic variations and their relationship with human health. It is maintained by the NCBI and is a resource of the national library of medicine (NLM) of the national institutes of health (NIH). ClinVar contains information on genetic variations (such as SNPs), small insertions and deletions, and structural variations), which have been reported to be associated with health-related phenotypes (observable traits or characteristics)[179]. ClinVar contains a list of AMT mutations (49 unique mutations, of which 37 are missense) presented per mutation consequence, per mutation's type and per clinical significance in Figures 8 A, B and C. Table in the Appendix4.

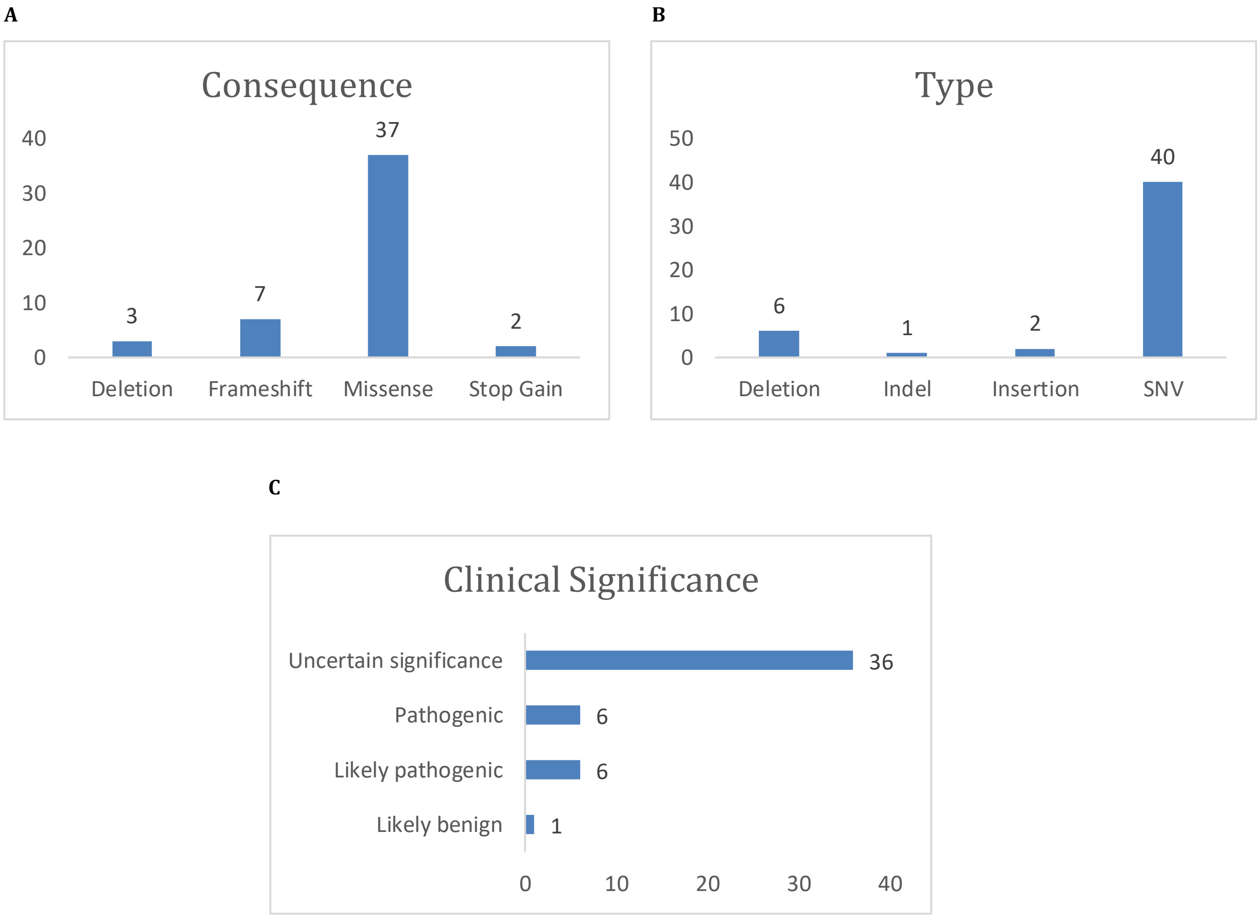


Figure 9. Mutation consequence (A), Type (B) and clinical significance (C) listed on the ClinVar website.

PubTator is a tool developed by the NCBI to annotate and extract information from biomedical literature. It is a web-based platform that allows users to upload texts in various formats (e.g., PubMed Central articles, PubMed abstracts, and full-text articles) and perform text-mining tasks such as entity recognition, relation extraction, and concept normalization[180]. PubTator contains a list of AMT mutations (nine unique mutations, of which five are missense) presented per mutation consequence, per mutation's type and per clinical significance in Figures 10 A, B and C. Table in the Appendix 5.

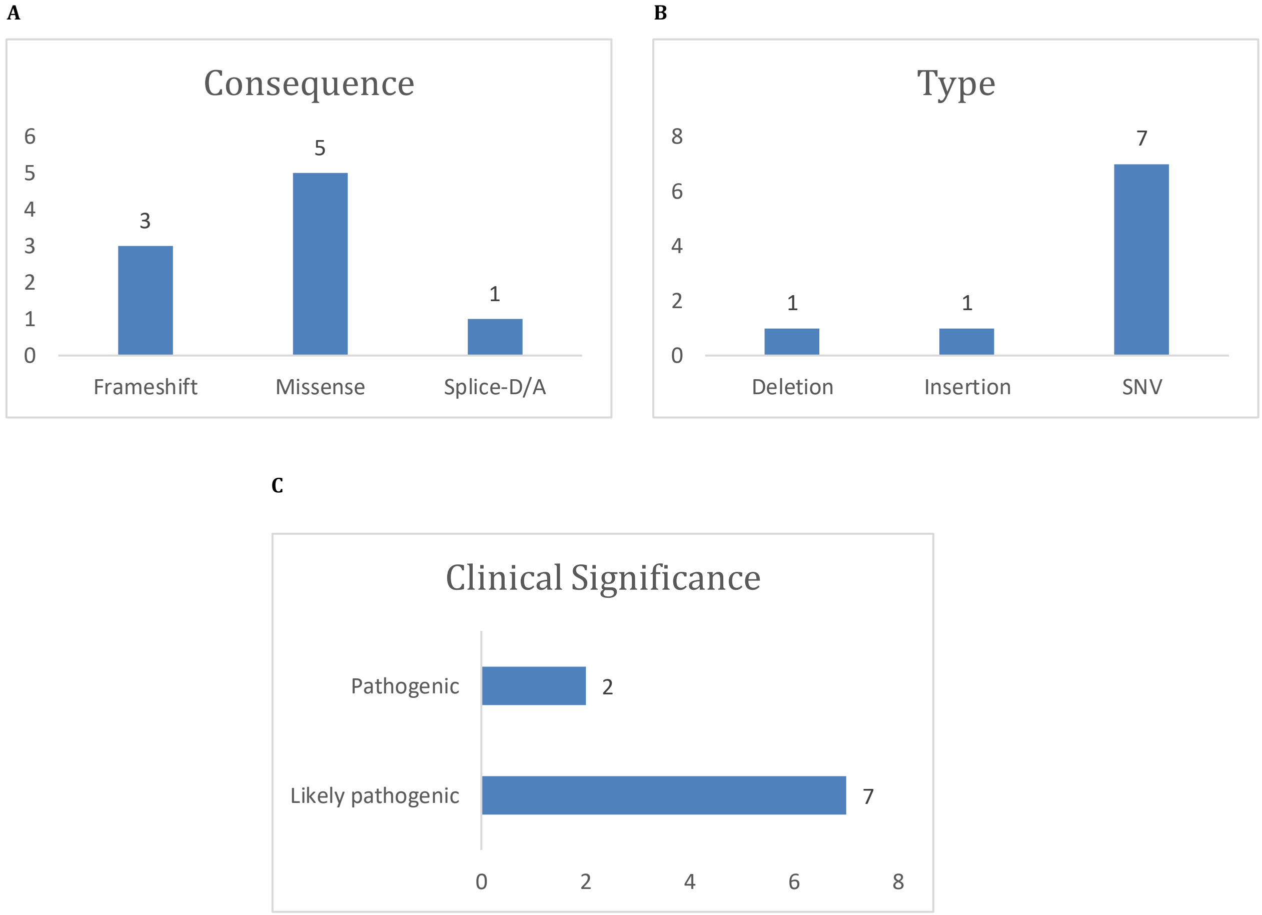


Figure 10. Mutation's Consequence (A), Type (B) and Clinical Significance (C) in Pubtator.

These four sources provided 335 unique mutations (Table 1) that were a combination of DNA and protein mutations, of which 149 were protein missense mutations divided in Figure 11 per data source (A) and consequence (B) for all 4 sources.

Source	Mutations Count
ClinVar	48
ClinVar/GenMed	1
GenMed	45
PubTator	6
Simple ClinVar	205
Simple ClinVar/ClinVar/GenMed	1
Simple ClinVar/GenMed	26
Simple ClinVar/GenMed/PubTator	1
Simple ClinVar/PubTator	2

Table 1. Total T-protein Mutations Count per Data Source.

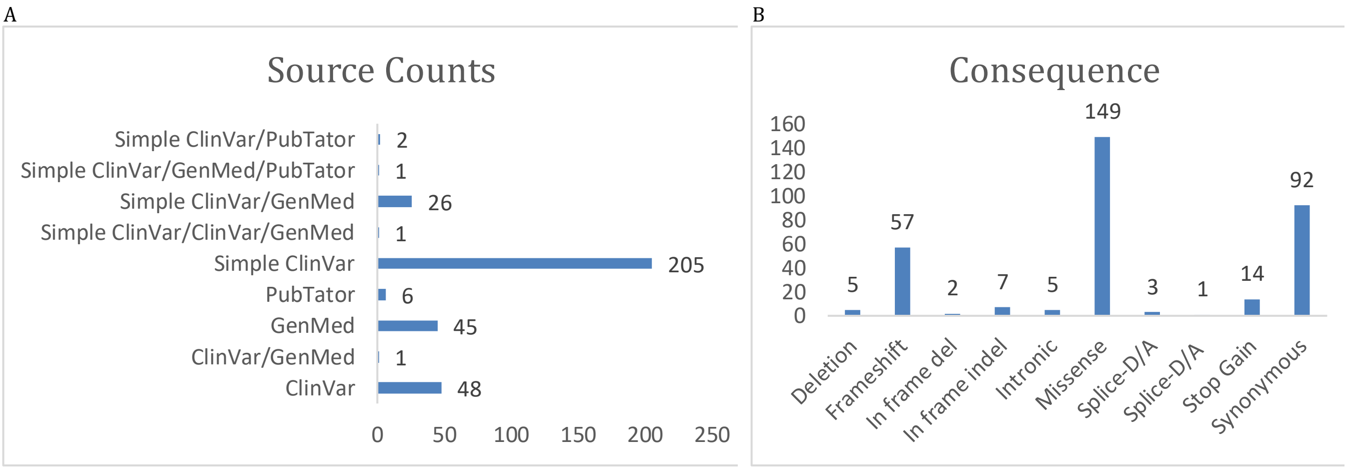


Figure 11. Total T-protein mutations per data source (A) and consequence (B) for all 4 sources.

4.2 Homology Model of T-protein

The values in the Table 2 represent scores for five different protein homology models. The MolPDF[181,182] score was given in the first column, DOPE[31] in the second column, and RMSD[183] in the third column.

Model filename	MolPDF	DOPE Score	RMSD Score
model1.pdb	277241.9	-85082.9	4.017
model2.pdb	235714.1	-84464.8	4.009
model3.pdb	328029.6	-84360.4	4.007
model4.pdb	285355.3	-84355.6	4.141
model5.pdb	231499.2	-84264.4	4.167

Table 2. Top scored Homology Models.

The first model had the highest MolPDF[184] score of 277241.9; therefore, it was likely to have the highest quality prediction among the five models. The first model also had the lowest DOPE score of -85,082.9 and was likely to have the highest quality prediction. The DOPE alignment between model1.pdb and the template for T-protein is shown in Figure 12.

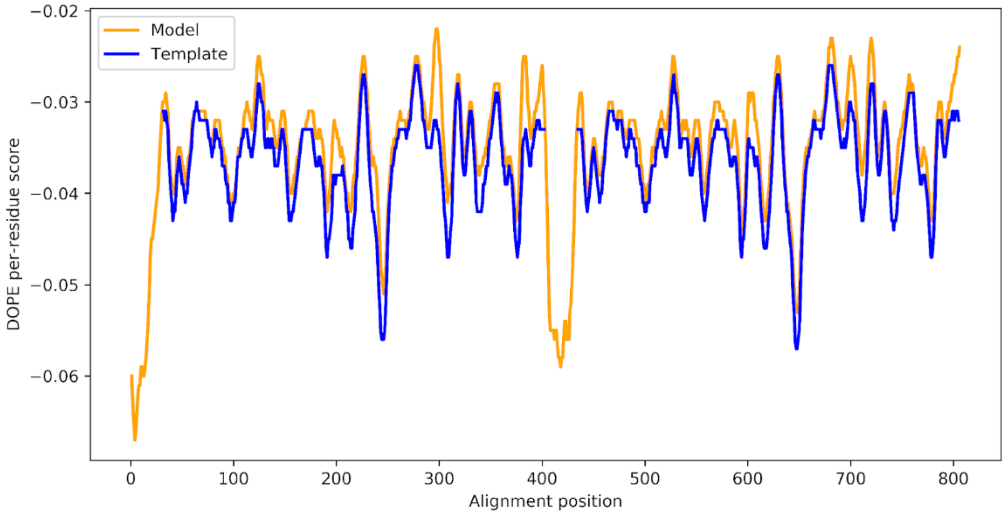


Figure 12. DOPE alignment of model1.pdb.

The third model had the lowest RMSD score of 4.007 Å and was likely to have the most accurate prediction. Additionally, the models were visually inspected using PyMOL software to confirm the quality of the prediction and to assess other important features such as the location and orientation of ligands, the presence of any anomalies or errors, and the overall structural stability of the protein.

The best model (Figure 13), which provided the most accurate and reliable prediction of the protein structure based on a combination of evaluation metrics and visual inspection, was the first model in the table model1.pdb.

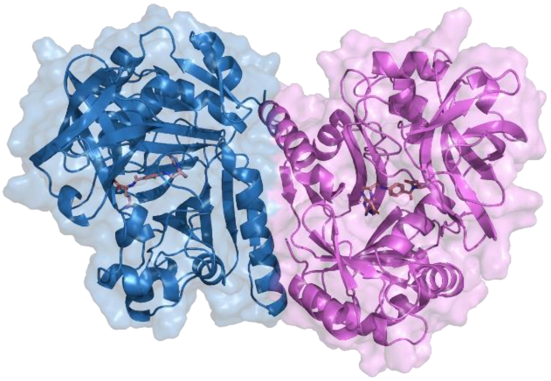


Figure 13. Structural representation of the T-protein best model.

4.3 Folate and T-protein Interactions

4.3.1 BINANA Close Contacts

BINANA software was used to analyze the interactions between the T-protein and ligand THH in terms of close contacts, hydrogen bonds, and salt bridges within 4 Å of the T-protein model, as shown in Figure 14. The program also calculated the distance between the T-protein and THH cofactor.

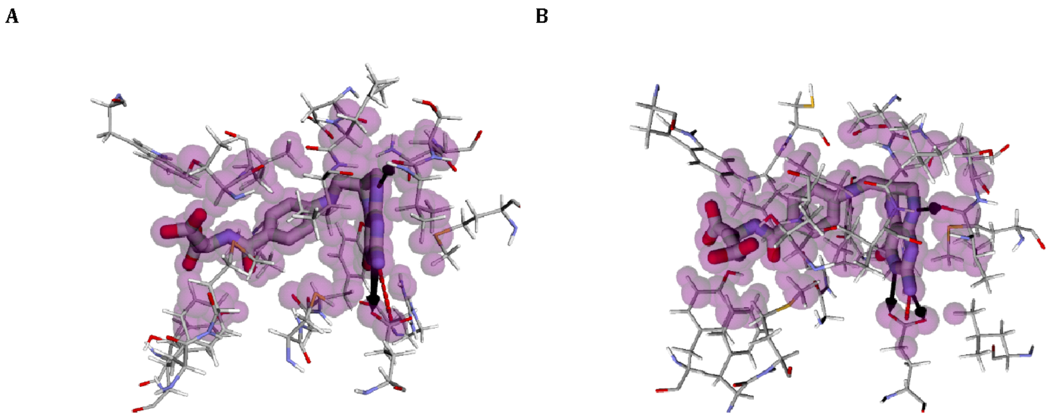


Figure 14. BINANA output of ChainA of model1.pdb (A) and BINANA output of ChainB of model1.pdb. Purple spheres (Close contact), Solid Black arrow from donor to acceptor (Hydroden bond), and Red dashed line (SaltBridge).

Table 3 lists the close contacts between residues and the cofactor in T-protein - Chain A and T-protein - Chain B. Close contacts are identified by the amino acids involved, their residues, and their names. These close contacts are likely to be involved in protein-ligand interactions that are crucial for the function of the T-protein. Table 3 also highlights the possible involvement of residues in hydrogen bonding, salt bridges, and other interactions that contribute to the stability and function of the protein.

T-protein CHAIN A	T-protein CHAIN B
PHE(F204)	PHE(F204)
TRP(W290)	MET(M205)
LEU(L116)	LEU(L116)
THR(T115)	TRP(W290)
MET(M205)	THR(T115)
LEU(L270)	LEU(L270)
THR(T397)	TYR(Y399)
TYR(Y399)	ILE(I131)
GLY(G224)	GLY(G224)
ILE(I131)	TYR(Y225)
TYR(Y225)	CYS(C271)
ASP(D129)	ASP(D129)
LEU(L130)	LEU(L130)
ASN(N145)	ASN(N145)

VAL(V143)	VAL(V143)
SER(S144)	SER(S144)
MET(M84)	MET(M84)
ARG(R261)	ARG(R261)
CYS(C223)	LEU(L179)
GLU(E232)	GLU(E232)

Table 3. List the close contacts between residues and the cofactor THH in model1.pdb ChainA and ChainB.

4.4 Pathogenicity Prediction

4.4.1 Evaluation of Software Tools for Predicting SAV Mutations Clinical Significance

SAV is an amino acid substitution in a protein sequence that can potentially influence the entire protein structure or function, as well as its binding affinity. Protein destabilization is related to disease. Identification and characterization of clinically significant SAV mutations are crucial for the diagnosis and management of genetic diseases. Several software tools are available to predict the potential pathogenicity of missense mutations in proteins. In this study, six software tools (LYRUS, PROVEAN, SNAP2, MUTAFRAME, SIFT, and SuSPect) were evaluated for their ability to predict the clinical significance of missense mutations in the T-protein.

A dataset of SAV mutations with confirmed clinical significance was obtained from the article reported previous by Coughlin et al. (2016)[6] published in Genetics in Medicine. The performances of the six software tools were evaluated by comparing their predictions with the confirmed clinical significance of the mutations listed in this article. The statistical significance of the predictions was then assessed.

4.4.2 LYRUS Result

LYRUS algorithm was used to predict the clinical significance of missense mutations in a dataset of 35 mutations. The results (Table 4) showed that the algorithm correctly predicted the clinical significance of 94.3% (33) of the mutations. However, it failed to predict a clinical significance of 5.7% (2) of the mutations. The LYRUS algorithm was effective in predicting the clinical significance of missense mutations and its accuracy was in line with that of previous studies in the field[185].

Source	Clinical Significance	Protein Change	Frequency	LYRUS Predicted Clinical Significance	Predict Score	Predict Probability
GenMed	Pathogenic	M1T	7	Failed to predict	0	0.1
GenMed	Pathogenic	G47R	4	Predicted	1	0.8
GenMed	Pathogenic	H71P	2	Predicted	1	0.9
GenMed	Pathogenic	R73C	8	Predicted	1	0.9

GenMed	Pathogenic	S77L	4	Predicted	1	0.8
GenMed	Pathogenic	H83R	4	Predicted	1	0.7
GenMed	Pathogenic	R94W	9	Predicted	1	0.7
GenMed	Pathogenic	M98R	2	Predicted	1	0.9
GenMed	Pathogenic	I106T	5	Predicted	1	0.5
GenMed	Pathogenic	S117L	4	Predicted	1	0.7
GenMed	Pathogenic	N145I	2	Predicted	1	0.8
GenMed	Pathogenic	L172P	1	Predicted	1	0.9
GenMed	Pathogenic	L182P	2	Predicted	1	1.0
GenMed	Pathogenic	Q183R	1	Predicted	1	0.9
GenMed	Pathogenic	V212A	1	Predicted	1	0.6
GenMed	Pathogenic	R222H	4	Predicted	1	0.5
GenMed	Pathogenic	R222C	12	Predicted	1	0.9
GenMed	Pathogenic	Y225C	1	Predicted	1	0.6
GenMed	Pathogenic	D229H	1	Predicted	1	1.9
GenMed	Pathogenic	R265H	5	Predicted	1	0.9
GenMed	Pathogenic	R265S	3	Predicted	1	0.9
GenMed	Pathogenic	R265C	4	Predicted	1	0.8
GenMed	Pathogenic	G269D	1	Predicted	1	1.0
GenMed	Pathogenic	D276H	4	Predicted	1	1.0
GenMed	Pathogenic	T282I	1	Predicted	1	0.9
GenMed	Pathogenic	R296H	7	Failed to predict	0	0.4
GenMed	Pathogenic	R296P	1	Predicted	1	0.9
GenMed	Pathogenic	R296C	4	Predicted	1	0.9
GenMed	Pathogenic	R320H	36	Predicted	1	0.9
GenMed	Pathogenic	R320G	2	Predicted	1	0.8
GenMed	Pathogenic	R320C	2	Predicted	1	0.9
GenMed	Pathogenic	R331Q	4	Predicted	1	0.8
GenMed	Pathogenic	R331P	1	Predicted	1	0.9
GenMed	Pathogenic	A361V	1	Predicted	1	0.5
GenMed	Pathogenic	G363R	1	Predicted	1	1.0

Table 4.LYRUS Prediction Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.3 PROVEAN Result

The PROVEAN algorithm was applied to a dataset of 35 missense mutations in order to predict their clinical significance. The results (Table 5) showed that the algorithm successfully predicted the clinical significance of 97.1% (34) of the mutations. However, it failed to predict the clinical significance of 2.9% (1) of the mutations. Its performance is in line with previous studies and provides further evidence of its usefulness in the interpretation of genetic test results and in guiding clinical decision-making. It should be noted that the limitations of the algorithm in predicting the clinical significance of a small number of mutations highlight the need for further research to improve its accuracy.

Source	Clinical Significance	Protein Change	Frequency	Provean Predict Clinical Significance	Provean Predict Effect	Provean Score
GenMed	Pathogenic	M1T	7	Failed to predict	0	-0.4
GenMed	Pathogenic	G47R	4	Predicted	1	-6.3
GenMed	Pathogenic	H71P	2	Predicted	1	-8.6
GenMed	Pathogenic	R73C	8	Predicted	1	-7.2

GenMed	Pathogenic	S77L	4	Predicted	1	-5.3
GenMed	Pathogenic	H83R	4	Predicted	1	-7.4
GenMed	Pathogenic	R94W	9	Predicted	1	-5.2
GenMed	Pathogenic	M98R	2	Predicted	1	-4.4
GenMed	Pathogenic	I106T	14	Predicted	1	-3.9
GenMed	Pathogenic	S117L	4	Predicted	1	-3.7
GenMed	Pathogenic	N145I	2	Predicted	1	-8.1
GenMed	Pathogenic	L172P	1	Predicted	1	-4.6
GenMed	Pathogenic	L182P	2	Predicted	1	-5.3
GenMed	Pathogenic	Q183R	1	Predicted	1	-3.8
GenMed	Pathogenic	V212A	1	Predicted	1	-3.9
GenMed	Pathogenic	R222H	4	Predicted	1	-4.9
GenMed	Pathogenic	R222C	12	Predicted	1	-7.9
GenMed	Pathogenic	Y225C	1	Predicted	1	-8.7
GenMed	Pathogenic	D229H	1	Predicted	1	-6.6
GenMed	Pathogenic	R265H	5	Predicted	1	-4.9
GenMed	Pathogenic	R265S	3	Predicted	1	-5.9
GenMed	Pathogenic	R265C	4	Predicted	1	-7.8
GenMed	Pathogenic	G269D	1	Predicted	1	-6.8
GenMed	Pathogenic	D276H	4	Predicted	1	-6.8
GenMed	Pathogenic	T282I	1	Predicted	1	-5.5
GenMed	Pathogenic	R296H	7	Predicted	1	-4.3
GenMed	Pathogenic	R296P	1	Predicted	1	-6.5
GenMed	Pathogenic	R296C	4	Predicted	1	-7.5
GenMed	Pathogenic	R320H	36	Predicted	1	-4.5
GenMed	Pathogenic	R320G	2	Predicted	1	-6.2
GenMed	Pathogenic	R320C	2	Predicted	1	-7.3
GenMed	Pathogenic	R331Q	4	Predicted	1	-3.5
GenMed	Pathogenic	R331P	1	Predicted	1	-5.5
GenMed	Pathogenic	A361V	1	Predicted	1	-3.4
GenMed	Pathogenic	G363R	1	Predicted	1	-6.2

Table 5. Provean Prediction of Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.4 SNAP2 Result

SNAP2 is a software that uses multiple algorithms and databases to predict the functional impact of missense mutations. The results (Table 6) of this study showed that SNAP2 could predict the clinical significance of 94.3% (33) of the missense mutations tested in the dataset. However, it failed to predict a clinical significance of 5.7% (2) of the mutations.

Source	Clinical Significance	Protein Change	Frequency	Snap2 Predict Clinical Significance	Snap2 Predict Effect	Snap2 Score
GenMed	Pathogenic	M1T	7	Failed to predict	0	-96
GenMed	Pathogenic	G47R	4	Predicted	1	54
GenMed	Pathogenic	H71P	2	Predicted	1	60
GenMed	Pathogenic	R73C	8	Predicted	1	67
GenMed	Pathogenic	S77L	4	Predicted	1	39
GenMed	Pathogenic	H83R	4	Predicted	1	88
GenMed	Pathogenic	R94W	9	Failed to predict	0	-24
GenMed	Pathogenic	M98R	2	Predicted	1	57
GenMed	Pathogenic	I106T	5	Predicted	1	39
GenMed	Pathogenic	S117L	4	Predicted	1	67

GenMed	Pathogenic	N145I	2	Predicted	1	84
GenMed	Pathogenic	L172P	1	Predicted	1	8
GenMed	Pathogenic	L182P	2	Predicted	1	87
GenMed	Pathogenic	Q183R	1	Predicted	1	50
GenMed	Pathogenic	V212A	1	Predicted	1	36
GenMed	Pathogenic	R222H	4	Predicted	1	75
GenMed	Pathogenic	R222C	12	Predicted	1	60
GenMed	Pathogenic	Y225C	1	Predicted	1	44
GenMed	Pathogenic	D229H	1	Predicted	1	50
GenMed	Pathogenic	R265H	5	Predicted	1	48
GenMed	Pathogenic	R265S	3	Predicted	1	39
GenMed	Pathogenic	R265C	4	Predicted	1	42
GenMed	Pathogenic	G269D	1	Predicted	1	97
GenMed	Pathogenic	D276H	4	Predicted	1	61
GenMed	Pathogenic	T282I	1	Predicted	1	48
GenMed	Pathogenic	R296H	7	Predicted	1	34
GenMed	Pathogenic	R296P	1	Predicted	1	65
GenMed	Pathogenic	R296C	4	Predicted	1	39
GenMed	Pathogenic	R320H	36	Predicted	1	72
GenMed	Pathogenic	R320G	2	Predicted	1	78
GenMed	Pathogenic	R320C	2	Predicted	1	60
GenMed	Pathogenic	R331Q	4	Predicted	1	55
GenMed	Pathogenic	R331P	1	Predicted	1	71
GenMed	Pathogenic	A361V	1	Predicted	1	65
GenMed	Pathogenic	G363R	1	Predicted	1	74

Table 6. SNAP2 Prediction of Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.5 MUTAFRAME Result

Mutaframe (Table 7) accurately predicted the clinical significance of 77.1% (27) of the missense mutations tested. However, insufficient data are available to predict the clinical significance of the remaining 22.9% (n = 8) of missense mutations.

Source	Clinical Significance	Protein Change	Frequency	Mutaframe Predict Clinical Significance	Mutaframe Predict Effect	Mutaframe Score
GenMed	Pathogenic	M1T	7	NA		
GenMed	Pathogenic	G47R	4	Predicted	1	0.8
GenMed	Pathogenic	H71P	2	Predicted	1	0.9
GenMed	Pathogenic	R73C	8	Predicted	1	0.9
GenMed	Pathogenic	S77L	4	Predicted	1	0.8
GenMed	Pathogenic	H83R	4	Predicted	1	0.9
GenMed	Pathogenic	R94W	9	Predicted	1	0.9
GenMed	Pathogenic	M98R	2	NA		
GenMed	Pathogenic	I106T	5	Predicted	1	0.9
GenMed	Pathogenic	S117L	4	Predicted	1	0.8
GenMed	Pathogenic	N145I	2	Predicted	1	0.9
GenMed	Pathogenic	L172P	1	Predicted	1	0.9
GenMed	Pathogenic	L182P	2	NA		
GenMed	Pathogenic	Q183R	1	NA		
GenMed	Pathogenic	V212A	1	Predicted	1	0.8
GenMed	Pathogenic	R222H	4	Predicted	1	0.9
GenMed	Pathogenic	R222C	12	Predicted	1	0.9

GenMed	Pathogenic	Y225C	1	Predicted	1	0.9
GenMed	Pathogenic	D229H	1	Predicted	1	0.9
GenMed	Pathogenic	R265H	5	Predicted	1	0.9
GenMed	Pathogenic	R265S	3	NA		
GenMed	Pathogenic	R265C	4	Predicted	1	0.9
GenMed	Pathogenic	G269D	1	Predicted	1	0.9
GenMed	Pathogenic	D276H	4	Predicted	1	0.9
GenMed	Pathogenic	T282I	1	Predicted	1	0.9
GenMed	Pathogenic	R296H	7	Predicted	1	0.7
GenMed	Pathogenic	R296P	1	NA		
GenMed	Pathogenic	R296C	4	Predicted	1	0.8
GenMed	Pathogenic	R320H	36	Predicted	1	0.9
GenMed	Pathogenic	R320G	2	NA		
GenMed	Pathogenic	R320C	2	Predicted	1	0.9
GenMed	Pathogenic	R331Q	4	Predicted	1	0.9
GenMed	Pathogenic	R331P	1	NA		
GenMed	Pathogenic	A361V	1	Predicted	1	0.9
GenMed	Pathogenic	G363R	1	Predicted	1	0.9

Table 7. Mutaframe Prediction of Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.6 SIFT Result

A SIFT score predicts whether an amino acid substitution affects protein function. The SIFT score ranges from 0.0 (deleterious) to 1.0 (tolerated).

The results (Table 8) of the study suggest that the SIFT method had a high accuracy in predicting the clinical significance of missense mutations in the dataset, with 91.4% (32) of the mutations being correctly classified. However, it is important to note that no method is perfect, and SIFT, like any other computational method for predicting the effect of genetic variation on protein function, has limitations. The failure to predict the clinical significance of 8.6% (3) of missense T-protein mutations highlights the need for additional validation, as well as the importance of considering multiple sources of evidence when interpreting the clinical significance of a genetic variant.

Source	ClinicalSignificance	Protein Change	Frequency	SIFT_Predict_ClinicalSignificance	SIFT_predict_effect	SIFT_Score
GenMed	Pathogenic	M1T	7	Predicted	1	0.00
GenMed	Pathogenic	G47R	4	Predicted	1	0.01
GenMed	Pathogenic	H71P	2	Failed to predict	0	0.25
GenMed	Pathogenic	R73C	8	Predicted	1	0.00
GenMed	Pathogenic	S77L	4	Failed to predict	0	0.09
GenMed	Pathogenic	H83R	4	Predicted	1	0.00
GenMed	Pathogenic	R94W	9	Predicted	1	0.05
GenMed	Pathogenic	M98R	2	Predicted	1	0.00
GenMed	Pathogenic	I106T	5	Predicted	1	0.02
GenMed	Pathogenic	S117L	4	Predicted	1	0.01
GenMed	Pathogenic	N145I	2	Predicted	1	0.00
GenMed	Pathogenic	L172P	1	Predicted	1	0.02
GenMed	Pathogenic	L182P	2	Predicted	1	0.00

GenMed	Pathogenic	Q183R	1	Predicted	1	0.00
GenMed	Pathogenic	V212A	1	Predicted	1	0.01
GenMed	Pathogenic	R222H	4	Predicted	1	0.00
GenMed	Pathogenic	R222C	12	Predicted	1	0.00
GenMed	Pathogenic	Y225C	1	Predicted	1	0.00
GenMed	Pathogenic	D229H	1	Predicted	1	0.00
GenMed	Pathogenic	R265H	5	Predicted	1	0.00
GenMed	Pathogenic	R265S	3	Predicted	1	0.00
GenMed	Pathogenic	R265C	4	Predicted	1	0.00
GenMed	Pathogenic	G269D	1	Failed to predict	0	0.06
GenMed	Pathogenic	D276H	4	Predicted	1	0.00
GenMed	Pathogenic	T282I	1	Predicted	1	0.00
GenMed	Pathogenic	R296H	7	Predicted	1	0.01
GenMed	Pathogenic	R296P	1	Predicted	1	0.01
GenMed	Pathogenic	R296C	4	Predicted	1	0.00
GenMed	Pathogenic	R320H	36	Predicted	1	0.03
GenMed	Pathogenic	R320G	2	Predicted	1	0.03
GenMed	Pathogenic	R320C	2	Predicted	1	0.01
GenMed	Pathogenic	R331Q	4	Predicted	1	0.00
GenMed	Pathogenic	R331P	1	Predicted	1	0.00
GenMed	Pathogenic	A361V	1	Predicted	1	0.02
GenMed	Pathogenic	G363R	1	Predicted	1	0.00

Table 8. SIFT Prediction of Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.7 SuSPect Result

The results (Table 9) that the SusPect algorithm was only able to accurately predict the clinical significance of 68.6% (24) of missense mutations tested and failed to predict the significance of missense T-protein mutations (31.4%, n = 11).

Source	ClinicalSignificance	Protein Change	Frequency	SuSPect Predict ClinicalSignificance	SuSPect predict effect	SuSPect_Score
GenMed	Pathogenic	M1T	7	Failed to predict	0	12
GenMed	Pathogenic	G47R	4	Predicted	1	67
GenMed	Pathogenic	H71P	2	Predicted	1	75
GenMed	Pathogenic	R73C	8	Predicted	1	71
GenMed	Pathogenic	S77L	4	Predicted	1	78
GenMed	Pathogenic	H83R	4	Predicted	1	76
GenMed	Pathogenic	R94W	9	Predicted	1	79
GenMed	Pathogenic	M98R	2	Predicted	1	76
GenMed	Pathogenic	I106T	5	Predicted	1	73
GenMed	Pathogenic	S117L	4	Predicted	1	75
GenMed	Pathogenic	N145I	2	Predicted	1	74
GenMed	Pathogenic	L172P	1	Predicted	1	51
GenMed	Pathogenic	L182P	2	Predicted	1	81
GenMed	Pathogenic	Q183R	1	Failed to predict	0	46
GenMed	Pathogenic	V212A	1	Failed to predict	0	39
GenMed	Pathogenic	R222H	4	Predicted	1	65
GenMed	Pathogenic	R222C	12	Predicted	1	64
GenMed	Pathogenic	Y225C	1	Predicted	1	53
GenMed	Pathogenic	D229H	1	Failed to predict	0	48
GenMed	Pathogenic	R265H	5	Predicted	1	54

GenMed	Pathogenic	R265S	3	Failed to predict	0	32
GenMed	Pathogenic	R265C	4	Predicted	0	49
GenMed	Pathogenic	G269D	1	Predicted	1	91
GenMed	Pathogenic	D276H	4	Predicted	1	83
GenMed	Pathogenic	T282I	1	Predicted	1	52
GenMed	Pathogenic	R296H	7	Failed to predict	0	31
GenMed	Pathogenic	R296P	1	Failed to predict	0	40
GenMed	Pathogenic	R296C	4	Failed to predict	0	35
GenMed	Pathogenic	R320H	36	Predicted	1	73
GenMed	Pathogenic	R320G	2	Predicted	1	73
GenMed	Pathogenic	R320C	2	Predicted	1	70
GenMed	Pathogenic	R331Q	4	Failed to predict	0	42
GenMed	Pathogenic	R331P	1	Failed to predict	0	18
GenMed	Pathogenic	A361V	1	Predicted	1	90
GenMed	Pathogenic	G363R	1	Predicted	1	96

Table 9. SuSPect Prediction of Clinical Significance Data listed in Coughlin, C. R. et al., 2016.

4.4.8 Consensus Result

The results showed that the six software tools had varying levels of performance in predicting the clinical significance of the missense mutations. LYRUS, PROVEAN, SNAP2 MUTAFRAME, and SIFT showed in Figure 15 a statistically significant prediction of the clinical significance of missense mutations, whereas SuSPect showed low performance.

In conclusion, the results of this study highlight the importance of using multiple software tools when evaluating the potential pathogenicity of missense mutations in proteins. The use of multiple tools may provide a more comprehensive and reliable assessment of the clinical significance of missense mutations, allowing for more accurate diagnosis and management of NKH. Table 10 shows the Consensus Prediction of Clinical Significance for Simple ClinVar, ClinVar, and PubTator Missense Mutations.

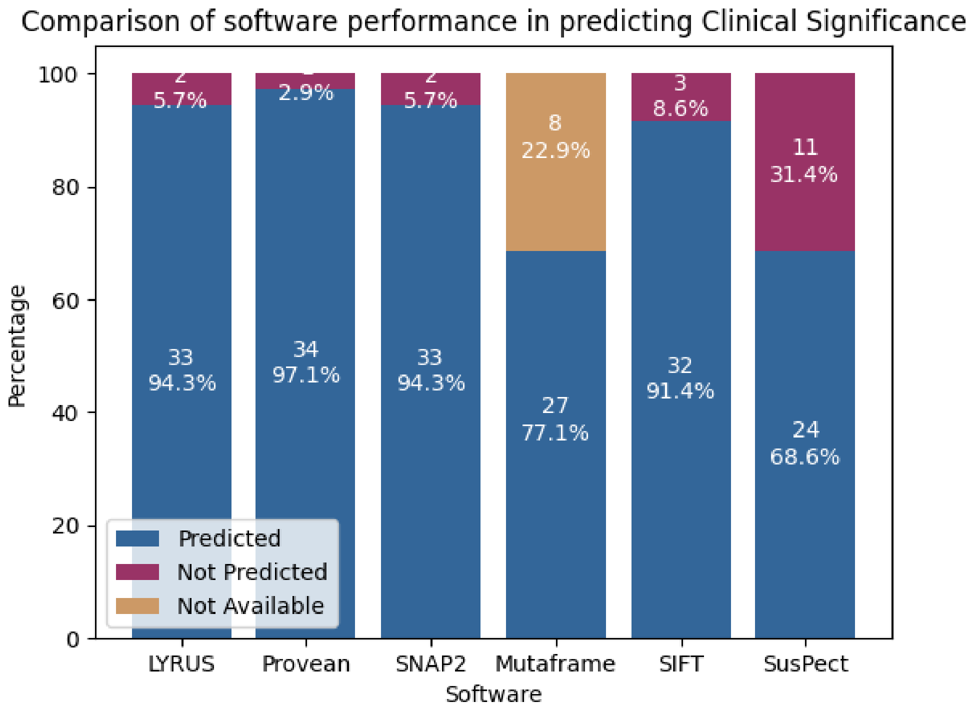


Figure 15. Comparison of software performance in predicting Clinical Significance of missense mutations listed in Coughlin, C. R. et al., 2016.

Source	ClinicalSignificance	Codon	SAV	LYRUS	Provean	SNAP2	Mutaframe	SIFT	Majority Vote
Simple ClinVar	Pathogenic	c.2T>A	M1K	neutral	neutral	neutral	NA	effect	Likely Benign
PubTator	Likely pathogenic		V7L	neutral	neutral	neutral	NA	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.59C>T	P20L	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.70C>T	R24C	neutral	neutral	neutral	effect	effect	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.86C>T	A29V	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.101G>A	R34H	neutral	neutral	effect	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.100C>T	R34C	neutral	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.110C>T	P37L	neutral	effect	neutral	effect	neutral	Likely Benign
Simple ClinVar	Pathogenic	c.125A>G	H42R	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.130G>A	A44T	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Likely pathogenic	c.139G>T	G47W	effect	neutral	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.139G>A	G47R	effect	neutral	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.148G>T	V50L	neutral	neutral	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.148G>C	V50L	neutral	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.152C>T	A51V	neutral	neutral	neutral	effect	effect	Likely Benign
ClinVar	Uncertain significance	c.155T>C	F52S	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.155T>G	F52C	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.158C>T	A53V	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.170T>C	L57P	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.173C>T	P58L	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.188A>G	D63G	neutral	effect	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.194A>G	H65R	neutral	neutral	effect	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.196A>G	T66A	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.203C>T	S68L	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathoge	c.212A>C	H71P	effect	effect	effect	effect	neutral	Likely Pathogenic

	nic								
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.217C>T	R73C	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.218G>A	R73H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.230C>T	S77L	effect	effect	effect	effect	neutral	Likely Pathogenic
ClinVar	Uncertain significance	c.266T>G	I89R	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.269_270del insCC	L90P	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.280C>T	R94W	effect	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.283G>A	V95M	neutral	neutral	effect	effect	effect	Likely Pathogenic
ClinVar/GenMed	Uncertain significance/Pathogenic	c.293T>G	M98R	effect	effect	effect	NA	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.298A>G	S100G	neutral	effect	neutral	effect	neutral	Likely Benign
Simple ClinVar	Likely pathogenic	c.311G>A	G104E	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.317T>C	I106T	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.335A>G	N112S	neutral	effect	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.338A>G	Q113R	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.341G>A	G114E	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar/ClinVar/GenMed	Pathogenic	c.350C>T	S117L	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.359C>T	T120I	neutral	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.362A>G	N121S	neutral	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.371G>A	G124E	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.401A>G	N134S	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.403A>G	T135A	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.425T>G	V142G	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.431C>T	S144F	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.434A>G	N145S	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Likely pathogenic	c.434A>T	N145I	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.436G>A	A146T	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.443G>T	C148F	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Likely benign	c.463C>T	L155F	neutral	neutral	neutral	neutral	neutral	Likely Benign
ClinVar	Uncertain significance	c.467T>C	M156T	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.476A>G	K159R	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.503G>C	R168T	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.515T>C	L172P	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.529A>G	N177D	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.566A>C	Q189P	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.584T>C	V195A	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar/Clinvar	Uncertain/conflicting/Likely benign	c.583G>A	V195M	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.589G>C	D197H	neutral	neutral	neutral	effect	effect	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.589G>A	D197N	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.593A>T	D198V	effect	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.613A>G	M205V	effect	effect	neutral	effect	neutral	Likely Pathogenic
ClinVar	Uncertain significance	c.614T>G	M205R	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.625G>A	V209M	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar/Clinvar	Uncertain/conflicting/ Benign/Likely benign	c.631G>A	E211K	neutral	neutral	effect	effect	neutral	Likely Benign
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.635T>C	V212A	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.638T>G	F213C	neutral	effect	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.643G>A	V215M	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.643G>C	V215L	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.656G>A	R219H	neutral	effect	neutral	effect	neutral	Likely Benign

ClinVar	Uncertain significance	c.658G>A	V220M	neutral	neutral	neutral	effect	effect	Likely Benign
ClinVar	Likely pathogenic	c.664C>A	R222S	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.665G>A	R222H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Likely pathogenic	c.665G>T	R222L	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.664C>T	R222C	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Likely pathogenic	c.674A>G	Y225C	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.677C>T	T226I	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.685G>C	D229H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.701C>T	S234L	effect	effect	neutral	effect	neutral	Likely Pathogenic
Simple ClinVar/Clinvar	Uncertain/conflicting/Likely benign	c.713C>T	A238V	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.719C>T	A240V	neutral	neutral	neutral	effect	effect	Likely Benign
ClinVar	Uncertain significance	c.721G>A	V241I	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.727C>A	L243M	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.752C>G	P251R	neutral	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed/PubTator	Pathogenic	c.794G>A	R265H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.793C>T	R265C	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Likely pathogenic	c.797T>C	L266P	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.805G>A	G269S	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.806G>A	G269D	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.825T>A	N275K	neutral	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.826G>C	D276H	effect	effect	effect	effect	effect	Likely Pathogenic
PubTator	Pathogenic	c.955G>C	D276H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.830T>C	I277T	neutral	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.830T>A	I277N	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Pathogenic	c.847C>T	P283S	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.850G>A	V284M	effect	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.856G>A	G286S	neutral	neutral	effect	neutral	effect	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.857G>A	G286D	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.883C>T	R295C	effect	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.887G>A	R296H	neutral	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Likely pathogenic	c.887G>T	R296L	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.886C>T	R296C	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.890G>A	R297Q	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.892G>A	A298T	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Benign/Likely benign	c.898A>G	M300V	neutral	neutral	neutral	neutral	neutral	Likely Benign
ClinVar	Uncertain significance	c.911G>A	G304E	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.944G>A	R315K	neutral	neutral	neutral	neutral	neutral	Likely Benign
ClinVar	Uncertain significance	c.950A>G	Q317R	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.953G>A	R318Q	neutral	neutral	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.952C>T	R318W	neutral	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Pathogenic	c.959G>A	R320H	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.959G>T	R320L	effect	effect	effect	effect	neutral	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.958C>T	R320C	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.962T>C	V321A	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.961G>A	V321M	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.972G>A	M324I	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.982G>A	A328T	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar/GenMed	Pathgenic/Likely pathogenic	c.992G>A	R331Q	effect	effect	effect	effect	effect	Likely Pathogenic

ClinVar	Uncertain significance	c.1001G>T	S334I	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.1021G>A	G341S	neutral	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1030A>T	I344F	neutral	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1058C>A	S353Y	effect	effect	effect	effect	effect	Likely Pathogenic
ClinVar	Uncertain significance	c.1076A>C	N359T	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.1082C>T	A361V	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1087G>C	G363R	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1097C>T	P366L	neutral	effect	neutral	effect	neutral	Likely Benign
ClinVar	Uncertain significance	c.1111C>T	R371C	neutral	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1112G>A	R371H	neutral	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1136A>C	E379A	neutral	effect	neutral	effect	neutral	Likely Benign
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.1138G>A	V380M	effect	neutral	effect	NA	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1141C>T	R381W	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1145G>A	R382Q	neutral	neutral	neutral	effect	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.1157T>C	M386T	neutral	neutral	neutral	neutral	neutral	Likely Benign
Simple ClinVar	Uncertain/conflicting	c.1190C>A	T397K	effect	effect	neutral	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1199A>G	Y400C	effect	effect	effect	effect	effect	Likely Pathogenic
Simple ClinVar	Uncertain/conflicting	c.1202C>A	T401N	neutral	neutral	neutral	neutral	neutral	Likely Benign

Table 10. Consensus Prediction of Clinical Significance of the LYRUS, PROVEAN, SNAP2 MUTAFRAME, and SIFT softwares for Simple ClinVar, ClinVar and Pubtator Missense Mutations.

