COMPUTATIONAL APPROACHES TO STUDY THE IMPACT OF MUTATIONS ON DISEASE AND DRUG RESISTANCE

EDITED BY: Nir Ben-Tal, Daisuke Kihara and Arun Prasad Pandurangan

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-212-7
DOI 10.3389/978-2-88974-212-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL APPROACHES TO STUDY THE IMPACT OF MUTATIONS ON DISEASE AND DRUG RESISTANCE

Topic Editors:
Nir Ben-Tal, Tel Aviv University, Israel
Daisuke Kihara, Purdue University, United States
Arun Prasad Pandurangan, University of Cambridge, United Kingdom

Citation: Ben-Tal, N., Kihara, D., Pandurangan, A. P., eds. (2022). Computational Approaches to Study the Impact of Mutations on Disease and Drug Resistance. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-212-7

Table of Contents

frontiers in Molecular Biosciences

EDITORIAL
published: 09 December 2021
doi: 10.3389/fmolb.2021.813552

Editorial: Computational Approaches to Study the Impact of Mutations on Disease and Drug Resistance

Nir Ben-Tal[1], *Daisuke Kihara*[2] *and Arun Prasad Pandurangan*[3*†]

[1]Tel Aviv University, Tel Aviv, Israel, [2]Purdue University, West Lafayette, IN, United States, [3]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

Keywords: mutation, genetic disorder, antimicrobial drug resistance (AMR), protein structure, computational tools and databases

Editorial on the Research Topic

Computational Approaches to Study the Impact of Mutations on Disease and Drug Resistance

Advances in next generation sequencing technologies provide wealth of data on genome variations. Understanding missense mutations is crucial to tackling global health problems related to inherited diseases and the emergence of drug resistance in cancers and infectious diseases. Advancement in research at systems and molecular level, is required to study the impact of mutations that affect both the regulation of gene expression and protein function through changes in protein stability and affinity towards other proteins, nucleic acids, biomolecules and small molecule ligands. High-quality experimental data on protein structure, mutant stability, functional annotations and phenotype-genotype associations in combination with the state of art techniques in artificial intelligence and machine will revolutionise development of highly accurate predictive computational models to study the impact of genetic mutations on human health and disease.

Predictive computational models offer an effective alternative to expensive experimental studies of genetic variations. These models can identify potential mutations linked to disease conditions and the emergence of antimicrobial drug resistance. At the molecular level proteins, via their interactions with other proteins and biomolecules, play an important role in many biological processes. The growing data on protein three-dimensional structure, along with variations observed in sequence data, will enable the development of new computational methods and tools to predict the impact of mutations on protein function, stability and interaction thereby aiding in the understanding of the basic mechanisms that govern disease conditions.

This research topic highlights the recent developments in computational approaches to analysis and predict the impact of mutation on protein stability, function and interaction. Development of accurate protein mutant stability requires the availability of properly curated high quality experimental thermodynamic dataset. To facilitate this, Turina et al. developed a semi-automatic text-mining tool to extract protein mutant thermostability data from the scientific literature. Feng et al. studied the role of phosphatase and tensin (PTEN) homolog gene mutation in low grade gliomas progression and prognosis. Using patient's RNA sequencing data, differential gene expression and gene ontology analysis they showed that PTEN mutation promote tumorigenesis and immune cell infiltration. Tan et al. trained a predictor using saturation mutagenesis data to access the impact of point mutations on protein stability and function. Mutants are scored using a statistical potential energy function derived from protein structural data in combination with evolutionary sequence conservation and substitution scores. Using the physicochemical properties of amino acids Savojardo et al. grouped variants linked to human genetic diseases into four types and established mapping between mutations, diseases, and phenotypes through the protein family

OPEN ACCESS

Edited and reviewed by:
Francesco Luigi Gervasio,
University College London,
United Kingdom

****Correspondence:***
Arun Prasad Pandurangan
app41@cam.ac.uk

†Present address:
Department of Biochemistry,
University of Cambridge, Cambridge,
United Kingdom

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *11 November 2021*
Accepted: *24 November 2021*
Published: *09 December 2021*

Citation:
Ben-Tal N, Kihara D and
Pandurangan AP (2021) Editorial:
Computational Approaches to Study
the Impact of Mutations on Disease
and Drug Resistance.
Front. Mol. Biosci. 8:813552.
doi: 10.3389/fmolb.2021.813552

domains. Tunstall et al. designed an *in silico* framework to understand pyrazinamide resistance mutation in the clinical isolates of four main *M. tuberculosis* lineages. Using a combination of genomic features and computational mutant stability and drug affinity predictors to explain differences in modern and ancient lineages within the context of drug resistance. Using molecular dynamic simulations, Nangraj et al. investigated the mutation in *pnc*A gene of *Mycobacterium tuberculosis* that confer resistance to the first line drug pyrazinamide and explained resistance in term of the changes in protein stability and drug binding site. Prabantu et al. modelled protein structures as network where nodes and the edges correspond to residues and interaction between residues respectively. They showed that the differences between wildtype and disease mutants can be explained by their respective changes in the network both locally at the site of mutation and globally that relate to protein allosteric effects. Birolo et al. analysed both pathogenic and benign variants in haploinsufficient genes and reported that variants significantly perturbing stability (both the stabilising and destabilising) correlate with pathogenicity. Mahlich et al. performed mutational analysis using variant effect predictor on human proteins and its orthologous from 20 species. They analysed the impact of common and rare variants in terms of conservation and also suggested that cross-species variants (CSVs) might be more often neutral than non-CSVs. Bhasin and Varadarajan used large scale mutational scanning dataset to study the mutational sensitivity and substitution preferences at buried and exposed positions. They used mutational sensitivity data and predicted sequence-based accessibility values to identify buried, active-site and exposed non active-site residues. Soto-Ospina et al. aimed to understand the impact of pathogenic mutations in amyloid precursor protein Presenilin 1 that are known to cause Alzheimer's disease. They used molecular modelling and dynamic simulations to explain the impact of mutations in terms of structural modifications of active site mutant residues found at the catalytic pore. In a focused review, Grace et al. explored the use of molecular docking and dynamics to study resistance mutation in *Mycobacterium tuberculosis* within the context of anti-tuberculosis drugs.

Understanding the impact of genetic mutation is critical to tackle disease and drug resistance. Experimental structures of biomolecules are becoming available at a rapid pace due to the recent developments in the field of cryo electron microscopy. In parallel, the technology development in computing hardware and software has enabled development of robust machine learning models to predict the structure of proteins and its interactions. Both these recent developments complement each other to provide high quality structural data of biological macromolecules and small molecules including drugs. The timing of these recent developments will enable decoding the complex mutational landscape and enable our understanding of the genotype to phenotype relationship, paving way to the achievement of precision medicine.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We thank all the authors that have contributed to this Research Topic.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 10 December 2020
doi: 10.3389/fmolb.2020.562416

Phosphatase and Tensin Homolog Mutation in Immune Cell Infiltration and Clinicopathological Features of Low-Grade Gliomas

Peng Feng[1], *Zhenqing Li*[2], *Yuchen Li*[3] *and Yuelin Zhang*[1]*

[1] Xi'an Medical University, Xi'an, China, [2] Research Center of Clinical Medicine, Affiliated Hospital of Nantong University, Nantong, China, [3] Hengyang Medical College, University of South China, Hengyang, China

The mutation of phosphatase and tensin homolog (*PTEN*) genes frequently occur in low-grade gliomas (LGGs) and are deeply associated with a poor prognosis and survival rate. In order to identify the crucial signaling pathways and genes associated with the *PTEN* mutation, we performed bioinformatics analysis on the RNA sequencing results, which were obtained from The Cancer Genome Atlas database. A total of 352 genes were identified as differentially expressed genes (DEGs). The gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis suggested that the DEGs were significantly enriched in categories associated with cell division and multiple metabolic progressions. The histological stage was significantly associated with *PTEN* expression levels. In addition, the *PTEN* mutation was associated with an abundance of *B* cells, neutrophils, macrophages, dendritic cells, and *CD8+ T* cells during tumor infiltration. The results showed that patients with LGGs harboring the *PTEN* mutation had a poor prognosis and more serious immune cell infiltration occurred depending on the mRNA expression level. These results demonstrated that multiple genes and signaling pathways play a key role in LGG from low grade to high grade, and are associated with *PTEN* mutations. In this study, we outlined an approach to assess the influence of *PTEN* mutations on prognosis, overall survival, and messenger RNA (mRNA) expression. Our results provided alternative strategies for the personalized treatment of patients with LGGs harboring the *PTEN* mutation.

Keywords: phosphatase and tensin homolog, prognosis, mutation, low-grade gliomas, gene

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
University of Cambridge,
United Kingdom

Reviewed by:
Antonella Di Pizio,
Technical University of Munich,
Germany
Yuan Pan,
Washington University in St. Louis,
United States

****Correspondence:***
Yuelin Zhang
xyyxyfp@163.com

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *21 May 2020*
Accepted: *20 November 2020*
Published: *10 December 2020*

Citation:
Feng P, Li Z, Li Y and Zhang Y
(2020) Phosphatase and Tensin
Homolog Mutation in Immune Cell
Infiltration and Clinicopathological
Features of Low-Grade Gliomas.
Front. Mol. Biosci. 7:562416.
doi: 10.3389/fmolb.2020.562416

INTRODUCTION

Gliomas, as a kind of common craniocerebral tumor, can be divided into four grades based on the 2007 World Health Organization classification of tumors. Grade I and II are low-grade gliomas (LGGs), while Grade III and IV are high-grade gliomas (Louis et al., 2007). A clinical investigation into malignant LGGs found that the overall survival of LGGs is significantly higher

Abbreviations: DEG, Differentially expressed genes; FDR, False discovery rate; GABA, Gamma-aminobutyric acid; GSEA, Gene set enrichment analysis; LGG, Low-grade gliomas; mRNA, messenger ribonucleic acid; PPI, Protein-protein interaction; STRING, Search Tool for the Retrieval of Interacting Genes; TCGA, The Cancer Genome Atlas.

(2–4 years) than high-grade gliomas (∼15 months), which are highly aggressive tumors and exhibit significant aggression (Buckner et al., 2016). Reported by a variety of literature, too many genes anticipate the signal pathway of a drug resistance response (Andersson et al., 2004; Calatozzolo et al., 2012). To precisely treat LGGs patients, it is of importance to provide personalized genetic information and expression correlations, because personnel treatments can provide precise therapeutic strategies based on specific genetic conditions. Consequently, the identification of the underlying biomarkers of disease progression and the underlying target gene are the prerequisite for personalized treatments.