4.4.9 Evaluation of Software Tools for Predicting Indels and Duplication Mutations Clinical Significance

Insertions and deletions (indels) are additions or deletions of one or more nucleotides in a DNA sequence. Indels are highly abundant in human genomes, second only to SNP, and comprise 15–21% of human polymorphisms[186].

Several software tools designed to predict and analyze these types of mutations, such as duplications, and indels, are available. This study included the PROVEAN and SIFT.

4.4.10 PROVEAN Result

The PROVEAN algorithm was applied to a dataset of 15 indels and duplication mutations to predict their clinical significance. Based on the cut-off of -2.5, for the PROVEAN score specific to the NCBI nr protein database released in August 2011, we evaluated the provided variation scores to identify potentially deleterious (damaging) and neutral (tolerated) variants. Variations with PROVEAN scores lower than or equal to -2.5 can be considered potentially deleterious, while variations with scores higher than -2.5 can be considered potentially neutral. The results (Table 11) showed that the algorithm predicted clinical significance likely to be pathogenic, suggesting that they may have a significant impact on protein function. These included T87_Q113del (score of -100.1), G114_Q157del (score of -173.2), and Y369_S370delinsTer (score of -13.9).

Other variations had higher scores, suggesting that they may be less likely to affect protein function. These included P20dup (score of -1.9), and L402del (score of -2.0).

Source	Type	Clinical Significance	Protein Change	Frequency	Provean Predict Clinical Significance	Provean Predict effect	Provean Score
Simple ClinVar	Duplication	Uncertain/conflicting	P20dup	1	Likely Benign	0	-1.9
Simple ClinVar	Deletion	Pathogenic	G54_W55insTer	1	Likely Pathogenic	1	-11.1
GenMed	Deletion	Pathogenic	T87_Q113del	1	Likely Pathogenic	1	-100.1
ClinVar	Deletion	Uncertain significance	G91del	1	Likely Pathogenic	1	-14.9
GenMed	Deletion	Pathogenic	G114_Q157del	1	Likely Pathogenic	1	-173.2
GenMed	Deletion	Pathogenic	V132_N134delinsD	1	Likely Pathogenic	1	-15.0
GenMed	Deletion	Pathogenic	K151_L155del	6	Likely Pathogenic	1	-22.8
GenMed	Indel	Pathogenic	G184E	2	Likely Pathogenic	1	-7.6
Simple ClinVar	Duplication	Uncertain/conflicting	P236_V237dup	1	Likely Pathogenic	1	-9.2
GenMed	Indel	Pathogenic	L270_C271delinsRG	1	Likely Pathogenic	1	-17.5
Simple ClinVar	Deletion	Likely Pathogenic	M324del	1	Likely Pathogenic	1	-7.8
ClinVar	Deletion	Uncertain significance	K358del	1	Likely Pathogenic	1	-10.2
Simple ClinVar	Deletion	Pathogenic	Y369_S370delinsTer	1	Likely Pathogenic	1	-13.9
Simple ClinVar	Microsatellite	Uncertain/conflicting	Q385del	1	Likely Pathogenic	1	-6.9
ClinVar	Deletion	Uncertain significance	L402del	1	Likely Benign	0	-2.0

Table 11. Provean Prediction of Clinical Significance of Indels and Duplication mutations of T-protein.

4.4.11 SIFT Result

This dataset of amino acids insertions, deletions and duplications was further analyzed using Mutalyzer (https://mutalyzer.nl/)[41,187]. to normalize the variant description format to use it as input of SIFT web server.

The results showed that the algorithm predicted clinical significance likely to be pathogenic, suggesting that they cause damage to protein function. These included the 10 SNPs described in Table12.

Other variations were neutral, suggesting that they were less likely to affect protein function. These included L402del and G184E. Other variations in SIFT, such as P20dup, T87_Q113del, and G114_Q157del, were not able to predict any effect on the protein.

Source	Type	Clinical Significance	Protein Change	Frequency	SIFT Predict Clinical Significance	SIFT Predict effect	SIFT effect
Simple ClinVar	Duplication	Uncertain/conflicting	P20dup	1			NA
Simple ClinVar	Deletion	Pathogenic	G54_W55insTer	1	Likely Pathogenic	1	damaging
GenMed	Deletion	Pathogenic	T87_Q113del	1			NA
ClinVar	Deletion	Uncertain significance	G91del	1	Likely Pathogenic	1	damaging
GenMed	Deletion	Pathogenic	G114_Q157del	1			NA
GenMed	Deletion	Pathogenic	V132_N134delinsD	1	Likely Pathogenic	1	damaging
GenMed	Deletion	Pathogenic	K151_L155del	6	Likely Pathogenic	1	damaging
GenMed	Indel	Pathogenic	G184E	2	Likely Benign	0	neutral
Simple ClinVar	Duplication	Uncertain/conflicting	P236_V237dup	1	Likely Pathogenic	1	damaging
GenMed	Indel	Pathogenic	L270_C271delinsRG	1	Likely Pathogenic	1	damaging

Simple ClinVar	Deletion	Likely Pathogenic	M324del	1	Likely Pathogenic	1	damaging
ClinVar	Deletion	Uncertain significance	K358del	1	Likely Pathogenic	1	damaging
Simple ClinVar	Deletion	Pathogenic	Y369_S370delinsTer	1	Likely Pathogenic	1	damaging
Simple ClinVar	Microsatellite	Uncertain/conflicting	Q385del	1	Likely Pathogenic	1	damaging
ClinVar	Deletion	Uncertain significance	L402del	1	Likely Benign	0	neutral

Table 12. SIFT Prediction of Clinical Significance of Indels and Duplication mutations of T-protein.

4.4.12 Frameshift and Stop Gain Mutations Clinical Significance

Frameshift and stop-gain mutations are two types of mutations that can significantly impact protein function. Frameshift mutations occur when one or more nucleotides are added or deleted from a DNA sequence, causing a shift in the reading frame of the genetic code. This can result in the incorporation of incorrect amino acids and the premature termination of protein synthesis. Stop-gain mutations, also known as nonsense mutations, occur when a point mutation, can be a single nucleotide variation (SVN) creates a premature stop codon in the DNA sequence, resulting in truncated and often non-functional proteins. The ability to predict the clinical significance of these mutations is crucial for proper diagnosis and treatment of genetic disorders.

Several software tools are available to predict the clinical significance of frameshift and stop-gain mutations. These tools use various algorithms and data sources to predict the functional impact of mutations.

4.4.13 SIFT Result

This dataset of amino acids frameshift and stop gain was further analyzed using Mutalyzer (https://mutalyzer.nl/)[187] to normalize the variant description format and to use it as the input of the SIFT web server. For example, Table 13 shows that the Y273Ter and S77Ter mutations are predicted to be damaging because they create a premature stop codon that prematurely truncates the protein. In contrast, the mutations D229Gfs*10, and S6Wfs*89 are predicted to be neutral, indicating that they are unlikely to have a significant effect on protein function. The A328Gfs*22, Q2Ter, and S6KfsTer22 mutations were not predicted.

Source	Type	Consequence	Condon	Protein Change	Frequency	SIFT Predict Clinical Significance	SIFT Predict effect	SIFT effect
ClinVar	Indel	Frameshift	c.381_383 delinsGG	D128fs	1	Likely pathogenic	1	damaging
ClinVar	Deletion	Frameshift	c.383del	D128fs	1	Likely pathogenic	1	damaging
ClinVar	SVN	Stop Gain	c.819T>A	Y273Ter	1	Pathogenic	1	damaging

ClinVar	SVN	Frameshift	c.230C>A	S77Ter	1	Likely pathogenic	1	damaging
ClinVar	Insertion	Frameshift	c.270_271insCACCC	G91fs	1	Pathogenic	1	damaging
PubTator	SNV	Frameshift		A328Gfs*22	1	Likely pathogenic		NA
GenMed	SNV	Stop gain	c.1101C>A	C367Ter	1	Likely pathogenic	1	damaging
GenMed	SNV	Stop gain	c.4C>T	Q2Ter	1	Likely pathogenic		NA
ClinVar	Insertion	Frameshift	c.847_848insA	S283fs	1	Likely pathogenic	1	damaging
ClinVar	SVN	Stop Gain	c.889C>T	R297Ter	1	Pathogenic	1	damaging
GenMed	Deletion	Stop gain	c.395_400del6	V132_N134delinsD	1	Pathogenic	1	damaging
ClinVar	Deletion	Frameshift	c.221del	Q74R	1	Pathogenic	1	damaging
PubTator	Deletion	Frameshift	c.977delA	E326Gfs*12	1	Likely pathogenic	1	damaging
ClinVar	Deletion	Frameshift	c.224del	H75R	1	Pathogenic	1	damaging
Simple ClinVar/GenMed	SNV	Stop gain	c.496C>T	Q166Ter	3	Likely pathogenic	1	damaging
Simple ClinVar/GenMed	SNV	Stop gain	c.574C>T	Q192Ter	1	Likely pathogenic	1	damaging
Simple ClinVar	SNV	Stop gain	c.870G>A	W290Ter	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Stop gain	c.1107_1108del	Y369_S370delinsTer	1	Pathogenic	1	damaging
Simple ClinVar	SNV	Stop gain	c.164G>A	W55Ter	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Stop gain	c.165del	G54_W55insTer	1	Likely pathogenic	1	damaging
Simple ClinVar	SNV	Stop gain	c.1153C>T	Q385Ter	1	Uncertain/conflicting	1	damaging
Simple ClinVar	SNV	Stop gain	c.178C>T	Q60Ter	1	Pathogenic	1	damaging
Simple ClinVar	SNV	Stop gain	c.256C>T	Q86Ter	1	Pathogenic	1	damaging
PubTator	Insertion	Frameshift	c.14_15insT	S6KfsTer22	1	Likely pathogenic		NA
GenMed	Deletion	In frame del	c.452_466del15	K151_L155del	6	Likely pathogenic	1	damaging
GenMed	Deletion	In frame del	c.1107_1108delAC	Y369Ter	4	Likely pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.695_696+33del35insGGCTGTACAGA	D229Gfs*10	1	Likely pathogenic	0	neutral
GenMed	Deletion	Frameshift	c.178_181delCAGT	Q60Tfs*35	2	Pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.999_1000delCA	H333Qfs*16	2	Likely pathogenic	1	damaging
GenMed	Duplication	Frameshift	c.998dup	H333Qfs*17	1	Likely pathogenic	1	damaging
GenMed	Insertion	Frameshift	c.996dup	H333Tfs*17	1	Likely pathogenic	1	damaging
GenMed	Insertion	Frameshift	c.1034+2T>C	IVS8+2insT	1	Pathogenic	1	damaging
GenMed	Insertion	Frameshift	c.534_535insCC	L179Pfs*3	2	Likely pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.451_466del16	K151Cfs*25	2	Likely pathogenic	1	damaging

GenMed	Deletion	Frameshift	c.452_465del14	K151Nfs*22	2	Pathogenic	1	damaging
GenMed	SNV	Frameshift	c.878-1G>A	K294L	1	Likely pathogenic		NA
GenMed	Deletion	Frameshift	c.987delC	M330Cfs*8	3	Likely pathogenic	1	damaging
GenMed	Duplication	Frameshift	c.14dupT	S6Kfs*22	2	Likely pathogenic		NA
Simple ClinVar	Deletion	In frame indel	c.452_466del	K151_L155del	1	Pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.1063delT	S355Lfs*2	1	Likely pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.16delA	S6Vfs*90	2	Likely pathogenic		NA
Simple ClinVar	Duplication	Frameshift	c.657dup	V220fs	1	Pathogenic	1	damaging
Simple ClinVar	Duplication	Frameshift	c.849dup	V284fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.478del	V160fs	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.734_735del	T245fs	1	Pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.13_16delGTAA	S6Wfs*89	1	Likely pathogenic	0	neutral
Simple ClinVar	Indel	Frameshift	c.1199_1202delinsTAT	P400fs	1	Uncertain/conflicting	1	damaging
Simple ClinVar	Duplication	Frameshift	c.14dup	S6fs	1	Pathogenic		NA
Simple ClinVar	Duplication	Frameshift	c.257dup	T87fs	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.348_349del	S117fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.1056del	S353fs	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.16del	S6fs	1	Pathogenic		NA
Simple ClinVar	Deletion	Frameshift	c.15_18del	S6fs	1	Pathogenic/Likely pathogenic	0	neutral
GenMed	Deletion	Frameshift	c.15_18delAAGT	S6Wfs*89	6	Likely pathogenic	0	neutral
GenMed	Duplication	Frameshift	c.731dupC	T245Nfs*32	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.908del	P303fs	1	Pathogenic	1	damaging
Simple ClinVar	Duplication	Frameshift	c.609dup	F204fs	1	Pathogenic	1	damaging
GenMed	Deletion	Frameshift	c.1040_1041delTG	V347Dfs*2	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.59del	P20fs	1	Likely pathogenic		NA
Simple ClinVar	Deletion	Frameshift	c.987del	M330fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.1209del	K403fs	1	Uncertain/conflicting	0	neutral
Simple ClinVar	Deletion	Frameshift	c.144_148del	K48fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.535del	L179fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.602_603del	K201fs	1	Pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.63del	L22fs	1	Likely pathogenic		NA
Simple	Deletion	Frameshift	c.875del	L292fs	1	Likely	1	damaging

ClinVar						pathogenic		
Simple ClinVar	Deletion	Frameshift	c.168_171 del	L57fs	1	Pathogenic	0	neutral
Simple ClinVar	Duplication	Frameshift	c.996dup	H333fs	1	Pathogenic/Likely pathogenic	1	damaging
Simple ClinVar	Duplication	Frameshift	c.982dup	A328fs	1	Pathogenic/Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.982del	A328fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Indel	Frameshift	c.982_983 delinsT	A328fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Deletion	Frameshift	c.148del	V50fs	1	Likely pathogenic	1	damaging
Simple ClinVar	Microsatellite	Frameshift	c.20_21del	V7fs	1	Pathogenic		NA
Simple ClinVar	Deletion	Frameshift	c.61del	A21fs	1	Likely pathogenic		NA

Table 13. SIFT Prediction of Clinical Significance of Frameshift and Stop Gain mutations of T-protein.

4.5 InterPro Results

The input fasta[188] file for UniProt ID P48728[189–192] (GCST_Human Aminomethyltransferase - https://www.uniprot.org/uniprotkb/P48728/entry) was analysed using InterPro[193] webserver. T-protein has a total length of 403 amino acids, and the InterPro software predicted (Figure 14) the following main regions:

- The N-terminal region of a signal peptide (SIGNAL_PEPTIDE_N_REGION) consists of 1-3 amino acids.
- The hydrophobic region of the signal peptide (SIGNAL_PEPTIDE_H_REGION) consists of 4–15 amino acids.
- C-terminal region of a signal peptide (SIGNAL_PEPTIDE_C_REGION) consisting of 16-23 amino acids.
- The signal peptide region (PHOBIUS SIGNAL_PEPTIDE) consists of 1–23 amino acids.
- COILS Coil, consisting of 146–166 amino acids.

InterPro software identified the following two domains:

- Aminomethyltransferase folate-binding domain (GCV_T_N) 38–291 amino acids in sequence.
- Glycine cleavage T-protein C-terminal barrel domain (GCV_T_C) 318-395 amino acids in the sequence.

These domains are important for protein structure, function, and cellular localization within the cell.

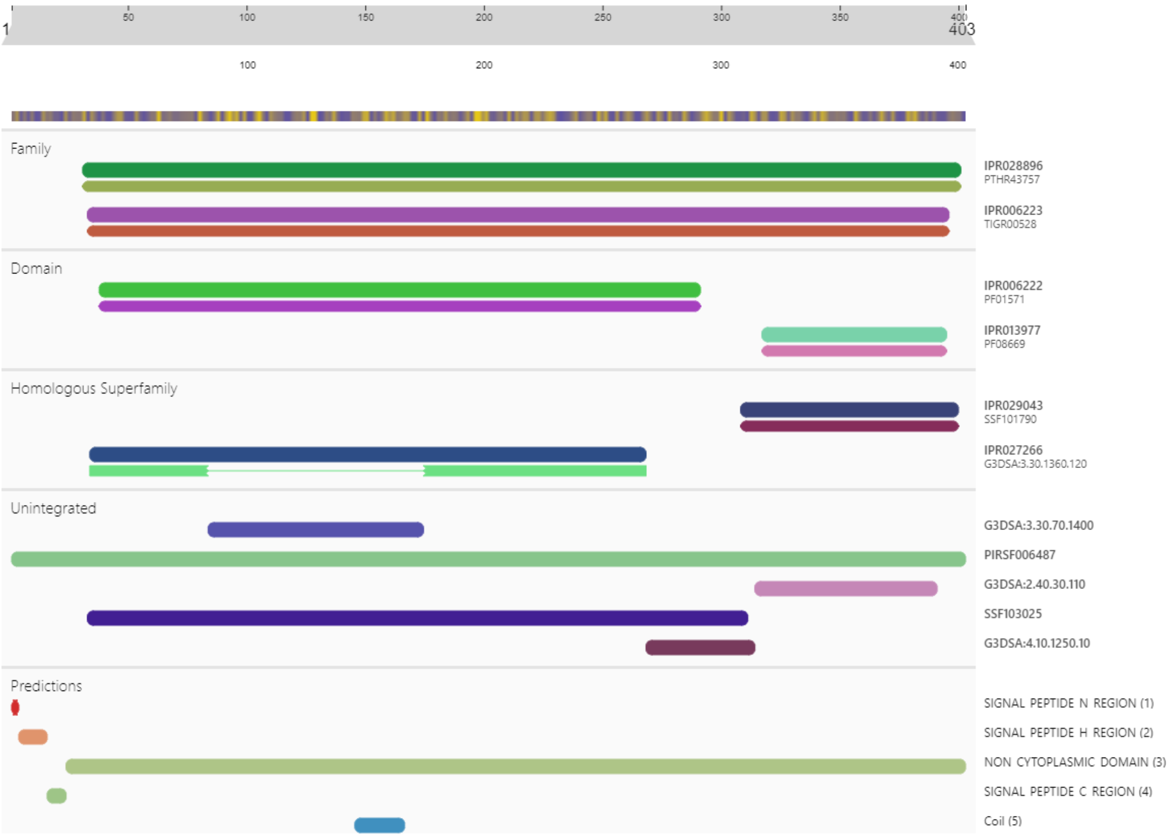


Figure 16. InterPro results of domains, and functional sites of T-protein.

Aminomethyltransferase-like (IPR028896)[194] - Short name: GCST/YgfZ/DmdA

Dimethylsulfoniopropionate (DMS) is catabolized in marine bacterioplankton through a pathway in which the initial step involves demethylation to methylmercaptopropionate (MMPA), which is then further catabolized to methane thiol and acetate. The enzyme responsible for the first step is dimethylsulfoniopropionate demethylase, DmdA[190,191]. The overall fold of DmdA is not similar to that of other enzymes that typically utilize tetrahydrofolate (THF). Instead, DmdA has a triple-domain structure similar to that observed for glycine cleavage T protein[195]. The glycine cleavage T protein is an aminomethyltransferase 2.1.2.10, which is part of the glycine cleavage complex responsible for the reversible oxidation of glycine[196] . This entry also includes YgfZ, a folate-binding protein involved in regulating ATP-dnaA levels and in the modification of some tRNAs. It is likely a key factor in regulatory networks that act via tRNA modifications, such as the initiation of chromosomal replication[197].

Glycine cleavage system T protein (IPRO06223)[198] - The glycine cleavage system (GCS)

Short name: GCS_T

This multi-enzyme system is composed of proteins P, H, T, and L, which catalyze the reversible oxidation of glycine. T protein is an aminomethyl transferase 2.1.2.10[199] that catalyzes the following reaction: (6S) tetrahydrofolate + S-aminomethyldihydrolipoprotein = (6R)-5,10-methylenetetrahydrofolate + NH3 + dihydrolipoylprotein. The glycine cleavage system is found in bacteria and mitochondria of eukaryotes. Mutations in the human T-protein gene are known to cause nonketotic hyperglycinemia[199].

Aminomethyltransferase, folate-binding domain (IPR006222)[200]-Short name: GCV_T_N

This domain is found at the N-terminus of glycine cleavage T-proteins, which are part of the glycine cleavage multi-enzyme complex (GCV) found in bacterial and eukaryotic mitochondria. GCV catalyses the catabolism of glycine in eukaryotes. T-protein (aminomethyltransferase, 2.1.2.10[192])) is a folate-dependent enzyme that catalyzes the release of ammonia and the transfer of the methylene carbon unit (C1 unit) to tetrahydrofolate (H4folate) from the aminomethyl intermediate attached to the lipoate cofactor of H-protein[9,201]. This domain is also found in YgfZ proteins. YgfZ in E. coli is a folate-binding protein involved in RNA modification and the regulation of chromosomal replication initiation[197]. YgfZ is not an aminomethyltransferase but is likely a folate-dependent regulatory protein[196]. This domain represents the folate-binding domain.

Glycine cleavage T-protein, C-terminal barrel domain (IPR013977)[202]-Short name: GCV_T_C

This entry represents the C-terminal β-barrel domain of glycine cleavage T-proteins, which is a part of the glycine cleavage multi-enzyme complex (GCV) found in bacteria and eukaryotes [203,204]. GCV catalyses the catabolism of glycine in eukaryotes. T-protein is an aminomethyl transferase.

Glycine cleavage T-protein/YgfZ, C-terminal (IPR029043)[122]-Short name: GcvT/YgfZ_C

This superfamily contains a β-barrel domain found at the C-terminus of the glycine cleavage T-protein (aminomethyltransferase)[77] and tRNA-modifying protein YgfZ. YgfZ binds folate and is likely a folate-dependent regulatory protein[197,205].

GTP-binding protein TrmE/Aminomethyltransferase GcvT, domain 1 (IPR027266)[206]-

Short name: TrmE/GcvT_dom1

This entry represents an alpha/beta domain found in the GTP-binding protein TrmE (N-terminal domain) and domain 1 of the glycine cleavage T protein (also known as aminomethyltransferase)[203]. TrmE is a guanine nucleotide-binding protein conserved between bacteria and eukaryotes. It is involved in the modification of uridine bases at the first anticodon (wobble) of tRNAs. The N-terminal portion of the protein is necessary for mediating dimer formation within the protein[205]. Glycine cleavage T-protein is part of the glycine cleavage multi-enzyme complex (GCV) found in the bacteria and mitochondria of eukaryotes[207]. GCV catalyses the oxidative decarboxylation of glycine. The T-protein is an aminomethyl transferase 2.1.2.10[192]. The N-terminal region (residues 14-35) of domain 1 plays a crucial role in H-protein interaction[79].

To further investigate the potential impact of NKH-causing mutations, a consensus approach was employed using five software tools with proven performance (LYRUS, PROVEAN, SNAP2, MUTAFRAME, and SIFT) described previously in the Pathogenicity Prediction section. The mutations predicted to be likely pathogenic were selected from three sources (Simple ClinVar, ClinVar, and PubTator), and were supplemented with the mutations reported in a previous study by Coughlin et al. (2016). The resulting list of mutations was then distributed per domain, and the functional sites were predicted by InterPro. To understand the association of NKH-causing mutations listed on Appendix 10 (InterPro results– Distribution of mutations NKH-causing per Domains and Functional Sites in the T-protein) with its associated functional sites predicted by InterPro and to contextualize these mutations within a structurally related family of proteins, and domains lollipop plot (Figure 17) was generated using react-mutation-plot (GitHub - thehyve/react-mutation-plot: A light weight adaptation React based mutation lollipop plot from cBioPortal frontend.).

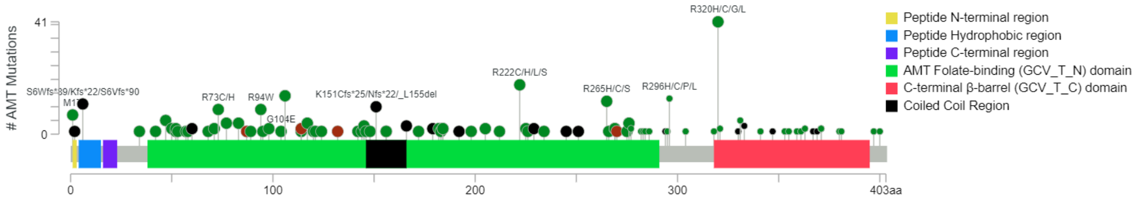


Figure 17. NKH-causing mutations distribution per InterPro results of domains, and functional sites of T-Protein.

4.6 Survey of NKH-causing missense mutations in T-protein

After classifying the various variants of AMT using different prediction software to identify SAV, indels, and duplication mutations predicted to cause NKH disease across three different data sources (ClinVar, SimpleClinVar, and PubTator), and including the mutations listed in the article by (Coughlin, C. R. et al., 2016)[6] . Table 14 shows the top ten most frequently observed NKH-causing mutations in AMT.

Frequency	Ref AA	Alt AA	Pos AA
36	Arg	His	320
12	Arg	Cys	222
9	Arg	Trp	94
8	Arg	Cys	73
7	Met	Thr	1
7	Arg	His	296
6	Lys	_Leu155del	151
6	Ser	Trpfs*89	6
5	Ile	Thr	106
5	Arg	His	265

Table 14. Top ten most frequently observed mutations in AMT.

A Python Matplotlib plot (Figure 18) was created to show the distribution of all known NKH-causing predicted missense, insert, delete, and duplication mutations across the entire length of the T-protein.

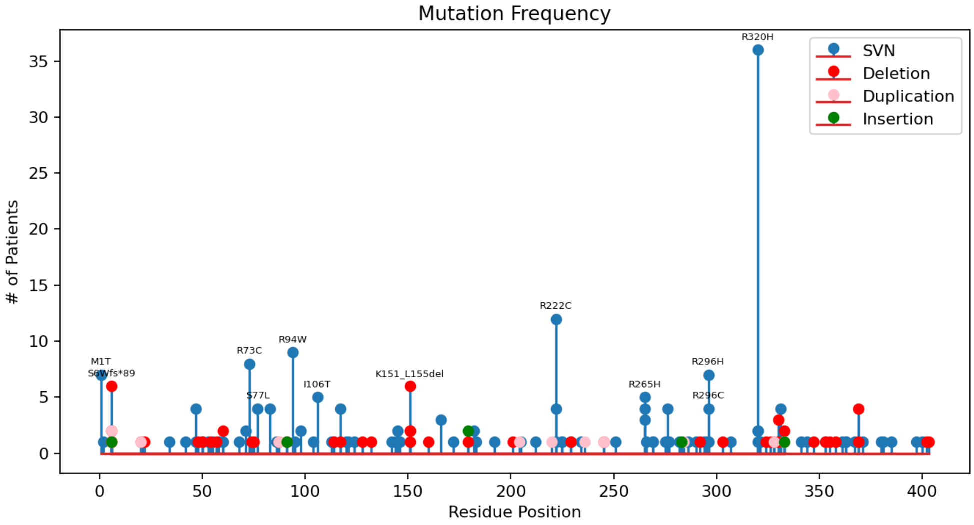


Figure 18. NKH-causing Mutation's frequency across the T-protein length.

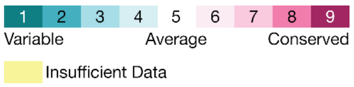
Based on the data, the most frequent types of damaging mutations were SNVs and deletions, with a total of 77 and 28 occurrences, respectively. Other mutations include duplications, insertions, microsatellites, and indels. The details of the mutation data are provided in Appendix 11. In terms of specific mutation sites, several mutations occur more frequently than others, including the following:

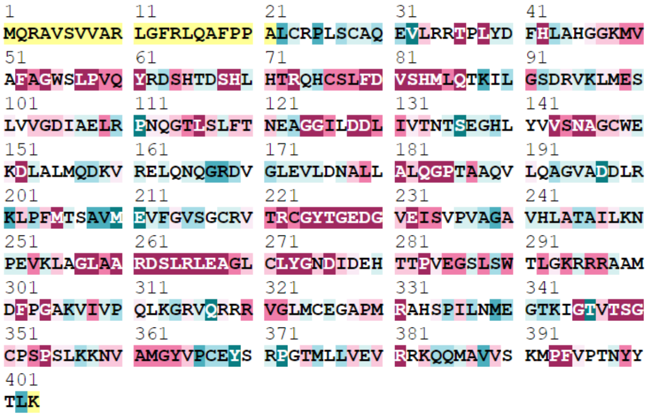
- Position 320: This site has 36 SNVs and 4 SNVs in separate occurrences.
- Position 222: This site has 12 SNVs, 4 SNVs, and 4 SNVs in separate occurrences, making it the second most frequently mutated site in the dataset.
- Position 94: This site has 9 SNVs.
- Position 1: This site has 7 SNVs.
- Position 6: This site has 6 deletions.
- Position 77: This site has 4 SNVs and 1 SNV in separate occurrences.
- Position 369: This site has 4 deletions.

4.7 ConSurf Result

To study the role of amino acid residues in the evolution of T-proteins, the ConSurf web server was used. The input to ConSurf[146]was the best model model1.pdb of the T-protein was obtained from the performance of the MODELLER software. The results of this analysis show the importance of the conservation of amino acid residues. The model1.pdb model was submitted to the ConSurf web server, which calculated the evolutionary conservation of each residue in the protein. ConSurf algorithm uses multiple sequence alignments of homologous proteins to determine the conservation of each residue. The results of ConSurf analysis showed a scale of conservation importance per amino acid residue in the T-protein. Some residues were highly conserved, indicating that they play critical roles in protein structure and function. Other residues were less conserved, suggesting that they may have undergone more evolutionary changes and may not be as important for protein function.

4.7.1 Missense mutations distribution over conserved residues





```
1          11         21         31         41
MQRAVSVVAR LGFRLQAFPP ALCRPLSCAQ EVLRRTPLYD FHLAHGGKMV
51         61         71         81         91
AFAGWSLPVQ YRDSHTDSHL HTRQHCSLFD VSHMLQTKIL GSDRVKLMES
101        111        121        131        141
LVVGDIAELR PNQGTLSLFT NEAGGILDDL IVTNTSEGHL YVVSNAGCWE
151        161        171        181        191
KDLALMQDKV RELQNQGRDV GLEVLDNALL ALQGPTAAQV LQAGVADDLR
201        211        221        231        241
KLPFMTSAVM EVFGVSGCRV TRCGYTGEDG VEISVPVAGA VHLATAILKN
251        261        271        281        291
PEVKLAGLAA RDSLRLEAGL CLYGNDIDEH TTPVEGSLSW TLGKRRRAAM
301        311        321        331        341
DFPGAKVIVP QLKGRVQRRR VGLMCEGAPM RAHSPILNME GTKIGTVTSG
351        361        371        381        391
CPSPSLKKNV AMGYVPCEYS RPGTMLLVEV RRKQQMAVVS KMPFVPTNYY
401
TLK
```

Figure 19. ConSurf color scale of T-Protein. The colour scale representing the conservation scores (9 - conserved, 1 - variable).

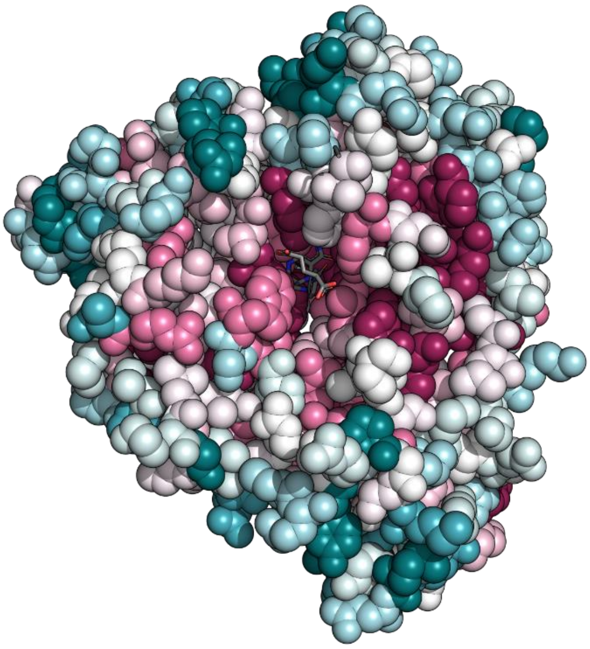


Figure 20. Conservation scores for chain B. PyMOL-generated image.

The ConSurf scale represents the degree of conservation of the amino acid residues in a protein. In Figure 19 a score of 9 indicates that the residue is highly conserved across evolutionary time, whereas a score of 1 indicates that the residue is highly variable. Figure 20 shows the conservation scores for Chain B generated by PyMOL. The output of ConSurf (Figure 21) showed that 37 residues (37.4%) had a conservation score of 9, 22 residues (22.2%) had a score of 8, 16 residues (16.2%) had a score of 7, five residues (5.1%) had a score of 6, 7 residues (7.1%) had a score of 5, and 10 residues (10.1%) had a score of 4. The "Insufficient Data" category suggests that the conservation score of that residue was not able to be calculated.

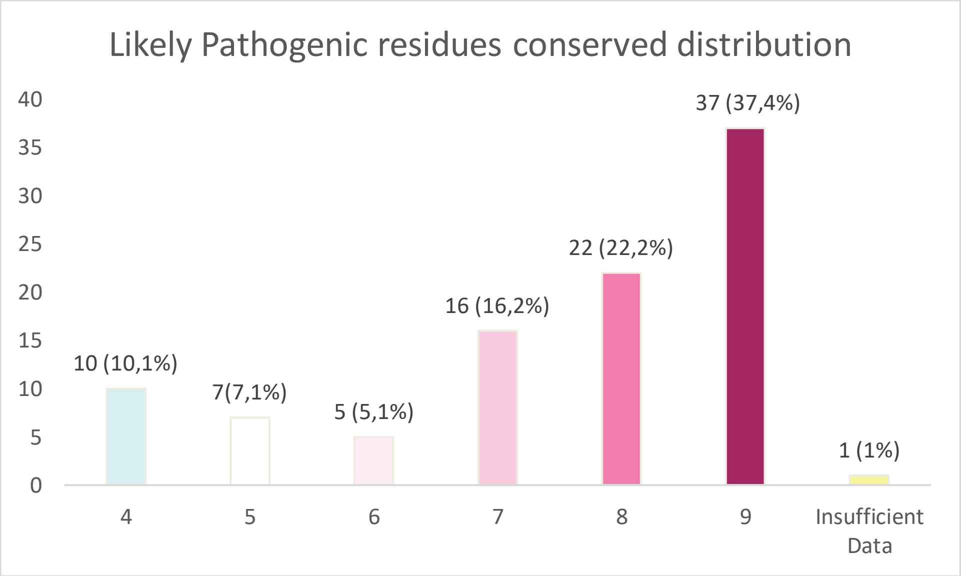


Figure 21. Likely Pathogenic mutations residues conserved distribution.

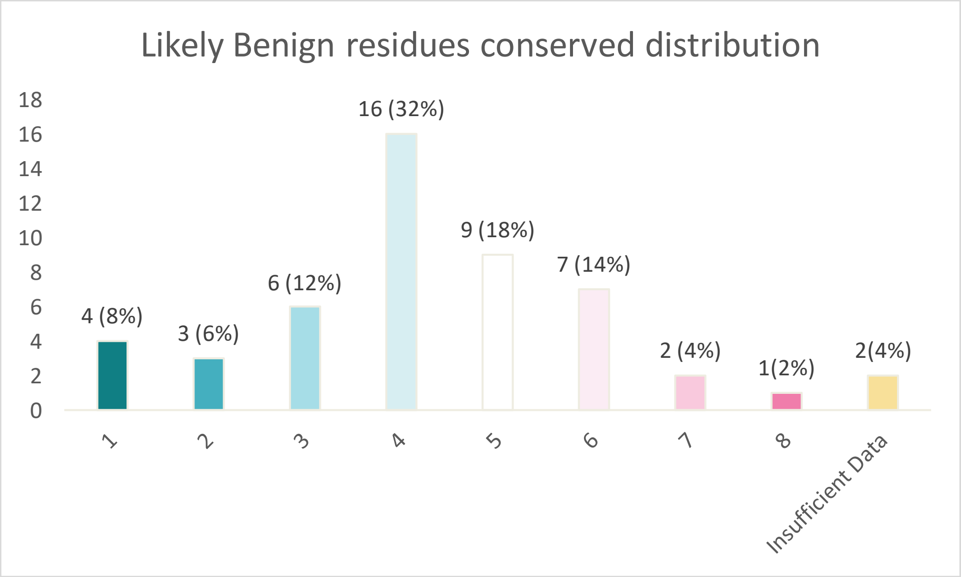


Figure 22. Likely Benign mutations residues conserved distribution.

The Consurf results indicated that most residues had a conservation score of 4 or lower, with some having scores of 5, 6, or 7. This suggests that these residues may be less conserved and may potentially tolerate changes better. Therefore, benign missense mutations (Figure 22) in these residues may be less likely to result in negative effects on protein function than pathogenic mutations in highly conserved residues (scores of 8 or 9). However, it is important to note that the functional impact of a mutation depends on many factors and should be evaluated on a case-by-case basis. The use of the ConSurf web server provided valuable insights into the evolutionary importance of amino acid residues in T-proteins. The results of this analysis can be used to guide future studies on T-proteins, including structural and functional analyses, as well as drug design. Conservation information can also be used to predict potential sites for mutagenesis studies, helping to further our understanding of the role of individual residues in T-proteins.

4.8 Molecular Dynamics Simulations

MD simulations were performed to understand the structure-function relationship of the T-protein, for which we selected five key mutations: R320H, R73C, H83R, R222C, and E211K. These mutations present a spectrum of disease severity, with R320H manifesting as severe, R222C as mild, and H83R as moderate. Our survey identified R320H, R73C, and R222C as the four most frequently observed NKH-causing mutations. The ML predictions from our study corroborate the benign nature of E211K[9,208].

For each system (T-protein WT, one homozygote (H83R and H83R) and two heterozygotes (R73C and R320H) and (R222C and E211K)) three replicas of MD simulations were performed with a protein-run distance of 500ns and a periodic image of 1.2 nm, see Appendix 12.

To delineate the molecular consequences of these mutations, we employed a suite of analytical techniques, namely the Density Distribution of RMSD, to determine the structural deviations of the mutated proteins relative to the WT T-protein. H-bond analysis was performed to determine the role of pivotal hydrogen bonds in protein conformation and stability. RMSF analysis highlighted regions with increased flexibility and potential functional deviations. MMPBSA Analysis to provide insights into potential alterations in protein stability and binding affinity.

4.8.1 Density Distribution of RMSD Analysis

RMSD provides valuable insights into the structural stability and conformational changes in proteins. By analyzing the RMSD values of regions proximate (within 5 Å) to the ligand THH, distinctions emerged between the homozygote WT T-protein and homozygote H83R, heterozygote R73C and R320H, and heterozygote R222C and E211K mutant proteins in Chains A and B. We selected the following regions: β-sheets 5, 6, 7, 9, 11, and 12, and α-helices

8 and 9 in proximity (within 5 Å) to the THH ligand using the PyMOL tool for both chains A and B, respectively, as shown in Figure 23 A and B, and studied their RMSD density.

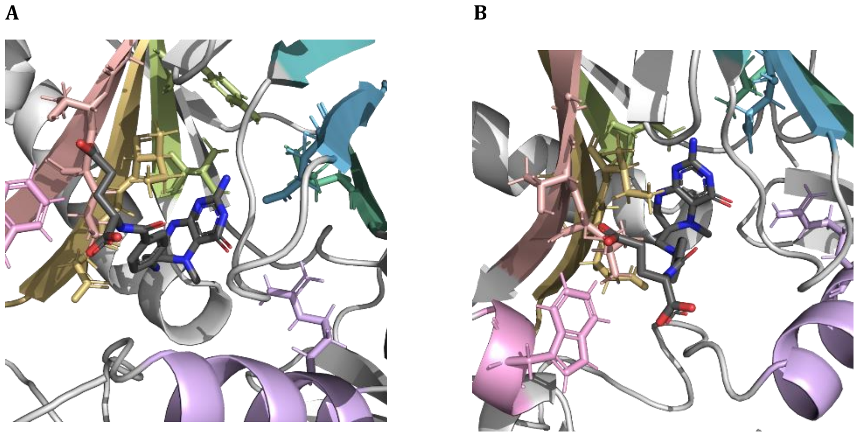


Figure 23. T-protein WT regions in proximity with ligand: Chain A with ligand THH(A), Chain B with ligand THH(B)

For WT Chain A (Figure 24 A), we observed a particular RMSD peak density within the β-sheets region, varying between 0.23nm for β-sheet 7 to a higher peak of density 0.29nm for β-sheet 11. In addition, β-sheet 11 recorded the highest variance, as indicated by its standard deviation (SD). Additionally, α-helix 8 presents an RMSD peak at a density of 0.31nn.

The WT Chain B analog (Figure 24 B) showed a more heterogeneous RMSD values. For instance, the RMSD peak value for β-sheet 5 reached 0.50nn, which considerably exceeded that of Chain A. Similarly, α-helix 8 had an RMSD peak of 0.43nn.

On the other hand, the H83R mutant, Chain A, did not exhibit elevated RMSD values, especially in the β-sheet regions. The range values are between 0.20nm (observed in β-sheet 6) to a peak of 0.37nm (captured in β-sheet 11). An interesting observation arises in α-helix 9 of Chain A (highlighted in Figure 24 C), which reveals a remarkable RMSD peak of 0.66nn, in stark contrast to the equivalent region in the WT.

Meanwhile, the H83R mutant Chain B maintained consistent RMSD values. α-helix 8, for instance, demonstrated an RMSD mode of 0.31nm (Figure 24 D). It is important to mention that the H83R mutant, especially Chain B, showed pronounced variability. This can be especially seen in regions such as β-sheet 11.

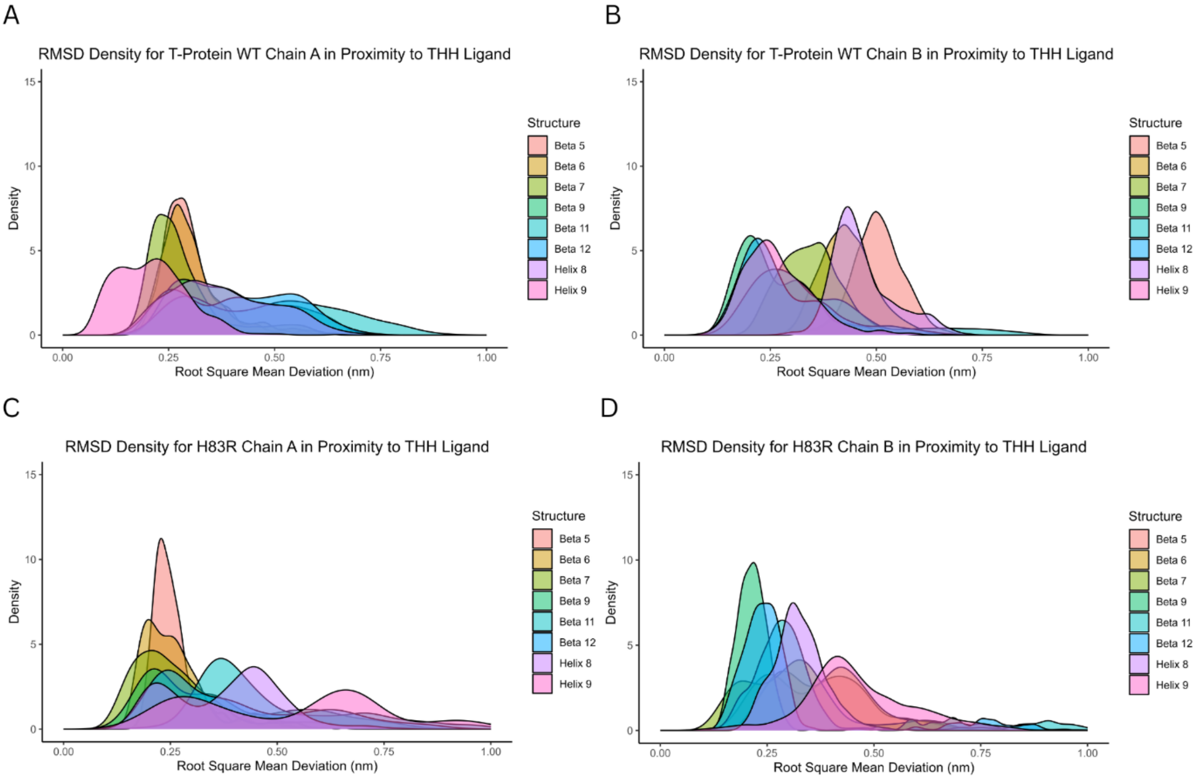
Regarding the R73C mutant on Chain A, we observe a diverse range of RMSD values. The higher peak of density was found in β-sheet 5, with an RMSD of 0.63nm (Figure 24 E) and the lowest RMSD peak occurred in α-helix 9 at 0.16nm, showing a potentially stable region in comparison to its WT. It is also important to note the variance present in this R73C mutant, with α-helix 9 presenting a high standard deviation.

The R320H mutant for Chain B showed β-sheet 5 and β-sheet 6 regions that displayed RMSD peaks at 0.25nm and 0.29nm, respectively (as highlighted in Figure 24 F). α-helix 8 in R320H Chain B had a high RMSD value of 0.48nm. While this R320H mutant chain B does not exhibit any radical peaks, its α-helix 8 suggests a potential shift in dynamic behavior compared to the WT.

The R222C Chain A mutant (Figure 24 G) did not exhibit elevated RMSD values, especially in the β-sheet regions. β-sheet 5 had an RMSD peak of 0.30nm, and α-helix 9 recorded 0.29nn. An interesting observation arises in α-helix 8 of this chain which reveals a remarkable RMSD peak of 0.61nn, in stark contrast to the equivalent region in the WT.

Regarding the E211K Chain B (Figure 24 H), the most pronounced RMSD peak was found in α-helix 8 with a value of 0.79nm, considerably exceeding the WT's peak in the same region. Also, β-sheet 5 in E211K showed RMSD value of 0.68nm.

From these results, analysis of the both chains and their respective mutants, and the WT T-protein, mutations induced specific effects on the RMSD distributions. These alterations can be indicative of changes in protein behavior, stability, and interaction with the THH ligand.



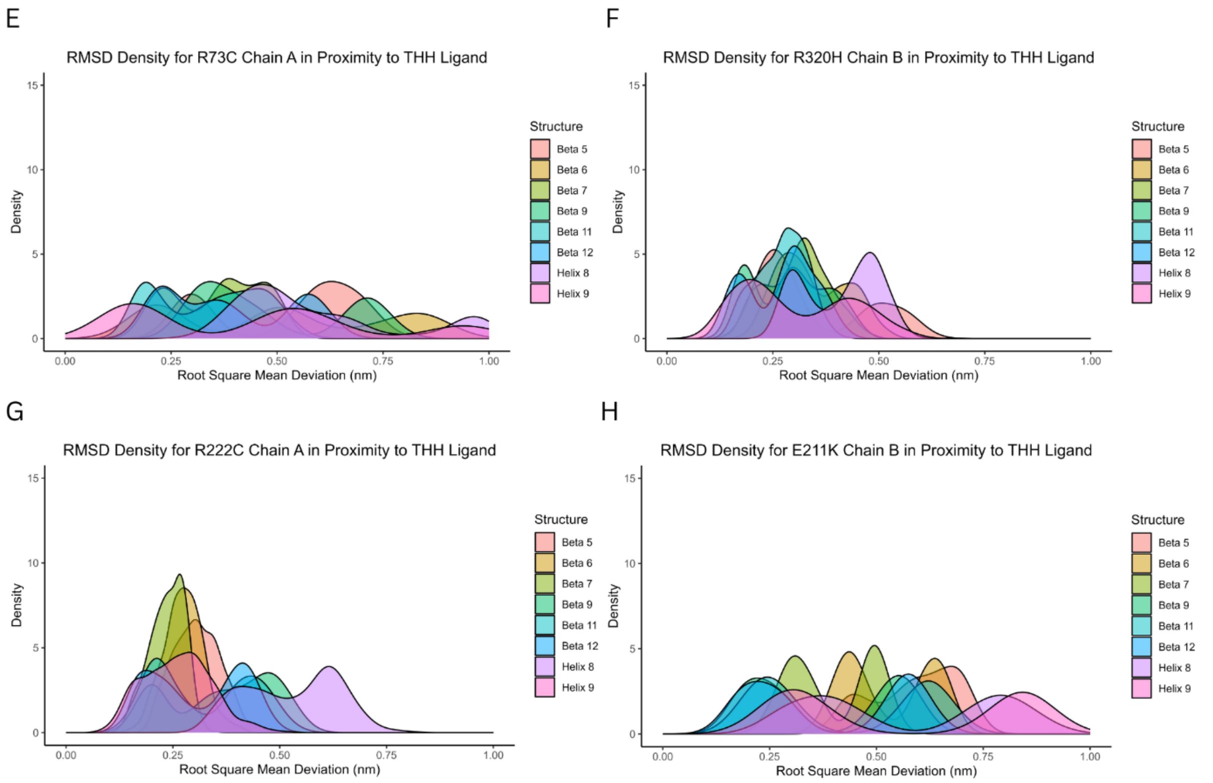


Figure 24. RMSD Density Distribution in Proximity to THH ligand: homozygote WT T-Protein with Chain A (A) and Chain B (B), homozygote H83R mutant with Chain A (C), and Chain B (D), heterozygote R73C Chain A (E) and R320H Chain B (F), and heterozygote R222C Chain A (G) and E211K Chain B (H)

4.8.2 MMPBSA Analysis

This method is particularly useful for estimating the free energy of binding between a ligand and a protein, providing insights into the strength and nature of binding. The MMPBSA analysis of Equation 6 Δ between WT Chain A and B and their ligand THH are shown in Appendices 13 and 14.

$$\Delta\text{G Complex} - \text{Receptor} = \Delta\text{G (Protein Chain– ligand THH)}$$

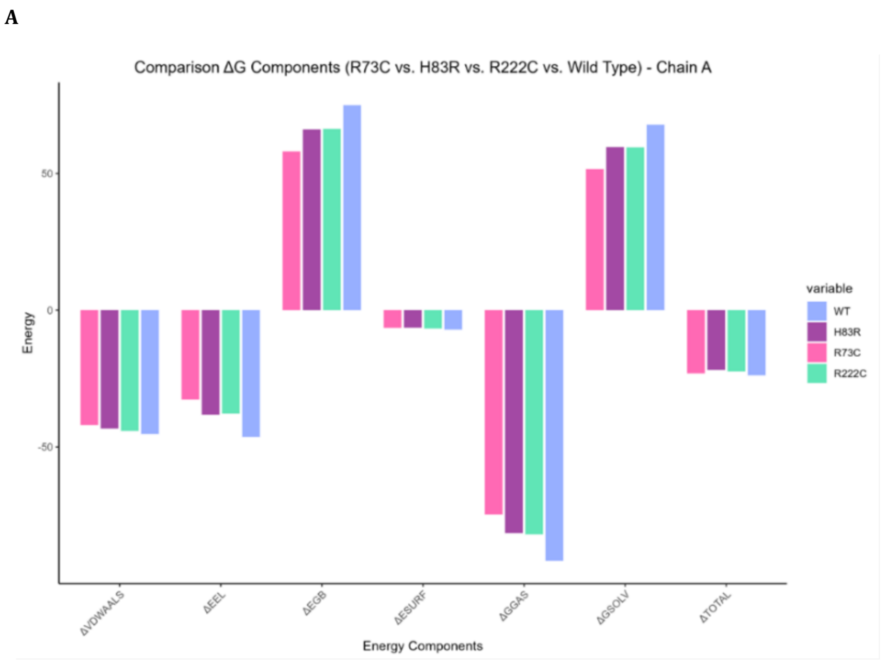
Equation 6. ΔG Complex-Receptor

Similarly, descriptions of the H83R, R73C, R320H, R222C and E211K mutant strains Chains A and B are described in appendices 15, 16, 17, 18, 19 and 20 respectively.

Figure 25 A, shows that the H83R variant (chain A) displays an increase in the van der Waals energy (ΔVDW), with an average value of -43.29 kcal/mol, in comparison to the WT's more stable -45.29 kcal/mol. A shift in the electrostatic energy (ΔEEL) of the H83R variant reached -38.21 kcal/mol, comparing to the WT ΔEEL -46.36 kcal/mol. The gas binding free energy (ΔGGAS) of the H83R variant has a value of -81.50 kcal/mol, compared with the increase from the WT's -91.65 kcal/mol. The total energy (ΔTOTAL) for the H83R variant was -21.81 kcal/mol, indicative of a decrease in the binding energy compared to that of the WT (-23.78

kcal/mol). The R222C variant exhibited a ΔTOTAL value of -22.38 kcal/mol, in proximity to the H83R variant's energy. In contrast, the R73C mutant displayed a ΔTOTAL value of -23.12 kcal/mol for Chain A, positioning it between the H83R variant and the WT.

In Figure 25, Panel B, representing the Chain B results, significant variations are evident. The ΔVDW for the H83R variant was -46.71 kcal/mol, compared to WT's -41.58 kcal/mol. ΔEEL showed a reading of -39.20 kcal/mol for H83R, with the WT registering at -31.77 kcal/mol. The ΔGGAS for the H83R variant was -85.91 kcal/mol, different from that of the WT (-73.34 kcal/mol). In terms of ΔTOTAL for chain B, the H83R variant average was -24.72 kcal/mol, while the WT was -24.46 kcal/mol. The E211K mutant presented a ΔTOTAL value of -20.23 kcal/mol, suggesting an increase in binding energy compared to the WT but closer to the R320H variant. The ΔTOTAL of the R320H variant for Chain B was -19.79 kcal/mol, suggesting a distinct increase in the binding energy from both the WT and H83R variants.



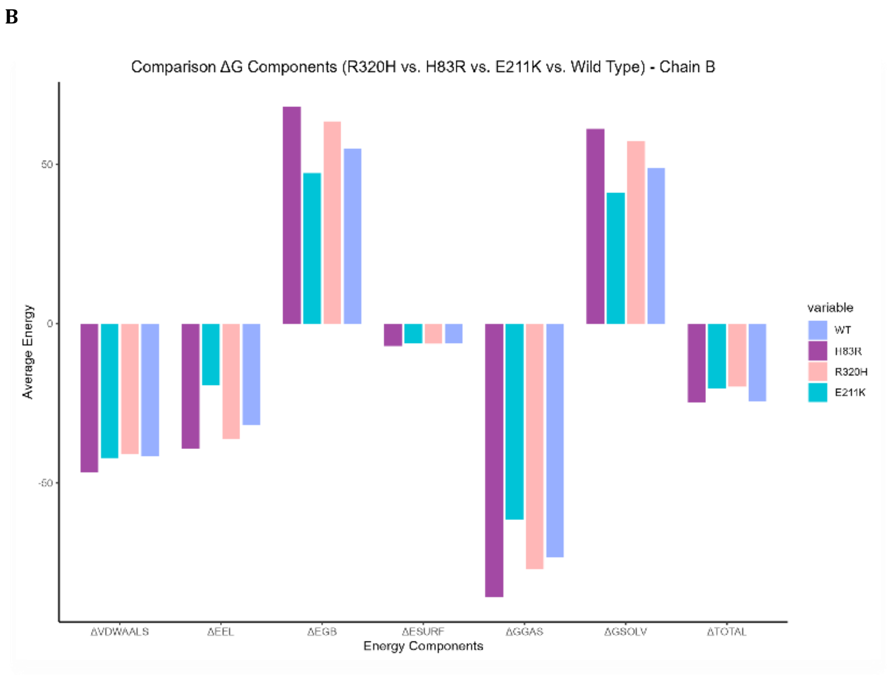


Figure 25. Calculate ΔG WT vs mutants R73C, R320H, R222C, E211K and H83R – Ligand of ΔG Components for Chain A (A) and for Chain B (B).

To determine the effects of the interaction energy between Chain A and Chain B, we compared the total free energy differences between both Chains of WT, H83R, (R73C and R320H) and (R222C and E211K) mutants. As shown in Figure 26 in panels A, B, C and D the residues ARG:34, ARG:35, GLN:157, ASP:158, ARG:161, ASN:165 and ASP:176 show favorable binding energies. This suggests that these residues contribute to stronger binding interactions between chains A and B on WT and their mutants' systems. On the other hand, in Figure 26 A, LEU:33 stands out, with an apparent binding contribution in the WT, but an absence in the mutants.

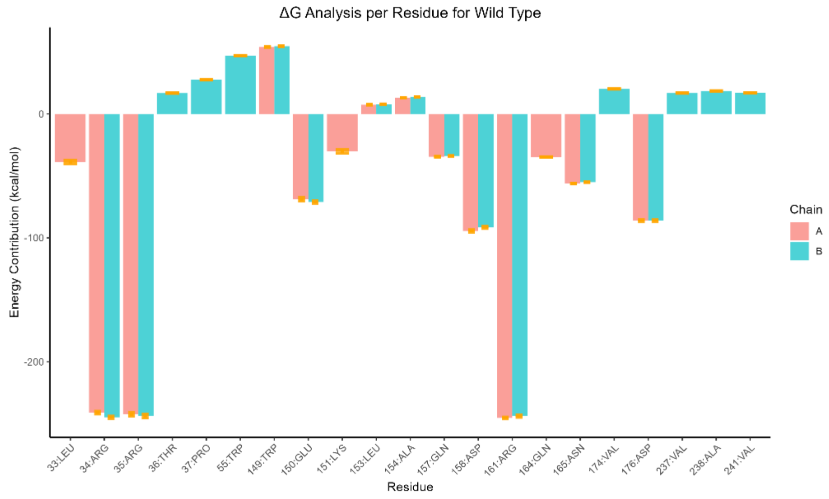
For H83R LYS:159 Chain A showed in Figure 26 B favorable binding energy that that originally had negligible binding in the WT. Also, for H83R GLY:167 and LEU:175 showed unfavorable binding energies that were negligible in WT.

Chain A (R73C) LEU:85 and CYS:148 showed in Figure 26 C unfavorable binding energies for residues that originally had negligible binding in the WT. For GLU:162 Chain B (R320H) show favorable binding energies that were negligible in WT. Also, it showed a decrease in binding energies for TRP:149 in Chain B (R320H). This suggests that the R73C and R320H mutations introduces new interaction sites or alters the conformation to favor interactions at these residues.

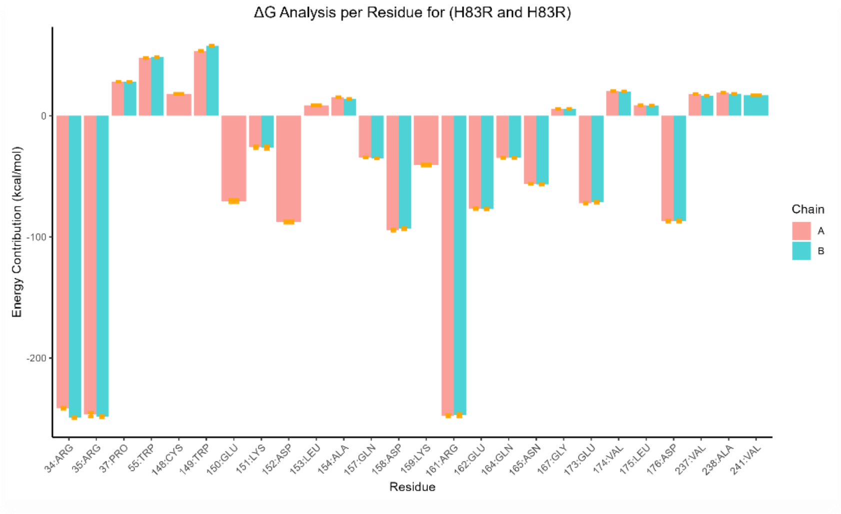
Figure 26 D shows that residues such as ARG:34 and ASP:161 in both chains (R222C and

E211K) exhibited enhanced binding energies, and the residue ARG:35 in Chain A (R222C) did not exhibited energy.

A



B



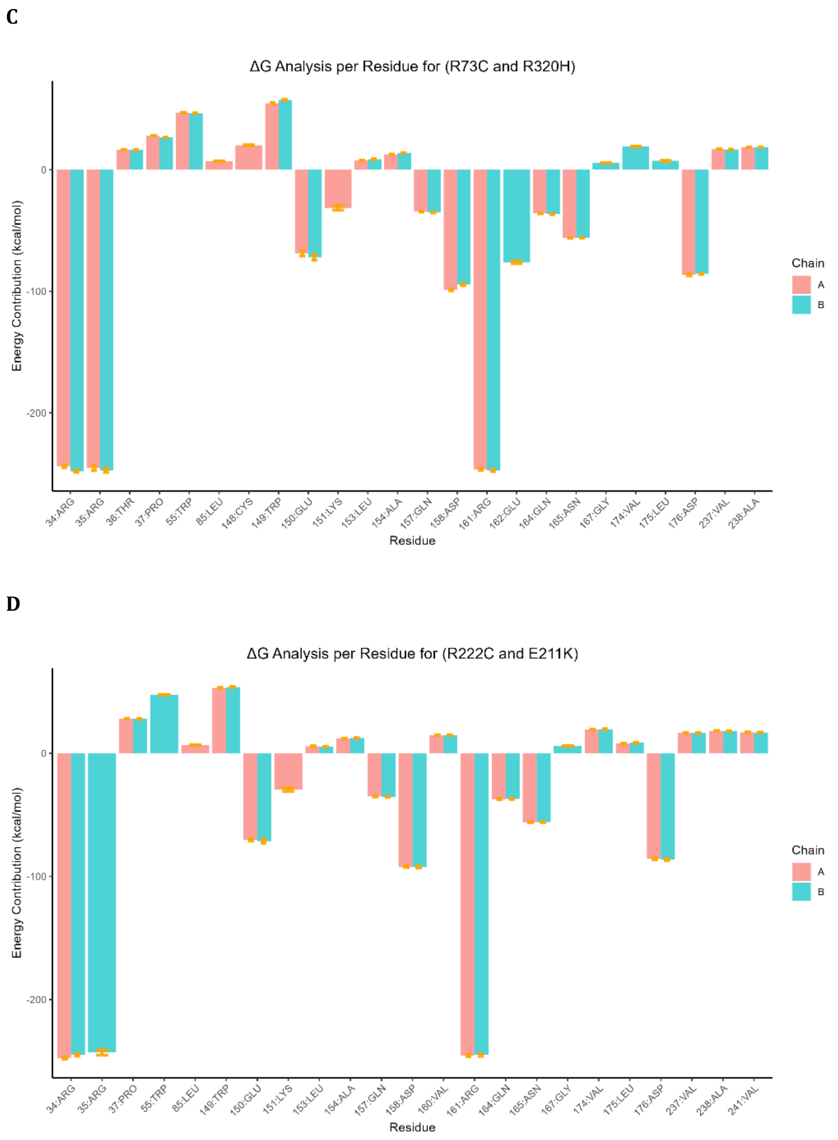


Figure 26. ΔG TOTAL energy analysis per residue between Chains (Chain A and Chain B) for WT (A), H83R (B), R73C and R320H (C) and R222C and E211K (D)

In Figure 27 residues THR:36, GLU:150, LYS:151, LEU:153, and GLN:164 showed contributions that were significant in one variant but not in the other. In Figure 27 A, residues such as ASP:152, GLU:162, and GLU:173 exhibited higher ΔG value than in WT. This suggests that these residues are particularly crucial to the interactions in H83R. In Figure 27 B, while TRP:55 in Chain A (R73C) showed a decrease, it was less evident in Chain B (R320H). Also, Figure 27 B shows elevated interaction energy changes in GLU:162 and GLN:164. In Figure 27 C, some residues, such as ARG:35 in Chain A (R222C) show large changes but not in Chain B (E211K). Two residues THR:36 and GLN:164 had favorable ΔG value in Chain B (E211K).

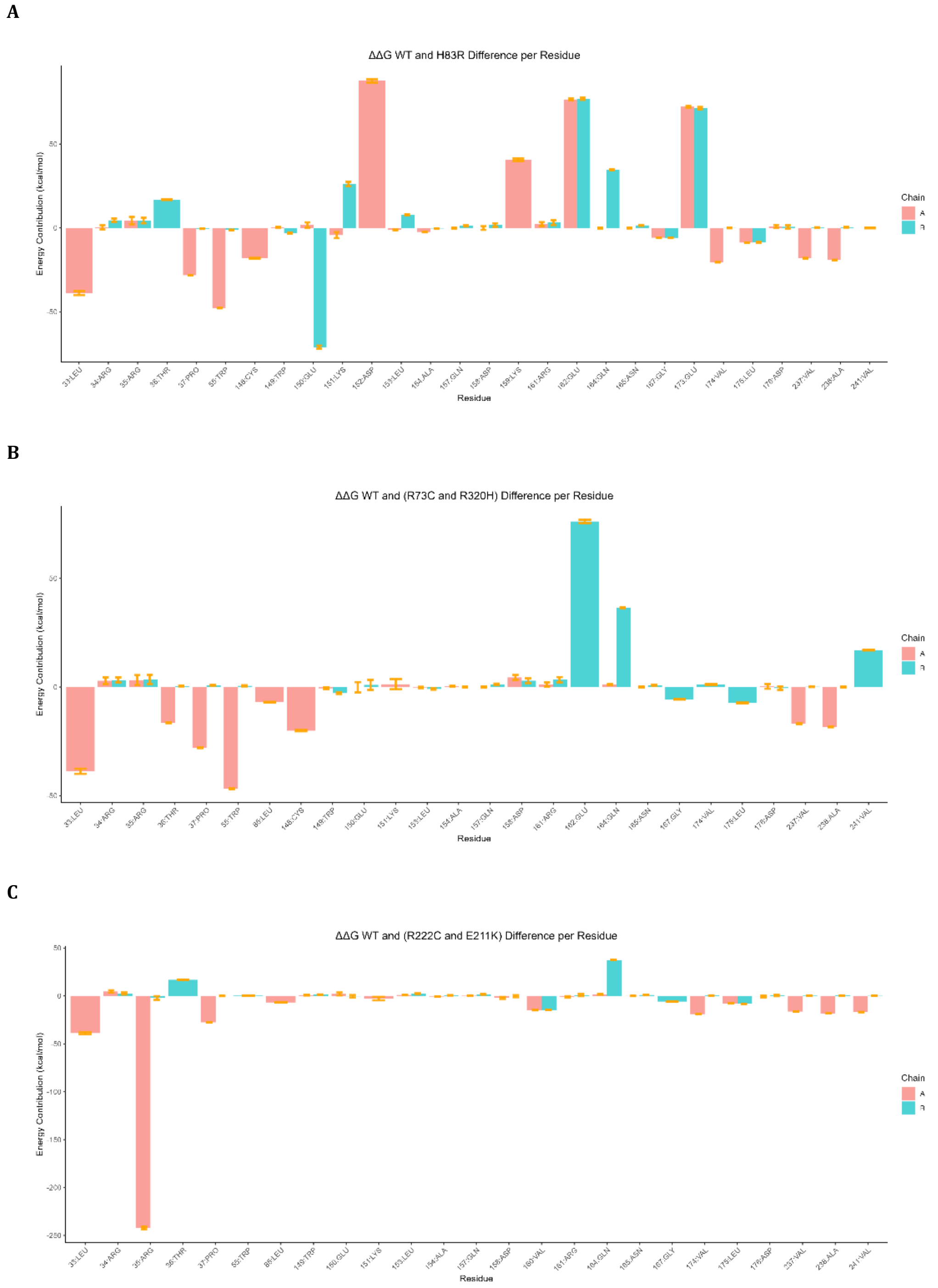


Figure 27. ΔΔG TOTAL energy analysis per residue between Chains (Chain A and Chain B) for WT and H83R (A) and for WT and (R73C and R320H) (B) and for WT and (R222C and E211K) (C)

4.8.3 H-Bonds Analysis

Hydrogen bonds are fundamental for determining the structural and functional dynamics of proteins. A detailed analysis of the interactions between protein chains and ligands can provide profound insights into the perturbations caused by mutations. In this context, we performed a comprehensive analysis of hydrogen-bonding interactions between a designated ligand and both the WT protein and its mutants variants. Figure 28 shows a comparison of H-Bonds between Chains and the Ligand THH of the WT T-protein and the ChainA mutants. From this figure, it can be highlighted that the H83R mutant has a median hydrogen bond count of 2, with hydrogen bonds ranging from 2 to 3. This distribution spanned from a minimum of 0 to a maximum of six hydrogen bonds throughout the study duration. The R73C mutation in ChainA demonstrated a median bond count of 3, with 50% of observations ranging from three to four bonds. The entire range of bond counts spanned from a minimum of 0 to a maximum of 7. The R222C mutation in ChainA reflected a median bond count of 3, and its range for the central 50% of observations was from 2 to 4, spanning a complete range from 0 to 6. In contrast, the WT ChainA showed a median hydrogen bond count of 4. Its interquartile range, which encompasses the central 50% of the observations, extends from 3 to 5 hydrogen bonds. The hydrogen bond count in WT ChainA varied from 0 to 7.

For ChainB the H83R mutant formed a median of 3 hydrogen bonds with the ligand, with half of its observations lying between 2 and 3 bonds. These data were consistent over a range of stretching from zero to six hydrogen bonds. The R320H mutation in ChainB exhibited a median bond count of 2, with half of the observations lying between 1 and 3. The full range extends from a minimum of zero bonds to a maximum of five. WT ChainB also showed a median of three hydrogen bonds, but its central 50% of observations were slightly more spread out, ranging from two to four bonds. The extremities of the hydrogen bonding distribution are marked by 0 and 7 bonds. The E211K mutation in ChainB showed a median of 2 bonds, with half of the observations being between 2 and 3 bonds.

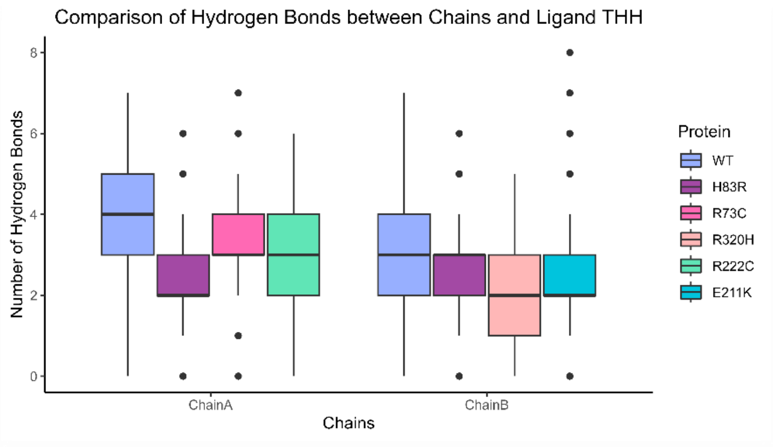


Figure 28. Comparison of Hbonds between Chains and the Ligand THH of WT T-protein and the mutants: H83R, R73C, R320H, R222C and E211K

Figure 29 shows the distribution of hydrogen bond counts between chains A and B for both

WT proteins and their mutant variants: (H83R and H83R), (R73C and R320H) and (R222C and E211K), revealing that the median hydrogen bond count for the WT was 7, in comparison to the (H83R and H83R) and (R73C and R320H) and (R222C and E211K) mutants, which had higher median values of 11. This immediate distinction underscores the pronounced predisposition of the mutant protein to engage in a greater degree of hydrogen bonding between its chains than that of the WT. The interquartile range of WT was condensed and ranged from 5 to 9. On the other hand, the (H83R and H83R) mutants exhibited a more expansive spread, ranging from 9 to 14. This distinction indicates a greater variability in the hydrogen bond formation of the mutant but also indicates potential shifts in its structural or functional dynamics. While the WT whiskers ranged from 1 to 15, with occasional outliers at 16 and 17, the H83R and H83R mutants demonstrated a broader spectrum, starting at 1.5, extending to 21.5, and presenting multiple outliers, especially in the lower spectrum. Similarly, the R73C and R320H mutant's system spread from 4 to 21.5 and shows numerous outliers at the higher end, with values of 22 and 23. The R222C and E211K mutant's system also showed an interquartile range from 9 to 12, with whiskers ranging from 4.5 to 16.5, and several outliers, notably at values like 1, 3, 4, 17, 18, and 19.

In summary, mutant's systems (H83R and H83R), (R73C and R320H) and (R222C and E211K), generally exhibited higher median hydrogen bond counts and broader ranges of hydrogen bond formation than the WT protein. This suggests potential alterations in the structural or functional dynamics of the mutant proteins, potentially driven by changes in chain-chain interactions. The presence of more outliers, especially at the lower end for the H83R and H83R mutant's system, at the higher end for the R73C and R320H mutant's system and varied for the R222C and E211K mutant's system, suggests sporadic instances where the number of hydrogen bonds is significantly reduced or increased, respectively.

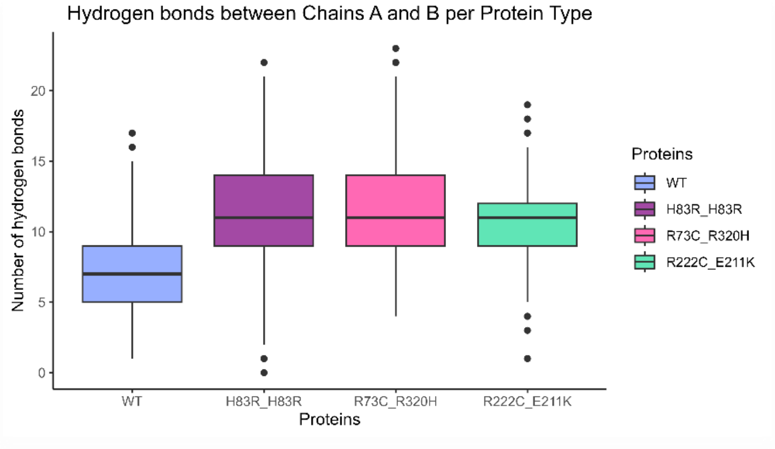


Figure 29. Comparison of H-bonds between both Chains A and B of WT T-Protein and the mutants

4.8.4 Root Mean Squared Fluctuations and Cross Correlations Analysis

RMSF values indicate the flexibility of a particular residue in a protein. Higher values typically suggest greater flexibility, whereas lower values indicate greater rigidity.

Analyzing the RMSF (Figure 30) for WT Chain A (A) and for WT Chain B (B) shows residues in α-helices 2, 4, 6, and 10 showed heightened mobility and β-sheets 3, 9, 11, 12, 13, 14, and 15 showed reduced mobility, implying increased stability in these zones.

Dynamic cross-correlation analysis (CCA) provides pairwise residue correlations. Positive values mean that residues move together, negative values mean that residues move opposite to each other, and values close to zero mean that residues move independently. The final visual output was an image showing the dynamic cross-correlation of the residues in the protein chain. Figure 30 show correlations between different residues for WT Chain A (C) and WT Chain B (D). Both chains show similar strengths and directions of these correlations, as represented by color intensity and hue. However, there's an increased correlation in the regions from loop 1 to β-sheet 3 and within the C-terminal β-barrel in Chain B. Although, a negative correlation is seen within α-helix 5, 6, and β-sheets 10, 11, and 12, especially when compared to the region from loop 1 to β-sheet 3. Additionally, the C-terminal β-barrel region has a negative correlation relative to the area from loop 1 to α-helix 8. This is useful for understanding how different parts of the protein move with respect to each other during simulation.

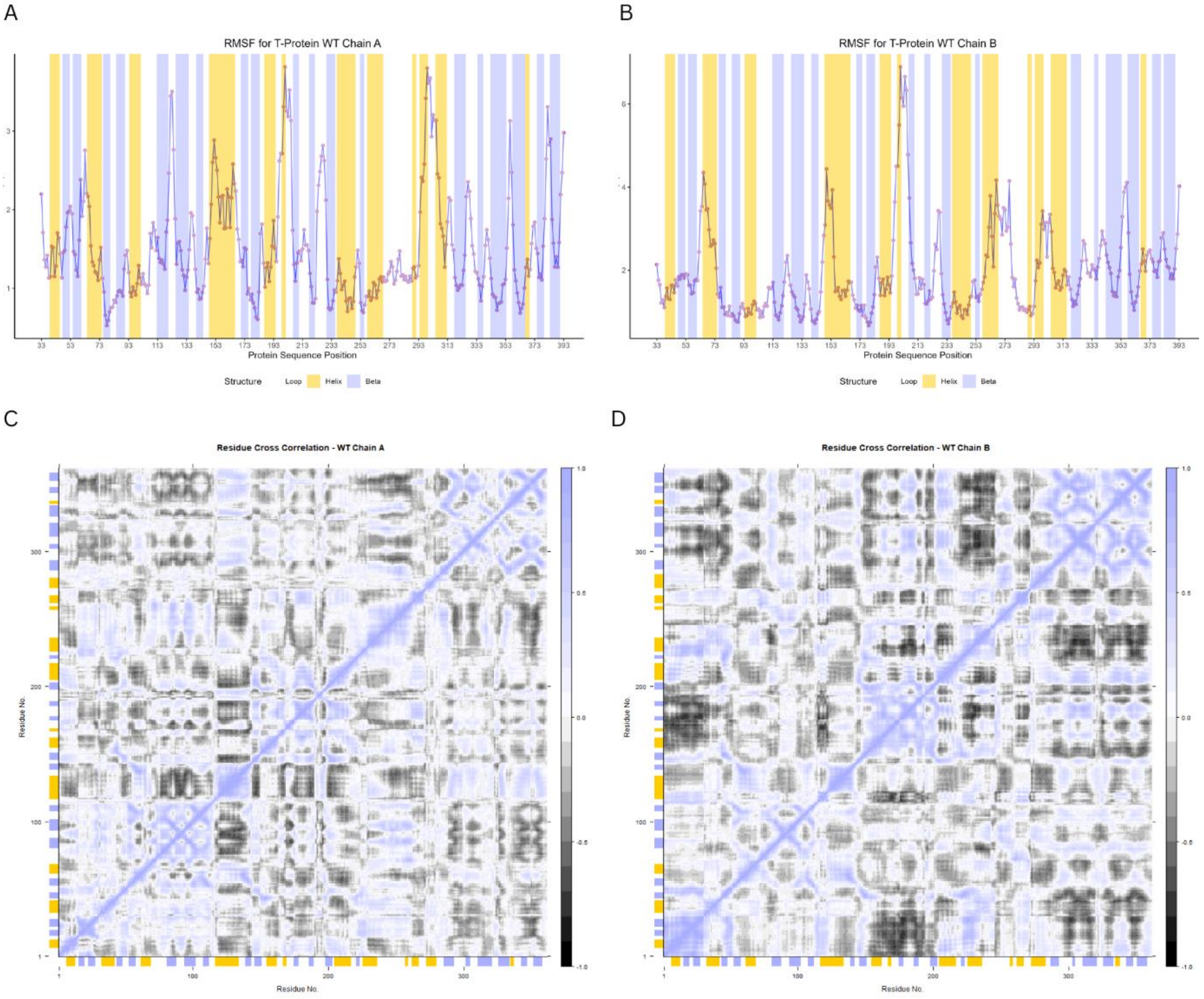


Figure 30. RMSF Analysis per residue: WT Chain A (A), WT Chain B (B) and Cross Correlation Analysis per residue - WT Chain A (C) and WT Chain B (D)

When comparing the RMSF values of the H83R variant (Figures 31 A and B) to those of the WT (Figures 30 A and B), it becomes evident that several RMSF values consistently exceeded those of the WT in multiple regions. In the H83R vs. WT RMSF analysis, residues within α-helices 2, 5, 8, and 9 of the H83R variant exhibited heightened mobility compared with the WT. β-sheet 6 displayed similar dynamics in both H83R and WT. Chain A demonstrated increased flexibility compared to the WT. In contrast, loops 9 and 10 exhibited increased flexibility in WT compared to H83R. α-helix 7 demonstrated nearly uniform flexibility in both chains. Distinct dynamics were observed in Chain B, with escalating fluctuations in H83R, especially in α-helices 2, 11, and 4, when compared with the WT. The reduced flexibility of α-helix 4 in the H83R variant emphasizes increased rigidity compared to that of the WT, potentially influencing its molecular interactions. A comparative analysis spanning positions 280–393 revealed that most residues in the H83R variant within Chain B exhibited amplified

RMSF values, indicating heightened flexibility. Lastly, enhanced stability in H83R Chain B was evident in residues spanning from α-helix 8 to first half of turn 22, showing reduced mobility and implying increased stability within these regions.

Comparing the dynamic cross-correlation of residues in the protein between Figure 31 H83R chain A (C) and H83R chain B (D) it shows in chain A more strength and direction of correlations between residues in region from loop 1 to β-sheet 4 and a negative correlation between residues in region from β-sheets 14 to 19 relative to the region from loop 1 to α-helix 8.

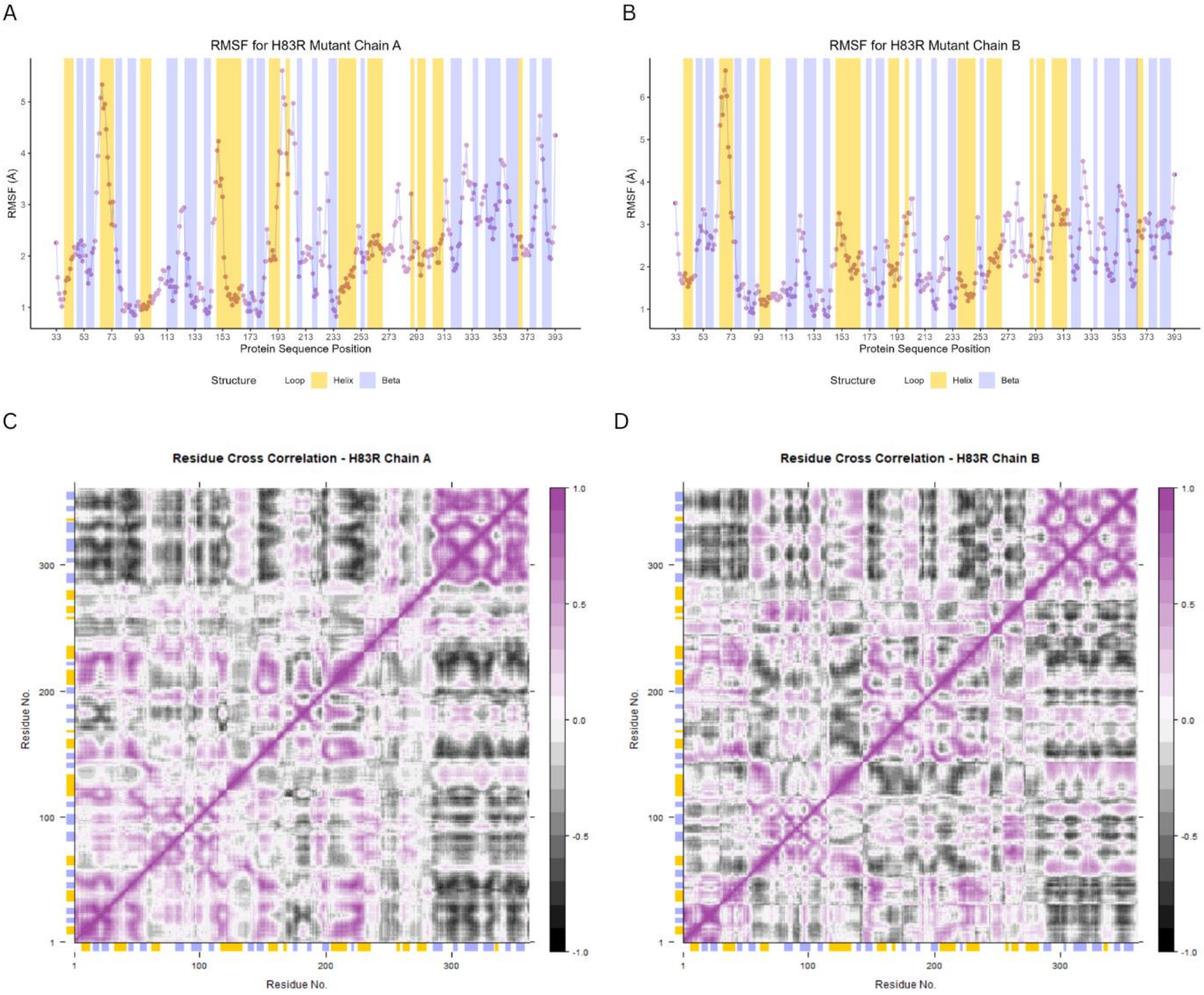


Figure 31. RMSF Analysis per residue: H83R Chain A (A), H83R Chain B (B) and Cross Correlation Analysis per residue - H83R - Chain A (C) and H83R Chain B (D)

The RMSF analysis for R73C in Figure 32 A), showed in general an increase in the values of RMSF, especially in α-helices 2, 4, 5, 8 and loop 22 and C-terminal β-barrel. The R73C mutation appears to have a pronounced effect on the mobility of the regions in the protein.

The R73C mutant predominantly displays increased flexibility, indicating they are more flexible than the corresponding regions in the wild type.

R320H Figure 32 B) Chain B, many positions in R320H mutation showed reduced fluctuation compared to the WT. However, a peak of flexibility in R320H are in residue positions in α-helix 2 show increased flexibility compared to the WT. The regions with reduced flexibility in R320H are in α-helix 8, first half of loop 22, indicating that the R320H mutation results in reduced flexibility in these regions compared to the WT.

Comparing the dynamic cross-correlation of residues in the protein (Figure 32), R73C chain A (C) shows a vivid color and black more strength and direction of correlations between residues across all protein length than Figure 32 and R320H chain B (D) and comparing also with Chain A WT.