The phosphatase and tensin homolog (*PTEN*) gene is a multifunctional tumor suppressor, which contains a catalytic domain and a tensin-like domain (Li et al., 1997; Helseth et al., 2010). Owing to the tumor suppressing functions of the *PTEN* gene, it has been found to mutate with high frequency in several types of carcinomas, including LGGs (Helseth et al., 2010; Johnson and O'Neill, 2012). To promote the cell proliferation of cancer, the key node target is Akt, in which the PTEN protein inhibits PI3K/Akt signaling and then activates the P21 protein (Mu et al., 2020). Although the variation in *PTEN* expression levels may correlate with the LGG tumorigenesis (Wiencke et al., 2007), these expression levels have clinical significance and can be used as prognostic biomarkers. Some studies have reported that *PTEN* is activated through AKT-independent (protein kinase B) mechanisms (McGirt et al., 2005). Patients harboring the *PTEN* mutation exhibit increasing alterations of multiple signaling pathways and cellular metabolism compared with those harboring the wild-type *PTEN* gene (Steck et al., 1997). Thus, a variation in the *PTEN* status may affect the tumor progression by regulating the immune microenvironment (Best et al., 2018; Wu et al., 2018). And, the disease prognosis and immune cell infiltration are highly associated with the immune microenvironment (Ino et al., 2013), as well as resistance or sensitivity to treatment measures (Norton et al., 2019). However, the importance of *PTEN* status in LGG progression and the molecular mechanism is still unclear.

In this study, we analyzed the RNA sequencing data of LGG patients, obtained from The Cancer Genome Atlas (TCGA) database. By performing the identification of differentially expressed genes, the molecular functions and correlation with LGG progression were analyzed using GSEA analysis. After the enrichment of differentially expressed genes, the association between differentially expressed genes and immune cell infiltration was further analyzed using the TIMER database, which elucidated the effect of the *PTEN* mutation on the tumor-related genes and signaling pathways.

MATERIALS AND METHODS

Gene Set Enrichment Analysis

The RNA-seq database of LGG patients was obtained from The Cancer Genome Atlas (TCGA) database[1], which included 516 cases. After the classification of differentially expressed genes, gene set enrichment analysis (GSEA) was used to identify the biological functions of the differentially expressed genes (DEGs) based on their biological status. Furthermore, the enriched signal pathways of LGG patients with or without the *PTEN* mutation were obtained. Enrichment results with a cut-off value of false discovery rate (FDR) < 0.25 and a p-value < 0.05 were identified to be as significant. The hazard ratio of LGG patients, including age, gender, PTEN status, and grade, were performed using the Cox proportional-hazards model of the R software.

Identification of Differentially Expressed Genes

In this study, the R software (version 3.5.2) containing the bioconductor software package (EdgeR) was used to identify the differential gene expression in LGG patients harboring various *PTEN* mutations compared with wild-type patients (Robinson et al., 2010; McCarthy et al., 2012). The identification criteria for the DEGs were as follows: P-value and FDR < 0.05; $|\text{log2FoldChange}| \geq 1.0$.

Pathway Enrichment Analysis of Differentially Expressed Genes

GSEA analysis was performed to ascertain the effect of differentially expressed genes on signaling utilizing the Hallmark gene sets[2]. Gene oncology (GO) annotations are the collaborative effort of developing and using ontologies to support biologically meaningful annotations of genes and their products, which include the biological process (BP), cellular component (CC), and molecular function (MF). Commonly, GO can be used to describe the annotation of the enriched genes in related signaling pathways and confirm the biological characteristics at the transcriptomic level. DEGs were classified using the clusterProfiler package. GO and KEGG were enriched based on the hypergeometric distribution of the GO concepts and KEGG pathways. To avoid high FDRs in multiple tests, the q-values of FDR control were also calculated.

Protein-Protein Interaction Network and Module Analysis

The Search Tool for the Retrieval of Interacting Genes (STRING)[3] (Bader and Hogue, 2003) was used for creating the protein-protein interaction (PPI) network of the DEGs and further attribute these genes to their specific biological functions, e.g., cellular component, biological process, and molecular function annotations (Dennis et al., 2003; Szklarczyk et al., 2015). Then the Cytoscape software (v3.0)[4] was used to visualize the PPI network and identify the core DEGs in the biological regulating process. Then the KEGG pathway was analyzed for the enrichment of DEGs in the top-ranked three modules.

[1] https://portal.gdc.cancer.gov

[2] https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

[3] https://string-db.org/

[4] https://cytoscape.org/

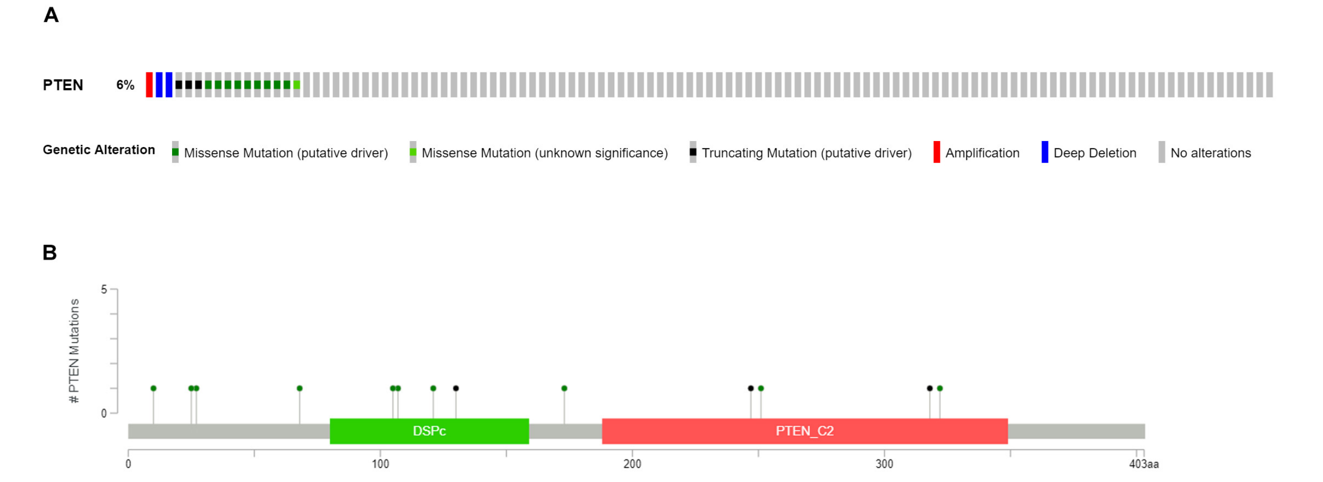


FIGURE 1 | Mutation frequency **(A)** and types **(B)** of *PTEN* mutations in patients with LGG obtained from The Cancer Genome Atlas (TCGA) database.

TIMER Database Analysis

Immune cell infiltration analysis was performed using TIMER 2.0[5] (Li et al., 2020). The association between the *PTEN* status of different cancers and the abundance of immune cell infiltrations were analyzed using the TIMER database to conclude the abundance of tumor-infiltrating immune cells, including B cells, CD8$^+$ T cells, macrophages, dendritic cells, CD4$^+$ T cells, and neutrophils.

Statistical Analysis

All statistical analyses were conducted using Graphpad and R 3.3.0. Student's *t*-test was used to analyze *PTEN* mRNA expression levels in cancer tissues with different *PTEN* statuses. The Benjamini-Hochberg procedure was used to adjust FDR in limma and GSEA (Mootha et al., 2003; Subramanian et al., 2005). A p-value < 0.05 was considered as significant. The survival curve was obtained using the cBioPortal website[6].

RESULTS

Data Information

Clinical patient information of LGGs, including the cancer tissue RNA-seq database and complete follow-up profiles, were obtained from the TCGA database. The LGG cases were divided into two groups, LGG with *PTEN* mutation (18 patients) and LGG without *PTEN* mutation (as shown in **Figure 1A**). Among these patients, 6% of LGG patients had mutated genes, which included missense mutations, nonsense mutations, amplifications, and deep deletions. For the *PTEN* mutation (**Figure 1B**), there were 13 amino acid sites of the PTEN protein that were identified as the commonly mutated sites, located at the DSPc and PTEN_C2 domains.

Clinical Impact of Low-Grade Glioma Progression and Prognosis

Clinical information for the LGG patients can provide a profile of related characteristics. Before further bioinformatics analysis, we studied the clinical information of the included patients, as shown in **Table 1**. The average age (54.44, 35–74 years old) of patients with the *PTEN* mutation was higher than the wild-type patient (42.52, 14–87 years old), indicating that age may promote the mutation of the *PTEN* gene. Moreover, the histological grade (G4:G3 = 16:2) of LGG patients harboring the *PTEN* mutation was higher than the wild-type group (G4:G3 = 247:250), indicating that LGG with a *PTEN* mutation is more serious.

Initially, the *PTEN* mRNA expression level of the wild-type *PTEN* and *PTEN* mutation groups were identified. As shown in **Figure 2A**, the *PTEN* expression level of the PTEN wild-type group was significantly higher than the *PTEN* mutated group. Meanwhile, the *PTEN* expression dependence on the PTEN status (as shown in **Figure 2B**) showed that the expression level of shallow deletion and diploid was significantly higher than the gain and deep deletion status.

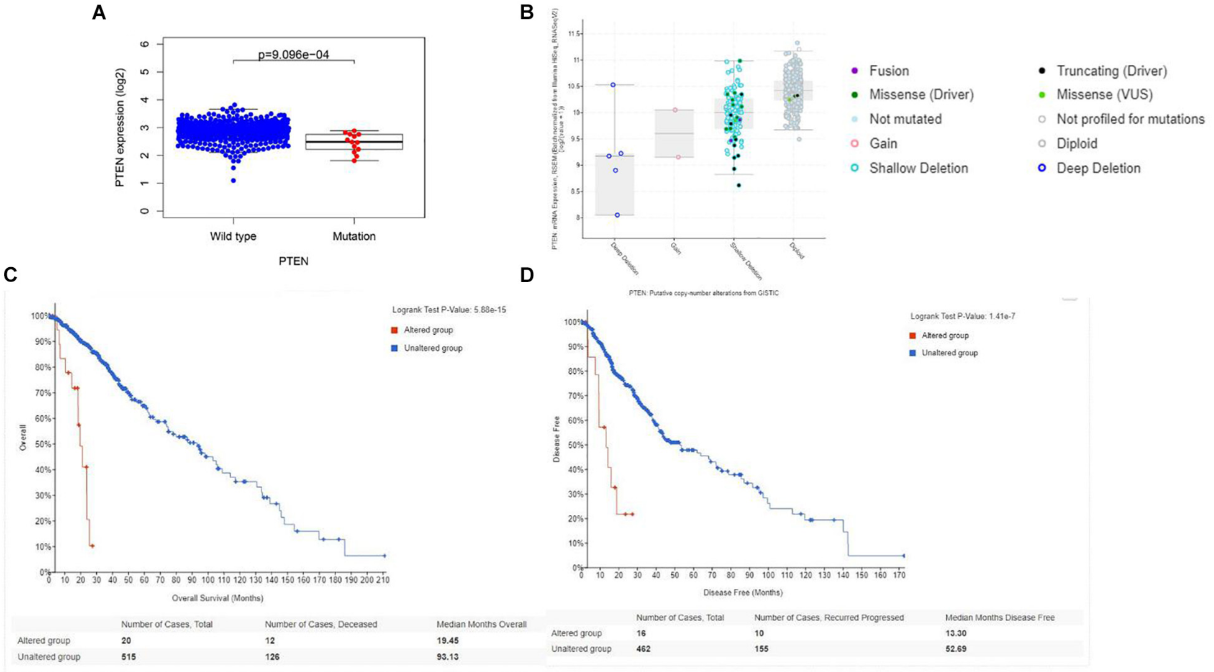
The PTEN gene is known as the tumor suppressor gene, while *PTEN* mutation can decrease the inhibition of tumorigenesis. In

TABLE 1 | Clinical characteristics of patients with low-grade glioma and their *PTEN* status obtained from the Cancer Genome Atlas database.

Characteristics	PTEN status	
	Wild-type	Mutated
Age, years	42.52	54.44
Range	14–87	35–74
Gender		
Female	224	6
Male	273	12
Histologic grade		
G3	247	2
G4	250	16

[5] http://timer.cistrome.org/

[6] https://www.cbioportal.org



	Number of Cases, Total	Number of Cases, Deceased	Median Months Overall
Altered group	**20**	**12**	**19.45**
Unaltered group	**515**	**126**	**93.13**

	Number of Cases, Total	Number of Cases, Recurred Progressed	Median Months Disease Free
Altered group	**16**	**10**	**13.30**
Unaltered group	**462**	**155**	**52.69**

FIGURE 2 | (A) Correlation between the *PTEN* mutation and mRNA expression; **(B)** transcriptional expression of *PTEN* dependence on the *PTEN* status. **(C)** Overall survival of LGG patient dependence on the PTEN status (alteration and wild-type). **(D)** Disease-free survival of LGG patient dependence on the PTEN status (alteration and wild-type).

a previous investigation, a patient with LGG recurrence suffered a poor prognosis (Liu et al., 2020). Thus, early treatment may be helpful for precise therapy in LGG patients harboring the *PTEN* mutation. To perform the Cox regression analysis of multiple factors and *PTEN* mRNA expression including tumor grade and patient age, the static results revealed that *PTEN* mRNA expression levels may affect the prognosis of LGG patients, which is independent of tumor grade, patient age, and gender (**Figure 3**).

PTEN Status Is Correlated With Immune Cell Infiltration Levels in Low-Grade Glioma

The correlation between *PTEN* status and immune cell infiltration (including B cells, CD8^{+} T cells, CD4^{+} T cells, macrophages, neutrophils, and dendritic cells) in LGG patients were evaluated using the TIMER database. The results showed that the *PTEN* mutation is significantly and positively correlated with the infiltration of B cells, macrophages, neutrophils, CD8^{+} T cells, and dendritic cells in LGG patients (**Figure 4**), but not CD4^{+} T cells. Among these differential groups, the immune cell infiltration of *PTEN* mutation was significantly higher than the wild-type group.

Gene Set Enrichment Analysis

To explore the effect of the DEGs on molecular function signaling, the GSEA analysis was employed. By performing the GSEA analysis, we identified eight significant biological function annotations, e.g., unfolded protein response, cholesterol homeostasis, epithelial mesenchymal transition, interferon alpha response, interferon gamma response, and angiogenesis (**Figure 5**). These annotations are the critical components in cancer cell proliferation. The enrichment results indicated that the *PTEN* mutation may play a pivotal role in various pathways involved in cancer cell migration, metabolism, and immune response regulation.

Identification of Differentially Expressed Genes

DEGs were identified by querying the RNA-seq datasets from the *PTEN* mutation (n = 18) or wild-type *PTEN* groups (n = 498). Here, 352 genes were identified as DEGs based on the criteria of $|$ log2FoldChange $|$ $\geq$ 1.0 and P $<$ 0.05 (as shown in **Figure 6A**). Among these DEGs, 91 genes were upregulated and 261 genes were downregulated. Meanwhile, we also explored the correlation between PTEN expression and tumor-related biomarker expression (including Nf1, H3F3A, CDKN2A, IDH1, and FGFR1/2) as shown in **Table 2**. The NF1 expression level was highly positively correlated with PTEN expression (Spearman's efficiency $R =$ 0.405).

GO and KEGG Analyses of Differentially Expressed Genes

In order to explore the biological effect of the dependence of these 352 DEGs on *PTEN* status, we performed GO and KEGG pathway analyses. The GO analysis of the DEGs (**Figure 6B**)

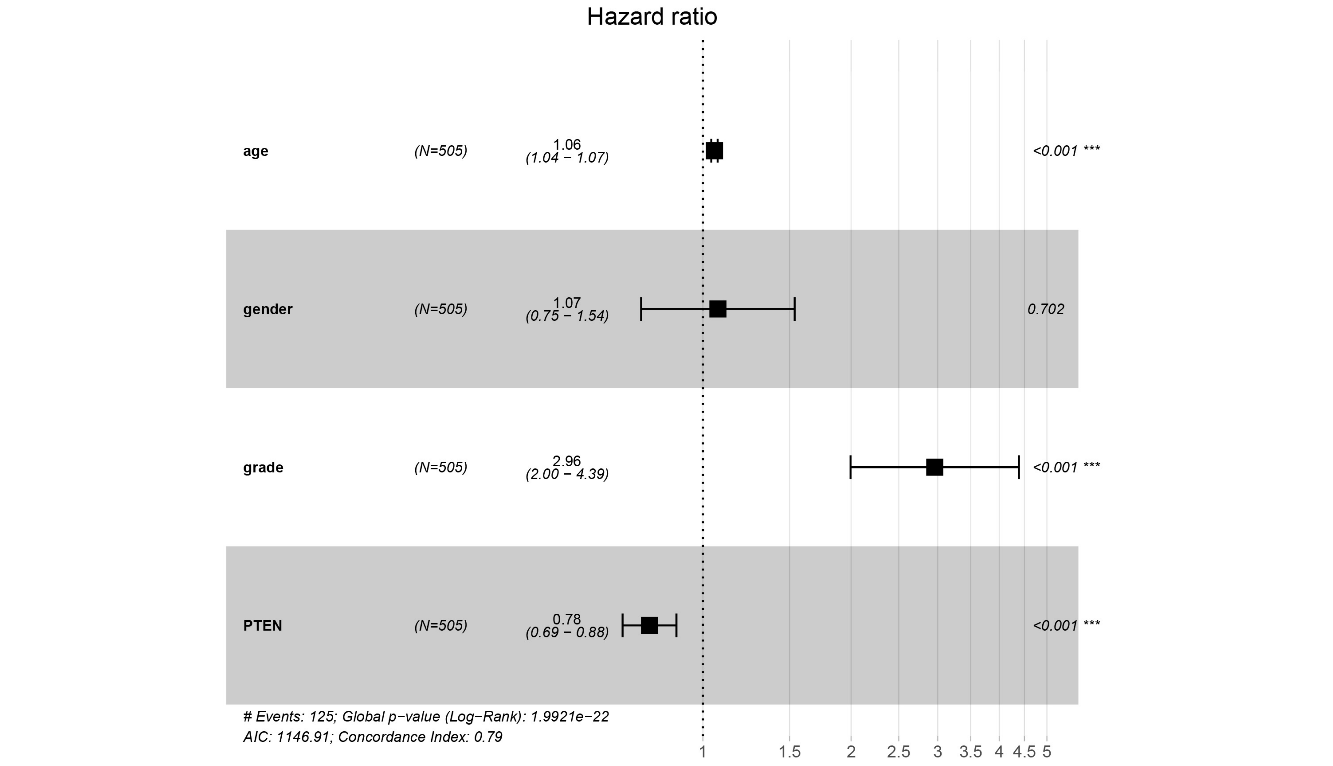


FIGURE 3 | *PTEN* expression levels affected the prognosis of patients with low-grade glioma independently of tumor stage and patient age and gender.

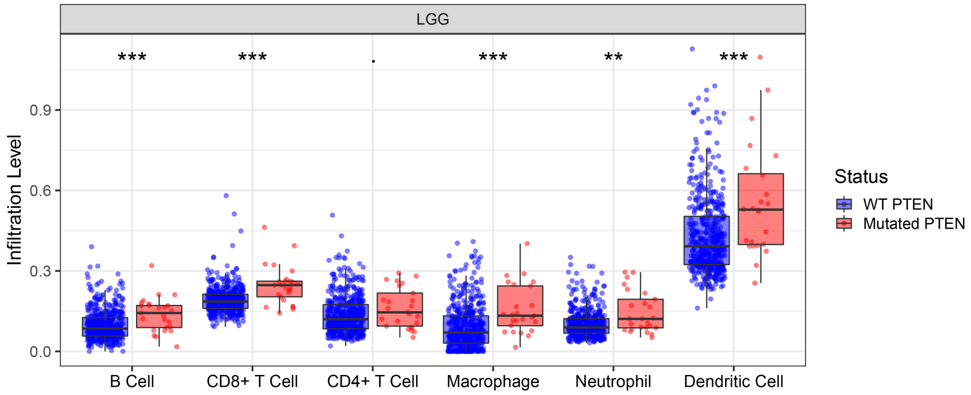


FIGURE 4 | *PTEN* mutation significantly correlates with immune cell infiltration. **$p \leq 0.01$, ***$p \leq 0.001$.

suggested that they were enriched during the regulation of the postsynaptic membrane potential, transsynaptic signaling, the regulation of membrane potential, modulation of the chemical synaptic transmission, synapse organization, synaptic membrane, postsynaptic membrane, pre-synapse, integral components of the synaptic membrane, regulation of the neurotransmitter receptor activity involved in the regulation of the postsynaptic membrane potential, postsynaptic neurotransmitter receptor activity, and ligand-gated ion channel activity. Moreover, the DEGs of the KEGG analysis were enriched in nicotine addiction, morphine addiction, the cyclic adenosine monophosphate (cAMP) signaling pathway, and neuroactive ligand-receptor interaction (**Figure 6C**).