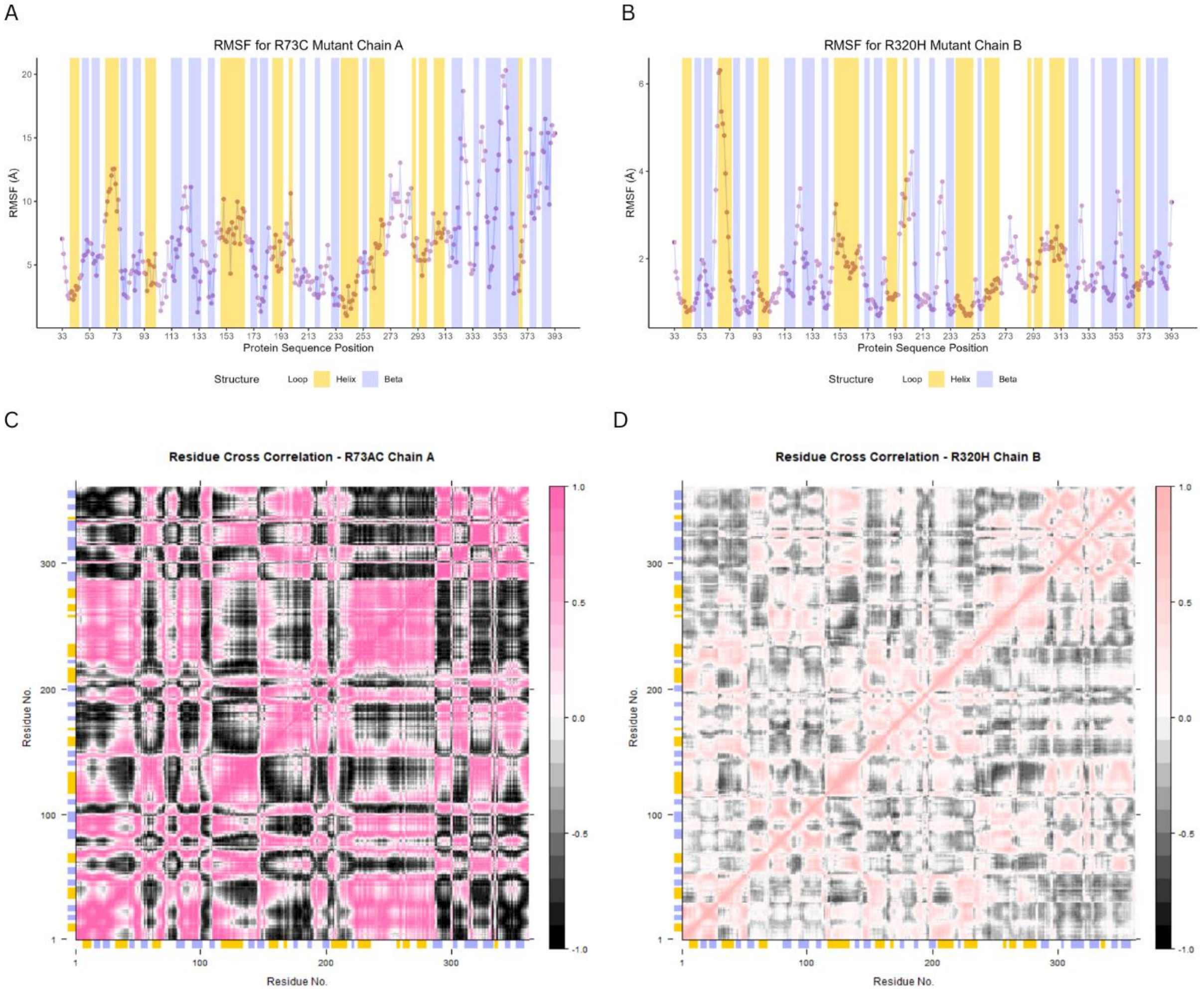


Figure 32. RMSF Analysis per residue: R73C Chain A (A), R320H Chain B (B) and Cross Correlation Analysis per residue - R73C - Chain A (C) and R320H Chain B (D)

Comparing the RMSF between R222C (Figure 33 A) and WT Chain A, many position in α-helices 4, 10 and Loop 9 and 16 and several others show negative RMSF differences, suggesting that the mutation might be causing these regions to be less flexible than in the WT. α-helix 2 and 8, show an increased flexibility in these regions in the mutant compared to the WT.

For Chain B, we observed several positions where the RMSF values of the E211K (Figure 33 B) mutant exceeded those of the WT, suggesting an increased flexibility or dynamic nature in these regions of the mutant protein. These peaks in RMSF difference, particularly noticeable around positions α-helices 2, 4 and 5 and Loop 22 signify regions in the E211K mutant that may experience greater motion or flexibility than the corresponding regions in the WT.

Comparing the dynamic cross-correlation of residues in the protein E211K chain B (Figure 33 D), it shows a vivid color and black more strength and direction of correlations between residues across all protein length than R222C chain A (Figure 33 C) and also comparing it with Chain B WT.

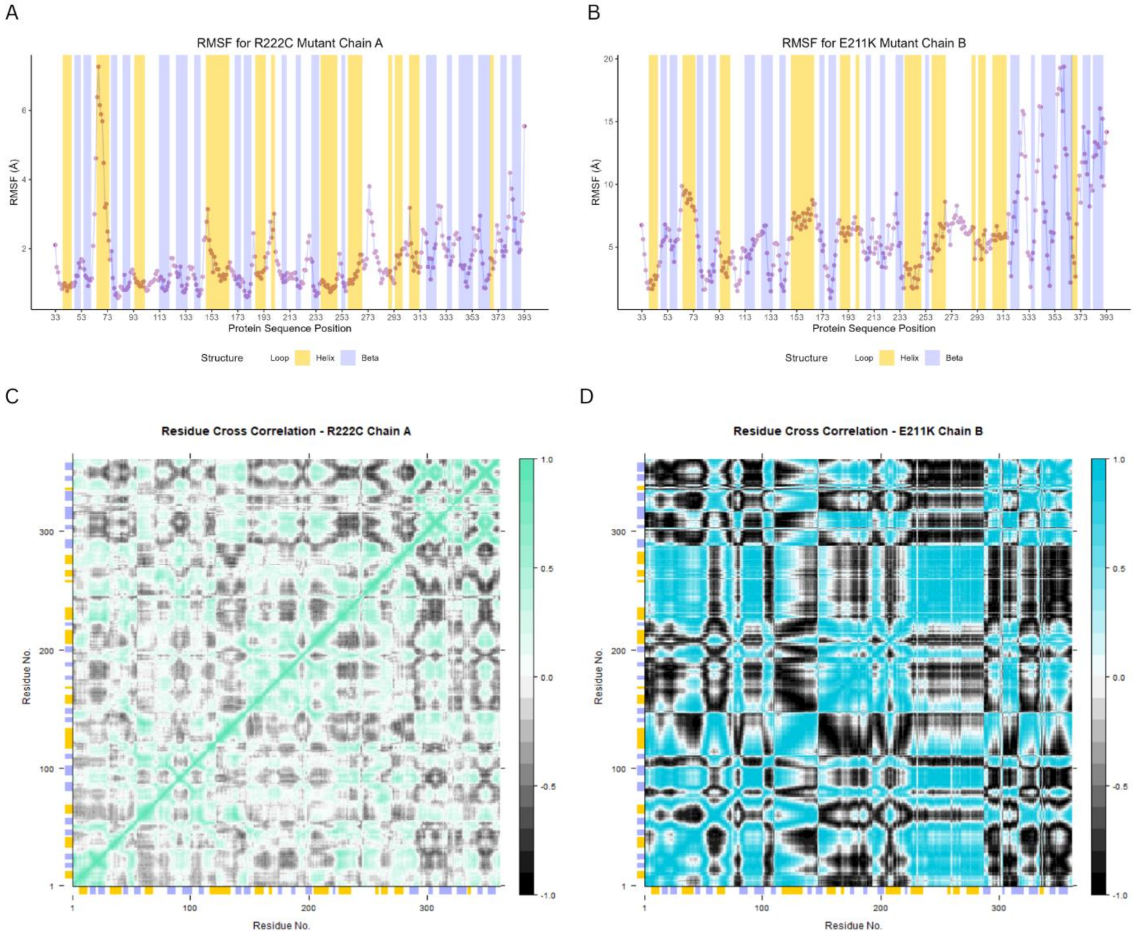


Figure 33. RMSF Analysis per residue: R222C Chain A (A), E211K Chain B (B) and Cross Correlation Analysis per residue -

R222C - Chain A (C) and E211K Chain B (D)

4.8.5 Dynamic Contact Extraction from MD Simulations

The heatmap presented in Figure 34 illustrates the frequency of H-bond contacts between chains per residue for the WT (A), H83R (B), R73C and R320H (C) and R222C and E211K (D). Vivid colors indicate more frequent interactions between the two residues during the MD simulation trajectories.

In Figure 34 A, this heatmap focuses on the WT Chains per residue. A total of 70 unique interactions were identified. Among these, the interaction that distinctly stands out is between residues A:ASP:158 of Chain A B:ARG:34 of Chain B with frequency of 0.75% and A:ARG:161 of Chain A and B:ARG:35 of Chain B with frequency of 0.34% highlighting a significant hydrogen bonding event in these residues in WT Chains.

In Figure 34 B, the H83R heatmap shows to be richer in terms of interaction density, surpassing that of the WT. The most intense interaction spots on this heatmap can be attributed to residues A:ASP:158 of Chain A and B:ARG:34 of Chain B. This signifies a strong or more frequent hydrogen-bonding event specific to the H83R mutation.

In Figure 34 C, heatmap for R73C and R320H shows that the interaction landscape appears less dense in comparison to its WT counterpart. The most noteworthy interaction, with the highest frequency, was observed between residues A:ASP:158 of Chain A and B:ARG:34 of Chain B. Comparing with WT the h-bonds interactions between A:ASP:158 and B:ARG:34 appear with frequencies 0.99% for R73C and R320H and 0.75% for WT. Some interactions become more frequent such as A:ARG:161 of Chain A and B:ARG:35 of Chain B and A:GLU:162 of Chain A and B:ARG:35 of Chain B than in WT.

In Figure 34 D, the heatmap corresponding to the mutation system R222C and E211K stands out interactions between residues A:ASP:176 of Chain A and B:GLN:164 of Chain B, indicating an almost constant interaction during the MD simulation trajectories. The Second highest interaction are between A:ASP:158 of Chain A and B:ARG:34 of Chain B.

The mutation R222C and E211K does introduce changes in the hydrogen bond contacts frequency compared to the WT. Some interactions become more frequent such as A:ASP:176 of Chain A and B:GLN:164 of Chain B than in WT. Others increased interaction are A:ASP:158 and B:ARG:34, and A:ASP:176 and B:ARG:161, and some decrease interactions frequency between A:ARG:161 and B:ARG:35 , and A:GLU:162 and B:ARG:35 than in WT.

The H83R mutation's system appears to enhance both the number and intensity of interactions, hinting at potentially stronger or more frequent hydrogen bonding events with residues, the (R73C and R320H) and (R222C and E211K) mutations seem to reduce the number of unique interactions compared to WT.

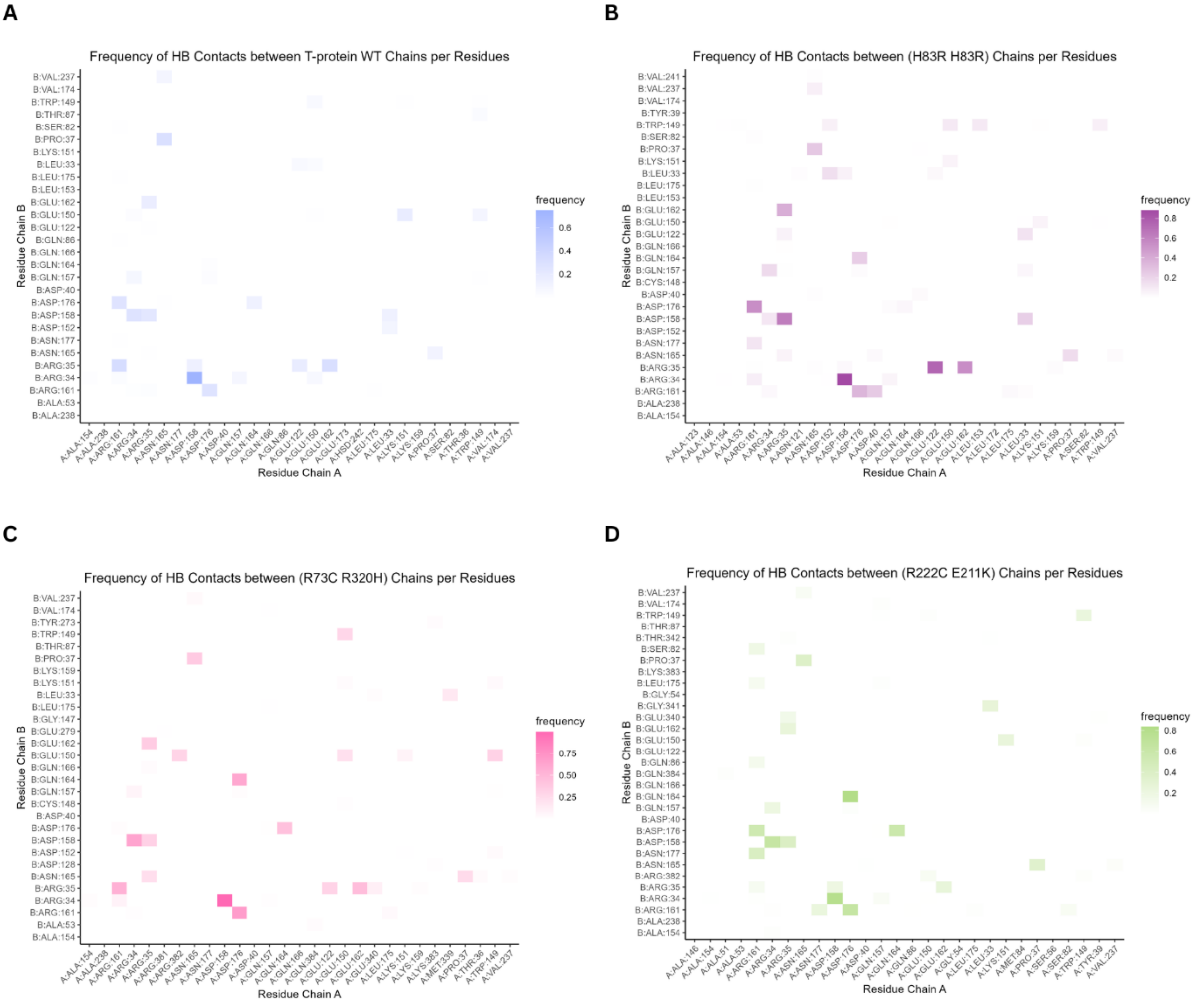


Figure 34. Frequency of HB Contacts between Chains per Residues: WT T-protein (A), H83R (B), R73C and R320H (C) and R222C and E211K (D)

Figure 35 shows the H-bond contacts between ligand THH and various residues in the protein structures: WT, H83R, R73C, R320H, R222C and E211K per chain A and B. The residue GLU:232 in both chain A and chain B forms hydrogen bonds with THH, highlighting its relevance in ligand recognition in the WT protein. Also, the interaction of CYS:271 with THH in WT Chain A suggests its importance.

For R222C, the key residues interacting through H-bonds were GLY:224, TYR:225 and ARG:261.

The H-bond key residue interactions for R73C were GLU:232, GLY:224, ARG:261 and

ARG:222.

The H-bond key residue interactions for R320H were GLU:232, ARG:261 and TYR:225.

The key residue interactions in H83R are TYR:225 in Chain A and ARG:261 and GLU:232 in both Chain A and Chain B.

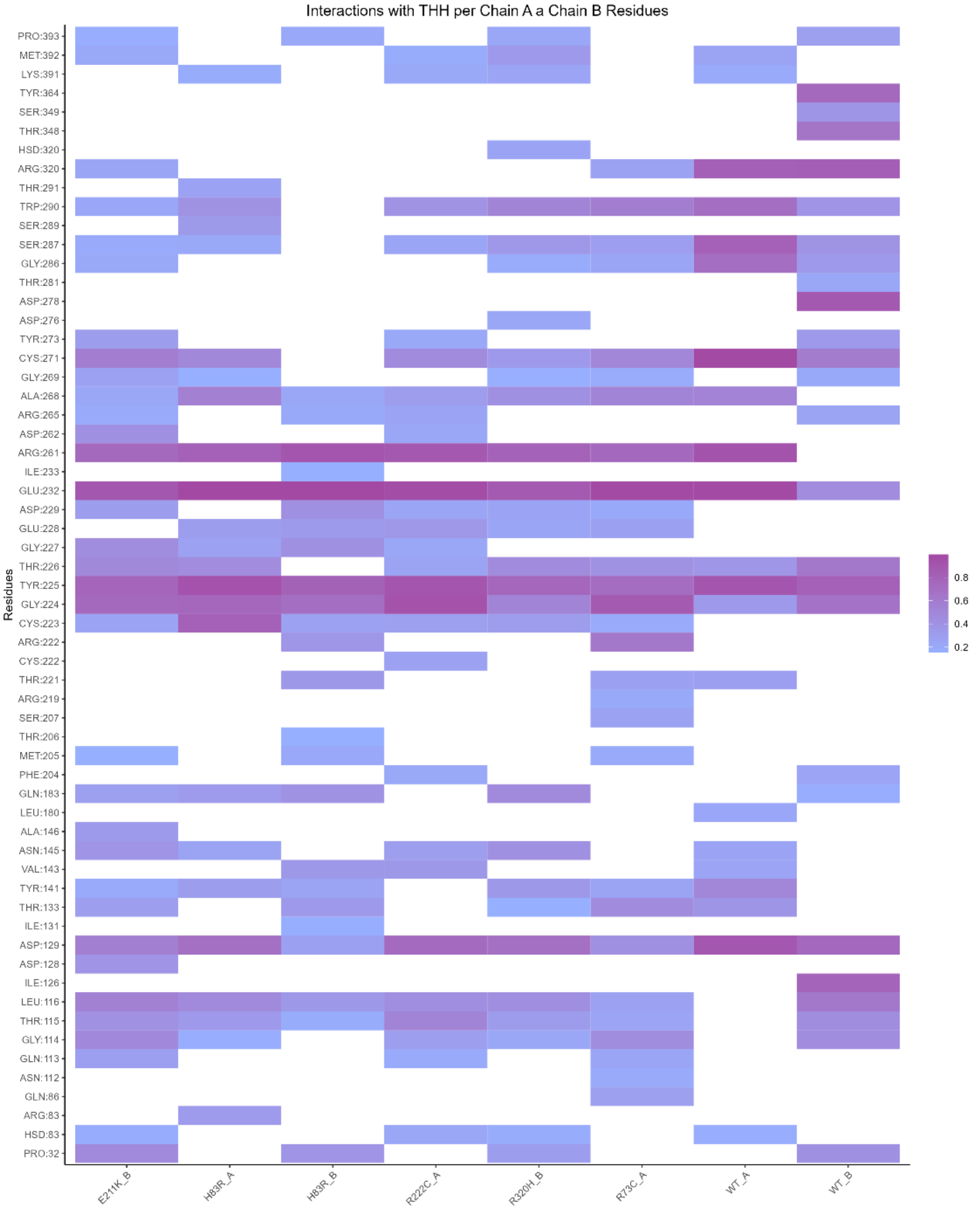


Figure 35. Frequency of HB Contacts with THH per Residues for WT and the protein mutants

4.9 Machine learning for T-protein pathogenicity prediction

The ESM-2 model from Facebook AI Research (FAIR) at Meta AI[123] is a pre-trained language model for proteins that can be used to predict the effects of protein mutations. Using this model, we generated features for the amino acid sequence of the T protein to predict pathogenicity when SAV mutations occur.

4.9.1 Features Exploration Phase

In the feature exploration phase of the ML model, the ESM-2 model was used.

Attention contacts sum: Figure 36 shows the attention contacts map for the WT T-protein, which served as an input feature for the ML model for predicting the attention contacts sum of each mutation position.

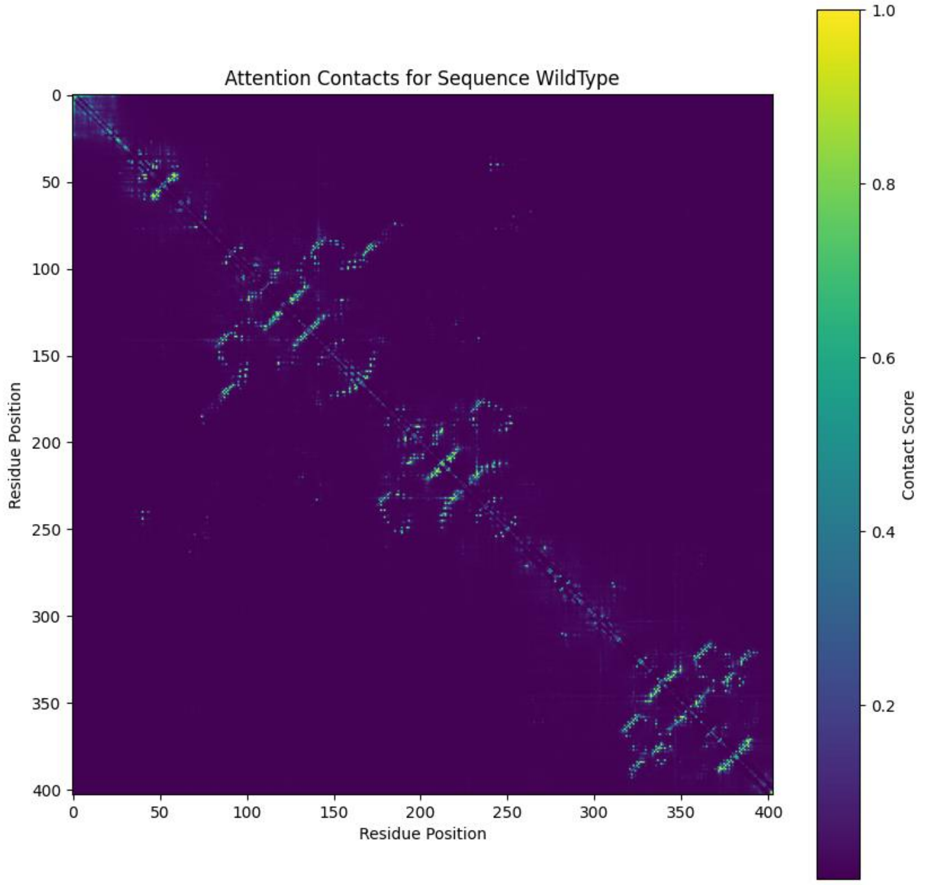


Figure 36. ESM-2 Attention Contacts for the WT T-protein

Amino Acid Predictions

Figure 37 shows *Softmax* Probabilities for Each Amino acid in the WT T-protein sequence length, where yellow represents a high probability, indicating that the model considers this amino acid highly likely at that position, while blue signifies a low probability. This visualization of amino acid prediction probabilities enhances our understanding of the T-protein and its potential for mutation.

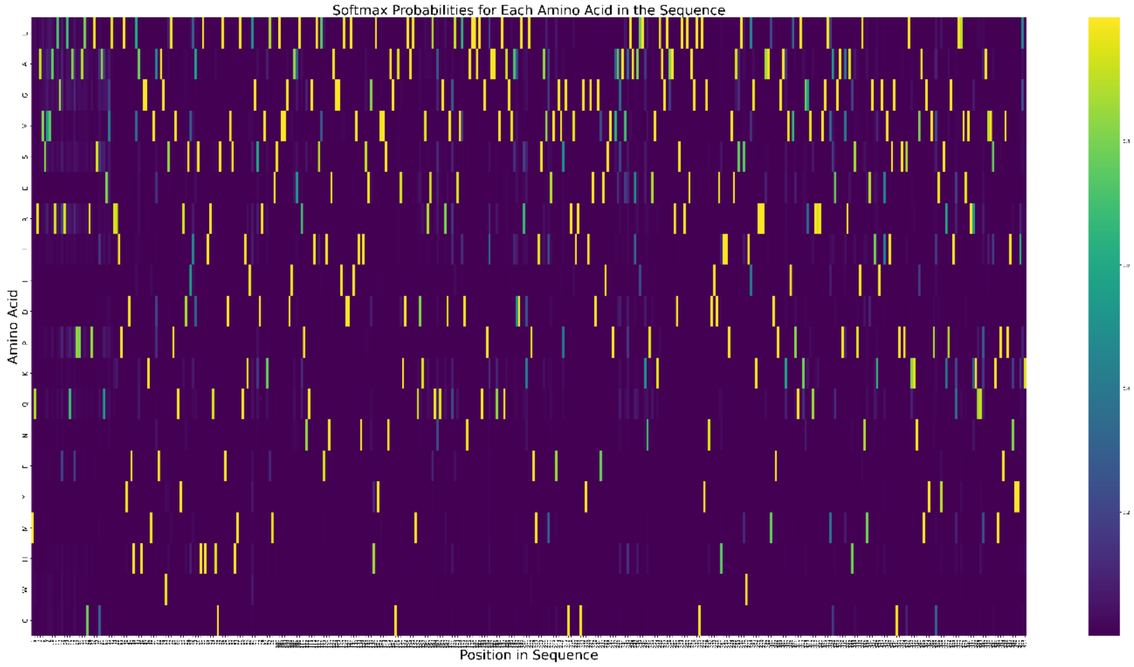


Figure 37. Softmax Probabilities for Each Amino acid in the WT T-protein sequence length

Normalization of Amino Acid Predictions

In Figure 38, each column represents a position in the sequence and each row represents one of the 20 common amino acids. The color of each cell in the matrix represents the normalized probability of the corresponding position of the respective amino acid. Yellow cells indicate a probability of zero (the maximum original probability at that position), whereas dark blue cells represent a value of -20 (the maximum possible difference after normalization). This visualization provides a clear view of the relative likelihood of different amino acids at each position, enhancing our understanding of the structure and function of the protein.

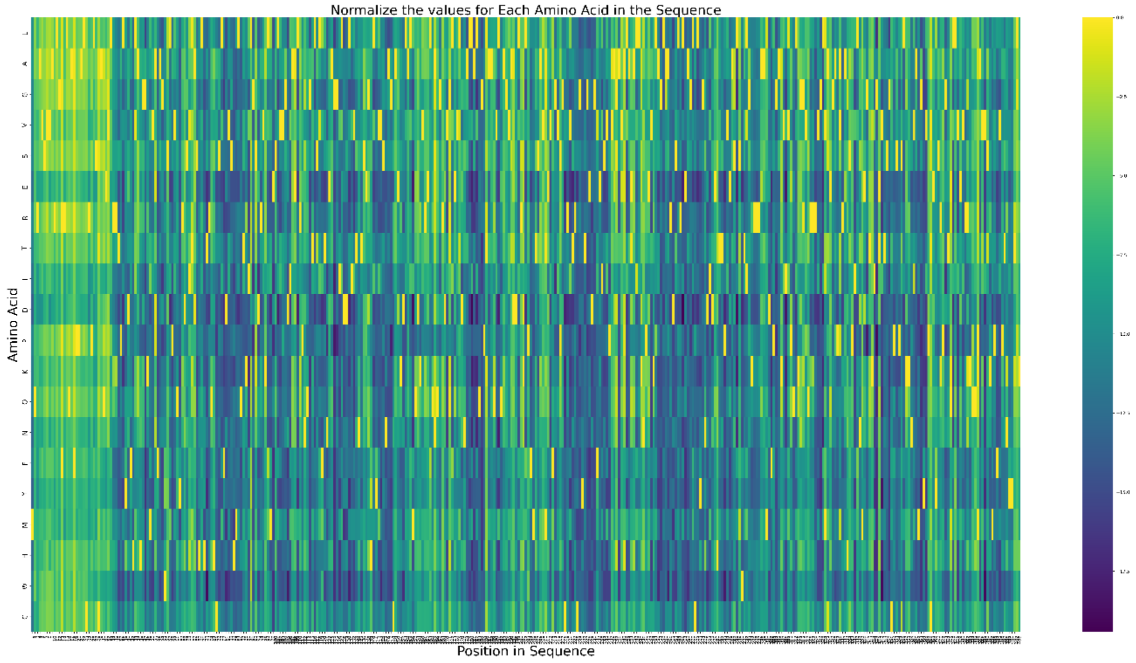


Figure 38. Normalized logits for each amino acid in the WT T-protein sequence length

Entropy Calculation

This feature can identify the positions with the lowest and highest entropy values, as the Figure 39 shows positions such as (82, 98, 123, 182, 183, 222, 223, 226, 227, 228, 229, 256, 257, 261, 271, 303, 319, 321, 344, 349) with lowest entropy indicating the most conserved and the positions (6, 12, 17, 27, 29, 31, 89, 164, 170, 185, 196, 200, 209, 238, 241, 248, 366, 374, 385, 401) with high entropy values are least conserved positions, respectively, in the protein sequence.

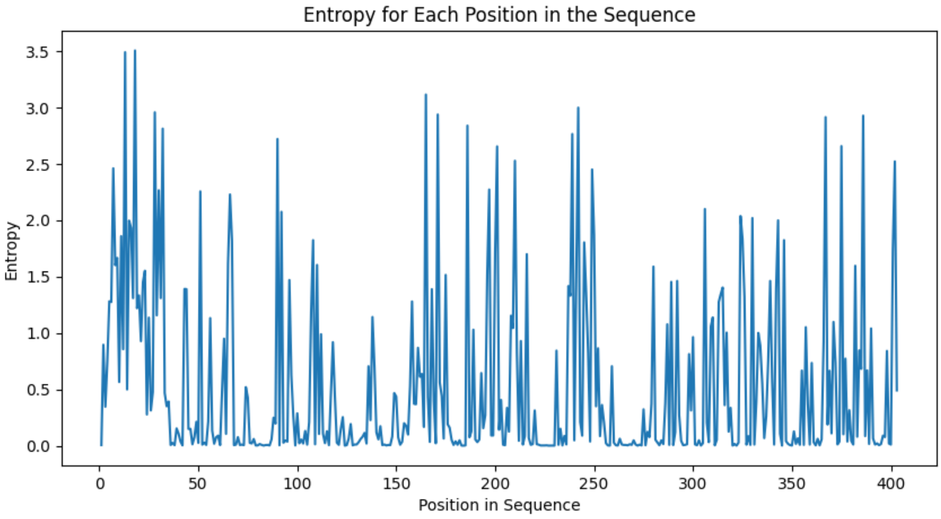


Figure 39. Entropy for each position in the WT T-protein sequence

Amino Acid Counts

As part of the feature exploration phase, we also counted each amino acid in the WT T-protein sequence. This analysis provides an overview of the protein composition, revealing which amino acids are the most common and less prevalent (Figure 40). Analysis showed that the most common amino acid in the protein sequence was leucine (L), appearing 48 times. Valine (V), Glycine (G), and Alanine (A) are also quite common, appearing 40, 35, and 34 times respectively. In contrast, Tryptophan (W) is the least common, with only 3 occurrences. This distribution of amino acids may provide important biological insights, as the prevalence of certain amino acids may be related to the function and structure of the protein. However, these amino acid counts were not directly used as features in the ML model to predict pathogenicity. This is because the model primarily focuses on the impact of mutations, which depend more on the locations and types of amino acids being replaced rather than on the overall count of each amino acid in the protein sequence.

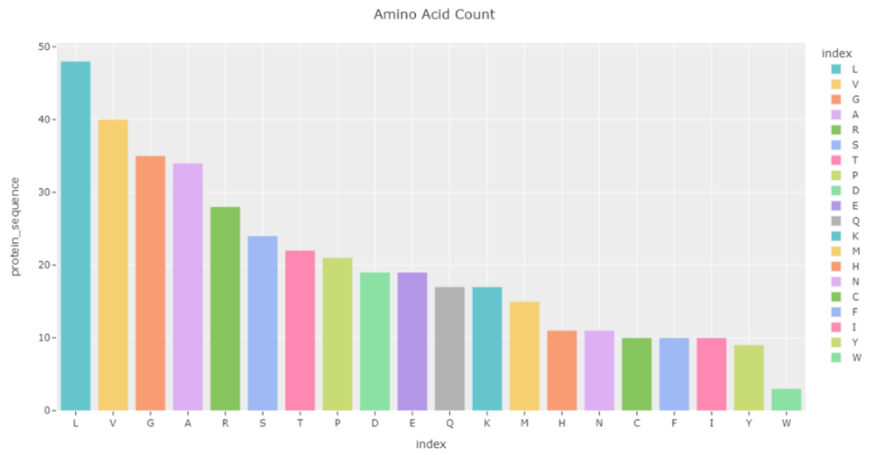


Figure 40. Amino acid count in the WT T-protein

4.9.2 Validation Phase

In the validation phase of the ML model for predicting the pathogenicity of mutations in the AMT protein, we assessed the performance of the different models using the MAE metric and Spearman correlations (Table 16).

Metrics	XGBRegressor	XGBClassifier	CatBoostRegressor	CatBoostClassifier	RandomForestREgressor	RandomForestClassifier
MAE on Train	0.0001	0.0	0.0039	0.0	0.0	0.01
MAE on Validation	0.0001	0.0	0.034	0.0	0.0	0.03

Table 15. MAE of Train and Validation stages for the Models: XGBRegressor, XGBClassifier, CatBoostRegressor,

CatBoostClassifier, RandomForestRegressor and RandomForestClassifier

Additionally, Spearman correlation values provided insights into the relationship between the predicted and actual pathogenicity scores. The correlations between the training and validation sets for each model are as follows:

Metrics	XGBRegressor	XGBClassifier	CatBoostRegressor	CatBoostClassifier	RandomForestRegressor	RandomForestClassifier
Spearman correlation on Train	1.0 (p=0.000)	1.0(p=0.000)	0.512(p=0.003)	1.0 (p=0.000)	1.0 (p=0.000)	0.701(p=1.125)
Spearman correlation on Validation	0.999 (p-value =0.000)	0.999 (p-value =0.000)	0.577(p-value=0.134)	0.999 (p-value= 0.000)	0.999 (p-value: 0.000)	0.607(p-value=0.111)

Table 16. Spearman of Train and Validation stages for the Models: XGBRegressor, XGBClassifier, CatBoostRegressor, CatBoostClassifier, RandomForestRegressor and RandomForestClassifier

Based on the MEA and Spearman results (Table 16) and the analysis of feature importance (Figure 41, panels A, B, C, D, E, and F), the following can be verified.

- The MAE for XGBRegressor, XGBClassifier, CatBoostRegressor, CatBoostClassifier, RandomForestRegressor, and RandomForestClassifier on both the training and validation data were very close to zero or exactly zero, indicating that the predictions made by these models were very accurate.
- For all the models, both the Training and Validation Correlation Values were extremely close to 1, indicating a very strong positive correlation between the predicted and actual values.
- For all models, the p-value was effectively zero for both the Training and Validation datasets, indicating that the correlation between the predicted and actual values was statistically significant.
- Feature importance analysis for the four models found that feature_probs was the most important feature, suggesting that it has the most significant impact on predicting the pathogenicity of mutations in the AMT protein.
- For XGBRegressor, only feature_probs is used, implying that it overwhelmingly dominates importance.
- For the XGBClassifier, after feature_probs, the entropy values and contact sum features were the next most important features.
- CatBoostRegressor found feature logits, feature_probs, entropy, and contacts_sum_inverse to be the most important features in that order.
- The CatBoostClassifier finds feature logits, feature_probs, entropy,

contacts_sum, and contacts_sum_inverse, which are the most important features in that order.

- For RandomForestRegressor, after feature_probs, the feature logits and entropy were the most important features in that order.
- For the RandomForestClassifier, after feature_probs, the feature logits, contact sum, and entropy are the most important features in that order.

However, caution should be exercised when interpreting these results. This could indicate overfitting, where the model has memorized the training data but may not perform well on new, unseen data. Therefore, further verification with more diverse data is necessary to provide a more comprehensive analysis of the performance of the models, draw conclusions on their effectiveness in predicting the pathogenicity of mutations in the AMT protein, and discuss the implications of the results.

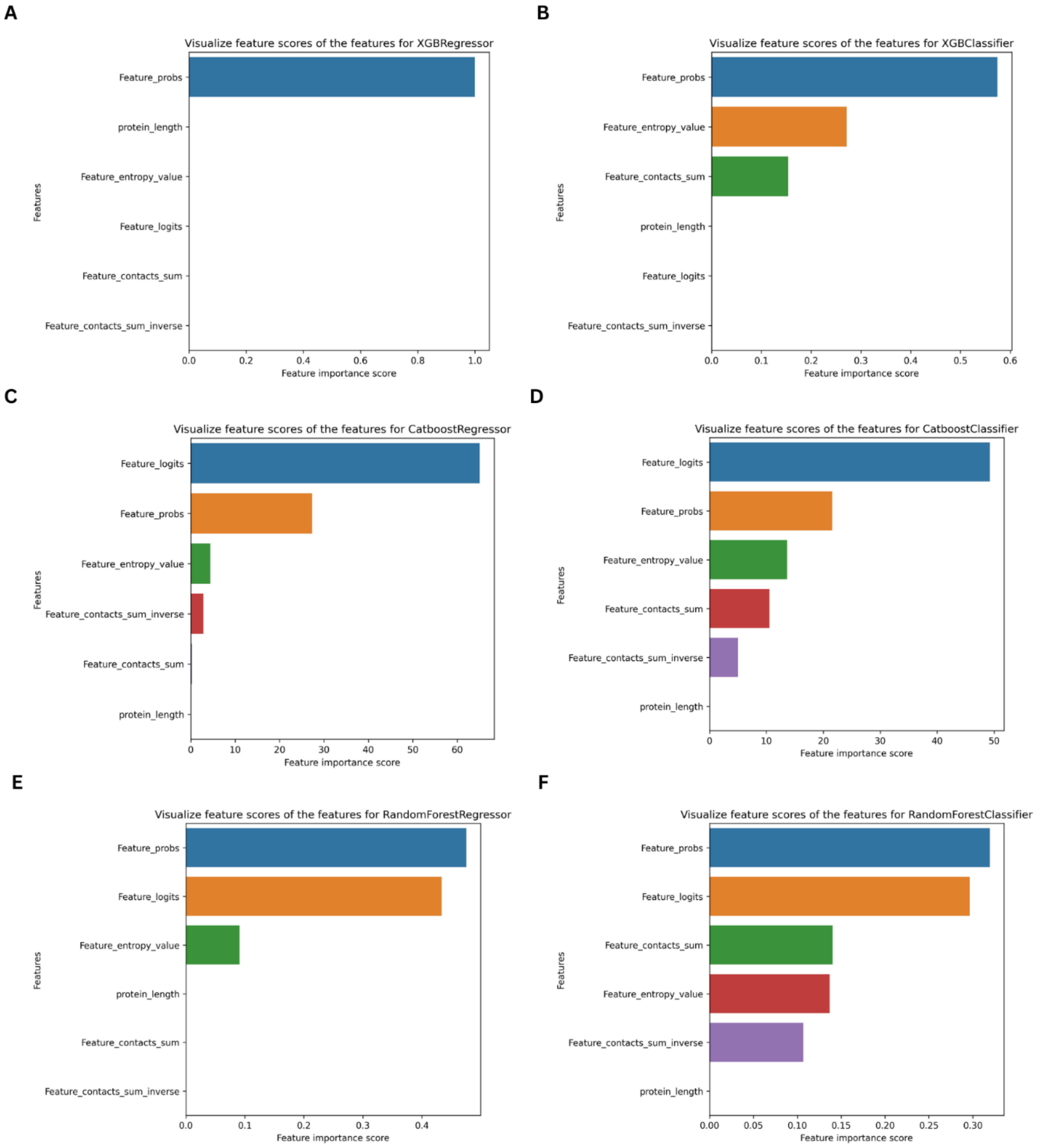


Figure 41. Feature Score for each Models: XGBRegressor (A), XGBClassifier (B), CatBoostRegressor (C) CatBoostClassifier (D), RandomForestRegressor (E) and RandomForestClassifier (F)

4.9.3 Test Results

Four models were selected to test the test dataset. The evaluation metrics used to assess the performance of the model on the test dataset are listed in Table 17.

Model	Accuracy	AUC-ROC	Precision	Recall	AUC-PR	F1-Score
XGBRegressor	1.0	1.0	1.0	1.0	1.0	1.0
XGBClassifier	1.0	1.0	1.0	1.0	1.0	1.0
CatBoostRegressor	1.0	1.0	1.0	1.0	1.0	1.0
CatBoostClassifier	1.0	1.0	1.0	1.0	1.0	1.0

RandomForestClassifier	1.0	1.0	1.0	1.0	1.0	1.0
RandomForestClassifier	1.0	1.0	1.0	1.0	1.0	1.0

Table 17. Metrics of Test stage for the Models: XGBRegressor, XGBClassifier, CatBoostRegressor, CatBoostClassifier, RandomForestRegressor and RandomForestClassifier

Overall, all four models demonstrated outstanding performance on the test dataset, achieving perfect accuracy and excellent performance in terms of AUC-ROC, precision, recall, AUC-PR, and F1-score. For this reason, we tested the consensus list dataset to provide a more comprehensive analysis of its performance, draw conclusions on its effectiveness in predicting the pathogenicity of mutations in the AMT protein, and discuss the implications of the results. All four models were used to predict the pathogenicity of the list of SAV mutations based on the consensus table results, which served as an independent test set. The results obtained from the four models for the consensus dataset are presented in Table 18.

Model	Accuracy	AUC-ROC	Precision	Recall	AUC-PR	F1-Score
XGBRegressor	0.89	0.856	0.867	0.977	0.862	0.919
XGBClassifier	0.904	0.885	0.902	0.954	0.89	0.927
CatBoostRegressor	0.897	0.871	0.884	0.996	0.876	0.923
CatBoostClassifier	0.897	0.871	0.884	0.996	0.876	0.923
RandomForestRegressor	0.897	0.866	0.876	0.997	0.871	0.924
RandomForestClassifier	0.89	0.86	0.875	0.966	0.867	0.918

Table 18. Metrics of Consensus Mutations datset for the Models: XGBRegressor, XGBClassifier, CatBoostRegressor, CatBoostClassifier, RandomForestRegressor and RandomForestClassifier

These results indicate that the XGBClassifier model is the most effective in predicting the pathogenicity of missense variants in the AMT protein, according to the results. This model had the highest F1-Score (0.927), accuracy (0.904), and AUC-ROC (0.885), which measures the ability of the model to distinguish between pathogenic and non-pathogenic mutations, making it a good overall performance metric, especially for imbalanced datasets[163].

However, understanding why these features are important and how they interact with each other might require more in-depth analysis, such as exploring partial dependence plots or using more advanced interpretability techniques such as SHapley Additive exPlanations (SHAP)[163]. The SHAP is a unified measure of feature importance that allocates the contribution of each feature Figure 42 A to the prediction for each instance. The SHAP values provide insights into the extent to which each feature changes its output from the baseline prediction. This useful to interpret how the protein features you have computed Figure 42 A, like 'Feature_probs, contribute to the final prediction. XGBoost, on the other hand, is an optimized distributed gradient boosting library designed to be highly efficient and flexible. The XGBClassifier is a specific implementation of the gradient boosting method for classification tasks. Together, the use of SHAP with XGB Classifier helps uncover the contributions of each feature to the predictions of the model, thereby elucidating how different properties of the protein sequences (such as attention contacts) and their interactions contribute to the prediction of the effects of mutations.

It is important to understand the data distribution in Figures 42 A, B, and C because the model's predictions and interpretations are dependent on it. Biases or imbalances in the training data can lead to skewed predictions and interpretations. Therefore, it is important to explore and understand the distribution of the training data when interpreting the model.