Module Screening

Data created by the STRING database were filtered, and the mutual effect and central genes within the DEGs were studied. The top 10 genes were confirmed to be central genes. These were confirmed as hub genes and included *PSSTR2*, *GABBR1*, *SSTR1*, *CXCL10*, *CCL4*, *ANXA1*, *SAA1*, *CCL4L1*, and *HRH3*. *SSTR2* exhibited the highest degree of nodes among those genes with nine. In the PPI network, the modules of genes

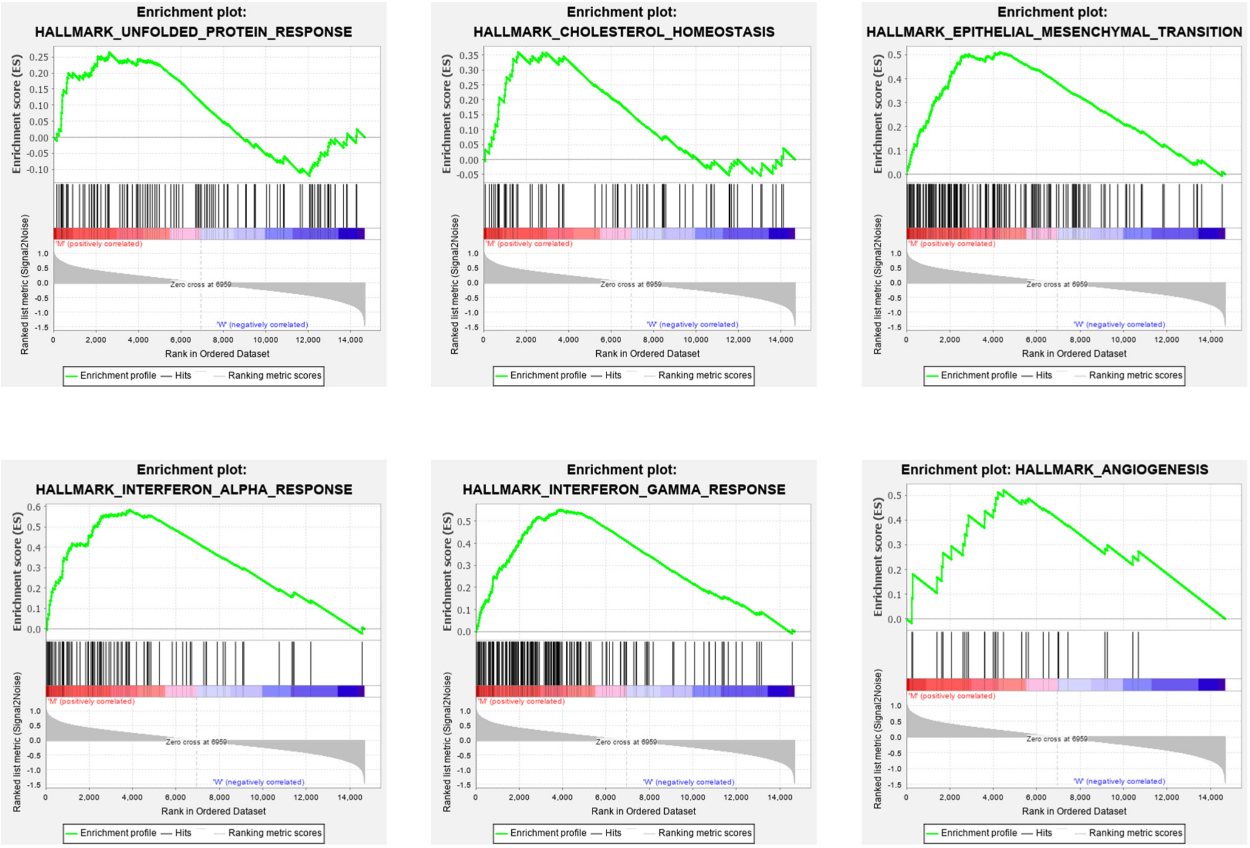


FIGURE 5 | Gene set enrichment analysis results for high *PTEN* expression levels in patients with low-grade glioma.

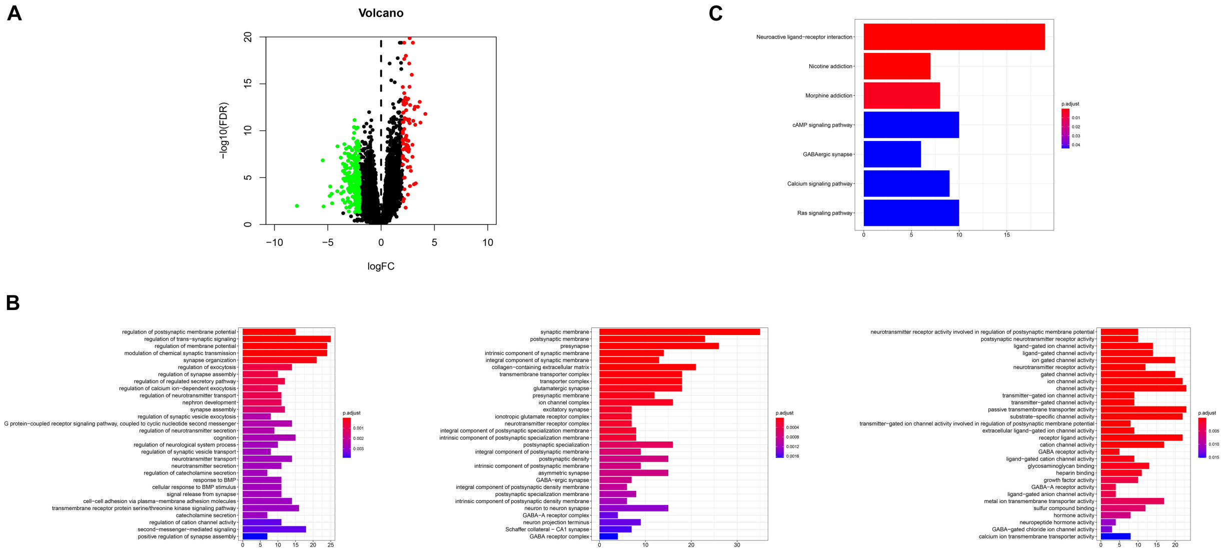


FIGURE 6 | (A) Volcano plot for the differentially expressed genes (DEGs); **(B)** GO enrichment terms of the DEGs, **(C)** KEGG pathway analysis of the DEGs.

were confirmed using the MCODE plug-in in Cytoscape. The top three modules of the GO and KEGG pathways were chosen for analysis (**Figure 7**). The enrichment results suggested that the genes in modules 1–3 were predominantly associated with the G protein-coupled receptor signaling pathway, coupled to the cyclic nucleotide second messenger,

TABLE 2 | Correlation between PTEN expression and tumor-related biomarkers.

Gene	Spearmen's correlation	*p*-value	*q*-value
NF1	0.405	9.26*e* − 22	2.72*e* − 20
H3F3A	−0.172	8.753*e* − 5	2.215*e* − 4
CDKN2A	−0.131	2.944*e* − 3	5.793*e* − 3
IDH1	−0.102	0.0212	0.0354
FGFR1	−0.0845	0.0555	0.0845
FGFR2	−0.151	5.952*e* − 4	1.319*e* − 3

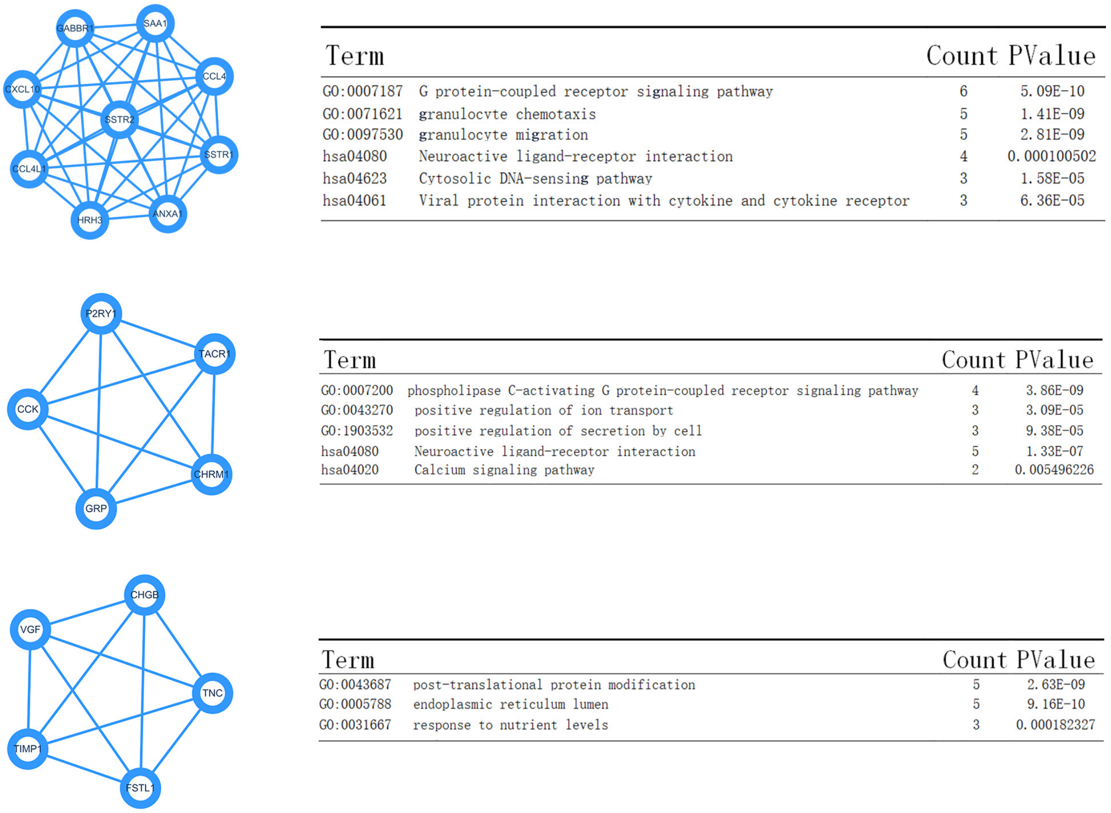
endothelial cell activation, gamma-aminobutyric acid (GABA) receptor activity, cAMP signaling pathway, phospholipase C-activating G protein-coupled receptor signaling pathway, neuroactive ligand-receptor interaction, and post-translational protein modification.

DISCUSSION

The PTEN protein acts as a tumor suppressor, which inhibits the down-stream proteins when performing its suppressing function (Maehama et al., 2001; McGirt et al., 2005). The bioactivity of the PTEN protein is highly dependent on the subsequent antagonism of the PI3K/AKT pathway. However, some literature has reported that PTEN can function through AKT-independence (Freeman et al., 2003). Consequently, it is necessary to explore the biological functions associated with the *PTEN* status. In this study, we carefully evaluated the critical role of *PTEN* mutation in LGG progression and prognosis, which may provide therapeutic scope for precise LGG therapy.

Firstly, the clinical analysis (**Figure 1A**) results showed that 6% of patients with LGG harbored *PTEN* mutations, including four types of mutations (missense mutations, nonsense mutations, amplifications, and deep deletions). Furthermore, the survival curve clearly revealed that patients with the *PTEN* mutation suffered a poorer prognosis, a lower survival rate, and greater disease recurrence than the wild-type group (**Figures 2C,D**). On the basis of the clinical results, early clinical intervention for LGG *PTEN* mutation groups would be helpful for improving the patient survival period.

Second, the Cox analysis revealed that the mRNA expression level of the *PTEN* mutation group was lower than the wild-type group ($P < 0.01$). And shallow deletion and normal diploid types of *PTEN* mRNA level were also higher than the deep deletion and gain status (**Figure 2B**). Considering the *PTEN* mutation types (missense), a mutation of *PTEN* led to the dysfunction of tumorigenesis suppression. Meanwhile, we also found that the PTEN expression level was related with some tumor-related biomarkers (**Table 2**).



Term		Count	PValue
GO:0007187	G protein-coupled receptor signaling pathway	6	5.09E-10
GO:0071621	granulocvte chemotaxis	5	1.41E-09
GO:0097530	granulocyte migration	5	2.81E-09
hsa04080	Neuroactive ligand-receptor interaction	4	0.000100502
hsa04623	Cytosolic DNA-sensing pathway	3	1.58E-05
hsa04061	Viral protein interaction with cytokine and cytokine receptor	3	6.36E-05

Term		Count	PValue
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	4	3.86E-09
GO:0043270	positive regulation of ion transport	3	3.09E-05
GO:1903532	positive regulation of secretion by cell	3	9.38E-05
hsa04080	Neuroactive ligand-receptor interaction	5	1.33E-07
hsa04020	Calcium signaling pathway	2	0.005496226

Term		Count	PValue
GO:0043687	post-translational protein modification	5	2.63E-09
GO:0005788	endoplasmic reticulum lumen	5	9.16E-10
GO:0031667	response to nutrient levels	3	0.000182327

FIGURE 7 | Top three modules from the Pixels Per Inch (PPI) network—**(A,B)** PPI network and GO and KEGG analyses of module 1; **(C,D)** PPI network and GO and KEGG analyses of module 2; **(E,F)** PPI network and GO and KEGG analyses of module 3.

Immune cell infiltration can affect the tumor occurrence, progression, and prognosis, owing to the effect of the tumor microenvironment (Chaffer and Weinberg, 2011; Deng et al., 2020). The tumor microenvironment is highly associated with immune cell infiltration. Consequently, exploring the immune cell infiltration level may provide more scope on the tumor progression and immune status. Here, we identified the role of immune cell infiltration in *PTEN* status using the TIMER database (**Figure 4**). By further analysis, B cells, neutrophils, CD4+ T cells, macrophages, CD8+ T cells, and dendritic cells were more significantly abundant in the *PTEN* mutation group than the wild-type *PTEN* group. Higher immune cell infiltration meant that the complex immune microenvironment may induce a serious progression status. Therefore, these results revealed that some specific genes or signaling may lead to immune cell infiltration in the *PTEN* mutation group.

Finally, we explored the role of critical molecular annotations that led to the poor prognosis of *PTEN* mutation LGG patients (**Figures 5**, **6**). The top six annotations of GSEA were associated with various cancer-related pathways, e.g., epithelial mesenchymal transition, interferon gamma response, interferon alpha response, cholesterol homeostasis, unfolded protein response, and angiogenesis. These molecular functions promoted tumorigenesis (epithelial mesenchymal transition) and enhanced drug resistance (unfolded protein response). By deeply affecting these signal pathways, LGGs with *PTEN* mutations can lead to a higher tumor grade and poor survival (**Figures 2C,D**).

After the identification of DEGs (**Figure 6A**), the GSEA analysis on the biological function levels were carefully studied (**Figures 6B,C**). The GO annotations showed that the top five ranking annotations were mostly associated with the signal transduction process, for example, the regulation of postsynaptic membrane potential, synaptic membrane, and ligand-gated ion channel activity. These annotations were "neuron"-related signaling and these molecular functions may affect cell metastasis. Among them, epithelial mesenchymal transition, a complex biological process, contributed to metastasis, wherein the genetic and epigenetic events caused the epithelial cells to acquire a mesenchymal gene activity signature and phenotype (Subramanian et al., 2005; Li et al., 2020). The GO analysis results showed that 352 DEGs could be attributed to the top three GO annotations: growth factor activity, GABA receptor activity, and neurotransmitter secretion. The enriched annotations of growth factor activity and neurotransmitter secretion were consistent with immune cell infiltration.

Moreover, the KEGG enrichment results confirmed the GO analysis, because the top ranking pathways were neuron-related molecular functions and tumor-related pathways (Ras signaling pathway). The PPI network analysis also provided four key nodes by STRING analysis (**Figure 7**). These node networks were highly associated with the protein modification process.

CONCLUSION

In summary, we systematically investigated the *PTEN* mutation condition with LGG poor prognosis and immune cell infiltration. These results revealed that a PTEN mutation can promote the tumorigenesis process and lead to more immune cell infiltration. Thus, our results showed the importance of *PTEN* status in disease progression and revealed that it may become a useful biomarker for diagnosing LGGs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA.

AUTHOR CONTRIBUTIONS

All authors contributed to the literature investigation, data collection, writing the manuscript, providing useful discussion of its content, and undertaking reviews or revising the manuscript before submission.

FUNDING

This study was supported by a grant (No. 2018ZDXM-SF-040) from the Key Project of Shaanxi Province-the field of Social Development.

REFERENCES

Andersson, U., Malmer, B., Bergenheim, A. T., Brannstrom, T., and Roger, H. (2004). Heterogeneity in the expression of markers for drug resistance in brain tumors. *Clin. Neuropathol.* 23, 21–27.

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Best, S. A., De Souza, D. P., Kersbergen, A., Antonia, N. P., Saravanan, D., and Dedreia, T. (2018). Synergy between the KEAP1/NRF2 and PI3K pathways drives non-small-cell lung cancer with an altered immune microenvironment. *Cell Metab.* 27, 935–943.e4.

Buckner, J. C., Shaw, E. G., Pugh, S. L., Arnab, C., Mark, G. R., and Geoffrey, R. B. (2016). Radiation plus procarbazine, CCNU, and vincristine in low-grade glioma. *N. Engl. J. Med.* 374, 1344–1355.

Calatozzolo, C., Pollo, B., Botturi, A., Loredana, D., Mariantonia, C., and Andrea, S. (2012). Multidrug resistance proteins expression in glioma patients with epilepsy. *J. Neurooncol.* 110, 129–135. doi: 10.1007/s11060-012-0946-9

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Deng, X., Lin, D., Zhang, X., Xuchao, S., Zelin, Y., and Liang, Y. (2020). Profiles of immune-related genes and immune cell infiltration in the tumor microenvironment of diffuse lower-grade gliomas. *J. Cell. Physiol.* 235, 7321–7331. doi: 10.1002/jcp.29633

Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:R60. doi: 10.1186/gb-2003-4-9-r60

Freeman, D. J., Li, A. G., Wei, G., Li, H. H., Kertesz, N., Lesche, R., et al. (2003). TEN tumor suppressor regulates p53 protein levels and activity through phosphatase-dependent and independent mechanisms. *Cancer Cell* 3, 117–130. doi: 10.1016/S1535-6108(03)00021-7

Helseth, R., Helseth, E., Johannesen, T. B., Langberg, C. W., Lote, K., Rønning, P., et al. (2010). Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. *Acta Neurol. Scand.* 122, 159–167. doi: 10.1111/j.1600-0404.2010.01350.x

Ino, Y., Yamazaki-Itoh, R., Shimada, K., Iwasaki, M., Kosuge, T., and Kanai, Y. (2013). Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer. *Br. J. Cancer* 108, 914–923. doi: 10.1038/bjc.2013.32

Johnson, D. R., and O'Neill, B. P. (2012). Glioblastoma survival in the United States before and during the temozolomide era. *J. Neurooncol.* 107, 359–364. doi: 10.1007/s11060-011-0749-4

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S. I., et al. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 275, 1943–1947. doi: 10.1126/science.275.5308.1943

Li, T., Fu, J., Zeng, Z., David, C., Jing, L., and Qianming, C. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514.

Liu, W., Xu, Z., Zhou, J., Xing, S., Li, Z., Gao, X., et al. (2020). High levels of HIST1H2BK in low-grade glioma predicts poor prognosis: a study using CGGA and TCGA Data. *Front. Oncol.* 10:627. doi: 10.3389/fonc.2020.00627

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109. doi: 10.1007/978-94-007-1399-4_10

Maehama, T., Taylor, G. S., and Dixon, J. E. (2001). PTEN and myotubularin: novel phosphoinositide phosphatases. *Annu. Rev. Biochem.* 70, 247–279. doi: 10.1146/annurev.biochem.70.1.247

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

McGirt, M. J., Woodworth, G. F., Coon, A. L., Frazier, J. M., Amundson, E., Garonzik, I., et al. (2005). Independent predictors of morbidity after image-guided stereotactic brain biopsy: a risk assessment of 270 cases. *J. Neurosurg.* 102, 897–901. doi: 10.3171/jns.2005.102.5.0897

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273. doi: 10.1038/ng1180

Mu, M., Niu, W., Zhang, X., Hu, S., and Niu, C. (2020). LncRNA BCYRN1 inhibits glioma tumorigenesis by competitively binding with miR-619-5p to regulate CUEDC2 expression and the PTEN/AKT/p21 pathway. *Oncogene* 39, 6879–6892. doi: 10.1038/s41388-020-01466-x

Norton, K. A., Gong, C., Jamalian, S., and Popel, A. S. (2019). Multiscale agent-based and hybrid modeling of the tumor immune microenvironment. *Processes* 7, 37. doi: 10.3390/pr7010037

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Steck, P. A., Pershouse, M. A., Jasser, S. A., Yung, W. K., Lin, H., Ligon, A. H., et al. (1997). Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nat. Genet.* 15, 356–362. doi: 10.1038/ng0497-356

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Wiencke, J. K., Zheng, S., Jelluma, N., Tarik, T., Scott, V., Tanja, T., et al. (2007). Methylation of the PTEN promoter defines low-grade gliomas and secondary glioblastoma. *Neuro Oncol.* 9, 271–279. doi: 10.1215/15228517-2007-003

Wu, J., Xu, W. H., Wei, Y., Qu, Y. Y., Zhang, H. L., and Ye, D. W. (2018). An Integrated score and nomogram combining clinical and immunohistochemistry factors to predict high ISUP grade clear cell renal cell carcinoma. *Front. Oncol.* 8:634. doi: 10.3389/fonc.2018.00634

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

BRIEF RESEARCH REPORT
published: 01 February 2021
doi: 10.3389/fmolb.2021.620793

Protein Stability Perturbation Contributes to the Loss of Function in Haploinsufficient Genes

Giovanni Birolo[1], *Silvia Benevenuta*[1], *Piero Fariselli*[1]*, *Emidio Capriotti*[2]*, *Elisa Giorgio*[3] and *Tiziana Sanavia*[1]

[1]*Department of Medical Sciences, University of Torino, Italy,* [2]*Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Italy,* [3]*Department of Molecular Medicine, University of Pavia, Italy*

Missense variants are among the most studied genome modifications as disease biomarkers. It has been shown that the "perturbation" of the protein stability upon a missense variant (in terms of absolute ΔΔG value, i.e., |ΔΔG|) has a significant, but not predictive, correlation with the pathogenicity of that variant. However, here we show that this correlation becomes significantly amplified in haploinsufficient genes. Moreover, the enrichment of pathogenic variants increases at the increasing protein stability perturbation value. These findings suggest that protein stability perturbation might be considered as a potential cofactor in diseases associated with haploinsufficient genes reporting missense variants.