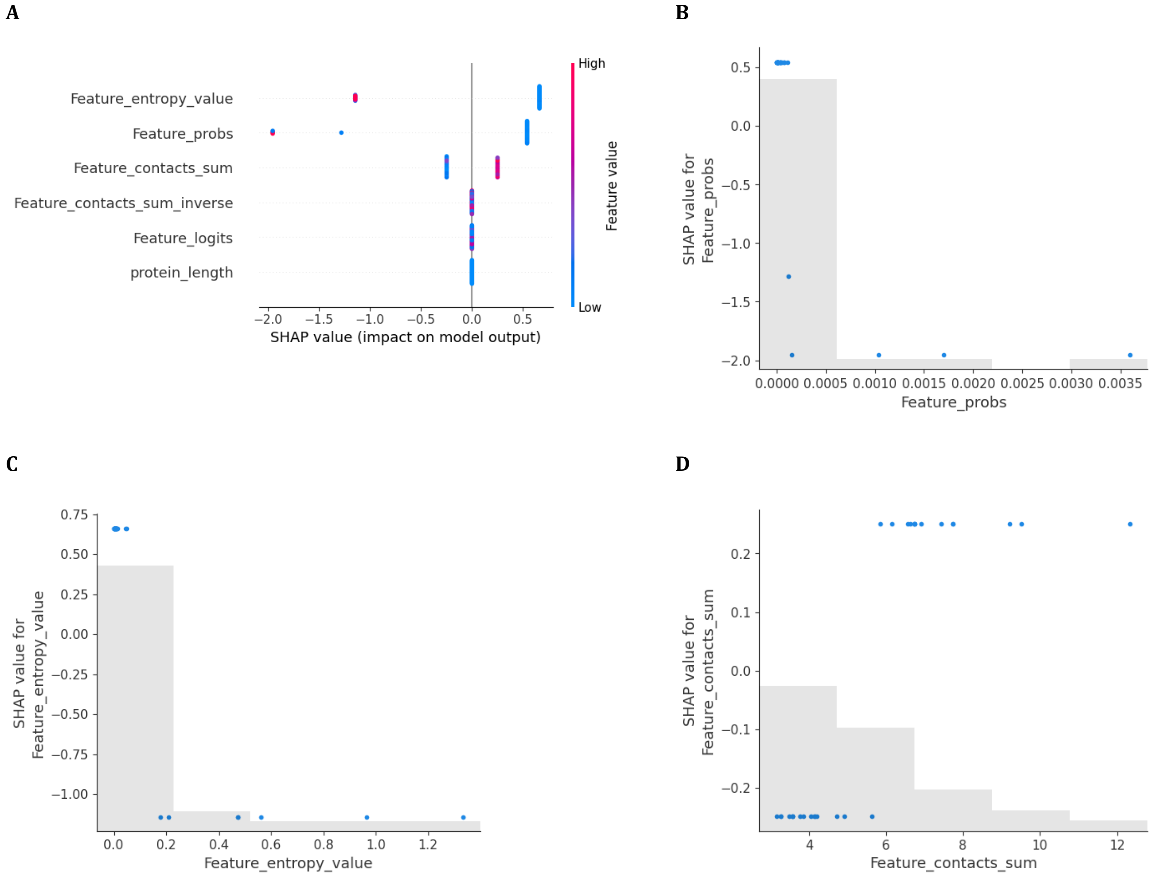


Figure 42. SHAP Impact Features on the XGBClassifier Model (A), Distribution of the Probabilities values on Training stage (B), Distribuion of the Entropy Values on Training stage (C) and Contacts Sum on the Training stage (D)

The details of the ML results for the test dataset using the XGBClassifier model can be found in Appendix 23. Detailed information regarding the ML process for the consensus dataset using the XGBClassifier model can be found in Appendix 24. As shown in Table 18, the RandomForestRegressor model was the second most effective in predicting the pathogenicity of missense variants in the AMT protein. This model has an F1-Score (0.924), accuracy (0.897), and AUC-ROC (0.866), which measures the model's ability to classify the outcome with continuous values, making it a good predictor of pathogenicity severity. We tested the ML for the full protein length and for all amino acid substitutions and obtained the heatmap in the Figure 43, which shows the Variant Pathogenicity Prediction using the RandomForestRegressor model.

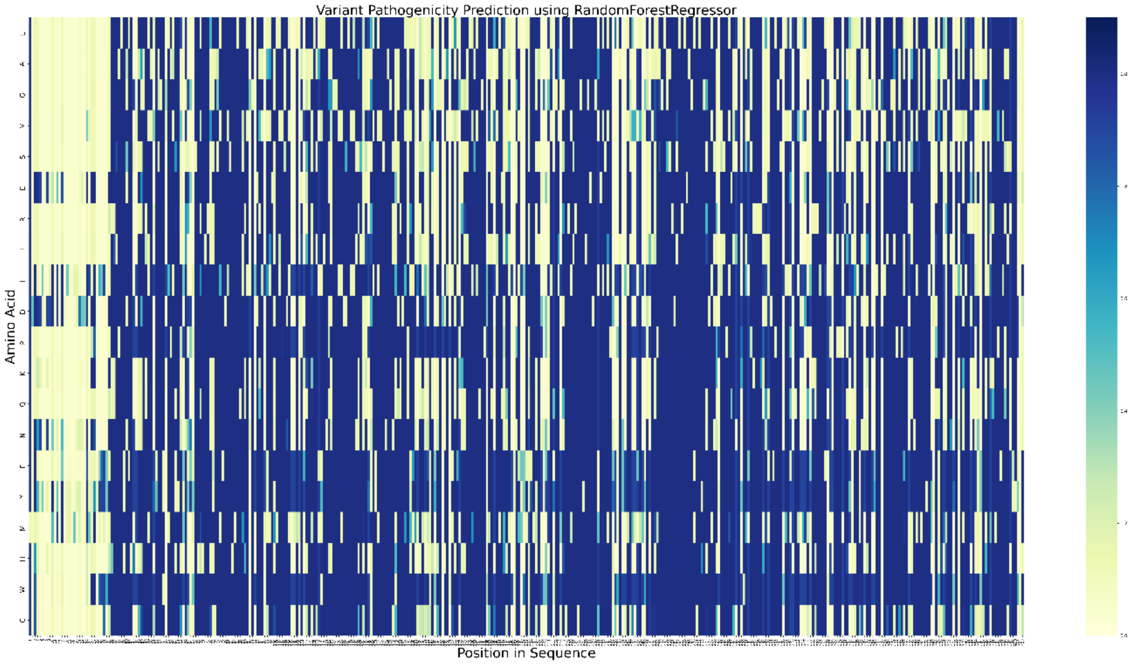


Figure 43. Variant Pathogenicity Prediction using the RandomForestRegressor model for T-protein

4.10 Assessment of COS in NKH patient due T-protein mutations

We reviewed 47 patient records from 14 publications over the years to identify the severity of neurobehavioral disease Appendix 25. In the assessment of patient COS, we did not assess COS for records where patients had died, because death can occur due to a single acute event that may not reflect multi-symptom disease severity, corresponding to 4 records. As shown in Appendix 26, four major domains emerged (in order of frequency): brain malformations, seizures, cognitive disorders, and muscle and movement dysfunctions. Most of these symptoms were predominantly observed in pediatric patients and were associated with a range of mutations (Appendix 26). Of the 43 patients, 10 had severe disease (COS>5), 6 had attenuated intermediate disease (COS ≤ 5 and >COS=1), and 2 had attenuated good disease (COS=1). 29 patients did not have sufficient information for scoring.

5 CONCLUSION AND FUTURE PERSPECTIVES

Reduced activity of the T-protein is caused by loss-of-function mutations in AMT. Several levels of reduced activity have been associated with attenuated or severe NKH phenotypes. Despite some progress, more research is needed to obtain a comprehensive understanding of genotype-phenotype interrelationships in NKH. Thus, this study performed a large-scale computational analysis of mutational effects on protein function by investigating a set of T-protein features generated from protein sequence, structure, and binding analyses and their relationship with NKH disease.

In this study, we collected a total a collection of 335 unique T-protein mutations, from four data sources ClinVar, SimpleClinVar, PubTator, and a key journal article published in Genetics in Medicine[6]. This collection contains 149 protein missense mutations, 57 frameshift mutations, 19 stop-gain mutations, 6 indels, 12 duplications, 5 deletions, with the remaining being splice-D/A and intronic DNA mutations.

Out of the 335 mutations, 131 had clinical significance that was either uncertain or conflicting. Properly determining the clinical relevance of these mutations is crucial for the accurate diagnosis and treatment of genetic disorders. To determine the clinical significance of these mutations, we adopted a consensus approach. This utilized five high-performance software tools (LYRUS, PROVEAN, SNAP2, MUTAFRAME, and SIFT) known for their proficiency in predicting the clinical significance of missense mutations, especially SAV mutations. For duplications, insertions, and deletions (indels) with uncertain/conflicting classifications, we used PROVEAN and SIFT. Lastly, the pathogenicity of frameshift and stop-gain mutations was predicted using SIFT.

For future research we can incorporate a broader range of software tools for a more comprehensive consensus approach, especially for duplications, insertions, deletions (indels), frameshift, and stop-gain mutations. Currently, we only employ SIFT for these.

Upon obtaining the list of pathogenic mutations, we distributed them per domain and predicted their functional sites via InterPro[193]webserver. The primary regions included: the signal peptide region, COILS Coil, Aminomethyltransferase folate-binding domain (GCV_T_N), and the Glycine cleavage T-protein C-terminal barrel domain (GCV_T_C). Notably, most of mutations are situated within the GCV_T_N domain.

All mutations previously classified as pathogenic in this study were compiled into a comprehensive list detailing T-protein NKH-causing disease mutations. We then ranked the ten most prevalent AMT mutations causing the disease. The mutations most frequently observed were R320H (with 36 frequency), followed by R222C (12 frequency), R94W (9 frequency), R73C (8 frequency), and M1T (7 frequency).

Furthering we selected 5 key mutations: R320H[3], R73C[3], H83R[3], R222C[3], and E211K to conduct MD simulations, aiming to learn insights of the structure-function relationship of the WT T-protein. These mutations represent a spectrum of disease severity: R320H is severe, R222C[209, 210] is mild, H83R[211] is moderate, and E211K[208] is seen as a polymorphism with benign nature. The mutations R320H, R73C, and R222C are among the four most recurrent NKH-inducing mutations of the survey performed in this study. The article published in Genetics in Medicine[6] also ranked them as the most frequent NKH-causing mutations. For our simulations, we utilized four systems: T-protein WT, one homozygote (H83R and H83R), and two heterozygotes (R73C and R320H, and R222C and E211K). We executed three replicas of MD simulations with a run distance of 500 ns for each system.

Our MD simulations were focused on studying the two main regions of the T-protein: firstly, the region proximal to the cofactor THH ligand, and secondly, the interactions between the T-protein chains A and B.

It was used a set of analytical techniques to help to understand the molecular consequences of these mutations, namely, Density Distribution of RMSD, H-bond Analysis, RMSF Analysis, Cross-Correlation of Residues, MMPBSA Analysis and finally Dynamic Contact Extraction from MD Simulations.

Density Distribution of RMSD helped to determine structural deviations in the mutated proteins compared to the WT T-protein, specifically focusing on regions within 5 Å to the THH ligand. The WT Chain A primarily showed RMSD peak densities in its β-sheet regions, especially notable in β-sheet 11. Chain B of the WT, however, demonstrated Increased RMSD variability, with elevated peaks in β-sheet 5 and α-helix 8. The H83R mutant in Chain A shows minimum RMSD deviations. However, in Chain B an elevated RMSD peak in β-sheet 11 and α-helix 9 comparing it with WT. The R222C mutant in Chain A showed elevated RMSD in α-helix 8 comparing it WT. The R73C mutant on Chain A has the β-sheet 5 as the most deviated region compared to the WT. The R320H mutant in Chain B, suggested and elevated in α-helix 8.

Lastly, the E211K mutant in Chain B displayed deviations, especially in α-helix 8 and β-sheet 5, from the WT.

ΔΔG analysis highlighted distinct residues interactions across different mutants compared to the WT. In the WT, LEU:33 has a pronounced binding contribution, which is absent in mutants. The H83R mutation due to their high ΔG values activates residues previously dormant in the WT, such as ASP:152, GLU:162, and GLU:173. The R320H mutation notably affects residues Chain B's GLN:164. Residue ARG:35 in R222C (Chain A) mutant indicates binding shifts, contrasting with THR: 36 which showed increased ΔG values in Chain B E211K.

MMPBSA Analysis provided clarity on alterations in stability and binding affinity with the THH ligand and interactions between the two chains.

Dynamic Contact Extraction from MD Simulations revealed the frequency of H-bond contacts between chains per residue for both the WT T-protein and its mutations. Dynamic Contact Extraction analysis showed H83R mutation seems to augment hydrogen bond interactions, both in terms of quantity and strength comparing with R73C, R320H, R222C, E211K and WT, particularly between residues B:ARG:34 of Chain A and A:ASP:158 of Chain B, indicating the H83R mutation enhances hydrogen bonding frequency. R73C and R320H H-Bonds interactions appear less dense than WT. Yet, the interactions between residues B:ARG:34 and A:ASP:158 remain consistent in both systems.

Additionally, we detailed the frequency of H-bond contacts between the T-protein per residue and the THH ligand during the MD simulation trajectories. This analysis revealed that WT GLU:232 in both chains consistently bonds with THH. The interaction between CYS:271 and THH in Chain A also stands out. R222C primary interacting residues are GLY:224 and ARG:261. R73C THH H-bonds form primarily with GLU:232, GLY:224, ARG:261, and ARG:222.

H83R TYR:225 in Chain A and ARG:261 in both chains show significant interactions with the ligand THH.

In this study we also explored machine learning approaches to predict the pathogenicity of T-protein mutations. We implemented and evaluated a methodology that compared six popular machine learning algorithms: XGB Classifier, XGB Regressor, Catboost Classifier, Catboost Regressor, Random Forest Classifier, and Random Forest Regressor that have been applied before with good pathogenicity predictions results[212,213].These algorithms were assessed based on their accuracy, F-measure, and AUC-ROC metrics. One key objective of our research was to identify features that enhance discriminative power and improve predictive performance regarding the severity of T-protein mutations. Our methodology focused on predicting the binary outcomes of disease presence (benign or pathological). The XGB

Classifier outperformed the other classifiers and regressors methods, achieving the best performance against the consensus mutations list with an F1-Score of 0.927, accuracy of 0.904, and AUC-ROC of 0.885. In future work, we aim to refine our machine learning approach to predict more nuanced severity levels of diseases, categorizing them as benign, attenuated good, attenuated intermediate, and severe.

One limitation we encountered was with our dataset. It was not only limited in size but also showed an imbalance, particularly with too few benign mutations. This directly impacted our feature engineering options. The restricted data obligated careful selection and utilization of the features, ensuring that the machine learning models predictive efficacy.

For upcoming research, it is necessary to amplify the datasets used for both training and testing. There's a need to include more benign mutations and to consider mutations in the T-protein sequence less pathogenic. This will potentially improve the algorithm and facilitate distinguishing between severity levels.

Another interesting work would be to evaluate if the ML model can accurately predict the COS of the clinical cases presented in this study. Further, we should determine its ability in forecasting the condition of patients categorized as attenuated good, attenuated intermediate, and severe. Such a machine learning model could serve as an invaluable tool, especially when a new patient is diagnosed with NKH disease. It would not only aid in predicting the disease progression but also provide insights into the potential outcomes for the patient.

6 REFERENCES

1. Homa Adle-Biassette, B. N. H. J. A. G. *Developmental Neuropathology*.
2. Hennermann, J. B., Berger, J.-M., Grieben, U., Scharer, G. & Van Hove, J. L. K. Prediction of long-term outcome in glycine encephalopathy: a clinical survey. *J Inherit Metab Dis* **35**, 253–261 (2012).
3. Swanson, M. A. *et al.* Biochemical and molecular predictors for prognosis in nonketotic hyperglycinemia. *Ann Neurol* **78**, 606–618 (2015).
4. Hoover-Fong, J. E. *et al.* Natural history of nonketotic hyperglycinemia in 65 patients. *Neurology* **63**, 1847–1853 (2004).
5. Dulac, O. Epileptic encephalopathy with suppression-bursts and nonketotic hyperglycinemia. in 1785–1797 (2013). doi:10.1016/B978-0-444-59565-2.00048-4.
6. Coughlin, C. R. *et al.* The genetic basis of classic nonketotic hyperglycinemia due to mutations in GLDC and AMT. *Genetics in Medicine* **19**, 104–111 (2017).
7. Zhang, H., Li, Y., Nie, J., Ren, J. & Zeng, A.-P. Structure-based dynamic analysis of the glycine cleavage system suggests key residues for control of a key reaction step. *Commun Biol* **3**, 756 (2020).
8. Kojima-Ishii, K. *et al.* Model Mice for Mild-Form Glycine Encephalopathy: Behavioral and Biochemical Characterizations and Efficacy of Antagonists for the Glycine Binding Site of N-Methyl D-Aspartate Receptor. *Pediatr Res* **64**, 228–233 (2008).
9. Okamura-Ikeda, K. *et al.* Crystal structure of human T-protein of glycine cleavage system at 2.0 A resolution and its implication for understanding non-ketotic hyperglycinemia. *J Mol Biol* **351**, 1146–59 (2005).
10. Markosian, C. *et al.* Analysis of impact metrics for the Protein Data Bank. *Sci Data* **5**, 180212 (2018).
11. Leung, K.-Y., De Castro, S. C. P., Galea, G. L., Copp, A. J. & Greene, N. D. E. Glycine Cleavage System H Protein Is Essential for Embryonic Viability, Implying Additional Function Beyond the Glycine Cleavage System. *Front Genet* **12**, (2021).
12. Zhang, X., Li, M., Xu, Y., Ren, J. & Zeng, A.-P. Quantitative study of H protein lipoylation of the glycine cleavage system and a strategy to increase its activity by co-expression of LplA. *J Biol Eng* **13**, 32 (2019).
13. Tan, Y.-L. *et al.* Tracing Metabolic Fate of Mitochondrial Glycine Cleavage System Derived Formate In Vitro and In Vivo. *Int J Mol Sci* **21**, 8808 (2020).
14. KIKUCHI, G., MOTOKAWA, Y., YOSHIDA, T. & HIRAGA, K. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proceedings of the Japan Academy, Series B* **84**, 246–263 (2008).
15. Döring, V., Darii, E., Yishai, O., Bar-Even, A. & Bouzon, M. Implementation of a Reductive Route of One-Carbon Assimilation in *Escherichia coli* through Directed Evolution. *ACS Synth Biol* **7**, 2029–2036 (2018).
16. Bang, J. & Lee, S. Y. Assimilation of formic acid and CO $_2$ by engineered *Escherichia coli* equipped with reconstructed one-carbon assimilation pathways. *Proceedings of the National Academy of Sciences* **115**, (2018).
17. Yishai, O., Bouzon, M., Döring, V. & Bar-Even, A. *In Vivo* Assimilation of One-Carbon *via* a Synthetic Reductive Glycine Pathway in *Escherichia coli*. *ACS Synth Biol* **7**, 2023–2028 (2018).
18. Tashiro, Y., Hirano, S., Matson, M. M., Atsumi, S. & Kondo, A. Electrical-biological hybrid system for CO2 reduction. *Metab Eng* **47**, 211–218 (2018).

19. Bar-Even, A. Formate Assimilation: The Metabolic Architecture of Natural and Synthetic Pathways. *Biochemistry* **55**, 3851–3863 (2016).

20. Bar-Even, A., Noor, E., Lewis, N. E. & Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences* **107**, 8889–8894 (2010).

21. Eswar, N. *et al.* Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci* **50**, (2007).

22. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, (2016).

23. Rosário-Ferreira, N. *et al.* In Silico End-to-End Protein-Ligand Interaction Characterization Pipeline: The Case of SARS-CoV-2. *ACS Synth Biol* **10**, (2021).

24. 1wsv pdb. *https://doi.org/10.2210/pdb1WSV/pdb*.

25. 1wsr pdb. *https://doi.org/10.2210/pdb1WSR/pdb*.

26. 1woo pdb. *https://doi.org/10.2210/pdb1WOO/pdb*.

27. 1wop pdb. *https://doi.org/10.2210/pdb1WOP/pdb*.

28. 1wor pdb. *https://doi.org/10.2210/pdb1WOR/pdb*.

29. Fiser, A. Template-Based Protein Structure Modeling. in 73–94 (2010). doi:10.1007/978-1-60761-842-3_6.

30. Meier, A. & Söding, J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Comput Biol* **11**, e1004343 (2015).

31. Shen, M. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Science* **15**, 2507–2524 (2006).

32. Reva, B. A., Finkelstein, A. V & Skolnick, J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? *Fold Des* **3**, 141–147 (1998).

33. Chowdhury, Md. H. U., Das, T. & Sikdar, S. In-silico structural insights of Dengue 4 NS3 protease: homology modeling and structural validation. *Journal of Phytomolecules and Pharmacology* **2**, 12–20 (2023).

34. Callaway, E. 'The entire protein universe': AI predicts shape of nearly every known protein. *Nature* **608**, 15–16 (2022).

35. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

36. PyMOL. *http://www.pymol.org*.

37. Durrant, J. D. & McCammon, J. A. BINANA: A novel algorithm for ligand-binding characterization. *J Mol Graph Model* **29**, (2011).

38. Hassan, M. M., Mogollón, D. C., Fuentes, O. & Sirimulla, S. DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities. *ChemRxiv* (2018).

39. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, (2021).

40. Lai, J., Yang, J., Gamsiz Uzun, E. D., Rubenstein, B. M. & Sarkar, I. N. LYRUS: a machine learning model for predicting the pathogenicity of missense variants. *Bioinformatics Advances* **2**, (2022).

41. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7**, e46688 (2012).

42. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, (2015).

43. Ancien, F., Pucci, F., Vranken, W. & Rooman, M. MutaFrame—an interpretative

visualization framework for deleteriousness prediction of missense variants in the human exome. *Bioinformatics* **38**, (2021).

44. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).

45. Ng, P. C. & Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res* **11**, 863–874 (2001).

46. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. E. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol* **426**, (2014).

47. Ben Chorin, A. *et al.* ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Science* **29**, (2020).

48. Celniker, G. *et al.* ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry* vol. 53 Preprint at https://doi.org/10.1002/ijch.201200096 (2013).

49. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**, 646–652 (2002).

50. Zhang, J., Zhang, H., Wu, T., Wang, Q. & van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *J Chem Theory Comput* **13**, 1034–1043 (2017).

51. Anandakrishnan, R., Drozdetski, A., Walker, R. C. & Onufriev, A. V. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophys J* **108**, 1153–1164 (2015).

52. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

53. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys* **112**, 8910–8922 (2000).

54. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J Phys Chem* **91**, 6269–6271 (1987).

55. Salo-Ahen, O. M. H. *et al.* Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. *Processes* **9**, 71 (2020).

56. Takada, S. Coarse-grained molecular simulations of large biomolecules. *Curr Opin Struct Biol* **22**, 130–137 (2012).

57. Kmiecik, S. *et al.* Coarse-Grained Protein Models and Their Applications. *Chem Rev* **116**, 7898–7936 (2016).

58. Hammonds, K. D. & Heyes, D. M. Shadow Hamiltonian in classical NVE molecular dynamics simulations involving Coulomb interactions. *J Chem Phys* **154**, 174102 (2021).

59. Narumi, T., Susukita, R., Ebisuzaki, T., McNiven, G. & Elmegreen, B. Molecular Dynamics Machine: Special-Purpose Computer for Molecular Dynamics Simulations. *Mol Simul* **21**, 401–415 (1999).

60. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* **112**, 6127–6129 (1990).

61. Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J Phys Chem A* **101**, 3005–3014 (1997).

62. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a

Dielectric Medium. *J Phys Chem* **100**, 19824–19839 (1996).

63. Schaefer, M. & Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J Phys Chem* **100**, 1578–1599 (1996).

64. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **55**, 383–394 (2004).

65. Padhi, A. K., Janežič, M. & Zhang, K. Y. J. Molecular dynamics simulations: Principles, methods, and applications in protein conformational dynamics. in *Advances in Protein Molecular and Structural Biology Methods* 439–454 (Elsevier, 2022). doi:10.1016/B978-0-323-90264-9.00026-X.

66. Collier, T. A., Piggot, T. J. & Allison, J. R. Molecular Dynamics Simulation of Proteins. in 311–327 (2020). doi:10.1007/978-1-4939-9869-2_17.

67. Patodia, S. Molecular Dynamics Simulation of Proteins: A Brief Overview. *J Phys Chem Biophys* **4**, (2014).

68. Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated minimization of steric clashes in protein structures. *Proteins: Structure, Function, and Bioinformatics* **79**, 261–270 (2011).

69. https://manual.gromacs.org/current/reference-manual/algorithms/energy-minimization.html. Gromacs Manual Reference.

70. Rowlinson *, J. S. The Maxwell–Boltzmann distribution. *Mol Phys* **103**, 2821–2828 (2005).

71. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* **19**, 120–127 (2009).

72. Adcock, S. A. & McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem Rev* **106**, 1589–1615 (2006).

73. Li, Y., Xu, J. & Li, D. Molecular dynamics simulation of nanoscale liquid flows. *Microfluid Nanofluidics* **9**, 1011–1031 (2010).

74. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73 (2017).

75. Foloppe, N. & MacKerell, Jr. , A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem* **21**, 86–104 (2000).

76. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **24**, 1999–2012 (2003).

77. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* **117**, 5179–5197 (1995).

78. Kollman, P. A. Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc Chem Res* **29**, 461–469 (1996).

79. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* **25**, 1656–1676 (2004).

80. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J Phys Chem B* **105**, 6474–6487 (2001).

81. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J Phys Chem B* **111**,

7812–7824 (2007).

82. Maisuradze, G. G., Senet, P., Czaplewski, C., Liwo, A. & Scheraga, H. A. Investigation of Protein Folding by Coarse-Grained Molecular Dynamics with the UNRES Force Field. *J Phys Chem A* **114**, 4471–4485 (2010).

83. Tsui, V. & Case, D. A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* **56**, 275–291 (2000).

84. Wagoner, J. & Baker, N. A. Solvation forces on biomolecular structures: A comparison of explicit solvent and Poisson-Boltzmann models. *J Comput Chem* **25**, 1623–1629 (2004).

85. González, M. A. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* **12**, 169–200 (2011).

86. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

87. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).

88. Acun, B. *et al.* Scalable molecular dynamics with NAMD on the Summit system. *IBM J Res Dev* **62**, 4:1-4:9 (2018).

89. Kumar, S., Seth, D. & Deshpande, P. A. Molecular dynamics simulations identify the regions of compromised thermostability in <scp>SazCA</scp>. *Proteins: Structure, Function, and Bioinformatics* **89**, 375–388 (2021).

90. Patel, S., Mackerell, A. D. & Brooks, C. L. CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J Comput Chem* **25**, 1504–1514 (2004).

91. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

92. Dequidt, A., Devémy, J. & Pádua, A. A. H. Thermalized Drude Oscillators with the LAMMPS Molecular Dynamics Simulator. *J Chem Inf Model* **56**, 260–268 (2016).

93. http://www.ks.uiuc.edu/Research/vmd/. VMD web site.

94. Chimera web site. *https://www.cgl.ucsf.edu/chimera/.*

95. https://www.chemcomp.com/Products.htm. MOE web site.

96. https://www.schrodinger.com/products/maestro. Maestro web site.

97. Jo, S. *et al.* CHARMM-GUI 10 years for biomolecular modeling and simulation. *J Comput Chem* **38**, 1114–1124 (2017).

98. https://www.charmm-gui.org/. CHARMM-GUI web site.

99. Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* **12**, 405–413 (2016).

100. Bentrop, D. *et al.* Three-Dimensional Structure of the Reduced C77S Mutant of the *Chromatium vinosum* High-Potential Iron−Sulfur Protein through Nuclear Magnetic Resonance: Comparison with the Solution Structure of the Wild-Type Protein ′. *Biochemistry* **35**, 5928–5936 (1996).

101. Wang, C., Greene, D., Xiao, L., Qi, R. & Luo, R. Recent Developments and Applications of the MMPBSA Method. *Front Mol Biosci* **4**, (2018).

102. Wang, F., Wan, H., Hu, J. & Chang, S. Molecular dynamics simulations of wild type and mutants of botulinum neurotoxin A complexed with synaptic vesicle protein 2C. *Mol Biosyst* **11**, 223–231 (2015).

103. Wang, D. D., Ou-Yang, L., Xie, H., Zhu, M. & Yan, H. Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine

learning methods. *Comput Struct Biotechnol J* **18**, 439–454 (2020).

104. Piao, L. *et al.* Molecular Dynamics Simulations of Wild Type and Mutants of SAPAP in Complexed with Shank3. *Int J Mol Sci* **20**, 224 (2019).

105. Fukuyoshi, S. *et al.* Molecular Dynamics Simulations to Investigate the Influences of Amino Acid Mutations on Protein Three-Dimensional Structures of Cytochrome P450 2D6.1, 2, 10, 14A, 51, and 62. *PLoS One* **11**, e0152946 (2016).

106. Immadisetty, K., Alenciks, J. & Kekenes-Huskey, P. M. Modulation of P2X4 pore closure by magnesium, potassium, and ATP. *Biophys J* **121**, 1134–1142 (2022).

107. Gromacs. gmx-trjcat manual. *https://manual.gromacs.org/documentation/2018/onlinehelp/gmx-trjcat.html*.

108. get dynamic contacts GIT. *https://github.com/getcontacts/getcontacts/blob/master/get_dynamic_contacts.py*.

109. get contact frequencies GIT. *https://github.com/getcontacts/getcontacts/blob/master/get_contact_frequencies.py*.

110. Google Colab website. *https://colab.google/*.

111. Reeb, J., Wirth, T. & Rost, B. Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* **21**, 107 (2020).

112. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun* **12**, 510 (2021).

113. Chandra, A., Tünnermann, L., Löfstedt, T. & Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife* **12**, (2023).

114. Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences* **115**, 4164–4169 (2018).

115. Simple Clinvar website. *https://simple-clinvar.broadinstitute.org/*.

116. Clinvar website. *https://www.ncbi.nlm.nih.gov/clinvar/*.

117. Towardsdatascience website. Evaluation metrics. *https://towardsdatascience.com/the-f1-score-bec2bbc38aa6*.

118. Qureshi, A. A. *et al.* Performance evaluation of machine learning models on large dataset of android applications reviews. *Multimed Tools Appl* (2023) doi:10.1007/s11042-023-14713-6.

119. Jeremyjordan website. Evaluating a machine learning model. *https://www.jeremyjordan.me/evaluating-a-machine-learning-model/*.

120. Towardsdatascience website. A (Much) Better Approach to Evaluate Your Machine Learning Model. *https://towardsdatascience.com/good-approach-to-evaluate-your-machine-learning-model-e2e1fd6aa6bb*.

121. Huggingface website. esm2_t33_650M_UR50D. *https://huggingface.co/facebook/esm2_t33_650M_UR50D*.

122. Github facebookresearch. ESM. *https://github.com/facebookresearch/esm*.

123. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)* **379**, 1123–1130 (2023).

124. pypi.org website. fair-esm. *https://pypi.org/project/fair-esm*.

125. Towardsdatascience website. How Huge Protein Language Models Could Disrupt Structural Biology. *https://towardsdatascience.com/how-huge-protein-language-models-could-disrupt-structural-biology-6b98193f880b*.

126. Ramith website. ESM-2 (evolutionary-scale prediction of atomic level protein structure

with a language model). *https://ramith.fyi/esm-2-evolutionary-scale-prediction-of-atomic-level-protein-structure-with-a-language-model/*.

127. ESMAtlas website. *https://esmatlas.com/*.
128. ai.meta.com blog. *https://ai.meta.com/blog/protein-folding-esmfold-metagenomics/*.
129. Huggingface ESM model documentation. *https://huggingface.co/docs/transformers/model_doc/esm*.
130. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, (2021).
131. Protein Sequence Design by Entropy-based Iterative Refinement. *https://www.biorxiv.org/content/biorxiv/early/2023/02/04/2023.02.04.527099.full.pdf* (2023).
132. Wikipedia website. Entropy_(information_theory). *https://en.wikipedia.org/wiki/Entropy_(information_theory)*.
133. SciPy website. Entropy. *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html*.
134. Sun, J., Yu, L., Dong, P., Lu, B. & Zhou, B. Adversarial Inverse Reinforcement Learning With Self-Attention Dynamics Model. *IEEE Robot Autom Lett* **6**, 1880–1886 (2021).
135. Wikipedia website. Protein contact map. *https://en.wikipedia.org/wiki/Protein_contact_map*.
136. Chen, C., Wu, T., Guo, Z. & Cheng, J. Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. *Proteins: Structure, Function, and Bioinformatics* **89**, 697–707 (2021).
137. SciPy Softmax. *https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.softmax.html*.
138. Zeng, X. *et al.* Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat Mach Intell* **4**, 1004–1016 (2022).
139. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, (2021).
140. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
141. Ng, P. C. & Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu Rev Genomics Hum Genet* **7**, 61–80 (2006).
142. Gou, X. *et al.* PPVED: A machine learning tool for predicting the effect of single amino acid substitution on protein function in plants. *Plant Biotechnol J* **20**, 1417–1431 (2022).
143. ElAbd, H. *et al.* Amino acid encoding for deep learning applications. *BMC Bioinformatics* **21**, 235 (2020).
144. Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat Methods* **19**, 730–739 (2022).
145. Meier J, R. R. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems. *Protein Science* **29**, 247–257 (2021).
146. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344–W350 (2016).

147. Seaborn website. *https://pypi.org/project/seaborn/*.

148. Seaborn website. Heatmap. *https://seaborn.pydata.org/generated/seaborn.heatmap.html*.

149. SciPy documentation. *https://docs.scipy.org/doc/scipy/index.html*.

150. Towardsdatascience website. Protein Sequence Classification. *https://towardsdatascience.com/protein-sequence-classification-99c80d0ad2df*.

151. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).

152. Gupta, A. *et al.* MP4: a machine learning based classification tool for prediction and functional annotation of pathogenic proteins from metagenomic and genomic datasets. *BMC Bioinformatics* **23**, 507 (2022).

153. Jiang, R., Shang, L., Wang, R., Wang, D. & Wei, N. Machine Learning Based Prediction of Enzymatic Degradation of Plastics Using Encoded Protein Sequence and Effective Feature Representation. *Environ Sci Technol Lett* **10**, 557–564 (2023).

154. Diaz, D. J., Kulikova, A. V., Ellington, A. D. & Wilke, C. O. Using machine learning to predict the effects and consequences of mutations in proteins. *Curr Opin Struct Biol* **78**, 102518 (2023).

155. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J Chem Inf Model* **60**, 2773–2790 (2020).

156. Horne, J. & Shukla, D. Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. *Ind Eng Chem Res* **61**, 6235–6245 (2022).

157. Sun, T., Chen, Y., Wen, Y., Zhu, Z. & Li, M. PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. *Commun Biol* **4**, 1311 (2021).

158. Alves, P. *et al.* Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis. *Mult Scler J Exp Transl Clin* **8**, 205521732211086 (2022).

159. Horne, J. & Shukla, D. Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. *Ind Eng Chem Res* **61**, 6235–6245 (2022).

160. Dunham, A. S., Beltrao, P. & AlQuraishi, M. High-throughput deep learning variant effect prediction with Sequence UNET. *Genome Biol* **24**, 110 (2023).

161. Wu, T.-H., Lin, P.-C., Chou, H.-H., Shen, M.-R. & Hsieh, S.-Y. Pathogenicity Prediction of Single Amino Acid Variants with Machine Learning Model Based on Protein Structural Energies. *IEEE/ACM Trans Comput Biol Bioinform* 1–1 (2021) doi:10.1109/TCBB.2021.3139048.

162. XGBoost website. *https://xgboost.readthedocs.io/en/stable/*.

163. Schmidt, A. *et al.* Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *Bioinformatics* **39**, (2023).

164. CatBoost website. *https://catboost.ai/*.

165. Pandurangan, A. P. & Blundell, T. L. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Science* **29**, 247–257 (2020).

166. Pellegrino, E. *et al.* Machine learning random forest for predicting oncosomatic variant NGS analysis. *Sci Rep* **11**, 21820 (2021).

167. Li, Y. & Fang, J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. *PLoS One* **7**, e47247 (2012).

168. Shuyu Wang and Hong-ling Tang and Peng Shan and Lei Zuo. ProS-GNN: Predicting effects of mutations on protein stability using graph neural networks. *bioRxiv* (2021).

169. Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput Biol* **16**, e1008291 (2020).

170. Zhou, Y., Pan, Q., Pires, D. E. V, Rodrigues, C. H. M. & Ascher, D. B. DDMut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Res* **51**, W122–W128 (2023).

171. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat Commun* **14**, 2787 (2023).

172. SVDS website. *https://www.svds.com/the-basics-of-classifier-evaluation-part-1/*.

173. Teng, S., Srivastava, A. K. & Wang, L. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* **11**, S5 (2010).

174. Machinelearningmastery website. *https://www.svds.com/the-basics-of-classifier-evaluation-part-1/*.

175. Semanticscholar website. *https://www.semanticscholar.org/paper/Classification-of-Fire-and-Smoke-Images-using-Tree-Reddy-K./0ddf22f41d4b2cb8ab89b1c8db8287a29fc46674*.

176. Deepchecks website. *https://deepchecks.com/f1-score-accuracy-roc-auc-and-pr-auc-metrics-for-models/*.

177. Farris, J., Calhoun, B., Alam, Md. S., Lee, S. & Haldar, K. Large scale analyses of genotype-phenotype relationships of glycine decarboxylase mutations and neurological disease severity. *PLoS Comput Biol* **16**, e1007871 (2020).

178. Pérez-Palma, E., Gramm, M., Nürnberg, P., May, P. & Lal, D. Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. *Nucleic Acids Res* **47**, (2019).

179. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, (2014).

180. Wei, C. H., Kao, H. Y. & Lu, Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* **41**, (2013).

181. molecular PDF. *https://salilab.org/modeller/tutorial/cryoem/assess.html#:~:text=There%20are%20various%20methods%20provided,in%20each%20generated%20PDB%20file.*

182. Mora Lagares, L. *et al.* Homology Modeling of the Human P-glycoprotein (ABCB1) and Insights into Ligand Binding through Molecular Docking Studies. *Int J Mol Sci* **21**, 4058 (2020).

183. Carugo, O. Statistical validation of the root-mean-square-distance, a measure of protein structural proximity. *Protein Engineering, Design and Selection* **20**, 33–37 (2007).

184. Argaman, N. & Makov, G. Density functional theory: An introduction. *Am J Phys* **68**, 69–79 (2000).

185. Banerjee, A., Saha, S., Tvedt, N. C., Yang, L.-W. & Bahar, I. Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods. *Curr Opin Struct Biol* **78**, 102517 (2023).

186. Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci Rep* **7**, 9313 (2017).

187. Lefter, M. *et al.* Mutalyzer 2: next generation HGVS nomenclature checker. *Bioinformatics* **37**, 2811–2817 (2021).

188. Pearson, W. R. [5] Rapid and sensitive sequence comparison with FASTP and FASTA. in 63–98 (1990). doi:10.1016/0076-6879(90)83007-V.

189. P48728. *https://www.uniprot.org/uniprotkb/P48728/entry*.

190. Reisch, C. R. *et al.* Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine bacteria. *Nature* **473**, 208–11 (2011).

191. Reisch, C. R., Moran, M. A. & Whitman, W. B. Bacterial Catabolism of Dimethylsulfoniopropionate (DMSP). *Front Microbiol* **2**, 172 (2011).

192. 2.1.2.10. *https://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=2.1.2.10*.

193. interpro. *https://www.ebi.ac.uk/interpro/*.

194. IPR028896. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR028896/*.

195. Schuller, D. J., Reisch, C. R., Moran, M. A., Whitman, W. B. & Lanzilotta, W. N. Structures of dimethylsulfoniopropionate-dependent demethylase from the marine organism Pelagabacter ubique. *Protein Sci* **21**, 289–98 (2012).

196. Teplyakov, A. *et al.* Crystal structure of the YgfZ protein from Escherichia coli suggests a folate-dependent regulatory role in one-carbon metabolism. *J Bacteriol* **186**, 7134–40 (2004).

197. Ote, T. *et al.* Involvement of the Escherichia coli folate-binding protein YgfZ in RNA modification and regulation of chromosomal replication initiation. *Mol Microbiol* **59**, 265–75 (2006).

198. IPR006223. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR006223/*.

199. Okamura-Ikeda, K. *et al.* Crystal structure of aminomethyltransferase in complex with dihydrolipoyl-H-protein of the glycine cleavage system: implications for recognition of lipoyl protein substrate, disease-related mutations, and reaction mechanism. *J Biol Chem* **285**, 18684–92 (2010).

200. IPR006222. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR006222/*.

201. McNeil, J. B. *et al.* Cloning, and molecular characterization of the GCV1 gene encoding the glycine cleavage T-protein from Saccharomyces cerevisiae. *Gene* **186**, 13–20 (1997).

202. IPR013977. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR013977/*.

203. Lokanath, N. K., Kuroishi, C., Okazaki, N. & Kunishima, N. Crystal structure of a component of glycine cleavage system: T-protein from Pyrococcus horikoshii OT3 at 1.5 A resolution. *Proteins* **58**, 769–73 (2005).

204. IPR029043. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR029043/*.

205. Scrima, A., Vetter, I. R., Armengod, M. E. & Wittinghofer, A. The structure of the TrmE GTP-binding protein and its implications for tRNA modification. *EMBO J* **24**, 23–33 (2005).

206. IPR027266. *https://www.ebi.ac.uk/interpro/entry/InterPro/IPR027266/*.

207. Douce, R., Bourguignon, J., Neuburger, M. & Rébeillé, F. The glycine decarboxylase system: a fascinating complex. *Trends Plant Sci* **6**, 167–76 (2001).

208. Toone, J. R., Applegarth, D. A., Coulter-Mackie, M. B. & James, E. R. Biochemical and Molecular Investigations of Patients with Nonketotic Hyperglycinemia. *Mol Genet Metab* **70**, 116–121 (2000).

209. Azize, N. A. A. *et al.* Mutation analysis of glycine decarboxylase, aminomethyltransferase and glycine cleavage system protein-H genes in 13 unrelated families with glycine encephalopathy. *J Hum Genet* **59**, 593–597 (2014).

210. Zaganas, I. *et al.* Genetic cause of epilepsy in a Greek cohort of children and young adults with heterogeneous epilepsy syndromes. *Epilepsy Behav Rep* **16**, 100477 (2021).

211. Veríssimo, C. *et al.* Nonketotic Hyperglycinemia. *J Child Neurol* **28**, 251–254 (2013).

212. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63**, 3–42

(2006).

213. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813 (2008).

214. Nanao, K. *et al.* Identification of the mutations in the T-protein gene causing typical and atypical nonketotic hyperglycinemia. *Hum Genet* **93**, (1994).

215. Toone, J. R., Applegarth, D. A., Coulter-Mackie, M. B. & James, E. R. Identification of the first reported splice site mutation (IVS7-1G?A) in the aminomethyltransferase (T-protein) gene (AMT) of the glycine cleavage complex in 3 unrelated families with nonketotic hyperglycinemia. *Hum Mutat* **17**, 76–76 (2001).

216. Swanson, M. A. *et al.* d-Glyceric aciduria does not cause nonketotic hyperglycinemia: A historic co-occurrence. *Mol Genet Metab* **121**, 80–82 (2017).

217. Toone, J. R., Applegarth, D. A., Levy, H. L., Coulter-Mackie, M. B. & Lee, G. Molecular genetic and potential biochemical characteristics of patients with T-protein deficiency as a cause of glycine encephalopathy (NKH). *Mol Genet Metab* **79**, 272–280 (2003).

218. Kure, S. *et al.* Comprehensive mutation analysis of *GLDC* , *AMT* , and *GCSH* in nonketotic hyperglycinemia. *Hum Mutat* **27**, 343–352 (2006).

219. Yilmaz, B. S. *et al.* Two Novel Missense Mutations in Nonketotic Hyperglycinemia. *J Child Neurol* **30**, 789–792 (2015).

220. Nonketotic hyperglycinemia: novel mutation in the aminomethyl transferase gene. Case report. *Arch Argent Pediatr* **114**, (2016).

221. Toone, J. R., Applegarth, D. A., Coulter-Mackie, M. B. & James, E. R. Recurrent Mutations in P- and T-Proteins of the Glycine Cleavage Complex and a Novel T-Protein Mutation (N145I): A Strategy for the Molecular Investigation of Patients with Nonketotic Hyperglycinemia (NKH). *Mol Genet Metab* **72**, 322–325 (2001).

222. Belcastro, V., Barbarini, M., Barca, S. & Mauro, I. A novel AMT gene mutation in a newborn with nonketotic hyperglycinemia and early myoclonic encephalopathy. *European Journal of Paediatric Neurology* **20**, 192–195 (2016).

7 APPENDIX

Appendix 1 – Detailed information on AMT crystal structures and their relevant residues.