Keywords: protein mutation, protein stability, haploinsuffciency, variant effect prediction, protein stability prediction

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom

Reviewed by:
Joost Schymkowitz,
VIB and KU Leuven Center for Brain
and disease Research, Belgium
Saraboji Kadhirvel,
SASTRA University, India

****Correspondence:***
Piero Fariselli
piero.fariselli@unito.it
Emidio Capriotti
emidio.capriotti@unibo.it

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *23 October 2020*
Accepted: *07 January 2021*
Published: *01 February 2021*

Citation:
Birolo G, Benevenuta S, Fariselli P, Capriotti E, Giorgio E and Sanavia T (2021) Protein Stability Perturbation Contributes to the Loss of Function in Haploinsufficient Genes. Front. Mol. Biosci. 8:620793. doi: 10.3389/fmolb.2021.620793

INTRODUCTION

Missense variations may cause loss-of-function by directly perturbing protein-protein interactions or ablating enzymatic activity or by inducing structural destabilization of the protein (Stein et al., 2019), which in turn may trigger protein misfolding and degradation. Many neurodegenerative diseases, such as Parkinson's disease, are also associated with destabilization of the corresponding proteins (Wilson et al., 2014). However, there are cases where missense variations increase protein stability while still being deleterious. As an example, the variation H101Q in the CLIC2 protein has been associated with a mental disorder and predicted to make the CLIC2 protein thermodynamically more stable and to interact more strongly with the ryanodine receptor, obstructing its transport to the cell membrane (Witham et al., 2011). Therefore, stability perturbations, rather than protein destabilization, can be linked with disease-causing variations.

Recently, Gerasimavicius et al. have highlighted an improvement in the identification of pathogenic variations using |ΔΔG| values (Gerasimavicius et al., 2020). However, very little is known about thermodynamic changes in human protein variants so far (Sanavia et al., 2020), and the processes establishing whether a variation perturbing the protein stability is or not disease-related are not clear yet. An extensive comparative analysis has proven that, on average, variations mostly involved in disease also associated with large effects on protein stability (Casadio et al., 2011). Although several studies tried to predict the functional or structural impacts of missense variations, the mechanism of the phenotypic impact through inheritance modes of the missense variations are still unclear. Indeed recessive variations are mainly observed in the buried region of protein structures and more likely associated with loss-of-function, whereas dominant variations are significantly enriched in the interfaces of molecular interactions and more difficult to be identified as disease-related (Guo et al., 2013; Martelli et al., 2016).

One of the most known pathogenic mechanisms for loss-of-function mutations is haploinsufficiency, a type of genetic dominance wherein a single functional copy of a gene is insufficient to maintain normal function. Different theories have been put forth to explain the cause of haploinsufficiency. One of them states that growth defects caused by changes in gene dosage are due to stoichiometric imbalances of protein complexes interfering with cellular functions (Veitia and Potier, 2015), whose interactions relying on the relative stoichiometry may be either cooperative or competitive. An example of this latter case is the cytotoxic T-lymphocyte-associated protein 4 (CTLA4), which competes for the same ligands with cluster of differentiation 28 (CD28), a T-cell activator. An inappropriate balance of CTLA4 and CD28 can result in T-cell overactivation by CD28 and autoimmune disease. Recently, it was observed a fatal heterozygous mutation in CTLA-4, predicted to decrease protein stability resulting in haploinsufficiency and decreased CTLA-4 expression in a patient reporting autoimmunity (Evan's syndrome), lymphoproliferation and severe infections (Moraes-Fontes et al., 2017).

In this brief report, we suggest that one possible contribution to the pathogenic mechanism in haploinsufficient genes can be related to missense variants perturbing protein stability.

METHOD

Dataset

Performance assessment of 13 computational stability predictors, i.e., FoldX 5.0 (Delgado et al., 2019), INPS3D (Savojardo et al., 2016), Rosetta (Alford et al., 2017), PoPMusic (Dehouck et al., 2011), I-Mutant (Capriotti et al., 2005), SDM (Worth et al., 2011), SDM2 (Pandurangan et al., 2017), mCSM (Pires et al., 2014a), DUET (Pires et al., 2014b), CUPSAT (Parthiban et al., 2006), MAESTRO (Laimer et al., 2016), ENCoM (Frappier et al., 2015), DynaMut (Rodrigues et al., 2018), was investigated for detecting pathogenicity in (Gerasimavicius et al., 2020), considering $|\Delta\Delta G|$ values obtained from each predictor on a dataset of 13,508 missense variations from 96 different high-resolution (<2 Å) crystal structures of disease-associated monomeric proteins encoded by 100 genes. The dataset includes 3,338 missense variants which are annotated in Clinvar (Landrum et al., 2018) as pathogenic or likely pathogenic, associated to proteins with at least 10 known pathogenic missense variations occurring at residues present in the structure. These pathogenic variants are compared against 10,170 "putatively benign" missense variants collected from gnomAD v2.1 (Karczewski et al., 2020) from the same genes as the pathogenic variants. In order to highlight whether the performance obtained by the protein stability predictors might be influenced by the inheritance mode of the related coding genes, we annotated them according to the curated lists of autosomal dominant/recessive and haploinsufficient genes reported by the MacArthur Lab (https://github.com/macarthur-lab/gene_lists). The number of variants for each inheritance mode, split by pathogenic/benign, are 1,217/1,252, 753/1,819, and 635/4,253 for haploinsufficient, dominant, and recessive genes, respectively.

Performance Evaluation

The assumption is that the $|\Delta\Delta G|$ values provided by the predictors can be used as a measure of pathogenicity, with lower values associated with neutral variations. The $|\Delta\Delta G|$ values are used to compute the area under the receiver operating characteristic curve (AUC) as the performance metric as in (Gerasimavicius et al., 2020). In this way, we do not need to select any specific threshold for the perturbation to define a pathogenicity score. However, to avoid biases due to the low proportion of pathogenic variants, here the AUC and the precision were calculated by averaging the results on balanced subsets. More precisely, the available pathogenic variants were matched with a random subset with the same number of benign variants for 100 times. This procedure was applied to the full dataset, for each gene separately and for the variants of each specific inheritance mode (i.e. haploinsufficient, autosomal dominant and recessive), along with their complement set. AUCs were always computed on $|\Delta\Delta G|$ values.

RESULTS

Figure 1 shows the AUCs obtained from each predictor and the mean output of the best two performing methods (FoldX 5.0 and INPS3D, orange bar in the figure). We also tested all combinations of the three best predictors, which performed slightly worse (**Supplementary Figure S1,S2**). The bars reported in **Figure 1** reflect the probability of a randomly chosen disease variant being assigned a higher-ranking score than a random benign one (Gerasimavicius et al., 2020). The barplots highlight the variability in terms of performance among the prediction stability-based methods, with FoldX 5.0 reaching the best AUC. It is worth noting that the combination of the scores from FoldX 5.0 and INPS3D increases the AUC performance of 2 percentage points over FoldX 5.0.

We then evaluated the scores by grouping the gene variants according to their inheritance mode (i.e. autosomal dominant/recessive or haploinsufficiency) in order to provide a biological interpretation. Interestingly, we found that the performance is significantly higher in haploinsufficient genes (**Figure 2**, top panel), while it is lower in not haploinsufficient dominant genes (**Figure 2**, central panel). Recessive genes show no significant differences from non-recessive genes. (**Figure 2**, central and bottom panels).

Since stability change is one of the possible disease mechanisms to be linked with potential pathogenicity, we do not expect a high predictive power for small $\Delta\Delta G$ values. However, we can expect an enrichment of pathogenic variants at increasing protein stability perturbations. This hypothesis is confirmed in **Figure 3**, where we observed that variants with very high $|\Delta\Delta G|$ values tend to be strongly enriched in pathogenic variants. In general this is valid for all genes, but much more for haploinsufficient genes.

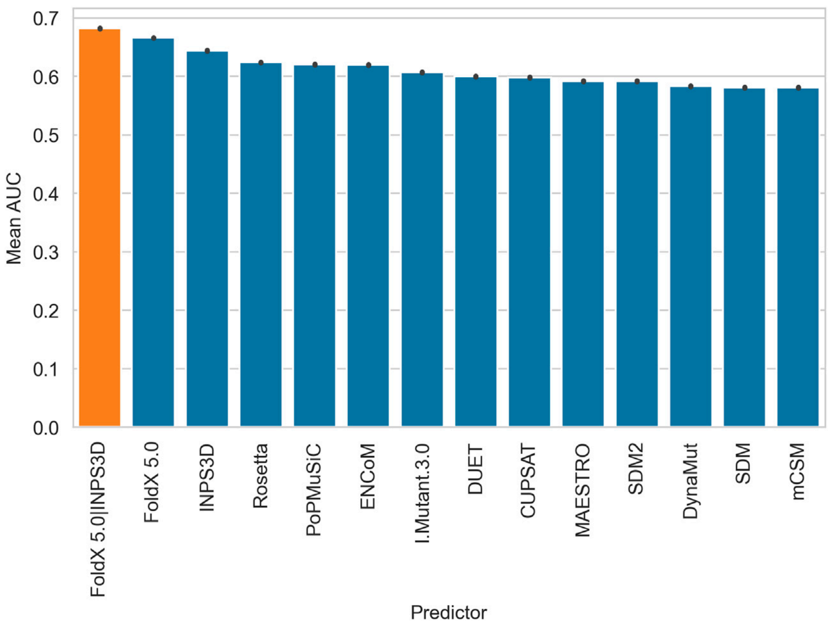


FIGURE 1 | Barplots displaying the performance (AUC) of all the ΔΔG predictors and the consensus (orange) of the best two performing methods (FoldX5.0 and INPS3D). The bars represent the mean AUC obtained by averaging balanced subsets (the available pathogenic variants were matched with a random sample with the same number of benign variants for one hundred times).

This result suggests that it is possible to generate a highly specific test for pathogenicity by selecting the variants according to a fixed threshold for the predicted |ΔΔG|. However, choosing the best |ΔΔG| threshold is highly dependent on the type of predictor used. When considering the best performing one, i.e., the mean between FoldX 5.0 and INPS3D |ΔΔG| values, we see that a threshold of 4.4 kcal/mol yields a precision (positive predictive value) of 96%.

Most of the variants are predicted to be destabilizing by the predictors, and this prevents us from analyzing the effect of the stabilizing variants separetely. Conversely, when only the predicted destabilizing variants are considered (**Supplementary Figure S3**), the trends are similar but slightly higher to those reported in **Figure 3**.

DISCUSSION

Genetic dominance originates from a variety of unrelated mechanisms (Veitia and Potier 2015). One of those is haploinsufficiency, namely the intolerance of a gene to the loss of one allele. As a consequence, the relative protein dosage is half of the normal level, which is not sufficient to ensure a normal function and consequently causes the pathological phenotype. Possible genetic causes are, for example, the deletion of one allele or protein-truncating variants that may induce nonsense-mediated decay of transcripts.

The better performance of ΔΔG predictors in haploinsufficient genes suggests that missense variants causing significant changes in protein stability may play a relevant role in disease development when genes are haploinsufficient. It does not seem far-fetched to argue that variants causing strong ΔΔG perturbations are likely to yield a non-functional protein, thus becoming loss-of-function variants, which are the main driver of pathogenicity in haploinsufficient genes. On the other hand, the lower performance on non-haploinsufficient dominant genes shows that this role does not extend to other dominance mechanisms, which are often activated by "gain-of-function" variants, where the mutated protein actively interferes with the gene function. This may suggest that ΔΔG perturbations are not predictive of "gain-of-function" effects.

Figure 3 shows that protein stability-based methods are able to predict pathogenic variants in haploinsufficient genes at high precision (>96%) using thresholds on |ΔΔG| values above 4.4 kcal/mol. However, since ΔΔG perturbation is only one of the many molecular mechanisms affecting pathogenicity, we do not expect to gain in sensitivity by decreasing the |ΔΔG| threshold: missense variants predicted to cause only modest ΔΔG changes may cause disease by other mechanisms like compromising the protein interaction capabilities. On the other hand, significant ΔΔG perturbations can shift the protein far from its dynamically active state, making the protein non-functional. Indeed, we confirmed that perturbing variants (predicted to be either very destabilizing or stabilizing) have a high probability of being pathogenic. Thus, by choosing an appropriate |ΔΔG| threshold (which is dependent on the specific ΔΔG predictor), we can turn ΔΔG predictors into highly precise pathogenicity predictors for haploinsufficient genes.

While the absolute value of the ΔΔG was used for all analyses, it would have been interesting to analyze variants predicted to increase or decrease stability separately. This would have allowed

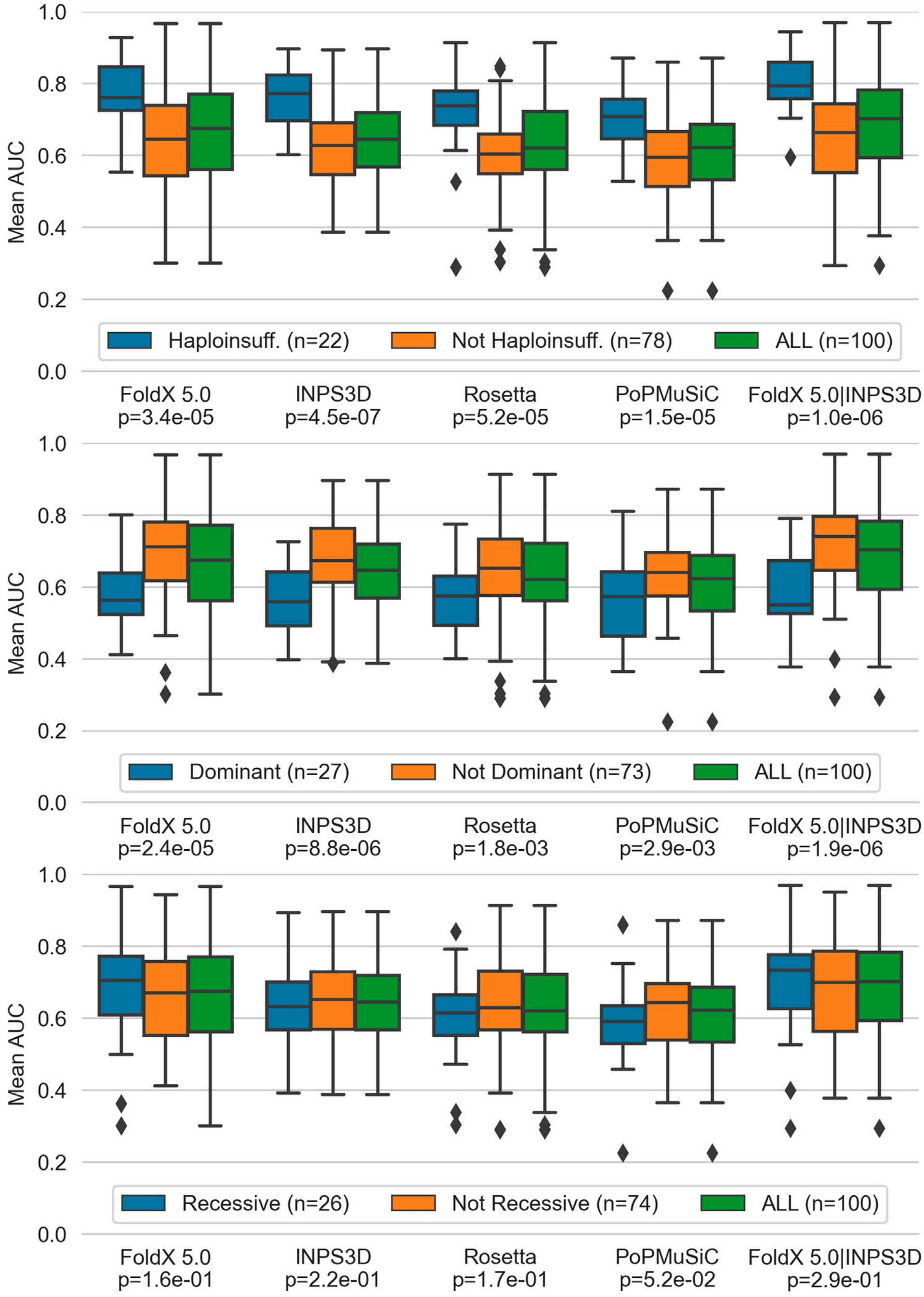


FIGURE 2 | Performance of top performing predictors, (i.e. FoldX 5.0 and INPS3D, Rosetta and PoPMuSiC along with the combined scores of the first two) split by haploinsufficient, dominant without haploinsufficiency and recessive genes. P-values of the pairwise comparison between each gene group and its complement by the Mann-Whitney *U* test are reported at the bottom of the *x*-axis.

us to check if stabilizing variants could be associated for instance with gain-of-function mechanisms, differently from destabilizing variants. However, a high proportion of the variants in our dataset were predicted to be destabilizing, leaving an insufficient number of stabilizing and especially highly stabilizing variants for a robust statistical analysis. This interesting question should be addressed in the next future when more data will be available by correctly mapping annotated variants to protein structures.

In conclusion, large ΔΔG perturbations in haploinsufficient gene products appear to be a significant factor in the pathogenicity assessment of the missense variants. Therefore, we recommend complementing the state-of-the-art pathogenicity

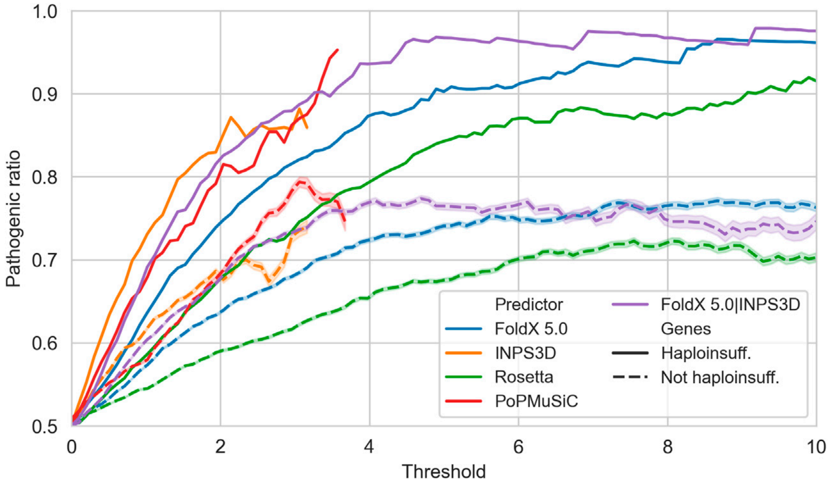


FIGURE 3 | Precision (*y*-axis) of the protein stability-based methods in predicting pathogenicity at different |ΔΔG| values, defined as the ratio of truly pathogenic over all the variants reporting predicted |ΔΔG| values above a specific threshold (*x*-axis). Solid and dashed lines are computed on variants in haploinsufficient and non-haploinsufficient genes, respectively. INPS3D and PoPMuSiC lines stop earlier since the methods do not provide predictions with |ΔΔG| values greater than the reported thresholds.

predictions with one of the best performing ΔΔG predictors, at least for haploinsufficient genes, when looking for possible disease causes. High |ΔΔG| values indicate that protein stability perturbation is a reasonable cause of the observed pathological condition.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://doi.org/10.1038/s41598-020-72404-w.

AUTHOR CONTRIBUTIONS

TS, PF and EC designed the research. GB retrieved the data and the annotations, ran the analyses and drafted the manuscript. All the authors interpreted the results and contributed to the submitted version.

FUNDING

This work was supported by the PRIN project, “Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation” of the Ministero Istruzione, Università e Ricerca (201744NR8S). We thank the EU projects GenoMed4All (H2020 RIA n.101017549) and BRAINTEASER (H2020 RIA n.101017598).

ACKNOWLEDGMENTS

We thank the Research for the MIUR program “Dipartimenti di Eccellenza 20182022D15D18000410001”.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.620793/full#supplementary-material.

REFERENCES

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O’Meara, M. J., DiMaio, F. P., Park, H., et al. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theor. Comput.* 13 (6), 3031–3048. doi:10.1021/acs.jctc.7b00125

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310. doi:10.1093/nar/gki375

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* 32 (10), 1161–1170. doi:10.1002/humu.21555

Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinf.* 12, 151. doi:10.1186/1471-2105-12-151

Delgado, J., Radusky, L. G., Cianferoni, D., and Serrano, L. (2019). FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* 35 (20), 4168–4169. doi:10.1093/bioinformatics/btz184

Frappier, V., Chartier, M., and Najmanovich, R. J. (2015). ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res.* 43 (W1), W395–W400. doi:10.1093/nar/gkv343

Gerasimavicius, L., Liu, X., and Marsh, J. A. (2020). Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep.* 10 (1), 15387. doi:10.1038/s41598-020-72404-w

Guo, Y., Wei, X., Das, J., Grimson, A., Lipkin, S. M., Clark, A. G., et al. (2013). Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. *Am. J. Hum. Genet.* 93 (1), 78–89. doi:10.1016/j.ajhg.2013.05.022

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581 (7809), 434–443. doi:10.1038/s41586-020-2308-7

Laimer, J., Hiebl-Flach, J., Lengauer, D., and Lackner, P. (2016). MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics* 32 (9), 1414–1416. doi:10.1093/bioinformatics/btv769

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46 (D1), D1062–D1067. doi:10.1093/nar/gkx1153

Martelli, P. L., Fariselli, P., Savojardo, C., Babbi, G., Aggazio, F., and Casadio, R. (2016). Large scale analysis of protein stability in OMIM disease related human protein variants. *BMC Genom.* 17 (Suppl. 2), 397. doi:10.1186/s12864-016-2726-y

Moraes-Fontes, M. F., Hsu, A. P., Caramalho, I., Martins, C., Araújo, A. C., Lourenço, F., et al. (2017). Fatal CTLA-4 heterozygosity with autoimmunity and recurrent infections: a de novo mutation. *Clin Case Rep.* 5 (12), 2066–2070. doi:10.1002/ccr3.1257

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45 (W1), W229–W235. doi:10.1093/nar/gkx439