Crystal Structure	Relevant Details	Relevant Information	Relevant residues
1WSV	**Organism(s)**	**Homo sapiens**	
	UniProtKB	**P48728**	
	Chain IDs	**A, B**	
	Sequence Lenght	**375**	
	X-Ray Resolution	**2.60 Å**	
	Number of Ligands	**2**	
	Ligand 1	**SO4 - SULFATE ION**	**12,34,37,43,172,214,226,260,262,278,27,291,303,353**
	Ligand 1 Name / **Formula /**	**SULFATE ION** **O4 S** **QAOWNCQODCNURD-UHFFFAOYSA-L**	

	InChI Key		
	Ligand 2	THH	56,88,102,103,115,117,176,177,196,197,204,233,242,262,371
	Ligand 2 Name / Formula / InChI Key	N-[4-({[(6S)-2-AMINO-4-HYDROXY-5-METHYL-5,6,7,8-TETRAHYDROPTERIDIN-6-YL]METHYL}AMINO)BENZOYL]-L-GLUTAMIC ACID C20 H25 N7 O6 ZNOVTXRBGFNYRX-STQMWFEESA-N	
1WSR	Organism(s)	1WSR	
	UniProtKB	P48728	
	Chain IDs	A, B	
	Sequence Lenght	375	
	X-Ray Resolution	2.00 Å	
	Number of Ligands	1	
	Ligand 1	SO4 - SULFATE ION	17,34,37,43,137,172,214,226,26,261,262,278,287,291,303,324,353
	Ligand 1 Name / Formula / InChI Key	SULFATE ION O4 S QAOWNCQODCNURD-UHFFFAOYSA-L	
1WOO	Organism(s)	Thermotoga maritima	
	UniProtKB	Q9WY54	
	Chain IDs	A	
	Sequence Lenght	364	
	X-Ray Resolution	2.40 Å	
	Number of Ligands	1	
	Ligand 1	ligand/ THG ((6S)-5,6,7,8-tetrahydrofolic acid)	51, 83,96, 98, 100, 168, 169,187, 188, 195, 227,236, 256,362
	Ligand 1 Name / Formula / InChI Key	(6S)-5,6,7,8-TETRAHYDROFOLATE C19 H23 N7 O6 MSTNYGQPCMXVAQ-RYUDHWBXSA-N	
	Organism(s)	Thermotoga maritima	
	UniProtKB	Q9WY54	
	Chain IDs	A	
	Sequence Lenght	364	
	X-Ray Resolution	2.00 Å	
	Number of Ligands	1	
	Ligand 1	ligand/ FFO (FOLINIC	55,83, 96, 98,100,110, 112,168,169, 188, 195, 227, 236, 256,362

1WOP		ACID = L-glutamic acid)	
	Ligand 1 Name / Formula / InChI Key	N-[4-({[(6S)-2-amino-5-formyl-4-oxo-3,4,5,6,7,8-hexahydropteridin-6-yl]methyl}amino)benzoyl]-L-glutamic acid C20 H23 N7 O7 VVIAGPKUTFNRDU-STQMWFEESA-N	
1WOR	Organism(s)	Thermotoga maritima	
	UniProtKB	Q9WY54	
	Chain IDs	A	
	Sequence Lenght	364	
	X-Ray Resolution	1.95 Å	
	Number of Ligands	1	
	Ligand 1	ligand/ RED (lipoic acid)	20,224,227, 228,
	Ligand 1 Name / Formula / InChI Key	DIHYDROLIPOIC ACID C8 H16 O2 S2	

Appendix 2 – Mutations in AMT gene in patients from Genetics in Medicine Article

Number	Origin	Protein	DNA mutation 1	Protein change	Parental origin M	DNA mutation 2	Protein change2	Parental origin P
B027		T	c.545T>C	p.L182P	Maternal	c.545T>C	p.L182P	Paternal
B029	UK	T	c.664C>T	p.R222C	Maternal	c.1-58C>T		
B055		T	c.551_552del2insAA	p.G184E	Unknown	c.551_552delins AA	p.G184E	Unknown
D025	USA	T	c.(1-55C>T;674A>G)	p.Y225C	Maternal	c.635T>C	p.V212A	Paternal
D041	USA	T	c.959G>A	p.R320H	Maternal	c.959G>A	p.R320H	Paternal
D073	USA	T	c.2T>C	p.M1T	Maternal	c.2T>C;p.M1T	p.M1T	Paternal
D075	USA	T	c.1034+2T>C	IVS8+2insT	Maternal	c.230C>T	p.S77L	Paternal
D137	USA	T	c.1107-1108delAC	p.Y369X	Maternal	c.1107-1108delAC	p.Y369X	Paternal
D140	NET	T	c.317T>C	p.I106T	Maternal	c.665G>C	p.R222H	Paternal
D142	USA	T	c.451_466del16	p.K151Cfs*25	Unknown	c.280C>T	p.R94W	Unknown
D157	USA	T	c.1107-1108delAC	p.Y369X	Maternal	c.280C>T	p.R94W	Paternal
D201	NWZ	T	c.794G>A	p.R265H	Maternal	c.794G>A	p.R265H	Paternal
D202	USA	T	c.1-64A>G		Maternal	c.959G>A	p.R320H	Paternal
D205	BRA	T	c.959G>A	p.R320H	Maternal	c.[217C>T;c.1082C>T]	p.[R73C;p.A361V]	Paternal
D228	USA	T	c.793C>A	p.R265S	Maternal	c.793C>A	p.R265S	Paternal
D231	USA	T	c.451_466del16	p.K151Cfs*25	Maternal	c.794G>A	p.R265H	Paternal
D236	ARA	T	c.664C>T	p.R222C	Maternal	c.664C>T	p.R222C	Paternal
D238	USA	T	c.230C>T	p.S77L	Unknown	c.959G>A	p.R320H	Unknown
D266	USA	T	c.1034+2T>C	IVS8+2T>C	Maternal	c.879G>A	IVS7-1G>A	Paternal
D282	USA	T	c.959G>A	p.R320H	Maternal	c.959G>A	p.R320H	Paternal
D392	JOR	T	c.958C>G	p.R320G	Maternal	c.958C>G	p.R320G	Paternal

D438	AUS	T	c.887G>A	p.R296H	Maternal	c.921G>A	p.V307V	Paternal
J001	BEL	T	c.731dupC	T245Nfs*32	Maternal	c.959G>A	p.R320H	Paternal
J005	NET	T	c.317T>C	p.I106T	Maternal	c.515T>C	p.L172P	Paternal
L005	ITA	T	c.987delC	p.M330Cfs*8	Maternal	c.794G>A	p.R265H	Paternal
L008	FRA	T	c.139G>A	p.G47R	Unknown	c.230C>T	p.S77L	Unknown
L011	GER	T	c.178_181delCAGT	p.Q60Tfs*35	Maternal	c.178_181delCAGT	p.Q60Tfs*35	Paternal
L013	GRE	T	c.992G>C	p.R331P	Maternal	c.217C>T	p.R73C	Paternal
L018	FRA	T	c.959G>A	p.R320H	Unknown	c.959G>A	p.R320H	Unknown
L019	FRA	T	c.877+5G>A	IVS7+5G>A	Maternal	c.350C>T	p.S117L	Paternal
L020	FRA	T	c.217C>T	p.R73C	Maternal	c.15_18del4	p.S6Wfs*89	Paternal
L023	CRO	T	c.959G>A	p.R320H	Unknown	c.959G>A	p.R320H	Unknown
L024	FRA	T	c.887G>A	p.R296H	Maternal	c.887G>A	p.R296H	Paternal
L025	FRA	T	c.1-55C>T		Maternal	c.471+2T>C	IVS4+2T>C	Paternal
L031	POR	T	c248A>G	p.H83R	Maternal	c248A>G	p.H83R	Paternal
L036	GER/RUS	T	c.959G>A	p.R320H	Maternal	c.496C>T	p.Q166X	Paternal
L037	TUR	T	c.793C>T	p.R265C	Maternal	c.959G>A	p.R320H	Paternal
L039	MAG	T	c.992G>A	p.R331Q	Maternal	c.992G>A	p.R331Q	Paternal
L040	NOR	T	c.794G>A	p.R265H	Maternal	c.959G>A	p.R320H	Paternal
L041	GER/CHIN	T	c.808_811del4insAGAG	p.L270_C271delinsRG	Maternal	c.959G>A	p.R320H	Paternal
L049	POR	T	c.248A>G	p.H83R	Maternal	c.248A>G	p.H83R	Paternal
L050	CRO	T	c.452_466del15	p.K151_L155del	Maternal	c.452_466del	p.K151_L155del	Paternal
L058	TUR	T	c.15_18delAAGT	p.S6Wfs*89	Maternal	c.15_18delAAGT	p.S6Wfs*89	Paternal
L059	GER	T	c.212A>C	p.H71P	Maternal	c.139G>A	p.G47R	Paternal
L061	GRE	T	c.992G>A	p.R331Q	Maternal	c.998dup	p.H333Qfs*17	Paternal
L064	FRA	T	c.452_466del15	p.K151_L155del	Maternal	c.452_466del15	p.K151_L155del	Paternal
M12	SPA	T	c.886C>T	p.R296C	Maternal	c.340-1G>A	p.G114_Q157del	Paternal
M17	SPA	T	c.878-1G>A	p.K294fs	Maternal	c.259-1>C	p.T87_Q113del	Paternal
M21	SPA	T	c.664C>T	p.R222C	Maternal	c.664C>T	p.R222C	Paternal
B028	IS	T	c.15_18delAAGT	p.S6Wfs*89	Maternal	c.15_18delAAGT	p.S6Wfs*89	Paternal

Appendix 3 – Unique Mutations in AMT gene in patients from Genetics in Medicine Article

Source	Gene	Type	Consequence	ClinicalSignificance	DNA mutation	Protein Change	Frequency
GenetMed	AMT	SNV	Synonymous	Likely pathogenic	c.339G>A	Q113Q	1
GenetMed	AMT	SNV	Synonymous	Likely pathogenic	c.921G>A	V307V	1
GenetMed	AMT	SNV	Stop gain	Pathogenic	c.4C>T	Q2X	1
GenetMed	AMT	Deletion	Stop gain	Pathogenic	c.395__400del6	V132_N134delinsD	1
GenetMed	AMT	SNV	Stop gain	Pathogenic	c.496C>T	Q166X	3
GenetMed	AMT	SNV	Stop gain	Pathogenic	c.574C>T	Q192X	1
GenetMed	AMT	SNV	Stop gain	Pathogenic	c.1101C>A	C367X	1
GenetMed	AMT	SNV	Splice-D/A	Pathogenic	c.471+2T>C	IVS4+2T>C	5
GenetMed	AMT	SNV	Splice-D/A	Pathogenic	c.878-1G>A	IVS7-1G>A	4
GenetMed	AMT	SNV	Splice-D/A	Pathogenic	c.259-1G>C	IVS2-1G>C	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.2T>C	M1T	7
GenetMed	AMT	SNV	Missense	Pathogenic	c.139G>A	G47R	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.212A>C	H71P	2

GenetMed	AMT	SNV	Missense	Pathogenic	c.217C>T	R73C	8
GenetMed	AMT	SNV	Missense	Pathogenic	c.230C>T	S77L	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.248A>G	H83R	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.280C>T	R94W	9
GenetMed	AMT	SNV	Missense	Pathogenic	c.293T>G	M98R	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.317T>C	I106T	5
GenetMed	AMT	SNV	Missense	Pathogenic	c.350C>T	S117L	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.434A>T	N145I	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.515T>C	L172P	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.545T>C	L182P	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.548A>G	Q183R	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.635T>C	V212A	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.664C>T	R222C	12
GenetMed	AMT	SNV	Missense	Pathogenic	c.665G>A	R222H	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.674G>C	Y225C	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.685G>C	D229H	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.794G>A	R265H	5
GenetMed	AMT	SNV	Missense	Pathogenic	c.793C>T	R265C	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.793C>A	R265S	3
GenetMed	AMT	SNV	Missense	Pathogenic	c.806G>A	G269D	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.826G>C	D276H	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.845C>T	T282I	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.887G>A	R296H	7
GenetMed	AMT	SNV	Missense	Pathogenic	c.886C>T	R296C	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.887G>C	R296P	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.959G>A	R320H	36
GenetMed	AMT	SNV	Missense	Pathogenic	c.958C>G	R320G	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.958C>T	R320C	2
GenetMed	AMT	SNV	Missense	Pathogenic	c.992G>A	R331Q	4
GenetMed	AMT	SNV	Missense	Pathogenic	c.992G>C	R331P	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.1082C>T	A361V	1
GenetMed	AMT	SNV	Missense	Pathogenic	c.1087G>C	G363R	2
GenetMed	AMT	SNV	Intronic	Pathogenic	c.340-1G>A	IVS3-1G>A	2
GenetMed	AMT	SNV	Intronic	Pathogenic	c.1034+1G>C	IVS8+1G>C	2
GenetMed	AMT	SNV	Intronic	Pathogenic	c.1034+2T>C	IVS8+2T>C	2
GenetMed	AMT	SNV	Intronic	Pathogenic	c.696+5G>C	IVS6+5G>C	1
GenetMed	AMT	SNV	Intronic	Pathogenic	c.877+5G>A	IVS7+5G>A	1
GenetMed	AMT	Indel	In frame indel	Pathogenic	c.551_552del2insAA	G184E	2
GenetMed	AMT	Indel	In frame indel	Pathogenic	c.808_811del4insAGAG	L270_C271delinsRG	1
GenetMed	AMT	Deletion	In frame del	Pathogenic	c.452_466del15	K151_L155del	6
GenetMed	AMT	Deletion	In frame del	Pathogenic	c.1107_1108delAC	Y369X	4
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.15_18delAAGT	S6Wfs*89	6
GenetMed	AMT	Duplication	Frameshift	Pathogenic	c.14dupT	S6Kfs*22	2
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.16delA	S6Vfs*90	2
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.13_16delGTAA	S6Wfs*89	1
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.178_181delCAGT	Q60Tfs*35	2
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.451_466del16	K151Cfs*25	2
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.452_465del14	K151Nfs*22	2
GenetMed	AMT	Insertion	Frameshift	Pathogenic	c.534_535insCC	L179Pfs*3	2
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.695_696+33del35insGGCTGTACAGA	D229Gfs*10	1
GenetMed	AMT	Duplication	Frameshift	Pathogenic	c.731dupC	T245Nfs*32	1
GenetMed	AMT	SNV	Frameshift	Pathogenic	c.878-1G>A	K294fs	1
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.987delC	M330Cfs*8	3

GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.999-1000delCA	H333Qfs*16	2
GenetMed	AMT	Duplication	Frameshift	Pathogenic	c.998dup	H333Qfs*17	1
GenetMed	AMT	Insertion	Frameshift	Pathogenic	c.996dup	H333Tfs*17	1
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.1040_1041delTG	V347Dfs*2	1
GenetMed	AMT	Deletion	Frameshift	Pathogenic	c.1063delT	S355Lfs*2	1
GenetMed	AMT	Insertion	Frameshift	Pathogenic	c.1034+2T>C	IVS8+2insT	1
GenetMed	AMT	Deletion	Deletion	Pathogenic	c.259-1>C	T87_Q113del	1
GenetMed	AMT	Deletion	Deletion	Pathogenic	c.340-1G>A	G114_Q157del	1

Appendix 4 – Unique Mutations in Simple Clinvar

Source	Gene	Type	Consequence	ClinicalSignificance	DNA mutation	Protein Change	Frequency
Simple ClinVar	AMT	SNV	Synonymous	Benign	c.510G>C	V170=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.522G>T	V174=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.705G>T	V235=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.705G>A	V235=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.852G>A	V284=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.948G>A	V316=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.663C>A	T221=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.843T>G	T281=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1038T>C	T346=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.108A>C	T36=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1092T>C	Y364=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1197C>T	Y399=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.198T>C	T66=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.216A>G	T72=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.231G>A	S77=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.621T>C	S207=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.789C>T	S263=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.867T>C	S289=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.204G>A	S68=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1005C>T	P335=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1062C>A	P354=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1062C>G	P354=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1062C>T	P354=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.111G>A	P37=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.174A>G	P58=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.357T>C	F119=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.555C>T	P185=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.609C>A	P203=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.60G>A	P20=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.939G>A	K313=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1029G>A	K343=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.123C>T	F41=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.477G>A	K159=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.546G>A	L182=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.774G>A	L258=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.798G>A	L266=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.78T>C	L26=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.99C>T	L33=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1128G>C	L376=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.303A>G	L101=	1

Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.354G>A	L118=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.381A>G	L127=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.393T>C	I131=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.441C>G	G147=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.459G>A	L153=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.43C>T	L15=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.513C>T	G171=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.516G>A	L172=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.582C>T	G194=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.651C>T	G217=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.681A>G	G227=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.771G>A	G257=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.807C>A	G269=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.822G>C	G274=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.858C>A	G286=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.879G>C	G293=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.126C>T	H42=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.135C>T	H45=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.195C>T	H65=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.519G>A	E173=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.576G>A	Q192=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.591T>C	D197=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.669T>C	C223=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.687T>C	D229=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely pathogenic	c.696G>A	E232=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.855G>A	E285=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.84C>T	C28=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1053C>T	C351=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1077T>C	N359=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.222G>A	Q74=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.240C>T	D80=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.363C>T	N121=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.435C>T	N145=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.750C>T	N250=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.42C>T	R14=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.657C>T	R219=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.666C>A	R222=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.891A>G	R297=	1
Simple ClinVar	AMT	SNV	Synonymous	Benign	c.954G>A	R318=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.1083G>A	A361=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.153G>A	A51=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.153G>T	A51=	1
Simple ClinVar	AMT	SNV	Synonymous	Uncertain/conflicting	c.159G>A	A53=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.159G>C	A53=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.285G>A	V95=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.534C>T	A178=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.561A>C	A187=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.579C>T	A193=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.579C>A	A193=	1
Simple ClinVar	AMT	SNV	Synonymous	Likely benign	c.714G>A	A238=	1
Simple ClinVar	AMT	SNV	Stop gain	Pathogenic	c.870G>A	W290X	1
Simple ClinVar	AMT	Deletion	Stop gain	Pathogenic	c.1107_1108 del	Y369_S370delins X	1
Simple ClinVar	AMT	SNV	Stop gain	Pathogenic	c.164G>A	W55Ter	1
Simple ClinVar	AMT	Deletion	Stop gain	Likely pathogenic	c.165del	G54_W55insX	1

Simple ClinVar/GenMed	AMT	SNV	Stop gain	Likely pathogenic	c.496C>T	Q16X	3
Simple ClinVar/GenMed	AMT	SNV	Stop gain	Likely pathogenic	c.574C>T	Q192X	1
Simple ClinVar	AMT	SNV	Stop gain	Uncertain/conflicting	c.1153C>T	Q385Ter	1
Simple ClinVar	AMT	SNV	Stop gain	Pathogenic	c.178C>T	Q60Ter	1
Simple ClinVar	AMT	SNV	Stop gain	Pathogenic	c.256C>T	Q86Ter	1
Simple ClinVar/GenMed	AMT	SNV	Splice-D/A	Likely pathogenic	c.259-1G>C	IVS2-1G>C	2
Simple ClinVar/GenMed	AMT	SNV	Splice-D/A	Pathogenic	c.471+2T>C	IVS4+2T>C	5
Simple ClinVar/GenMed	AMT	SNV	Splice-D/A	Pathogenic	c.878-1G>A	IVS7-1G>A	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.583G>A	V195M	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.584T>C	V195A	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.625G>A	V209M	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.635T>C	V212A	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.643G>C	V215L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.359C>T	T120I	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.403A>G	T135A	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.425T>G	V142G	1
Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.674A>G	Y225C	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1190C>A	T397K	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1199A>G	Y400C	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1202C>A	T401N	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.196A>G	T66A	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.230C>T	S77L	4
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.350C>T	S117L	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.431C>T	S144F	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.701C>T	S234L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1058C>A	S353Y	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.203C>T	S68L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.298A>G	S100G	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.752C>G	P251R	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1097C>T	P366L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.173C>T	P58L	1
Simple ClinVar	AMT	SNV	Missense	Pathogenic	c.2T>A	M1K	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.613A>G	M205V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.59C>T	P20L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.638T>G	F213C	1
Simple ClinVar	AMT	SNV	Missense	Benign/Likely benign	c.898A>G	M300V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1157T>C	M386T	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.476A>G	K159R	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.727C>A	L243M	1
Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.797T>C	L266P	1
Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.311G>A	G104E	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.317T>C	I106T	5
GenMed	AMT	SNV	Missense	Pathogenic	c.280C>T	I106T	9
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.341G>A	G114E	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.371G>A	G124E	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.515T>C	L172P	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.805G>A	G269S	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.806G>A	G269D	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.830T>C	I277T	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.857G>A	G286D	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.856G>A	G286S	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1021G>A	G341S	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1030A>T	I344F	1
Simple ClinVar	AMT	SNV	Missense	Pathogenic	c.125A>G	H42R	1

Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.139G>A	G47R	4
Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.139G>T	G47W	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.194A>G	H65R	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.212A>C	H71P	2
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.338A>G	Q113R	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.589G>A	D197N	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.589G>C	D197H	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.631G>A	E211K	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.685G>C	D229H	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.826G>C	D276H	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1136A>C	E379A	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.362A>G	N121S	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.401A>G	N134S	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.434A>G	N145S	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Likely pathogenic	c.434A>T	N145I	2
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.825T>A	N275K	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.959G>T	R320L	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.992G>A	R331Q	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.101G>A	R34H	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.100C>T	R34C	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1112G>A	R371H	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1141C>T	R381W	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1145G>A	R382Q	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.217C>T	R73C	8
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.280C>T	R94W	9
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.503G>C	R168T	1
Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.665G>T	R222L	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.665G>A	R222H	4
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.664C>T	R222C	12
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.70C>T	R24C	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.794G>A	R265H	5
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.793C>T	R265C	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.883C>T	R295C	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.887G>A	R296H	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.886C>T	R296C	4
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.86C>T	A29V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.944G>A	R315K	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.958C>T	R320C	2
Simple ClinVar/GenMed	AMT	SNV	Missense	Pathogenic	c.959G>A	R320H	36
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.982G>A	A328T	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.1082C>T	A361V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.130G>A	A44T	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.152C>T	A51V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.158C>T	A53V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.962T>C	V321A	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.961G>A	V321M	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.1138G>A	V380M	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.148G>T	V50L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.148G>C	V50L	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.283G>A	V95M	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.713C>T	A238V	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.892G>A	A298T	1
Simple ClinVar	AMT	SNV	Missense	Uncertain/conflicting	c.436G>A	A146T	1
Simple ClinVar/GenMed	AMT	SNV	Missense	Uncertain/conflicting	c.1087G>C	G363R	2

Simple ClinVar	AMT	Duplication	In frame indel	Uncertain/conflicting	c.705_710dup	P236_V237dup	1
Simple ClinVar	AMT	Duplication	In frame indel	Uncertain/conflicting	c.57_59dup	P20dup	1
Simple ClinVar	AMT	Deletion	In frame indel	Likely pathogenic	c.970_972del	M324del	1
Simple ClinVar	AMT	Deletion	In frame indel	Pathogenic	c.452_466del	K151_L155del	1
Simple ClinVar	AMT	Microsatellite	In frame indel	Uncertain/conflicting	c.1150CAG[1]	Q385del	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic	c.657dup	V220fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Likely pathogenic	c.849dup	V284fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.478del	V160fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.734_735del	T245fs	1
Simple ClinVar	AMT	Indel	Frameshift	Uncertain/conflicting	c.1199_1202delinsTAT	Y400fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic	c.14dup	S6fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic	c.257dup	T87fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.348_349del	S117fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.1056del	S353fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.16del	S6fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic/Likely pathogenic	c.15_18del	S6fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.908del	P303fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic	c.609dup	F204fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.59del	P20fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.987del	M330fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Uncertain/conflicting	c.1209del	K403fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.144_148del	K48fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.535del	L179fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.602_603del	K201fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.63del	L22fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.875del	L292fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.168_171del	L57fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic/Likely pathogenic	c.996dup	H333fs	1
Simple ClinVar	AMT	Duplication	Frameshift	Pathogenic/Likely pathogenic	c.982dup	A328fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.982del	A328fs	1
Simple ClinVar	AMT	Indel	Frameshift	Likely pathogenic	c.982_983delinsT	A328fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.148del	V50fs	1
Simple ClinVar	AMT	Microsatellite	Frameshift	Pathogenic	c.20_21del	V7fs	1
Simple ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.61del	A21fs	1

Appendix 5 – Unique Mutations in Clinvar (ClinVar (nih.gov))

Source	Gene	Type	Consequence	ClinicalSignificance	DNA mutation	Protein Change	Frequency
ClinVar	AMT	SNV	Stop Gain	Pathogenic	c.819T>A	Y273Ter	1
ClinVar	AMT	SNV	Stop Gain	Pathogenic	c.889C>T	R297Ter	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.335A>G	N112S	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.443G>T	C148F	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.467T>C	M156T	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.529A>G	N177D	1

ClinVar	AMT	SNV	Missense	Uncertain significance	c.566A>C	Q189P	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.593A>T	D198V	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.614T>G	M205R	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.643G>A	V215M	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.656G>A	R219H	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.658G>A	V220M	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.677C>T	T226I	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.719C>T	A240V	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.721G>A	V241I	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.830T>A	I277N	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.266T>G	I89R	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.269_270delinsCC	L90P	1
ClinVar/GenMed	AMT	SNV	Missense	Uncertain significance	c.293T>G	M98R	2
ClinVar	AMT	SNV	Missense	Uncertain significance	c.218G>A	R73H	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.188A>G	D63G	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.850G>A	V284M	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.890G>A	R297Q	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.911G>A	G304E	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.950A>G	Q317R	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.952C>T	R318W	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.953G>A	R318Q	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.972G>A	M324I	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.1001G>T	S334I	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.1076A>C	N359T	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.1111C>T	R371C	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.170T>C	L57P	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.155T>G	F52C	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.155T>C	F52S	1
ClinVar	AMT	SNV	Missense	Uncertain significance	c.110C>T	P37L	1
ClinVar	AMT	SNV	Missense	Pathogenic	c.847C>T	P283S	1
ClinVar	AMT	SNV	Missense	Likely pathogenic	c.664C>A	R222S	1
ClinVar	AMT	SNV	Missense	Likely pathogenic	c.887G>T	R296L	1
ClinVar	AMT	SNV	Missense	Likely benign	c.463C>T	L155F	1
ClinVar	AMT	Insertion	Frameshift	Pathogenic	c.270_271insCACCC	G91fs	1
ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.221del	Q74R	1
ClinVar	AMT	Deletion	Frameshift	Pathogenic	c.224del	H75R	1
ClinVar	AMT	Indel	Frameshift	Likely pathogenic	c.381_383delinsGG	D128fs	1
ClinVar	AMT	Deletion	Frameshift	Likely pathogenic	c.383del	D128fs	1
ClinVar	AMT	SNV	Frameshift	Likely pathogenic	c.230C>A	S77Ter	1
ClinVar	AMT	Insertion	Frameshift	Likely pathogenic	c.847_848insA	S283fs	1
ClinVar	AMT	Deletion	Deletion	Uncertain significance	c.271_273del	G91del	1
ClinVar	AMT	Deletion	Deletion	Uncertain significance	c.1068GAA[2]	K358del	1
ClinVar	AMT	Deletion	Deletion	Uncertain significance	c.1204_1206del	L402del	1

Appendix 6 – Unique Mutations in Pubtator (PubTator Central - NCBI - NLM - NIH)

Source	Gene	Type	Consequence	ClinicalSignificance	DNA mutation	Protein Change	Frequency
PubTator	AMT	SNV	Splice-D/A	Likely pathogenic	c.259-2A > T		1
PubTator	AMT	SNV	Splice-D/A	Likely pathogenic	c.878-1 G > A		1
PubTator	AMT	SNV	Splice-D/A	Likely pathogenic	c.IVS7-1G	IVS7-1G	1

PubTator/Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.752C>G	P251R	1
PubTator/Simple ClinVar/GenMed/	AMT	SNV	Missense	Pathogenic	c.794G>A	R265H	5
PubTator/Simple ClinVar	AMT	SNV	Missense	Likely pathogenic	c.1138G>A	V380M	1
PubTator	AMT	SNV	Missense	Likely pathogenic		V7L	1
PubTator	AMT	SNV	Missense	Likely pathogenic	c.955G>C	D276H	1
PubTator	AMT	Insertion	Frameshift	Likely pathogenic	c.982_983insG		1
PubTator	AMT	SNV	Frameshift	Likely pathogenic		A328Gfs*22	1
PubTator	AMT	Deletion	Frameshift	Likely pathogenic	c.977delA	E326Gfs*12	1
PubTator	AMT	Insertion	Frameshift	Likely pathogenic	c.14_15insT	S6KfsTer22	1
PubTator	AMT		Deletion	Likely pathogenic	c.183delC		1

Appendix 7 – Pathology Predict of LYRUS for the 149 missense mutations

SAV	Predict	Predict Probability	Variation_Number	dScore	Score1	evmutation	fathmm	Folding_Free_Energy	SASA(free sasa)	maestro	ANM_Mode_0	Mechanical_Stiffness	Effectiveness	Sensitivity	PyRosetta_mutant	PyRosetta_difference	active_site(p2rank)
M1T	0	0.07396401	0.456086287	1.189	-2.009	-8.331309432	-1.77	-0.1042183	192.83	-7.612894	5977.778075	1.940397898	0.021486965	66.35230207	196.40543	-0.5056	0
M1K	0	0.13380057	0.456086287	-0.1	-2.009	-8.331309432	-1.76	0.033569497	192.83	-7.657714	5977.778075	1.940397898	0.021486965	66.35230207	194.5391	-2.37193	0
V7L	0	0.03733494	0.816640986	0.086	-1.656	-3.085494962	-1.78	-0.146834667	129.08	-7.591105	1948.293909	2.587663186	0.074297648	24.38683392	195.92981	-0.98122	0
P20L	0	0.047179632	0.288135593	-0.29	-2.398	-6.404668824	-1.73	0.975295	109.42	-7.489787	317.8027331	3.19328581	0.099620755	0.537615642	200.21571	3.30468	0
R24C	0	0.022199262	0.340523883	0.735	-2.23	-6.865042401	-1.97	0.167522667	195.3	-7.52855	408.9725126	3.984732224	0.302070905	4.015832061	200.26056	3.34953	0
A29V	0	0.021103762	0.816640986	0.877	-1.871	-6.333708878	-1.83	0.413481333	83.7	-7.527444	353.0751635	5.442213969	0.42837414	4.513855751	198.22904	1.31801	0
R34H	0	0.046495974	0.354391371	1.244	-2.003	-7.30172557	-2.25	1.011628667	116.64	-7.574968	151.3153346	9.241264433	1.003427883	1.968162652	196.70512	-0.20591	0
R34C	0	0.1183836	0.354391371	3.114	-2.003	-8.194919026	-2.28	0.412377333	116.64	-7.597272	151.3153346	9.241264433	1.003427883	1.968162652	200.26056	3.34953	0
P37L	0	0.4445264	0.51155624	1.549	-1.763	-7.858537887	-2.87	1.26563	59.66	-7.744377	125.1899822	10.3386365	1.126318575	1.640460004	200.21571	3.30468	0
H42R	1	0.84704036	0.13559322	2.44	-1.032	-8.559824851	-1.92	4.578353333	5.51	-7.597859	127.657501	10.84948356	1.07127292	1.671953923	197.11694	0.20591	0
A44T	0	0.021211639	0.896764253	1.498	-2.054	-8.002037273	-1.93	0.604782	82.06	-7.52794	182.9248435	9.28678278	0.723386428	2.399359239	196.7381	-0.17293	0
G47R	1	0.7862637	0.174114022	1.798	-0.933	-8.215063867	-1.83	17.368	12.08	-7.574577	128.1262906	10.44526548	1.07445027	1.668052159	196.01813	-0.8929	0
G47W	1	0.9312644	0.174114022	4.271	-0.933	-8.215063867	-1.85	31.34963333	12.08	-7.632131	128.1262906	10.44526548	1.07445027	1.668052159	198.37386	1.46283	0
V50L	0	0.4845525	0.115562404	1.672	-1.131	-8.236468951	-1.79	0.446199333	75.1	-7.595024	132.2190522	8.97255399	1.147827249	1.71372384	195.92981	-0.98122	0
V50L	0	0.4845525	0.115562404	1.672	-1.131	-8.236468951	-1.79	0.446199333	75.1	-7.595024	132.2190522	8.97255399	1.147827249	1.71372384	195.92981	-0.98122	0
A51V	0	0.011965709	0.522342065	0.454	-2.5	-7.926926406	-1.65	-0.291164667	67.98	-7.537191	130.2033198	9.005581766	1.171735852	1.689317171	198.22904	1.31801	0
F52C	1	0.86979604	0.114021572	3.826	-0.991	-8.573161769	-1.92	1.342353333	75.23	-7.536931	91.40803749	9.757874228	1.668726627	1.185428911	198.94753	2.0365	1
F52S	1	0.89667	0.114021572	2.609	-0.991	-8.573161769	-1.87	1.95024	75.23	-7.440552	91.40803749	9.757874228	1.668726627	1.185428911	195.40305	-1.50798	1
A53V	1	0.51526487	0.160246533	2.062	-1.268	-8.506932417	-1.84	0.850439667	53.87	-7.624647	87.86699052	9.806542813	1.733128141	1.140624287	198.22904	1.31801	0
L57P	1	0.9203119	0.226502311	2.719	-1.283	-8.658643423	-1.83	5.969126667	1.19	-7.420805	96.82904044	10.98431562	1.545124838	1.259129422	193.60635	-3.30468	0
P58L	1	0.92555875	0.130970724	2.77	-0.772	-8.283627381	-3.42	9.219223333	11.59	-7.786745	95.99405386	11.11701542	1.515651248	1.246363239	200.21571	3.30468	0
D63G	0	0.13526922	0.405238829	1.157	-1.424	-7.830280403	-1.68	0.369354333	60.75	-7.600916	100.5878874	9.764694145	0.985860398	1.293794704	199.85493	2.9439	0
H65R	0	0.23507959	0.151001541	1.806	-2.181	-8.108077717	-1.68	-0.064346133	82.32	-7.560159	63.69532667	10.53005427	1.584030175	0.809510488	197.11694	0.20591	0
T66A	0	0.05871112	0.385208012	-0.412	-2.459	-8.14671028	-1.6	0.058896033	52.64	-7.661196	59.52857218	10.57322791	1.296482526	0.748604225	197.08396	0.17293	0
S68L	1	0.55547345	0.11248074	1.393	-1.103	-8.50994356	-1.76	-3.181113333	19.66	-7.750422	55.30919905	10.96137927	1.450957611	0.707605244	198.86219	1.95116	0
H71P	1	0.8556137	0.178736518	2.833	-1.112	-8.433870747	-1.75	4.059263333	29.94	-7.484407	53.98443122	10.70402312	1.206214969	0.693677323	195.56847	-1.34256	0

R73H	1	0.65513897	0.200308166	3.272	-1.016	-7.72325908	-2.33	2.00025	42.86	-7.565079	37.99965996	11.42378576	1.787234863	0.486067552	196.70512	-0.20591	0
R73C	1	0.856099	0.200308166	3.986	-1.016	-8.213042723	-2.35	2.15734	42.86	-7.603466	37.99965996	11.42378576	1.787234863	0.486067552	200.26056	3.34953	0
S77L	1	0.75184315	0.291217257	2.295	-1.019	-7.867368616	-1.76	1.579233	1.78	-7.700774	42.93144712	12.16197297	1.907962103	0.563912393	198.86219	1.95116	0
H83R	1	0.66190654	0.277349769	2.503	-0.939	-7.280072348	-1.23	1.191490333	31.47	-7.557783	61.39788378	10.79933143	2.448446197	0.797315512	197.11694	0.20591	1
I89R	1	0.9766177	0.167950693	2.652	-1.521	-7.657728246	-2.06	8.97243	0	-7.310485	54.56285812	11.34630449	1.339998362	0.678574536	194.51255	-2.39848	0
L90P	0	0.27959308	0.403697997	1.557	-2.605	-7.117941427	-0.86	2.258483333	54.84	-7.382609	74.22212741	10.5142983	0.968294617	0.924453786	193.60635	-3.30468	0
R94W	1	0.7105414	0.167950693	3.739	-1.734	-6.785001558	-1.03	5.506123333	46.6	-7.68494	63.94941272	11.27211612	0.97383491	0.791392377	199.26676	2.35573	0
V95M	0	0.06449509	0.312788906	1.43	-1.786	-6.112862425	-0.91	-0.489741667	38.15	-7.587235	57.90888168	11.10601742	1.136739616	0.718781655	195.92569	-0.98534	0
M98R	1	0.92106575	0.309707242	2.029	-2.054	-6.802121052	-0.88	7.459953333	0	-7.385766	42.96495728	11.99606263	1.363100703	0.533971431	195.15894	-1.75209	0
S100G	0	0.41751784	0.178736518	1.842	-1.361	-6.645624081	-0.82	-0.229128	55.93	-7.57924	58.68038563	10.26175162	1.019137604	0.741803565	197.99888	1.08785	0
G104E	1	0.83335507	0.271186441	2.034	-1.634	-6.835284908	-0.86	1.748146667	4.77	-7.477582	34.59217071	11.53064389	2.68237391	0.440249582	193.38834	-3.52269	0
I106T	1	0.53203666	0.26348228	1.141	-1.583	-7.031506891	-1.12	3.179026667	5.51	-7.448506	45.45814127	11.2769369	1.993859766	0.570582388	195.75904	-1.15199	0
N112S	0	0.26242673	0.201848998	1.62	-1.507	-6.694696577	-1.72	0.471946667	43.35	-7.593466	53.57420437	10.56399675	1.986787961	0.68627747	197.9616	1.05057	0
Q113R	0	0.4886157	0.092449923	1.588	-2.151	-2.915753189	-0.89	-0.015837667	55.99	-7.530702	43.72118668	10.89628348	2.787472072	0.558399453	198.26724	1.35621	0
G114E	1	0.97723114	0.169491525	2.447	-1.115	-6.778807283	-1.02	4.293096667	6.9	-7.366134	24.78628804	11.41047097	4.527754051	0.313006013	193.38834	-3.52269	0
S117L	1	0.72228926	0.146379045	0.962	-1.602	-7.012583461	-0.98	5.479726667	3.18	-7.824789	12.56520093	11.88298403	5.413372527	0.157392053	198.86219	1.95116	0
T120I	0	0.3845235	0.107858243	1.267	-1.061	-6.991530942	-0.87	-0.3140331	16.45	-7.765675	41.62542718	11.5056438	1.783789024	0.543313514	198.06302	1.15199	0
N121S	0	0.13863604	0.232665639	1.149	-1.375	-2.449382136	-1.28	0.170264267	59.87	-7.579426	62.99830652	10.29042348	1.254653543	0.82540739	197.9616	1.05057	0
G124E	1	0.98455703	0.032357473	2.821	-0.759	-7.266742574	-4.82	19.94116667	1.19	-7.494048	62.21955571	10.91443994	1.261535141	0.825063596	193.38834	-3.52269	0
N134S	0	0.032251842	0.329738059	0.458	-1.929	-2.760267908	-0.93	2.40523	18.18	-7.577705	42.90817386	11.39712976	2.014176948	0.538357264	197.9616	1.05057	0
T135A	0	0.11282194	0.160246533	0.577	-1.624	-5.12514003	-1.68	-0.801150333	4.31	-7.57161	52.69216816	10.60190631	1.645977186	0.667729404	197.08396	0.17293	0
V142G	1	0.93586737	0.234206471	2.415	-1.337	-6.890137805	-1.09	3.250876667	1.19	-7.326084	23.6100613	12.17600386	3.50384633	0.293754346	195.0665	-1.84453	0
S144F	1	0.9085527	0.038520801	2.901	-1.162	-8.082453075	-0.92	2.1716595	5.15	-7.735944	20.96775464	11.63633583	6.68144462	0.26607589	198.41901	1.50798	0
N145I	1	0.83490014	0.038520801	3.161	-0.827	-8.082271956	-1.15	0.80897	55.5	-7.738156	26.7225111	11.1050578	5.705967892	0.343676293	200.55503	3.644	1
N145S	1	0.8456092	0.038520801	2.086	-0.827	-8.082271956	-1.02	0.829042	55.5	-7.576531	26.7225111	11.1050578	5.705967892	0.343676293	197.9616	1.05057	1
A146T	1	0.58608264	0.038520801	1.446	-0.895	-8.082510914	-1.17	1.565733333	6.82	-7.533199	49.00369655	10.66387458	3.120042524	0.633661665	196.7381	-0.17293	0
C148F	1	0.83481693	0.038520801	2.522	-1.582	-8.082530749	-0.88	-0.417908667	15.69	-7.614859	53.94053991	10.4456049	2.768360004	0.698172172	194.87453	-2.0365	0
L155F	0	0.08255519	0.314329738	1.191	-1.775	-7.329429623	-0.84	0.210820333	49.41	-7.427789	68.98717623	10.62185023	1.530949613	0.881572124	196.46785	-0.44318	0
M156T	1	0.5619407	0.251155624	1.821	-1.681	-6.541367032	-0.95	3.03585	1.19	-7.340866	60.79575804	11.64348069	1.525057621	0.767765328	196.40543	-0.5056	0
K159R	0	0.021129733	0.432973806	0.39	-1.823	-1.414454181	-1.67	-0.835485	53.53	-7.61938	81.52939433	10.33481767	0.943012858	1.028559033	197.53087	0.61984	0
R168T	0	0.017741816	0.862865948	0.988	-2.652	-6.795087244	-0.9	0.989679667	89.45	-7.589308	121.2901038	9.389789951	0.497121891	1.506052805	198.15752	1.24649	0
L172P	1	0.89731497	0.291217257	2.724	-1.456	-6.817187034	-1.19	2.97763	33.84	-7.387367	75.56902445	10.38957073	1.212976578	0.94821189	193.60635	-3.30468	0
N177D	0	0.054040603	0.476117103	1.078	-2.607	-5.515073478	-0.62	0.871222	32.33	-7.514748	74.90927613	10.96721906	1.661119081	0.97560422	196.10555	-0.80548	0
L182P	1	0.9855642	0.17257319	2.763	-1.341	-7.435353033	-2.24	4.2827	0	-7.261655	32.88478732	12.32833658	2.668991761	0.437730035	193.60635	-3.30468	0

Q183R	1	0.90374285	0.087827427	2.215	-1.017	-7.760842856	-1.32	4.2555	4.92	-7.628162	24.87191477	12.34255665	3.178393076	0.328189149	198.26724	1.35621	0
Q189P	0	0.038211636	0.460708783	1.633	-2.281	-6.188201947	-0.95	-1.927583333	96.19	-7.44814	85.77565688	10.2868187	1.090365347	1.144608977	196.71877	-0.19226	0
V195M	0	0.07750378	0.184899846	1.415	-1.652	-3.229304564	-1.02	-1.70177	31.97	-7.611457	108.9543367	10.06405463	0.865116489	1.457015255	195.92569	-0.98534	0
V195A	0	0.15144137	0.184899846	1.39	-1.652	-7.074740463	-0.94	-0.188514	31.97	-7.548887	108.9543367	10.06405463	0.865116489	1.457015255	195.59302	-1.31801	0
D197N	0	0.054167658	0.388289676	1.391	-2.493	-3.063178425	-0.89	0.789419667	83.29	-7.590306	116.2933536	9.323399968	0.892990374	1.544136115	197.71651	0.80548	0
D197H	0	0.036545772	0.388289676	0.998	-2.493	-2.737977559	-0.91	0.847121	83.29	-7.613622	116.2933536	9.323399968	0.892990374	1.544136115	198.75612	1.84509	0
D198V	1	0.54639965	0.181818182	2.328	-1.35	-6.495366583	-2.15	1.55718	47.43	-7.398091	114.7281004	9.354408995	0.978899671	1.521540356	201.69946	4.78843	0
M205R	1	0.9057756	0.033898305	2.651	-1.109	-7.212022848	-0.99	1.47251	23.01	-7.457111	25.57989	11.27865264	5.341021152	0.330280962	195.15894	-1.75209	1
M205V	1	0.9281326	0.033898305	1.681	-1.109	-7.212022848	-0.89	4.87891	23.01	-7.605749	25.57989	11.27865264	5.341021152	0.330280962	197.89637	0.98534	1
V209M	0	0.03996482	0.429892142	1.037	-2.1	-3.397455534	-0.92	-0.558106333	78.07	-7.611287	72.31045949	10.67736649	1.271567621	0.947069241	195.92569	-0.98534	0
E211K	0	0.018308766	0.591679507	0.681	-2.137	-3.65216025	-1.73	-0.2724	100.48	-7.624884	104.6657696	10.0238458	0.893158663	1.390584645	198.92098	2.00995	0
V212A	1	0.64312017	0.115562404	1.699	-1.39	-7.211987455	-1.82	3.302136667	15.9	-7.453808	94.81319571	10.97339221	0.989208514	1.265775029	195.59302	-1.31801	0
F213C	0	0.43029442	0.064714946	0.714	-1.856	-6.182489723	-0.88	3.597296667	45.78	-7.569204	120.5938647	9.952561004	0.78033322	1.614034811	198.94753	2.0365	0
V215M	0	0.09333889	0.194144838	1.568	-1.481	-6.218198733	-0.91	-0.268886333	45.12	-7.603322	122.88834	10.24891231	0.809390792	1.632152867	195.92569	-0.98534	0
V215L	0	0.040416855	0.194144838	0.682	-1.481	-1.840122793	-0.88	-0.671759667	45.12	-7.648647	122.88834	10.24891231	0.809390792	1.632152867	195.92981	-0.98122	0
R219H	0	0.38496298	0.13559322	1.674	-1.168	-5.672778301	-0.98	-0.225028333	22.8	-7.606262	51.33107235	11.73080272	1.801905386	0.672966316	196.70512	-0.20591	0
V220M	0	0.39748344	0.192604006	2.727	-1.391	-6.617012689	-1.16	-1.240096667	13.98	-7.520962	43.16531797	11.94469341	2.16910738	0.570365637	195.92569	-0.98534	0
R222S	1	0.78810334	0.110939908	2.364	-1.011	-2.292336784	-2.28	2.576873333	12.88	-7.541955	27.09526513	12.05841986	4.111679392	0.356652675	196.71608	-0.19495	0
R222H	1	0.54085904	0.110939908	1.677	-1.011	-6.417855539	-2.35	1.907366667	12.88	-7.59721	27.09526513	12.05841986	4.111679392	0.356652675	196.70512	-0.20591	0
R222C	1	0.9418878	0.110939908	4.065	-1.011	-1.839283927	-2.37	2.48724	12.88	-7.761248	27.09526513	12.05841986	4.111679392	0.356652675	200.26056	3.34953	0
R222L	1	0.9149676	0.110939908	2.364	-1.011	-6.417855539	-2.35	0.728928667	12.88	-7.788622	27.09526513	12.05841986	4.111679392	0.356652675	198.66724	1.75621	0
Y225C	1	0.63658583	0.12788906	1.812	-1.043	-6.668585821	-1.33	1.163533333	59.71	-7.593008	3.177195822	12.4759014	24.01747209	0.038474136	199.58359	2.67256	1
Y225C	1	0.63658583	0.12788906	1.812	-1.043	-6.668585821	-1.33	1.163533333	59.71	-7.593008	3.177195822	12.4759014	24.01747209	0.038474136	199.58359	2.67256	1
T226I	1	0.8904121	0.152542373	2.719	-0.819	-6.357783018	-1.34	3.985033333	2.38	-7.658108	8.432350771	12.15329327	7.57296178	0.105261186	198.06302	1.15199	0
D229H	1	0.9701493	0.081664099	3.491	-0.879	-7.11741443	-1.11	8.24346	26.57	-7.703066	30.02332032	11.42025155	3.633374843	0.393819017	198.75612	1.84509	0
S234L	1	0.5560172	0.138674884	1.352	-1.073	-3.825025379	-0.73	0.435392067	6.93	-7.791731	46.79365876	11.94546071	2.252351542	0.613902781	198.86219	1.95116	0
A238V	0	0.007923348	0.677966102	0.665	-2.377	-5.768558807	-0.99	0.102198667	106.58	-7.536254	147.3258279	9.435379109	0.839235098	1.938487098	198.22904	1.31801	0
A240V	0	0.0327122	0.379044684	0.279	-1.516	-3.324288402	-1.19	0.575295333	0	-7.591003	102.0132246	11.17037185	1.185136111	1.347717146	198.22904	1.31801	0
V241I	0	0.01870064	0.369799692	0.718	-1.837	-4.791024429	-0.93	-0.747154333	72.72	-7.624056	124.2582117	10.56009227	0.974059973	1.644538168	196.57208	-0.33895	0
L243M	0	0.24001174	0.243451464	1.684	-1.337	-4.059797607	-1.23	0.259854667	8.34	-7.485555	98.88580629	11.42646432	1.059335512	1.317554811	196.90691	-0.00412	0
P251R	0	0.049273543	0.602465331	1.908	-1.962	-7.352159004	-0.86	1.165323333	120.11	-7.553972	111.9467671	9.750674833	0.709052314	1.484862717	198.4595	1.54847	0
R265S	1	0.9136018	0.083204931	2.407	-0.984	-8.215644502	-2.77	3.741473333	9.12	-7.491336	7.594415742	11.92287945	8.858951634	0.09139785	196.71608	-0.19495	0
R265H	1	0.94550323	0.083204931	3.315	-0.984	-8.215644502	-2.81	3.59588	9.12	-7.541891	7.594415742	11.92287945	8.858951634	0.09139785	196.70512	-0.20591	0
R265C	1	0.8305707	0.083204931	1.86	-0.984	-8.215644502	-2.83	3.05479	9.12	-7.638213	7.594415742	11.92287945	8.858951634	0.09139785	200.26056	3.34953	0

L266P	1	0.9847312	0.181818182	2.995	-1.082	-3.582214495	-1.37	9.100286667	4.77	-7.277118	14.01587191	12.10105512	5.218049544	0.172357299	193.60635	-3.30468	0
G269D	1	0.9878449	0.130970724	2.765	-0.735	-6.626229557	-1.71	11.5582	11.21	-7.494055	13.53318035	11.48776598	7.60939173	0.169684266	193.96713	-2.9439	0
G269S	1	0.91657066	0.130970724	2.4	-0.735	-2.487797672	-1.66	8.373946667	11.21	-7.53014	13.53318035	11.48776598	7.60939173	0.169684266	195.82318	-1.08785	0
N275K	0	0.29079404	0.365177196	1.941	-1.45	-6.371380431	-0.91	-0.570864667	67.78	-7.599311	39.24666118	10.53917326	3.175850262	0.513524517	197.53671	0.62568	0
D276H	1	0.9886728	0.11248074	3.668	-0.764	-7.010775544	-2.45	14.59786667	2.05	-7.567068	27.4570234	11.47771045	3.800569028	0.360863395	198.75612	1.84509	0
D276H	1	0.9886728	0.11248074	3.668	-0.764	-7.010775544	-2.45	14.59786667	2.05	-7.567068	27.4570234	11.47771045	3.800569028	0.360863395	198.75612	1.84509	0
I277N	1	0.97509265	0.13559322	2.825	-1.228	-6.76633209	-1.15	3.45644	1.19	-7.406435	28.82039533	11.22612017	3.554004015	0.382180234	193.26703	-3.644	0
I277T	0	0.4084365	0.13559322	1.369	-1.228	-6.76633209	-1.05	2.945466667	1.19	-7.471226	28.82039533	11.22612017	3.554004015	0.382180234	195.75904	-1.15199	0
T282I	1	0.92665064	0	2.459	-1.123	-7.467999642	-1.03	3.701566667	9.54	-7.713969	47.73445561	11.65434987	1.787273198	0.637488895	198.06302	1.15199	0
P283S	1	0.9595409	0.104776579	2.525	-0.877	-7.36891948	-3.82	4.05289	2.38	-7.479856	34.96703065	11.98356603	2.535814991	0.464708334	198.26455	1.35352	0
V284M	1	0.511185	0.155624037	2.909	-1.284	-7.274439434	-0.93	-0.656140333	10.4	-7.593857	51.66502876	11.35452949	1.921518781	0.687620864	195.92569	-0.98534	0
G286D	1	0.9627053	0.164869029	2.032	-1.602	-7.855086441	-0.88	7.28989	0	-7.49358	32.50375293	11.80532327	2.896470127	0.433193707	193.96713	-2.9439	0
G286S	0	0.44559294	0.164869029	1.409	-1.602	-7.648083364	-0.8	3.627266667	0	-7.547191	32.50375293	11.80532327	2.896470127	0.433193707	195.82318	-1.08785	0
R295C	1	0.5891814	0.101694915	1.678	-1.531	-7.05130566	-1.79	0.814938333	95.56	-7.502231	98.21886427	9.097721942	1.2523164	1.266368228	200.26056	3.34953	0
R296L	1	0.8911825	0.134052388	2.364	-1.011	-6.901170655	-1.75	-1.739	8.64	-7.665233	80.1474281	10.00572127	1.287189737	1.032954034	198.66724	1.75621	0
R296P	1	0.9072701	0.134052388	2.857	-1.011	-6.901170655	-1.73	-0.994383667	8.64	-7.511147	80.1474281	10.00572127	1.287189737	1.032954034	195.36256	-1.54847	0
R296C	1	0.91067064	0.134052388	4.065	-1.011	-6.901170655	-1.78	0.093125667	8.64	-7.619331	80.1474281	10.00572127	1.287189737	1.032954034	200.26056	3.34953	0
R296H	0	0.39437172	0.134052388	1.677	-1.011	-6.901170655	-1.72	0.998915333	8.64	-7.566592	80.1474281	10.00572127	1.287189737	1.032954034	196.70512	-0.20591	0
R297Q	1	0.6675152	0.06779661	1.469	-1.158	-1.412462568	-1.72	0.283762333	108.02	-7.555775	87.04375459	10.08812755	1.286792156	1.131998929	195.55482	-1.35621	0
A298T	0	0.00879054	0.636363636	0.458	-1.888	-1.993669544	-1.84	0.489482667	88.86	-7.51065	122.7725355	9.1092868	0.968887456	1.597080179	196.7381	-0.17293	0
M300V	0	0.020460898	0.326656394	0.217	-2.695	-5.889059257	-1.73	1.507643333	102.42	-7.554107	102.6296641	9.831317716	0.902207365	1.336707842	197.89637	0.98534	0
G304E	1	0.9942521	0.103235747	2.861	-0.759	-5.775611691	-5.16	21.4569	6.26	-7.489128	71.67861612	10.65258625	1.030706075	0.943215535	193.38834	-3.52269	0
R315K	0	0.005493832	0.268104777	-1.07	-2.802	-1.644091609	-0.97	-0.423796	200.42	-7.565733	129.4737462	8.80569011	0.77720609	1.728386402	196.29119	-0.61984	0
Q317R	0	0.018923683	0.627118644	1.434	-2.313	-2.695299002	-0.85	-0.139829767	138.96	-7.563041	107.1680215	9.884448038	0.882681891	1.434371212	198.26724	1.35621	0
R318W	0	0.2986974	0.275808937	1.579	-1.69	-6.498237284	-1.4	2.66879	82.19	-7.630134	82.71090991	10.85162519	1.151720111	1.104035585	199.26676	2.35573	0
R318Q	0	0.31422654	0.275808937	1.201	-1.69	-4.22576137	-1.32	0.954908	82.19	-7.544146	82.71090991	10.85162519	1.151720111	1.104035585	195.55482	-1.35621	0
R320G	1	0.84175473	0.058551618	2.381	-1.016	-7.10880561	-1.2	2.091463333	11.74	-7.585943	42.30856857	11.8400635	2.29946014	0.559328409	197.80393	0.8929	0
R320C	1	0.9242337	0.058551618	3.986	-1.016	-7.10880561	-1.2	1.068746667	11.74	-7.669337	42.30856857	11.8400635	2.29946014	0.559328409	200.26056	3.34953	0
R320H	1	0.85110253	0.058551618	3.272	-1.016	-7.10880561	-1.2	0.994686333	11.74	-7.637843	42.30856857	11.8400635	2.29946014	0.559328409	196.70512	-0.20591	0
R320L	1	0.80732626	0.058551618	1.37	-1.016	-7.10880561	-0.63	-1.70894	11.74	-7.732741	42.30856857	11.8400635	2.29946014	0.559328409	198.66724	1.75621	0
V321A	1	0.5637408	0.149460709	1.814	-1.116	-6.954957557	-1.73	0.659606333	15.5	-7.434305	45.54374425	11.89589646	2.202964074	0.594785063	195.59302	-1.31801	0
V321M	1	0.7781703	0.149460709	3.003	-1.116	-6.954957557	-1.77	-0.697364667	15.5	-7.521286	45.54374425	11.89589646	2.202964074	0.594785063	195.92569	-0.98534	0
M324I	0	0.007232552	0.440677966	-0.81	-2.952	-3.494590172	-0.95	0.881655333	35.78	-7.653768	63.13760263	11.6267942	1.296872291	0.797185599	197.55742	0.64639	0
A328T	0	0.2374376	0.132511556	1.696	-1.715	-6.828930485	-1.18	-0.751991	23.63	-7.579787	104.3392726	9.628207539	0.808797399	1.313013735	196.7381	-0.17293	0

R331P	1	0.9121143	0.06779661	2.826	-1.009	-2.700090403	-1.24	0.334572333	115.25	-7.563035	73.83104897	10.95602111	1.430668017	0.946174113	195.36256	-1.54847	0
R331Q	1	0.7604779	0.06779661	1.496	-1.009	-6.326954547	-1.22	0.257022667	115.25	-7.588451	73.83104897	10.95602111	1.430668017	0.946174113	195.55482	-1.35621	0
S334I	0	0.056540247	0.235747304	0.722	-2.131	-6.273776359	-1.08	0.721959667	14.31	-7.767393	99.39783022	10.51170443	1.038840804	1.291591108	199.50446	2.59343	0
G341S	0	0.22727709	0.18798151	1.336	-1.309	-1.904182056	-1.34	3.96087	42.69	-7.575715	186.0122592	9.07543138	0.487318089	2.434981206	195.82318	-1.08785	0
I344F	0	0.33149856	0.201848998	2.438	-1.374	-6.464975366	-1.4	1.511663333	29.63	-7.5181	122.5577288	10.18445246	0.757501509	1.620540255	195.82558	-1.08545	0
S353Y	1	0.9459498	0.200308166	3.128	-0.99	-7.249872737	-2.08	28.51716667	3.45	-7.672389	41.575334	11.47747224	1.53507183	0.513065455	197.78295	0.87192	0
N359T	1	0.51331896	0.13559322	1.316	-1.273	-6.543390473	-1.15	1.034588333	10.42	-7.643051	57.38914927	11.02139714	1.255503711	0.715287274	199.40304	2.49201	0
A361V	1	0.5286005	0.208012327	2.065	-1.107	-7.026646253	-1.8	0.419914567	0	-7.687794	39.44275691	12.23589884	2.107266073	0.500640706	198.22904	1.31801	0
G363R	1	0.9778556	0.106317411	3.004	-0.889	-7.805315657	-1.4	13.33453333	4.37	-7.50719	49.61902277	12.13151487	1.859787509	0.649496012	196.01813	-0.8929	0
G363R	1	0.9778556	0.106317411	3.004	-0.889	-7.805315657	-1.4	13.33453333	4.37	-7.50719	49.61902277	12.13151487	1.859787509	0.649496012	196.01813	-0.8929	0
P366L	0	0.06227743	0.405238829	1.673	-1.982	-7.032611025	-1.09	2.2112	35.76	-7.677987	105.2350201	10.25055023	0.927141406	1.398125202	200.21571	3.30468	0
R371C	0	0.11653404	0.281972265	2.666	-2.45	-7.240943329	-1.22	0.83337	172.38	-7.552795	107.6311897	9.002431609	1.123691892	1.405519237	200.26056	3.34953	0
R371H	0	0.022412749	0.281972265	0.298	-2.45	-7.240943329	-1.17	0.317232	172.38	-7.490661	107.6311897	9.002431609	1.123691892	1.405519237	196.70512	-0.20591	0
E379A	0	0.31529728	0.103235747	1.528	-1.665	-6.953422755	-1.16	-2.374793333	44.6	-7.705442	125.5058922	10.70589661	0.744917248	1.619614642	200.96024	4.04921	0
V380M	1	0.5256226	0.149460709	2.863	-1.397	-7.008763471	-1.17	0.044191	22.07	-7.541952	120.1207396	10.56147214	0.782614759	1.537643235	195.92569	-0.98534	0
R381W	1	0.62673587	0.11248074	1.735	-1.308	-7.555484724	-1.29	-0.272895	163.9	-7.520827	144.7980703	8.70175247	0.711732056	1.853332764	199.26676	2.35573	0
R382Q	0	0.1058353	0.140215716	0.071	-2.668	-7.095406495	-1.04	0.059158267	199.55	-7.607717	169.6196548	8.545603732	0.609898587	2.183916318	195.55482	-1.35621	0
M386T	0	0.012804236	0.516178737	0.467	-3.099	-6.428834066	-0.92	1.891506667	80.53	-7.476057	123.996296	9.954959985	0.671325338	1.583161968	196.40543	-0.5056	0
T397K	1	0.7537756	0.183359014	1.956	-1.644	-7.010773686	-0.99	-1.268826667	50.13	-7.586371	32.44968541	10.18705361	4.246389027	0.415476285	195.0447	-1.86633	0
Y400C	1	0.7486318	0.186440678	3.662	-1.227	-7.131815935	-1.08	2.131996667	87.19	-7.5904	55.3096902	10.14839126	2.524272012	0.723719687	199.58359	2.67256	0
T401N	0	0.027712168	0.605546995	0.476	-1.725	-3.131293716	-1.8	0.505384333	71.29	-7.614033	71.29285189	8.971241484	2.082453748	0.929809826	194.41902	-2.49201	0

Appendix 8 – Source and Consequence of Pathology Predict of LYRUS for the 149 missense mutations

Source	Clinical Significance	DNA mutation	Frequency	SAV	Predict	Predict Probability
Simple ClinVar	Pathogenic	c.2T>A	1	M1K	0	0.1
GenMed	Pathogenic	c.2T>C	7	M1T	0	0.1
PubTator	Likely pathogenic		1	V7L	0	0.0
Simple ClinVar	Uncertain/conflicting	c.59C>T	1	P20L	0	0.0
Simple ClinVar	Uncertain/conflicting	c.70C>T	1	R24C	0	0.0
Simple ClinVar	Uncertain/conflicting	c.86C>T	1	A29V	0	0.0
Simple ClinVar	Uncertain/conflicting	c.101G>A	1	R34H	0	0.0
Simple ClinVar	Uncertain/conflicting	c.100C>T	1	R34C	0	0.1
ClinVar	Uncertain significance	c.110C>T	1	P37L	0	0.4
Simple ClinVar	Pathogenic	c.125A>G	1	H42R	1	0.8
Simple ClinVar	Uncertain/conflicting	c.130G>A	1	A44T	0	0.0
Simple ClinVar	Likely pathogenic	c.139G>T	1	G47W	1	0.9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.139G>A	4	G47R	1	0.8
Simple ClinVar	Uncertain/conflicting	c.148G>T	1	V50L	0	0.4
Simple ClinVar	Uncertain/conflicting	c.148G>C	1	V50L	0	0.4
Simple ClinVar	Uncertain/conflicting	c.152C>T	1	A51V	0	0.0
ClinVar	Uncertain significance	c.155T>C	1	F52S	1	0.9
ClinVar	Uncertain significance	c.155T>G	1	F52C	1	0.9
Simple ClinVar	Uncertain/conflicting	c.158C>T	1	A53V	1	0.5
ClinVar	Uncertain significance	c.170T>C	1	L57P	1	0.9
Simple ClinVar	Uncertain/conflicting	c.173C>T	1	P58L	1	0.9
ClinVar	Uncertain significance	c.188A>G	1	D63G	0	0.1
Simple ClinVar	Uncertain/conflicting	c.194A>G	1	H65R	0	0.2
Simple ClinVar	Uncertain/conflicting	c.196A>G	1	T66A	0	0.1
Simple ClinVar	Uncertain/conflicting	c.203C>T	1	S68L	1	0.6
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.212A>C	2	H71P	1	0.9
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.217C>T	8	R73C	1	0.9
ClinVar	Uncertain significance	c.218G>A	1	R73H	1	0.7
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.230C>T	4	S77L	1	0.8
GenMed	Pathogenic	c.248A>G	4	H83R	1	0.7
ClinVar	Uncertain significance	c.266T>G	1	I89R	1	1.0
ClinVar	Uncertain significance	c.269_270delinsCC	1	L90P	0	0.3
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.280C>T	9	R94W	1	0.7
Simple ClinVar	Uncertain/conflicting	c.283G>A	1	V95M	0	0.1
ClinVar/GenMed	Uncertain significance/Pathogenic	c.293T>G	2	M98R	1	0.9
Simple ClinVar	Uncertain/conflicting	c.298A>G	1	S100G	0	0.4
Simple ClinVar	Likely pathogenic	c.311G>A	1	G104E	1	0.8
Simple ClinVar/GenMed	Pathogenic	c.317T>C	5	I106T	1	0.5
ClinVar	Uncertain significance	c.335A>G	1	N112S	0	0.3
Simple ClinVar	Uncertain/conflicting	c.338A>G	1	Q113R	0	0.4
Simple ClinVar	Uncertain/conflicting	c.341G>A	1	G114E	1	1.0
Simple ClinVar/ClinVar/GenMed	Pathogenic	c.350C>T	4	S117L	1	0.7
Simple ClinVar	Uncertain/conflicting	c.359C>T	1	T120I	0	0.4
Simple ClinVar	Uncertain/conflicting	c.362A>G	1	N121S	0	0.1
Simple ClinVar	Uncertain/conflicting	c.371G>A	1	G124E	1	1.0

Simple ClinVar	Uncertain/conflicting	c.401A>G	1	N134S	0	0.0
Simple ClinVar	Uncertain/conflicting	c.403A>G	1	T135A	0	0.1
Simple ClinVar	Uncertain/conflicting	c.425T>G	1	V142G	1	0.9
Simple ClinVar	Uncertain/conflicting	c.431C>T	1	S144F	1	0.9
Simple ClinVar	Uncertain/conflicting	c.434A>G	1	N145S	1	0.8
Simple ClinVar/GenMed	Likely pathogenic	c.434A>T	2	N145I	1	0.8
Simple ClinVar	Uncertain/conflicting	c.436G>A	1	A146T	1	0.6
ClinVar	Uncertain significance	c.443G>T	1	C148F	1	0.8
ClinVar	Likely benign	c.463C>T	1	L155F	0	0.1
ClinVar	Uncertain significance	c.467T>C	1	M156T	1	0.6
Simple ClinVar	Uncertain/conflicting	c.476A>G	1	K159R	0	0.0
Simple ClinVar	Uncertain/conflicting	c.503G>C	1	R168T	0	0.0
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.515T>C	1	L172P	1	0.9
ClinVar	Uncertain significance	c.529A>G	1	N177D	0	0.1
GenMed	Pathogenic	c.545T>C	2	L182P	1	1.0
GenMed	Pathogenic	c.548A>G	1	Q183R	1	0.9
ClinVar	Uncertain significance	c.566A>C	1	Q189P	0	0.0
Simple ClinVar	Uncertain/conflicting	c.584T>C	1	V195A	0	0.2
Simple ClinVar	Uncertain/conflicting	c.583G>A	1	V195M	0	0.1
Simple ClinVar	Uncertain/conflicting	c.589G>C	1	D197H	0	0.0
Simple ClinVar	Uncertain/conflicting	c.589G>A	1	D197N	0	0.1
ClinVar	Uncertain significance	c.593A>T	1	D198V	1	0.5
Simple ClinVar	Uncertain/conflicting	c.613A>G	1	M205V	1	0.9
ClinVar	Uncertain significance	c.614T>G	1	M205R	1	0.9
Simple ClinVar	Uncertain/conflicting	c.625G>A	1	V209M	0	0.0
Simple ClinVar	Uncertain/conflicting	c.631G>A	1	E211K	0	0.0
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.635T>C	1	V212A	1	0.6
Simple ClinVar	Uncertain/conflicting	c.638T>G	1	F213C	0	0.4
ClinVar	Uncertain significance	c.643G>A	1	V215M	0	0.1
Simple ClinVar	Uncertain/conflicting	c.643G>C	1	V215L	0	0.0
ClinVar	Uncertain significance	c.656G>A	1	R219H	0	0.4
ClinVar	Uncertain significance	c.658G>A	1	V220M	0	0.4
ClinVar	Likely pathogenic	c.664C>A	1	R222S	1	0.8
Simple ClinVar/GenMed	Pathogenic	c.665G>A	4	R222H	1	0.5
Simple ClinVar	Likely pathogenic	c.665G>T	1	R222L	1	0.9
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.664C>T	12	R222C	1	0.9
GenMed	Pathogenic	c.674G>C	1	Y225C	1	0.6
Simple ClinVar	Likely pathogenic	c.674A>G	1	Y225C	1	0.6
ClinVar	Uncertain significance	c.677C>T	1	T226I	1	0.9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.685G>C	1	D229H	1	1.0
Simple ClinVar	Uncertain/conflicting	c.701C>T	1	S234L	1	0.6
Simple ClinVar	Uncertain/conflicting	c.713C>T	1	A238V	0	0.0
ClinVar	Uncertain significance	c.719C>T	1	A240V	0	0.0
ClinVar	Uncertain significance	c.721G>A	1	V241I	0	0.0
Simple ClinVar	Uncertain/conflicting	c.727C>A	1	L243M	0	0.2
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.752C>G	1	P251R	0	0.0
Simple ClinVar/GenMed/PubTator	Pathogenic	c.794G>A	5	R265H	1	1.0
GenMed	Pathogenic	c.793C>A	3	R265S	1	0.9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.793C>T	4	R265C	1	0.8
Simple ClinVar	Likely pathogenic	c.797T>C	1	L266P	1	1.0
Simple ClinVar	Uncertain/conflicting	c.805G>A	1	G269S	1	0.9
Simple ClinVar/GenMed	Pathogenic	c.806G>A	1	G269D	1	1.0

Simple ClinVar	Uncertain/conflicting	c.825T>A	1	N275K	0	0.3
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.826G>C	4	D276H	1	1.0
PubTator	Pathogenic	c.955G>C	1	D276H	1	1.0
Simple ClinVar	Uncertain/conflicting	c.830T>C	1	I277T	0	0.4
ClinVar	Uncertain significance	c.830T>A	1	I277N	1	1.0
GenMed	Pathogenic	c.845C>T	1	T282I	1	0.9
ClinVar	Pathogenic	c.847C>T	1	P283S	1	1.0
ClinVar	Uncertain significance	c.850G>A	1	V284M	1	0.5
Simple ClinVar	Uncertain/conflicting	c.856G>A	1	G286S	0	0.4
Simple ClinVar	Uncertain/conflicting	c.857G>A	1	G286D	1	1.0
Simple ClinVar	Uncertain/conflicting	c.883C>T	1	R295C	1	0.6
Simple ClinVar/GenMed	Pathogenic	c.887G>A	7	R296H	0	0.4
GenMed	Pathogenic	c.887G>C	1	R296P	1	0.9
ClinVar	Likely pathogenic	c.887G>T	1	R296L	1	0.9
Simple ClinVar/GenMed	Pathogenic	c.886C>T	4	R296C	1	0.9
ClinVar	Uncertain significance	c.890G>A	1	R297Q	1	0.7
Simple ClinVar	Uncertain/conflicting	c.892G>A	1	A298T	0	0.0
Simple ClinVar	Benign/Likely benign	c.898A>G	1	M300V	0	0.0
ClinVar	Uncertain significance	c.911G>A	1	G304E	1	1.0
Simple ClinVar	Uncertain/conflicting	c.944G>A	1	R315K	0	0.0
ClinVar	Uncertain significance	c.950A>G	1	Q317R	0	0.0
ClinVar	Uncertain significance	c.953G>A	1	R318Q	0	0.3
ClinVar	Uncertain significance	c.952C>T	1	R318W	0	0.3
Simple ClinVar/GenMed	Pathogenic	c.959G>A	36	R320H	1	0.9
GenMed	Pathogenic	c.958C>G	2	R320G	1	0.8
Simple ClinVar	Uncertain/conflicting	c.959G>T	1	R320L	1	0.8
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.958C>T	2	R320C	1	0.9
Simple ClinVar	Uncertain/conflicting	c.962T>C	1	V321A	1	0.6
Simple ClinVar	Uncertain/conflicting	c.961G>A	1	V321M	1	0.8
ClinVar	Uncertain significance	c.972G>A	1	M324I	0	0.0
Simple ClinVar	Uncertain/conflicting	c.982G>A	1	A328T	0	0.2
Simple ClinVar/GenMed	Pathgenic/Likely pathogenic	c.992G>A	4	R331Q	1	0.8
GenMed	Pathogenic	c.992G>C	1	R331P	1	0.9
ClinVar	Uncertain significance	c.1001G>T	1	S334I	0	0.1
Simple ClinVar	Uncertain/conflicting	c.1021G>A	1	G341S	0	0.2
Simple ClinVar	Uncertain/conflicting	c.1030A>T	1	I344F	0	0.3
Simple ClinVar	Uncertain/conflicting	c.1058C>A	1	S353Y	1	0.9
ClinVar	Uncertain significance	c.1076A>C	1	N359T	1	0.5
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.1082C>T	1	A361V	1	0.5
GenMed	Pathogenic	c.992G>C	1	G363R	1	1.0
Simple ClinVar	Uncertain/conflicting	c.1087G>C	1	G363R	1	1.0
Simple ClinVar	Uncertain/conflicting	c.1097C>T	1	P366L	0	0.1
ClinVar	Uncertain significance	c.1111C>T	1	R371C	0	0.1
Simple ClinVar	Uncertain/conflicting	c.1112G>A	1	R371H	0	0.0
Simple ClinVar	Uncertain/conflicting	c.1136A>C	1	E379A	0	0.3
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.1138G>A	1	V380M	1	0.5
Simple ClinVar	Uncertain/conflicting	c.1141C>T	1	R381W	1	0.6
Simple ClinVar	Uncertain/conflicting	c.1145G>A	1	R382Q	0	0.1
Simple ClinVar	Uncertain/conflicting	c.1157T>C	1	M386T	0	0.0
Simple ClinVar	Uncertain/conflicting	c.1190C>A	1	T397K	1	0.8
Simple ClinVar	Uncertain/conflicting	c.1199A>G	1	Y400C	1	0.8
Simple ClinVar	Uncertain/conflicting	c.1202C>A	1	T401N	0	0.0

Appendix 9 – ConSurf results– The Evolutionary Importance of Amino Acid Residues in the T-Protein

Source	ClinicalSignificance	Codon	SAV	LYRUS	Provean	SNAP2	Mutaframe	SIFT	Majority Vote	ConSurf
Simple ClinVar	Pathogenic	c.2T>A	M1K	neutral	neutral	neutral	NA	effect		5*
GenMed	Pathogenic	c.2T>C	M1T	neutral	neutral	neutral	NA	effect		5*
PubTator	Likely pathogenic		V7L	neutral	neutral	neutral	NA	neutral	Likely benign	7*
Simple ClinVar	Uncertain/conflicting	c.59C>T	P20L	neutral	neutral	neutral	neutral	neutral	Likely benign	5*
Simple ClinVar	Uncertain/conflicting	c.70C>T	R24C	neutral	neutral	neutral	effect	effect	Likely benign	5
Simple ClinVar	Uncertain/conflicting	c.86C>T	A29V	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.101G>A	R34H	neutral	neutral	effect	effect	neutral	Likely benign	5
Simple ClinVar	Uncertain/conflicting	c.100C>T	R34C	neutral	effect	effect	effect	effect	Likely pathologic	5
ClinVar	Uncertain significance	c.110C>T	P37L	neutral	effect	neutral	effect	neutral	Likely benign	5
Simple ClinVar	Pathogenic	c.125A>G	H42R	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.130G>A	A44T	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Likely pathogenic	c.139G>T	G47W	effect	neutral	effect	effect	effect	Likely pathologic	7
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.139G>A	G47R	effect	neutral	effect	effect	effect	Likely pathologic	7
Simple ClinVar	Uncertain/conflicting	c.148G>T	V50L	neutral	neutral	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.148G>C	V50L	neutral	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.152C>T	A51V	neutral	neutral	neutral	effect	effect	Likely benign	5
ClinVar	Uncertain significance	c.155T>C	F52S	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.155T>G	F52C	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.158C>T	A53V	effect	effect	effect	effect	effect	Likely pathologic	8
ClinVar	Uncertain significance	c.170T>C	L57P	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.173C>T	P58L	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.188A>G	D63G	neutral	effect	neutral	effect	neutral	Likely benign	3
Simple ClinVar	Uncertain/conflicting	c.194A>G	H65R	neutral	neutral	effect	effect	neutral	Likely benign	7
Simple ClinVar	Uncertain/conflicting	c.196A>G	T66A	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.203C>T	S68L	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.212A>C	H71P	effect	effect	effect	effect	neutral	Likely pathologic	7
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.217C>T	R73C	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.218G>A	R73H	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.230C>T	S77L	effect	effect	effect	effect	neutral	Likely pathologic	8
GenMed	Pathogenic	c.248A>G	H83R	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.266T>G	I89R	effect	effect	effect	effect	effect	Likely pathologic	5
ClinVar	Uncertain significance	c.269_270 delinsCC	L90P	neutral	neutral	neutral	effect	neutral	Likely benign	3
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.280C>T	R94W	effect	effect	neutral	effect	effect	Likely pathologic	6
Simple ClinVar	Uncertain/conflicting	c.283G>A	V95M	neutral	neutral	effect	effect	effect	Likely	4

									pathologic	
ClinVar/GenMed	Uncertain significance/Pathogenic	c.293T>G	M98R	effect	effect	effect	NA	effect	Likely pathologic	6
Simple ClinVar	Uncertain/conflicting	c.298A>G	S100G	neutral	effect	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Likely pathogenic	c.311G>A	G104E	effect	effect	effect	effect	neutral	Likely pathologic	7
Simple ClinVar/GenMed	Pathogenic	c.317T>C	I106T	effect	effect	effect	effect	effect	Likely pathologic	6
GenMed	Pathogenic	c.280C>T	I106T	effect	effect	effect	effect	effect	Likely pathologic	6
ClinVar	Uncertain significance	c.335A>G	N112S	neutral	effect	neutral	effect	neutral	Likely benign	6
Simple ClinVar	Uncertain/conflicting	c.338A>G	Q113R	neutral	neutral	neutral	effect	neutral	Likely benign	6
Simple ClinVar	Uncertain/conflicting	c.341G>A	G114E	effect	effect	effect	effect	neutral	Likely pathologic	7
Simple ClinVar/ClinVar/GenMed	Pathogenic	c.350C>T	S117L	effect	effect	effect	effect	effect	Likely pathologic	7
Simple ClinVar	Uncertain/conflicting	c.359C>T	T120I	neutral	effect	effect	effect	neutral	Likely pathologic	7
Simple ClinVar	Uncertain/conflicting	c.362A>G	N121S	neutral	effect	effect	effect	neutral	Likely pathologic	7
Simple ClinVar	Uncertain/conflicting	c.371G>A	G124E	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.401A>G	N134S	neutral	neutral	neutral	effect	neutral	Likely benign	5
Simple ClinVar	Uncertain/conflicting	c.403A>G	T135A	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.425T>G	V142G	effect	effect	effect	effect	effect	Likely pathologic	5
Simple ClinVar	Uncertain/conflicting	c.431C>T	S144F	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.434A>G	N145S	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Likely pathogenic	c.434A>T	N145I	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.436G>A	A146T	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.443G>T	C148F	effect	effect	effect	effect	effect	Likely pathologic	7
ClinVar	Likely benign	c.463C>T	L155F	neutral	neutral	neutral	neutral	neutral	Likely benign	5
ClinVar	Uncertain significance	c.467T>C	M156T	effect	effect	effect	effect	effect	Likely pathologic	5
Simple ClinVar	Uncertain/conflicting	c.476A>G	K159R	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.503G>C	R168T	neutral	neutral	neutral	neutral	neutral	Likely benign	2
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.515T>C	L172P	effect	effect	effect	effect	effect	Likely pathologic	4
ClinVar	Uncertain significance	c.529A>G	N177D	neutral	neutral	neutral	effect	neutral	Likely benign	4
GenMed	Pathogenic	c.545T>C	L182P	effect	effect	effect	NA	effect	Likely pathologic	7
GenMed	Pathogenic	c.548A>G	Q183R	effect	effect	effect	NA	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.566A>C	Q189P	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.584T>C	V195A	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.583G>A	V195M	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.589G>C	D197H	neutral	neutral	neutral	effect	effect	Likely benign	1
Simple ClinVar	Uncertain/conflicting	c.589G>A	D197N	neutral	neutral	neutral	effect	neutral	Likely benign	1
ClinVar	Uncertain significance	c.593A>T	D198V	effect	effect	neutral	effect	effect	Likely pathologic	4
Simple ClinVar	Uncertain/conflicting	c.613A>G	M205V	effect	effect	neutral	effect	neutral	Likely pathologic	9
ClinVar	Uncertain significance	c.614T>G	M205R	effect	effect	effect	effect	neutral	Likely pathologic	9

Simple ClinVar	Uncertain/conflicting	c.625G>A	V209M	neutral	neutral	neutral	effect	neutral	Likely benign	2
Simple ClinVar	Uncertain/conflicting	c.631G>A	E211K	neutral	neutral	effect	effect	neutral	Likely benign	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.635T>C	V212A	effect	effect	effect	effect	effect	Likely pathologic	4
Simple ClinVar	Uncertain/conflicting	c.638T>G	F213C	neutral	effect	neutral	effect	neutral	Likely benign	3
ClinVar	Uncertain significance	c.643G>A	V215M	neutral	neutral	neutral	effect	neutral	Likely benign	3
Simple ClinVar	Uncertain/conflicting	c.643G>C	V215L	neutral	neutral	neutral	effect	neutral	Likely benign	3
ClinVar	Uncertain significance	c.656G>A	R219H	neutral	effect	neutral	effect	neutral	Likely pathologic	4
ClinVar	Uncertain significance	c.658G>A	V220M	neutral	neutral	neutral	effect	effect	Likely benign	6
ClinVar	Likely pathogenic	c.664C>A	R222S	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Pathogenic	c.665G>A	R222H	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Likely pathogenic	c.665G>T	R222L	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.664C>T	R222C	effect	effect	effect	effect	effect	Likely pathologic	9
GenMed	Pathogenic	c.674G>C	Y225C	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Likely pathogenic	c.674A>G	Y225C	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.677C>T	T226I	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.685G>C	D229H	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.701C>T	S234L	effect	effect	neutral	effect	neutral	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.713C>T	A238V	neutral	neutral	neutral	effect	neutral	Likely benign	3
ClinVar	Uncertain significance	c.719C>T	A240V	neutral	neutral	neutral	effect	effect	Likely benign	7
ClinVar	Uncertain significance	c.721G>A	V241I	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.727C>A	L243M	neutral	neutral	neutral	effect	neutral	Likely benign	4
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.752C>G	P251R	neutral	effect	neutral	effect	effect	Likely pathologic	4
Simple ClinVar/GenMed/ PubTator	Pathogenic	c.794G>A	R265H	effect	effect	effect	effect	effect	Likely pathologic	9
GenMed	Pathogenic	c.793C>A	R265S	effect	effect	effect	NA	effect	Likely pathologic	9
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.793C>T	R265C	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Likely pathogenic	c.797T>C	L266P	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.805G>A	G269S	effect	effect	effect	effect	neutral	Likely pathologic	9
Simple ClinVar/GenMed	Pathogenic	c.806G>A	G269D	effect	effect	effect	effect	neutral	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.825T>A	N275K	neutral	effect	effect	effect	effect	Likely pathologic	7
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	c.826G>C	D276H	effect	effect	effect	effect	effect	Likely pathologic	9
PubTator	Pathogenic	c.955G>C	D276H	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.830T>C	I277T	neutral	effect	effect	effect	effect	Likely pathologic	7
ClinVar	Uncertain significance	c.830T>A	I277N	effect	effect	effect	effect	effect	Likely pathologic	7
GenMed	Pathogenic	c.845C>T	T282I	effect	effect	effect	effect	effect	Likely pathologic	7
ClinVar	Pathogenic	c.847C>T	P283S	effect	effect	effect	effect	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.850G>A	V284M	effect	effect	neutral	effect	effect	Likely	6

									pathologic	
Simple ClinVar	Uncertain/conflicting	c.856G>A	G286S	neutral	neutral	effect	neutral	effect	Likely benign	8
Simple ClinVar	Uncertain/conflicting	c.857G>A	G286D	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.883C>T	R295C	effect	effect	neutral	effect	effect	Likely pathologic	4
Simple ClinVar/GenMed	Pathogenic	c.887G>A	R296H	neutral	effect	effect	effect	effect	Likely pathologic	8
GenMed	Pathogenic	c.887G>C	R296P	effect	effect	effect		effect	Likely pathologic	8
ClinVar	Likely pathogenic	c.887G>T	R296L	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar/GenMed	Pathogenic	c.886C>T	R296C	effect	effect	effect	effect	effect	Likely pathologic	8
ClinVar	Uncertain significance	c.890G>A	R297Q	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.892G>A	A298T	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Benign/Likely benign	c.898A>G	M300V	neutral	neutral	neutral	neutral	neutral	Likely benign	5
ClinVar	Uncertain significance	c.911G>A	G304E	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.944G>A	R315K	neutral	neutral	neutral	neutral	neutral	Likely benign	6
ClinVar	Uncertain significance	c.950A>G	Q317R	neutral	neutral	neutral	effect	neutral	Likely benign	1
ClinVar	Uncertain significance	c.953G>A	R318Q	neutral	neutral	neutral	effect	neutral	Likely benign	6
ClinVar	Uncertain significance	c.952C>T	R318W	neutral	effect	effect	effect	effect	Likely pathologic	6
Simple ClinVar/GenMed	Pathogenic	c.959G>A	R320H	effect	effect	effect	effect	effect	Likely pathologic	8
GenMed	Pathogenic	c.958C>G	R320G	effect	effect	effect	NA	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.959G>T	R320L	effect	effect	effect	effect	neutral	Likely pathologic	8
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.958C>T	R320C	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.962T>C	V321A	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.961G>A	V321M	effect	effect	effect	effect	effect	Likely pathologic	8
ClinVar	Uncertain significance	c.972G>A	M324I	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.982G>A	A328T	neutral	neutral	neutral	effect	neutral	Likely benign	6
Simple ClinVar/GenMed	Pathgenic/Likely pathogenic	c.992G>A	R331Q	effect	effect	effect	effect	effect	Likely pathologic	9
GenMed	Pathogenic	c.992G>C	R331P	effect	effect	effect	NA	effect	Likely pathologic	9
ClinVar	Uncertain significance	c.1001G>T	S334I	neutral	neutral	neutral	effect	neutral	Likely benign	6
Simple ClinVar	Uncertain/conflicting	c.1021G>A	G341S	neutral	effect	neutral	effect	effect	Likely pathologic	4
Simple ClinVar	Uncertain/conflicting	c.1030A>T	I344F	neutral	effect	effect	effect	effect	Likely pathologic	5
Simple ClinVar	Uncertain/conflicting	c.1058C>A	S353Y	effect	effect	effect	effect	effect	Likely pathologic	8
ClinVar	Uncertain significance	c.1076A>C	N359T	effect	effect	effect	effect	effect	Likely pathologic	7
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	c.1082C>T	A361V	effect	effect	effect	effect	effect	Likely pathologic	8
GenMed	Pathogenic	c.992G>C	G363R	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.1087G>C	G363R	effect	effect	effect	effect	effect	Likely pathologic	8
Simple ClinVar	Uncertain/conflicting	c.1097C>T	P366L	neutral	effect	neutral	effect	neutral	Likely benign	2
ClinVar	Uncertain significance	c.1111C>T	R371C	neutral	effect	neutral	effect	effect	Likely pathologic	4
Simple ClinVar	Uncertain/conflicting	c.1112G>A	R371H	neutral	effect	neutral	effect	effect	Likely	4

									pathologic	
Simple ClinVar	Uncertain/conflicting	c.1136A>C	E379A	neutral	effect	neutral	effect	neutral	Likely benign	4
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	c.1138G>A	V380M	effect	neutral	effect	NA	effect	Likely pathologic	7
Simple ClinVar	Uncertain/conflicting	c.1141C>T	R381W	effect	effect	effect	effect	effect	Likely pathologic	9
Simple ClinVar	Uncertain/conflicting	c.1145G>A	R382Q	neutral	neutral	neutral	effect	neutral	Likely benign	5
Simple ClinVar	Uncertain/conflicting	c.1157T>C	M386T	neutral	neutral	neutral	neutral	neutral	Likely benign	4
Simple ClinVar	Uncertain/conflicting	c.1190C>A	T397K	effect	effect	neutral	effect	effect	Likely pathologic	5
Simple ClinVar	Uncertain/conflicting	c.1199A>G	Y400C	effect	effect	effect	effect	effect	Likely pathologic	5
Simple ClinVar	Uncertain/conflicting	c.1202C>A	T401N	neutral	neutral	neutral	neutral	neutral	Likely benign	5

Appendix 10 – InterPro results– Distribution of mutations NKH-causing per Domains and Functional Sites in the T-Protein

T-Protein Change	Frequency	Domain	Funcional Sites
M1T	7		N-terminal region of a signal peptide
Q2X	1		N-terminal region of a signal peptide
S6Wfs*89	7		Hydrophobic region of a signal peptide.
S6Kfs*22	2		Hydrophobic region of a signal peptide.
S6Vfs*90	2		Hydrophobic region of a signal peptide.
R34C	1	GCV_T_N	Aminomethyltransferase folate-binding domain
H42R	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G47R	4	GCV_T_N	Aminomethyltransferase folate-binding domain
G47W	1	GCV_T_N	Aminomethyltransferase folate-binding domain
V50L	2	GCV_T_N	Aminomethyltransferase folate-binding domain
F52S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
F52C	1	GCV_T_N	Aminomethyltransferase folate-binding domain
A53V	1	GCV_T_N	Aminomethyltransferase folate-binding domain
L57P	1	GCV_T_N	Aminomethyltransferase folate-binding domain
P58L	1	GCV_T_N	Aminomethyltransferase folate-binding domain
Q60Tfs*35	2	GCV_T_N	Aminomethyltransferase folate-binding domain
S68L	1	GCV_T_N	Aminomethyltransferase folate-binding domain
H71P	2	GCV_T_N	Aminomethyltransferase folate-binding domain
R73C	8	GCV_T_N	Aminomethyltransferase folate-binding domain
R73H	1	GCV_T_N	Aminomethyltransferase folate-binding domain
S77L	4	GCV_T_N	Aminomethyltransferase folate-binding domain
H83R	4	GCV_T_N	Aminomethyltransferase folate-binding domain
T87_Q113del	1	GCV_T_N	Aminomethyltransferase folate-binding domain
I89R	1	GCV_T_N	Aminomethyltransferase folate-binding domain
R94W	9	GCV_T_N	Aminomethyltransferase folate-binding domain
V95M	1	GCV_T_N	Aminomethyltransferase folate-binding domain
M98R	2	GCV_T_N	Aminomethyltransferase folate-binding domain
G104E	1	GCV_T_N	Aminomethyltransferase folate-binding domain
I106T	14	GCV_T_N	Aminomethyltransferase folate-binding domain
G114E	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G114_Q157del	1	GCV_T_N	Aminomethyltransferase folate-binding domain
S117L	4	GCV_T_N	Aminomethyltransferase folate-binding domain
T120I	1	GCV_T_N	Aminomethyltransferase folate-binding domain
N121S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G124E	1	GCV_T_N	Aminomethyltransferase folate-binding domain
V132_N134delinsD	1	GCV_T_N	Aminomethyltransferase folate-binding domain

V142G	1	GCV_T_N	Aminomethyltransferase folate-binding domain
S144F	1	GCV_T_N	Aminomethyltransferase folate-binding domain
N145I	2	GCV_T_N	Aminomethyltransferase folate-binding domain
N145S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
A146T	1	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
C148F	1	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
K151Cfs*25	2	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
K151Nfs*22	2	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
K151_L155del	6	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
M156T	1	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
Q166X	3	GCV_T_N	Aminomethyltransferase folate-binding domain, Coil-COILS entry
L172P	1	GCV_T_N	Aminomethyltransferase folate-binding domain
L179Pfs*3	2	GCV_T_N	Aminomethyltransferase folate-binding domain
L182P	2	GCV_T_N	Aminomethyltransferase folate-binding domain
Q183R	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G184E	2	GCV_T_N	Aminomethyltransferase folate-binding domain
Q192X	1	GCV_T_N	Aminomethyltransferase folate-binding domain
D198V	1	GCV_T_N	Aminomethyltransferase folate-binding domain
M205V	1	GCV_T_N	Aminomethyltransferase folate-binding domain
M205R	1	GCV_T_N	Aminomethyltransferase folate-binding domain
V212A	1	GCV_T_N	Aminomethyltransferase folate-binding domain
R222C	12	GCV_T_N	Aminomethyltransferase folate-binding domain
R222H	4	GCV_T_N	Aminomethyltransferase folate-binding domain
R222L	1	GCV_T_N	Aminomethyltransferase folate-binding domain
R222S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
Y225C	2	GCV_T_N	Aminomethyltransferase folate-binding domain
T226I	1	GCV_T_N	Aminomethyltransferase folate-binding domain
D229H	1	GCV_T_N	Aminomethyltransferase folate-binding domain
D229Gfs*10	1	GCV_T_N	Aminomethyltransferase folate-binding domain
S234L	1	GCV_T_N	Aminomethyltransferase folate-binding domain
T245Nfs*32	1	GCV_T_N	Aminomethyltransferase folate-binding domain
P251R	1	GCV_T_N	Aminomethyltransferase folate-binding domain
R265H	5	GCV_T_N	Aminomethyltransferase folate-binding domain
R265C	4	GCV_T_N	Aminomethyltransferase folate-binding domain
R265S	3	GCV_T_N	Aminomethyltransferase folate-binding domain
L266P	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G269D	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G269S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
L270_C271delinsRG	1	GCV_T_N	Aminomethyltransferase folate-binding domain
N275K	1	GCV_T_N	Aminomethyltransferase folate-binding domain
D276H	4	GCV_T_N	Aminomethyltransferase folate-binding domain
I277T	1	GCV_T_N	Aminomethyltransferase folate-binding domain
I277N	1	GCV_T_N	Aminomethyltransferase folate-binding domain
T282I	1	GCV_T_N	Aminomethyltransferase folate-binding domain
P283S	1	GCV_T_N	Aminomethyltransferase folate-binding domain
V284M	1	GCV_T_N	Aminomethyltransferase folate-binding domain
G286D	1	GCV_T_N	Aminomethyltransferase folate-binding domain
K294fs	1		

R295C	1		
R296H	7		
R296C	4		
R296P	1		
R296L	1		
R297Q	1		
G304E	1		
R318W	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R320H	36	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R320C	2	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R320G	2	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R320L	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
V321A	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
V321M	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
M330Cfs*8	3	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R331Q	4	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R331P	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
H333Tfs*17	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
H333Qfs*17	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
H333Qfs*16	2	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
G341S	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
V347Dfs*2	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
S353Y	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
S355Lfs*2	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
N359T	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
A361V	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
G363R	2	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
C367X	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
Y369X	4	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R371C	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R371H	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
V380M	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
R381W	1	GCV_T_C	Glycine cleavage T-protein C-terminal barrel domain
T397K	1		
Y400C	1		

Appendix 11 – Survey results– Distribution of mutations NKH-causing in the T-Protein

Source	Protein Change	Frequency	Type	POS AA
GenMed	M1T	7	SNV	1
GenMed	Q2Ter	1	SNV	2
PubTator	S6KfsTer22	1	Insertion	6
GenMed	S6Wfs*89	1	Deletion	6
GenMed	S6Wfs*89	6	Deletion	6
Simple ClinVar	S6fs	1	Duplication	6
Simple ClinVar	S6fs	1	Deletion	6
Simple ClinVar	S6fs	1	Deletion	6
GenMed	S6Vfs*90	2	Deletion	6
GenMed	S6Kfs*22	2	Duplication	6
Simple ClinVar	V7fs	1	Microsatellite	7
Simple ClinVar	P20fs	1	Deletion	20
Simple ClinVar	A21fs	1	Deletion	21

Simple ClinVar	L22fs	1	Deletion	22
Simple ClinVar	R34C	1	SNV	34
Simple ClinVar	H42R	1	SNV	42
Simple ClinVar	G47W	1	SNV	47
Simple ClinVar/GenMed	G47R	4	SNV	47
Simple ClinVar	K48fs	1	Deletion	48
Simple ClinVar	V50L	1	SNV	50
Simple ClinVar	V50L	1	SNV	50
Simple ClinVar	V50fs	1	Deletion	50
ClinVar	F52S	1	SNV	52
ClinVar	F52C	1	SNV	52
Simple ClinVar	A53V	1	SNV	53
Simple ClinVar	G54_W55insTer	1	Deletion	54
Simple ClinVar	W55Ter	1	SNV	55
ClinVar	L57P	1	SNV	57
Simple ClinVar	L57fs	1	Deletion	57
Simple ClinVar	P58L	1	SNV	58
Simple ClinVar	Q60Ter	1	SNV	60
GenMed	Q60Tfs*35	2	Deletion	60
Simple ClinVar	S68L	1	SNV	68
Simple ClinVar/GenMed	H71P	2	SNV	71
Simple ClinVar/GenMed	R73C	8	SNV	73
ClinVar	R73H	1	SNV	73
ClinVar	Q74R	1	Deletion	74
ClinVar	H75fs	1	Deletion	75
Simple ClinVar/GenMed	S77L	4	SNV	77
ClinVar	S77Ter	1	SNV	77
GenMed	H83R	4	SNV	83
Simple ClinVar	Q86Ter	1	SNV	86
Simple ClinVar	T87fs	1	Duplication	87
GenMed	T87_Q113del	1	Deletion	87
ClinVar	I89R	1	SVN	89
ClinVar	G91fs	1	Insertion	91
ClinVar	G91del	1	Deletion	91
Simple ClinVar/GenMed	R94W	9	SNV	94
Simple ClinVar	V95M	1	SNV	95
ClinVar/GenMed	M98R	2	SNV	98
Simple ClinVar	G104E	1	SNV	104
Simple ClinVar/GenMed	I106T	5	SNV	106
GenMed	Q113Q	1	SNV	113
Simple ClinVar	G114E	1	SNV	114
GenMed	G114_Q157del	1	Deletion	114
Simple ClinVar/ClinVar/GenMed	S117L	4	SNV	117
Simple ClinVar	S117fs	1	Deletion	117
Simple ClinVar	T120I	1	SNV	120
Simple ClinVar	N121S	1	SNV	121
Simple ClinVar	G124E	1	SNV	124
ClinVar	D128fs	1	Indel	128
ClinVar	D128fs	1	Deletion	128
GenMed	V132_N134delinsD	1	Deletion	132
Simple ClinVar	V142G	1	SNV	142
Simple ClinVar	S144F	1	SNV	144
Simple ClinVar	N145S	1	SNV	145

Simple ClinVar/GenMed	N145I	2	SNV	145
Simple ClinVar	A146T	1	SNV	146
ClinVar	C148F	1	SNV	148
GenMed	K151_L155del	6	Deletion	151
Simple ClinVar	K151_L155del	1	Deletion	151
GenMed	K151Nfs*22	2	Deletion	151
GenMed	K151Cfs*25	2	Deletion	151
ClinVar	M156T	1	SVN	156
Simple ClinVar	V160fs	1	Deletion	160
Simple ClinVar/GenMed	Q166Ter	3	SNV	166
Simple ClinVar/GenMed	L172P	1	SNV	172
GenMed	L179Pfs*3	2	Insertion	179
Simple ClinVar	L179fs	1	Deletion	179
GenMed	L182P	2	SNV	182
GenMed	Q183R	1	SNV	183
GenMed	G184E	2	Indel	184
Simple ClinVar/GenMed	Q192Ter	1	SNV	192
ClinVar	D198V	1	SNV	198
Simple ClinVar	K201fs	1	Deletion	201
Simple ClinVar	F204fs	1	Duplication	204
Simple ClinVar	M205V	1	SNV	205
ClinVar	M205R	1	SNV	205
Simple ClinVar/GenMed	V212A	1	SNV	212
Simple ClinVar	V220fs	1	Duplication	220
ClinVar	R222S	1	SNV	222
Simple ClinVar/GenMed	R222H	4	SNV	222
Simple ClinVar	R222L	1	SNV	222
Simple ClinVar/GenMed	R222C	12	SNV	222
GenMed	Y225C	1	SNV	225
GenMed	Y225C	1	SNV	225
ClinVar	T226I	1	SNV	226
Simple ClinVar/GenMed	D229H	1	SNV	229
GenMed	D229Gfs*10	1	Deletion	229
Simple ClinVar	S234L	1	SNV	234
Simple ClinVar	P236_V237dup	1	Duplication	236
Simple ClinVar	T245fs	1	Deletion	245
GenMed	T245Nfs*32	1	Duplication	245
Simple ClinVar/PubTator	P251R	1	SNV	251
Simple ClinVar/GenMed/PubTator	R265H	5	SNV	265
GenMed	R265S	3	SNV	265
Simple ClinVar/GenMed	R265C	4	SNV	265
Simple ClinVar	L266P	1	SNV	266
Simple ClinVar	G269S	1	SNV	269
Simple ClinVar/GenMed	G269D	1	SNV	269
GenMed	L270_C271delinsRG	1	Indel	270
ClinVar	Y273Ter	1	SNV	273
Simple ClinVar	N275K	1	SNV	275
Simple ClinVar/GenMed	D276H	4	SNV	276
Simple ClinVar/GenMed	D276H	1	SNV	276
Simple ClinVar	I277T	1	SNV	277
ClinVar	I277N	1	SNV	277
GenMed	T282I	1	SNV	282
ClinVar	P283S	1	SNV	283

ClinVar	S283fs	1	Insertion	283
ClinVar	V284M	1	SVN	284
Simple ClinVar	V284fs	1	Duplication	284
Simple ClinVar	G286D	1	SNV	286
Simple ClinVar	W290Ter	1	SNV	290
Simple ClinVar	L292fs	1	Deletion	292
GenMed	K294fs	1	SNV	294
Simple ClinVar	R295C	1	SNV	295
Simple ClinVar/GenMed	R296H	7	SNV	296
GenMed	R296P	1	SNV	296
ClinVar	R296L	1	SNV	296
Simple ClinVar/GenMed	R296C	4	SNV	296
ClinVar	R297Q	1	SNV	297
ClinVar	R297Ter	1	SNV	297
Simple ClinVar	P303fs	1	Deletion	303
ClinVar	G304E	1	SNV	304
GenMed	V307V	1	SNV	307
ClinVar	R318W	1	SNV	318
Simple ClinVar/GenMed	R320H	36	SNV	320
GenMed	R320G	2	SNV	320
Simple ClinVar	R320L	1	SNV	320
Simple ClinVar/GenMed	R320C	2	SNV	320
Simple ClinVar	V321A	1	SNV	321
Simple ClinVar	V321M	1	SNV	321
Simple ClinVar	M324del	1	Deletion	324
PubTator	E326Gfs*12	1	Deletion	326
Simple ClinVar	A328fs	1	Duplication	328
Simple ClinVar	A328fs	1	Deletion	328
Simple ClinVar	A328fs	1	Indel	328
PubTator	R328Gfs*22	1	SNV	328
GenMed	M330Cfs*8	3	Deletion	330
Simple ClinVar	M330fs	1	Deletion	330
Simple ClinVar/GenMed	R331Q	4	SNV	331
GenMed	R331P	1	SNV	331
GenMed	H333Tfs*17	1	Insertion	333
Simple ClinVar	H333fs	1	Duplication	333
GenMed	H333Qfs*16	2	Deletion	333
GenMed	H333Qfs*17	1	Duplication	333
Simple ClinVar	G341S	1	SNV	341
Simple ClinVar	I344F	1	SNV	344
GenMed	V347Dfs*2	1	Deletion	347
Simple ClinVar	S353Y	1	SNV	353
Simple ClinVar	S353fs	1	Deletion	353
GenMed	S355Lfs*2	1	Deletion	355
ClinVar	K358del	1	Deletion	358
ClinVar	N359T	1	SVN	359
Simple ClinVar/GenMed	A361V	1	SNV	361
GenMed	G363R	1	SNV	363
GenMed	G363R	1	SNV	363
GenMed	C367Ter	1	SNV	367
Simple ClinVar	Y369_S370delinsTer	1	Deletion	369
GenMed	Y369Ter	4	Deletion	369
ClinVar	R371C	1	SVN	371

Simple ClinVar	R371H	1	SNV	371
Simple ClinVar/PubTator	V380M	1	SNV	380
Simple ClinVar	R381W	1	SNV	381
Simple ClinVar	Q385del	1	Microsatellite	385
Simple ClinVar	Q385Ter	1	SNV	385
Simple ClinVar	T397K	1	SNV	397
Simple ClinVar	Y400C	1	SNV	400
Simple ClinVar	P400fs	1	Indel	400
Simple ClinVar	K403fs	1	Deletion	403

Appendix 12 –Minimum Distance, RMSD and Radius Gyration for WT T-Protein

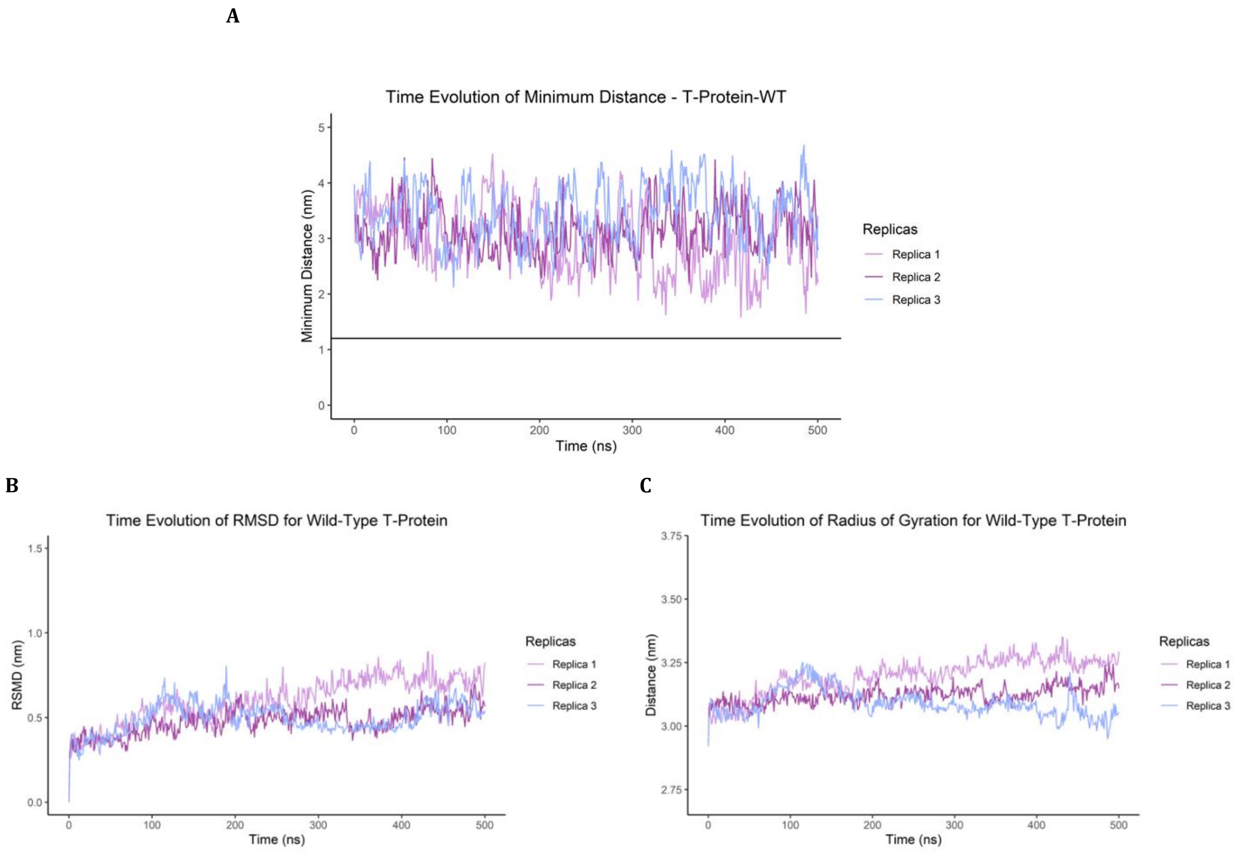


Figure 44 - Minimum Distance(A), RMSD (Root Mean Square Deviation) (B) and Radius of Gyration(C) of the replicas from the WT T-protein.

Appendix 13 – ΔG (Complex - Receptor - wild-type T-protein Chain A- ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	0.00	3.57	0.00	0.13	0.00
ΔANGLE	-0.00	4.97	0.00	0.18	0.00
ΔDIHED	0.00	2.71	0.00	0.10	0.00
ΔUB	0.00	0.89	0.00	0.03	0.00
ΔIMP	0.00	0.86	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-45.29	1.19	4.13	0.04	0.15
ΔEEL	-46.36	4.49	9.80	0.16	0.35
Δ1-4 VDW	0.00	2.26	0.00	0.08	0.00
Δ1-4 EEL	0.00	7.40	0.00	0.26	0.00
ΔEGB	74.95	1.79	8.76	0.06	0.31
ΔESURF	-7.08	0.04	0.42	0.00	0.02
ΔGGAS	-91.65	4.81	10.60	0.17	0.37
ΔGSOLV	67.87	1.79	8.61	0.06	0.30
ΔTOTAL	-23.78	5.13	6.09	0.18	0.22

Appendix 14 – ΔG (Complex - Receptor - wild-type T-protein Chain B - ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	-0.00	3.44	0.00	0.12	0.00
ΔANGLE	0.00	4.31	0.00	0.15	0.00
ΔDIHED	-0.00	2.65	0.00	0.09	0.00
ΔUB	-0.00	0.91	0.00	0.03	0.00
ΔIMP	0.00	0.84	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-41.58	0.79	4.01	0.03	0.14
ΔEEL	-31.77	3.61	10.57	0.13	0.37
Δ1-4 VDW	0.00	2.18	0.00	0.08	0.00
Δ1-4 EEL	-0.00	6.69	0.00	0.24	0.00
ΔEGB	55.06	0.08	10.05	0.00	0.36
ΔESURF	-6.18	0.02	0.55	0.00	0.02
ΔGGAS	-73.34	3.90	11.90	0.14	0.42
ΔGSOLV	48.88	0.08	9.75	0.00	0.34
ΔTOTAL	-24.46	3.90	4.45	0.14	0.16

Appendix 15 – ΔG (Complex - Receptor – H83R Chain A - ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	-0.00	3.62	0.00	0.12	0.00
ΔANGLE	-0.00	4.61	0.00	0.16	0.00
ΔDIHED	-0.00	2.42	0.00	0.08	0.00
ΔUB	0.00	0.87	0.00	0.03	0.00
ΔIMP	-0.00	0.81	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-43.29	1.24	3.59	0.04	0.12
ΔEEL	-38.21	0.28	11.58	0.01	0.40
Δ1-4 VDW	-0.00	2.22	0.00	0.08	0.00
Δ1-4 EEL	0.00	6.94	0.00	0.24	0.00
ΔEGB	66.10	2.06	10.61	0.07	0.36
ΔESURF	-6.41	0.07	0.47	0.00	0.02
ΔGGAS	-81.50	1.74	11.46	0.06	0.39
ΔGSOLV	59.69	2.06	10.54	0.07	0.36
ΔTOTAL	-21.81	2.70	5.08	0.09	0.17

Appendix 16 – ΔG (Complex - Receptor – H83R Chain B - ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	-0.00	3.55	0.00	0.13	0.00
ΔANGLE	0.00	4.43	0.00	0.15	0.00
ΔDIHED	-0.00	2.26	0.00	0.08	0.00
ΔUB	0.00	0.99	0.00	0.03	0.00

ΔIMP	0.00	0.84	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-46.71	1.48	3.84	0.03	0.13
ΔEEL	-39.20	2.68	13.55	0.05	0.46
Δ1-4 VDW	-0.00	1.89	0.00	0.07	0.00
Δ1-4 EEL	-0.00	7.24	0.00	0.27	0.00
ΔEGB	68.18	0.5	11.43	0.05	0.39
ΔESURF	-6.99	0.01	0.50	0.00	0.02
ΔGGAS	-85.91	3.33	14.45	0.08	0.50
ΔGSOLV	61.19	0.50	11.14	0.05	0.38
ΔTOTAL	-24.72	3.36	5.18	0.09	0.18

Appendix 17 – ΔG (Complex - Receptor – R73C Chain A - ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	-0.00	3.52	0.00	0.13	0.00
ΔANGLE	-0.00	4.46	0.00	0.17	0.00
ΔDIHED	-0.00	2.66	0.00	0.10	0.00
ΔUB	-0.00	0.92	0.00	0.03	0.00
ΔIMP	0.00	0.94	0.00	0.04	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-41.99	0.82	3.74	0.03	0.14
ΔEEL	-32.71	0.13	13.28	0.00	0.50
Δ1-4 VDW	-0.00	2.32	0.00	0.09	0.00
Δ1-4 EEL	-0.00	7.75	0.00	0.29	0.00
ΔEGB	58.03	4.76	15.23	0.18	0.58
ΔESURF	-6.46	0.03	0.53	0.00	0.02
ΔGGAS	-74.70	1.56	14.91	0.06	0.56
ΔGSOLV	51.58	4.76	14.89	0.18	0.56
ΔTOTAL	-23.12	5.01	4.78	0.19	0.18

Appendix 18 – ΔG (Complex - Receptor – R320H Chain B - ligand THH)

Energy Component	Average	SD(Prop.)	SD	SEM(Prop.)	SEM
ΔBOND	0.00	3.82	0.00	0.14	0.00
ΔANGLE	-0.00	4.95	0.00	0.19	0.00
ΔDIHED	-0.00	2.65	0.00	0.10	0.00
ΔUB	-0.00	0.94	0.00	0.04	0.00
ΔIMP	0.00	0.83	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-40.91	0.70	6.97	0.03	0.26
ΔEEL	-36.14	1.32	18.25	0.05	0.69
Δ1-4 VDW	-0.00	2.28	0.00	0.09	0.00
Δ1-4 EEL	0.00	7.22	0.00	0.27	0.00
ΔEGB	63.52	2.41	19.17	0.09	0.72
ΔESURF	-6.26	0.02	1.11	0.00	0.04

ΔGGAS	-77.05	1.94	23.18	0.07	0.88
ΔGSOLV	57.26	2.41	18.27	0.09	0.69
ΔTOTAL	-19.79	3.10	7.23	0.12	0.69

Appendix 19 – ΔG (Complex - Receptor – R222C Chain A - ligand THH)

Energy Component	**Average**	**SD(Prop.)**	**SD**	**SEM(Prop.)**	**SEM**
ΔBOND	-0.00	3.49	0.00	0.13	0.00
ΔANGLE	0.00	4.93	0.00	0.18	0.00
ΔDIHED	0.00	2.49	0.00	0.09	0.00
ΔUB	-0.00	0.94	0.00	0.03	0.00
ΔIMP	-0.00	0.80	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-44.17	1.23	3.39	0.04	0.12
ΔEEL	-37.80	1.17	15.50	0.04	0.57
Δ1-4 VDW	0.00	2.29	0.00	0.08	0.00
Δ1-4 EEL	-0.00	6.24	0.00	0.23	0.00
ΔEGB	66.27	0.35	16.98	0.01	0.62
ΔESURF	-6.68	0.05	0.50	0.00	0.02
ΔGGAS	-81.97	2.10	16.63	0.08	0.61
ΔGSOLV	59.59	0.35	16.62	0.01	0.61
ΔTOTAL	-22.38	2.12	4.24	0.08	0.15

Appendix 20 – ΔG (Complex - Receptor – E211K Chain B - ligand THH)

Energy Component	**Average**	**SD(Prop.)**	**SD**	**SEM(Prop.)**	**SEM**
ΔBOND	0.00	3.69	0.00	0.13	0.00
ΔANGLE	-0.00	4.84	0.00	0.18	0.00
ΔDIHED	-0.00	2.60	0.00	0.09	0.00
ΔUB	-0.00	0.95	0.00	0.03	0.00
ΔIMP	-0.00	0.86	0.00	0.03	0.00
ΔCMAP	0.00	0.00	0.00	0.00	0.00
ΔVDWAALS	-42.13	0.22	4.36	0.01	0.16
ΔEEL	-19.29	0.25	13.58	0.01	0.50
Δ1-4 VDW	-0.00	2.18	0.00	0.08	0.00
Δ1-4 EEL	-0.00	7.03	0.00	0.26	0.00
ΔEGB	47.34	1.21	11.60	0.04	0.42
ΔESURF	-6.16	0.21	0.60	0.01	0.02
ΔGGAS	-61.41	1.32	14.99	0.05	0.55
ΔGSOLV	41.18	1.23	11.30	0.04	0.41
ΔTOTAL	-20.23	1.80	7.10	0.07	0.26

Appendix 21 – Machine Learning– First Dataset

Source	Protein Change	Pathogenicity Effect	Pathogenicity Description
Genet	M1T	1	damaging
Genet	G47R	1	damaging
Genet	H71P	1	damaging
Genet	R73C	1	damaging
Genet	S77L	1	damaging
Genet	R94W	1	damaging
Genet	M98R	1	damaging
Genet	I106T	1	damaging
Genet	S117L	1	damaging
Genet	N145I	1	damaging
Genet	L172P	1	damaging
Genet	L182P	1	damaging
Genet	Q183R	1	damaging
Genet	V212A	1	damaging
Genet	R222H	1	damaging
Genet	R222C	1	damaging
Genet	Y225C	1	damaging
Genet	D229H	1	damaging
Genet	R265H	1	damaging
Genet	R265C	1	damaging
Genet	G269D	1	damaging
Genet	D276H	1	damaging
Genet	R296H	1	damaging
Genet	R296C	1	damaging
Genet	R320G	1	damaging
Simple Clinvar	H42R	1	damaging
Simple Clinvar	G47W	1	damaging
Simple Clinvar	G104E	1	damaging
Simple ClinVar/ClinVar	S117L	1	damaging
Clinvar	L155F	0	neutral
Simple ClinVar/GenMed	L172P	1	damaging
ClinVar	R222S	1	damaging
Simple ClinVar	R222L	1	damaging
Simple ClinVar	L266P	1	damaging
ClinVar	P283S	1	damaging
ClinVar	R296L	1	damaging
ClinVar	V195M	0	neutral
ClinVar	M300V	0	neutral
Simple ClinVar/Clinvar	A238V	0	neutral

Appendix 22 – Machine Learning– Second Dataset

Source	Protein Change	Pathogenicity Effect	Pathogenicity Description
Simple ClinVar/ClinVar	E211K	0	neutral
Genet	H83R	1	damaging
Genet	R265S	1	damaging
Genet	T282I	1	damaging
Genet	R296P	1	damaging

Genet	R320H	1	damaging
Genet	R320C	1	damaging
Genet	R331Q	1	damaging
Genet	R331P	1	damaging
Genet	A361V	1	damaging
Genet	G363R	1	damaging
Simple Clinvar	M1K	1	damaging

Appendix 23 – Machine Learning– Test Dataset Results

Source	Protein Change	Pathogenicity Effect	Pathogenicity Description	Machine Learning Prediction
Simple ClinVar/ClinVar	E211K	0	neutral	0
Genet	H83R	1	damaging	1
Genet	R265S	1	damaging	1
Genet	T282I	1	damaging	1
Genet	R296P	1	damaging	1
Genet	R320H	1	damaging	1
Genet	R320C	1	damaging	1
Genet	R331Q	1	damaging	1
Genet	R331P	1	damaging	1
Genet	A361V	1	damaging	1
Genet	G363R	1	damaging	1
Simple Clinvar	M1K	1	damaging	1

Appendix 24 – Machine Learning– Consensus Dataset Results

Source	ClinicalSignificance	SAV	LYRUS	Provean	SNAP2	Mutaframe	SIFT	Consensus Majority Vote	XGBClassifier Prediction
Simple ClinVar	Pathogenic	M1K	neutral	neutral	neutral	NA	effect	Likely Benign	1
PubTator	Likely pathogenic	V7L	neutral	neutral	neutral	NA	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	P20L	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	R24C	neutral	neutral	neutral	effect	effect	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	A29V	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	R34H	neutral	neutral	effect	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	R34C	neutral	effect	effect	effect	effect	Likely Pathogenic	0
ClinVar	Uncertain significance	P37L	neutral	effect	neutral	effect	neutral	Likely Benign	1
Simple ClinVar	Pathogenic	H42R	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	A44T	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Likely pathogenic	G47W	effect	neutral	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	G47R	effect	neutral	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V50L	neutral	neutral	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V50L	neutral	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	A51V	neutral	neutral	neutral	effect	effect	Likely Benign	0
ClinVar	Uncertain significance	F52S	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	F52C	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	A53V	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	L57P	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	P58L	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	D63G	neutral	effect	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	H65R	neutral	neutral	effect	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	T66A	neutral	neutral	neutral	neutral	neutral	Likely Benign	0

Simple ClinVar	Uncertain/conflicting	S68L	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	H71P	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	R73C	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	R73H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	S77L	effect	effect	effect	effect	neutral	Likely Pathogenic	1
ClinVar	Uncertain significance	I89R	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	L90P	neutral	neutral	neutral	effect	neutral	Likely Benign	1
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	R94W	effect	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V95M	neutral	neutral	effect	effect	effect	Likely Pathogenic	1
ClinVar/GenMed	Uncertain significance/Pathogenic	M98R	effect	effect	effect	NA	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	S100G	neutral	effect	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Likely pathogenic	G104E	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	I106T	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	N112S	neutral	effect	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	Q113R	neutral	neutral	neutral	effect	neutral	Likely Benign	1
Simple ClinVar	Uncertain/conflicting	G114E	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar/ClinVar/GenMed	Pathogenic	S117L	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	T120I	neutral	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	N121S	neutral	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	G124E	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	N134S	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	T135A	neutral	neutral	neutral	effect	neutral	Likely Benign	1
Simple ClinVar	Uncertain/conflicting	V142G	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	S144F	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	N145S	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Likely pathogenic	N145I	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	A146T	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	C148F	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Likely benign	L155F	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
ClinVar	Uncertain significance	M156T	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	K159R	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	R168T	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	L172P	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	N177D	neutral	neutral	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	Q189P	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	V195A	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar/Clinvar	Uncertain/conflicting/Likely benign	V195M	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	D197H	neutral	neutral	neutral	effect	effect	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	D197N	neutral	neutral	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	D198V	effect	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	M205V	effect	effect	neutral	effect	neutral	Likely Pathogenic	1
ClinVar	Uncertain significance	M205R	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V209M	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	V212A	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	F213C	neutral	effect	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	V215M	neutral	neutral	neutral	effect	neutral	Likely Benign	1
Simple ClinVar	Uncertain/conflicting	V215L	neutral	neutral	neutral	effect	neutral	Likely Benign	1
ClinVar	Uncertain significance	R219H	neutral	effect	neutral	effect	neutral	Likely Benign	1
ClinVar	Uncertain significance	V220M	neutral	neutral	neutral	effect	effect	Likely Benign	0

ClinVar	Likely pathogenic	R222S	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	R222H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Likely pathogenic	R222L	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	R222C	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Likely pathogenic	Y225C	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	T226I	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	D229H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	S234L	effect	effect	neutral	effect	neutral	Likely Pathogenic	1
ClinVar	Uncertain significance	A240V	neutral	neutral	neutral	effect	effect	Likely Benign	1
ClinVar	Uncertain significance	V241I	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	L243M	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	P251R	neutral	effect	neutral	effect	effect	Likely Pathogenic	0
Simple ClinVar/GenMed/PubTator	Pathogenic	R265H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	R265C	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Likely pathogenic	L266P	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	G269S	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	G269D	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	N275K	neutral	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic/Likely pathogenic	D276H	effect	effect	effect	effect	effect	Likely Pathogenic	1
PubTator	Pathogenic	D276H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	I277T	neutral	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	I277N	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Pathogenic	P283S	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	V284M	effect	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	G286S	neutral	neutral	effect	neutral	effect	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	G286D	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	R295C	effect	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	R296H	neutral	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Likely pathogenic	R296L	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	R296C	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	R297Q	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	A298T	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Benign/Likely benign	M300V	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
ClinVar	Uncertain significance	G304E	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	R315K	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
ClinVar	Uncertain significance	Q317R	neutral	neutral	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	R318Q	neutral	neutral	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	R318W	neutral	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Pathogenic	R320H	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	R320L	effect	effect	effect	effect	neutral	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	R320C	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V321A	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	V321M	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	M324I	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	A328T	neutral	neutral	neutral	effect	neutral	Likely Benign	1
Simple ClinVar/GenMed	Pathgenic/Likely pathogenic	R331Q	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	S334I	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	G341S	neutral	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	I344F	neutral	effect	effect	effect	effect	Likely Pathogenic	1

Simple ClinVar	Uncertain/conflicting	S353Y	effect	effect	effect	effect	effect	Likely Pathogenic	1
ClinVar	Uncertain significance	N359T	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar/GenMed	Uncertain/conflicting/Pathogenic	A361V	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	G363R	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	P366L	neutral	effect	neutral	effect	neutral	Likely Benign	0
ClinVar	Uncertain significance	R371C	neutral	effect	neutral	effect	effect	Likely Pathogenic	0
Simple ClinVar	Uncertain/conflicting	R371H	neutral	effect	neutral	effect	effect	Likely Pathogenic	0
Simple ClinVar	Uncertain/conflicting	E379A	neutral	effect	neutral	effect	neutral	Likely Benign	0
Simple ClinVar/PubTator	Uncertain/conflicting/Likely pathogenic	V380M	effect	neutral	effect	NA	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	R381W	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	R382Q	neutral	neutral	neutral	effect	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	M386T	neutral	neutral	neutral	neutral	neutral	Likely Benign	0
Simple ClinVar	Uncertain/conflicting	T397K	effect	effect	neutral	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	Y400C	effect	effect	effect	effect	effect	Likely Pathogenic	1
Simple ClinVar	Uncertain/conflicting	T401N	neutral	neutral	neutral	neutral	neutral	Likely Benign	0

Appendix 25 - Quantitative severity scale for major NKH disease domains

Domain and Associated Scoring Scale	
Cognitive Disorders	0–No disease
	1–Behavior al issues, learning disabilities, speech delay
	2–Mental disability, some words, global delay in developmental markers
	3–Severe mental disability, no cognitiveabilities
Seizures	0–No disease
	1–Hiccups (infants), Abnormal EEG, Seizures controlled by medication
	3–Intract able seizures (> 2AEDs)
Muscle/Movement Control	0–No disease
	1–Assisted locomotion
	2–Hypotonia, able to roll over or lift head, low muscle tone
Brain malformation	0–No disease
	3–Present
EEG =electroenc ephalogram	
AEDs =Anti-epileptic drugs	

Appendix 26 - COS for patients with NKH due T-Protein mutations

PatientID	Gender	Allele1	Allele2	Protein1	Protein2	Reference	Onset	Died	Ratio CSF: plasma Gly	Seizures	Cognition	Brain Malformations	Muscle/ Movement Control	COS	Severity
1	F	c.344+2+2insT	c.230C>T	IVS8+2insT	p.S77L	[3]	1 we	-	0.23	-	-	3(ACC)	-	INS	
2	M	c.16delA	c.16delA	p.S6Vfs*90	p.S6Vfs*90	[3]	1 we	-	0.31	-	-	-	-	INS	
3	F	c.731dupC	c.959G>A	A244Afs*32	p.R320H	[3]	1 we	-	0.11	3	3(≤ 6 we)	3(HCC)	-	9	**Severe**
4	M	c.959G>A	c.665G>A	p.R320H	p.R222H	[3,214,215]	1 we	-	0.09	3	3 (≤ 6 we)	0	-	6	**Severe**
5	M	c.887G>A	c.845C>T	p.R296H	p.T282I	[3]	1 we	-	0.19	1	3 (≤ 6 we)	0	-	4	**Att. Intermediate**
6	M	c.496C>T	c.696+5G>C	p.Q166X	IVS6+5G>C	[3]	6 we	-	0.14	3	3 (≤ 6 we)	0	-	6	**Severe**
7	M	c.826G>C	c.826G>C	p.D276H	p.D276H	[3]	1 we	-	0.21	3	3 (≤ 6 we)	3(HCC)	-	9	**Severe**
8	M	c.280C>T	c.471+2T>C	p.R94W	IVS4+2T>C	[3]	2 mo	-	0.3	1	3 (≤ 6 we)	3(ACC)	-	7	**Severe**
9	M	c.2T>C	c.2T>C	p.M1T	p.M1T	[3]	1 mo	-	0.14	-	-	0	-	INS	
10	M	c13_16delGTAA	c.350C>T	p.S5Vfs*27	p.S117L	[3, 216]	1 we	-	0.14	-	-	3(HCC)	-	INS	
11	F	c.2T>C	c.2T>C	p.M1T	p.M1T	[3]	1 we	-	0.22	-	-	-	-	INS	
12	F	c.674G>C	c.635T>C	p.Y225C	p.V212A	[3,217]	6 we	-	0.14	-	-	3(HCC, DW)	-	INS	
13	F	c.959G>A	c.959G>A	p.R320H	p.R320H	[3]	1 we	-	-	-	-	-	-	INS	
14	M	c.1040_1041delTG	c.230C>T	p.V347Dfs*2	p.S77L	[3]	1 we	-	-	-	-	3(BA)	-	INS	
15	M	c.959G>A	c.317T>C	p.R320H	p.I106T	[3]	1 we	-	0.13	1	2(21 DQ)	0	-	3	**Att. Intermediate**
16	F	c.317T>C	c.515T>C	p.I106T	p.L172P	[3]	1 we	-	0.22	0	2(23 DQ)	0	-	2	**Att. Intermediate**
17	F	c.317T>C	c.665G>C	p.I106T	p.R222H	[3]	2 mo	-	0.1	0	2(30 DQ)	0	-	2	**Att. Intermediate**
18	M	c.317T>C	c.959G>A	p.I106T	p.R320H	[3]	5 mo	-	0.11	0	1(50 DQ)	0	-	1	**Att. Good**
19	M	c.959G>A	c.664C>T	p.R320H	p.R222C	[3,209, 210]	3 yr	-	-	0	1(60 DQ)	0	-	1	**Att. Good**
20	-	c.230C>T	c.230C>T	p.S77L	p.S77L	[218, 210]	1 day	-	0.24	-	-	-	-	INS	
21	-	c.136G	c.230C>T	p.G47R	p.S77L	[218]	6 day	-	0.18	-	-	-	-	INS	
22	-	c.125A>G	c.125A>G	p.H42R	p.H42R	[218]	1 day	-	0.27	-	-	-	-	INS	
23	-	c.471+2T>C	c.887G>A	IVS4+2T>C	p.R296H	[218]	1 day	-	0.26	-	-	-	-	INS	
24	-	c.54delC	c.826G>C		p.D276H	[218]	1 day	-	0.34	-	-	-	-	INS	

25	-	c.147delG	c.970_972delATG			[218]	1 day	-	0.11	-	-	-	-	INS	
26	-	c.60delG	c.471+2T>C		IVS4+2T>C	[218]	2 day	-	0.16	-	-	-	-	INS	
27	-	c.982_972GC>T	c.452_466del15		p.K151_L155del	[218]	2 day	-	0.17	-	-	-	-	INS	
28	-	c.212A>C	c.217C>T	p.H71P	p.R73C	[218]	2 day	-	0.16	-	-	-	-	INS	
29	-	c.61delC	c.535delC			[218]	2 day	-	0.08	-	-	-	-	INS	
30	-	c.139G>T	-	p.G47W		[218]	6 day	-	0.08	-	-	-	-	INS	
31	-	c.136G>A	c.230C>T	p.G47R	p.S77L	[218], [214]	6 day	-	0.18	-	-	-	-	INS	
32	F	c.826G>C	c.982delG	p.A276H	p.A328fs	[209]	1 day	-	0.38	1	3	3(HCC)	-	7	**Severe**
33	F	c.664C>T	c.688G>C	p.R222C	p.G230R	[209]	2 day	-	0.11	1	1	3(HCC)	-	5	**Att. Intermediate**
34	M	c.664C>T	c.688G>C	p.R222C	p.G230R	[209]	4 day	-	0.12	3	2	-	-	INS	
35	M	c.794G>A	c.794G>A	p.R265H	p.R265H	[209]	3 day	yes	0.24	-	-	-	-	INS	
36	M	c.794G>A	c.794G>A	p.R265H	p.R265H	[209]	1 day	-	0.17	3	3	-	-	6	**Severe**
37	M	c.794G>A	c.794G>A	p.R265H	p.R265H	[209]	5 day	yes	0.1	-	-	-	-	INS	
38	F	c.248A>G	c.248A>G	p.H83R	p.H83R	[211]	2.5 mo	-	0.07	1	3	-	-	INS	
39	F	c.248A>G	c.248A>G	p.H83R	p.H83R	[211]	6 day	-	0.15	1	3	-	-	4	**Att. Intermediate**
40	F	c.339G>A	c.339G>A	p.Q113Q	p.Q113Q	[219]	1 day	-	0.29	3	-	3	-	6	**Severe**
41	M	c.565C>T	c.565C>T	p.Q189*	p.Q189*	[220]	1 day	-	0.17	3	3	3	3	12	**Severe**
42	-	c.434A>T		p.N145I		[221]	1 day	yes		-	-	-	-	INS	
43	-	c.574C>T		p.Q192X		[217]	1 day	yes		-	-	-	-	INS	
44	F	c.635T>C	c.674A>G	p.V212A	p.Y225C	[217]		-	0.17	-	-	-	-	INS	
45	M	c.793C>T	c.793C>T	p.R265C	p.R265C	[222]	1 day	-		1	3	3(HCC)	2	9	**Severe**
46	F	c.878-1G>A	c.959G>A	IVS7-1G>A	p.R320H	[221], [215]	1 day	-	>0.12	-	-	-	-	INS	
47	-	c.631G>A		p.E211K	p.S77L	[208]		-	0.23	-	-	-	-	INS	