Parthiban, V., Gromiha, M. M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 34, W239–W242. doi:10.1093/nar/gkl190

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014b). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* 42, W314–W319. doi:10.1093/nar/gku411

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014a). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30 (3), 335–342. doi:10.1093/bioinformatics/btt691

Rodrigues, C. H., Pires, D. E., and Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46 (W1), W350–W355. doi:10.1093/nar/gky300

Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput. Struct. Biotechnol. J.* 18, 1968–1979. doi:10.1016/j.csbj.2020.07.011

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 32 (16), 2542–2544. doi:10.1093/bioinformatics/btw192

Stein, A., Fowler, D. M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.*, 44(7), 575–588. doi:10.1016/j.tibs.2019.01.003

Veitia, R. A., and Potier, M. C. (2015). Gene dosage imbalances: action, reaction, and models. *Trends Biochem. Sci.* 40 (6), 309–317. doi:10.1016/j.tibs.2015.03.011

Wilson, G. R., Sim, J. C., McLean, C., Giannandrea, M., Galea, C. A., Riseley, J. R., et al. (2014). Mutations in RAB39B cause X-linked intellectual disability and early-onset Parkinson disease with α-synuclein pathology. *Am. J. Hum. Genet.* 95 (6), 729–735. doi:10.1016/j.ajhg.2014.10.015

Witham, S., Takano, K., Schwartz, C., and Alexov, E. (2011). A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins* 79 (8), 2444–2454. doi:10.1002/prot.23065

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* 39, W215–W222. doi:10.1093/nar/gkr363

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 10 March 2021
doi: 10.3389/fmolb.2020.620554

Influence of Disease-Causing Mutations on Protein Structural Networks

Vasam Manjveekar Prabantu[1], *Nagarajan Naveenkumar*[1,2,3] *and Narayanaswamy Srinivasan*[1]*

[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India,* [2]*National Centre for Biological Sciences, TIFR, Bangalore, India,* [3]*Bharathidasan University, Tiruchirappalli, India*

The interactions between residues in a protein tertiary structure can be studied effectively using the approach of protein structure network (PSN). A PSN is a node-edge representation of the structure with nodes representing residues and interactions between residues represented by edges. In this study, we have employed weighted PSNs to understand the influence of disease-causing mutations on proteins of known 3D structures. We have used manually curated information on disease mutations from UniProtKB/Swiss-Prot and their corresponding protein structures of wildtype and disease variant from the protein data bank. The PSNs of the wildtype and disease-causing mutant are compared to analyse variation of global and local dissimilarity in the overall network and at specific sites. We study how a mutation at a given site can affect the structural network at a distant site which may be involved in the function of the protein. We have discussed specific examples of the disease cases where the protein structure undergoes limited structural divergence in their backbone but have large dissimilarity in their all atom networks and vice versa, wherein large conformational alterations are observed while retaining overall network. We analyse the effect of variation of network parameters that characterize alteration of function or stability.

Keywords: disease-causing mutations, protein structure networks, allostery, network variability, protein function

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
University of Cambridge,
United Kingdom

Reviewed by:
Frank Eisenhaber,
*Bioinformatics Institute (A*STAR),*
Singapore
Igor N. Berezovsky,
*Bioinformatics Institute (A*STAR),*
Singapore

*****Correspondence:***
Narayanaswamy Srinivasan
ns@iisc.ac.in

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *23 October 2020*
Accepted: *17 December 2020*
Published: *10 March 2021*

Citation:
Prabantu VM, Naveenkumar N and Srinivasan N (2021) Influence of Disease-Causing Mutations on Protein Structural Networks. Front. Mol. Biosci. 7:620554. doi: 10.3389/fmolb.2020.620554

INTRODUCTION

The amino acid sequence determines the protein 3-D structure (Anfinsen, 1973) which is related to its function. An alteration in the amino acid sequence can bring about changes in the folding and stability of the protein (Lorch et al., 1999; Lorch et al., 2000), interaction of the protein with other molecules (Rignall et al., 2002; Ung et al., 2006) and change in functional levels (Tiede et al., 2006) or overall function of the protein as well. A mutation in the amino acid sequence may alter the structure of a protein but it does not necessarily alter its function, although, the mutation at specific sites such as conserved residues can bring about a change in the structure and function of the protein.

In humans, the most frequent genetic variants are single nucleotide polymorphisms (SNPs) which have been studied extensively (Buetow et al., 1999; Cargill et al., 1999; Collins et al., 1999; Halushka et al., 1999). SNPs could be non-synonymous which bring about a change in the amino acid sequence. Several such genetic variants are known to cause mutations in their gene product and their information is available in resources such as the SNPdb (Sherry et al., 2001) and 1000 Genomes project (Auton et al., 2015). Some of the mutations in a protein are known to enhance the susceptibility or predisposition to a disease and are referred to as disease causing mutations. A few resources are available that map the gene variants to the diseases they may cause. ClinVar is a public archive mapping sequence variants and human phenotype (Landrum et al., 2018), COSMIC is a large catalogue of mutations associated with cancer (Forbes et al., 2017) and SwissVar is a one stop database for the easy retrieval of amino acid polymorphisms and the phenotype information (Mottaz et al., 2010). All the information from the SwissVar is now directly available via the UniProt knowledgebase (Bateman, 2019). However, specific information of the gene variants is compiled as a catalogue and is available on the Humsavar knowledge base which is an index of manually curated human polymorphisms and disease mutations. (https://www.uniprot.org/docs/humsavar).

Mutations in the protein sequence can alter the structure that is natively conferred by the sequence of the wildtype (Taverna and Goldstein, 2002; Tokuriki and Tawfik, 2009). In several scenarios the site of mutation is distant from the site of function, and still one observes a loss of function or alteration in functional levels (Mitternacht and Berezovsky, 2011; Yang et al., 2016). Although, the conformation of the mutant protein may be highly similar to the conformation of the wildtype, there could be alterations in their topologies at sites distant from the site of mutation (Rajasekaran et al., 2017). This concept of alteration of the structure at distant sites from the site of perturbation has been well documented under the subject of allostery (Gunasekaran et al., 2004; Weinkam et al., 2013; Naganathan 2019). Without much change in the overall topology of the protein an allosteric signal can transmit the effect of a perturbation to a different site in the protein structure (Guarnera and Berezovsky, 2019a; Guarnera and Berezovsky, 2019b). The internal protein structural network defines the connectivity between atoms/residues (Vijayabaskar and Vishveshwara, 2010). When perturbations are bought into the system such as disease-causing mutations, it is seen that the variation in the connectivity of the elements within the system brings about allosteric changes in functional sites and elsewhere (Dubay et al., 2015; Guarnera et al., 2017; Tan et al., 2019; Tee, Guarnera and Berezovsky, 2019; Guarnera and Berezovsky, 2020).

In this study, we use the Humsavar knowledge base to identify disease-causing mutations in proteins and analyse the variability in protein structural networks between wildtype and disease-causing variant. We explore the possibility of mutations at a given site that can affect the structural network at a distant site which may be involved in the function of the protein.

MATERIALS AND METHODS

A Dataset of Disease-Causing Variants in Humans

The disease variant information provided in the Humsavar knowledge base is a manually curated subset of UniProtKB/Swiss-Prot protein data for human polymorphisms and disease mutations with their amino acid variations imported from Ensembl variation databases. Humsavar knowledge base has been screened to identify proteins that have X-ray crystal structures of the wildtype and associated disease-causing mutant available on the protein databank (PDB) (Berman et al., 2000; Berman et al., 2002). Of the 2,943 proteins reported on the knowledge base having disease causing variants, 1,316 of them have at least one crystal structure available. In the protein structural networks involved in our analysis we are looking into the geometry at local sites which are closer than 4.5 Å while constructing all atom networks (Yao et al., 2019). Hence, in our data set for analysis we have applied a resolution cut-off criterion of 3Å. Additional condition of a difference in refinement factors (R_{free}–R_{work}) of no more than 5% was also used. Protein structures available in the free form, without a bound ligand are chosen by screening them using the BioLip database (Yang et al., 2013). Disease cases are identified by pairwise alignment of the sequences obtained from uniport and PDB entries to obtain unique chains of disease-causing mutant and wildtype structures having the best resolution. 74 cases with crystal structures of the wildtype and corresponding disease-causing mutant are found. Details of these protein structure pairs are listed in **Supplementary Table S1**.

All Atom - Protein Structure Network Model

The Protein Structural Network (PSN) models residues as nodes and constructs edges between nodes that satisfy the proximity criteria. Atoms from a pair of non-adjacent residues that fall within a distance cut-off of 4.5 Å are considered to make atom contact and therefore form an edge between the corresponding residues in the PSN (Brinda and Vishveshwara, 2005). The network model is an all-atom based, weighted and non-directed graph where the edge weight is given by:

$$Edgeweight\left(I_{ij}\right) = \frac{number\ of\ atom\ contacts\ between\ the\ residues\ i,\ j}{Highest\ number\ of\ atom\ contacts\ between\ the\ amino\ acids\ i, j}$$

The highest number of atom contacts between any pair amino acids is generated from analysing all the structures in the dataset of high-resolution crystal structures. In this paper, the Cα-atom position is used to represent the position of a node corresponding to a residue and edges are represented using lines. A hub is a node in the network that is well connected to several other nodes (Cohen and Barabási, 2002). We identify the minimum number of edges necessary to define at least one hub in all the structures of the disease cases and hence defined any node in the PSN having equal to or greater than 11 edges as a hub. We represent the hubs using spheres.

Network Dissimilarity Score

The network dissimilarity score (NDS) iis used to compare two networks with identical number of nodes to generate a difference score that quantifies the dissimilarity in their spectra and the weight of edges (Gadiyaram et al., 2017; Ghosh et al., 2017). The adjacency matrix is a representation of a network which is generated as described in the *All Atom - Protein Structure Network Model*. Let us say we are comparing the networks of a proteins A and B. The adjacency matrices of PSN A and PSN B are compared to generate the edge difference score (EDS).

$$EDS = \frac{||A - B||_F}{\sqrt{\left(\sum edge\ weight_A \ \times \sum edge\ weight_B\right)}}$$

The edge difference score captures the difference in edge weights between corresponding edges of the networks. A Laplacian of the adjacency matrix is derived before their spectra (eigen values and eigen vectors) are generated. The spectral information is used in computing the correspondence score (CRS) and eigen value weighted cosine scores (EWCS).

$$CRS = 1 - \frac{6\sum\ (Index\ Evec_A - Index\ Evec_B)^2}{n\,(n^2 - 1)}$$

Where, n is number of nodes in the PSN. The index difference of eigen vectors, once arranged in ascending order of their eigen values, is used in the numerator.

$$EWCS = \frac{\sum\left(1 - cosine\left(\theta_{ij}\right)\right)^2 |1 - Eval_A||1 - \ Eval_B|}{\sum|1 - \ Eval_A||1 - Eval_B|}$$

where, $Eval_A$ and $Eval_B$ are eigen values of PSN A and PSN B. The cosine between a pair of nodes is generated using the ratio between the dot product of their eigenvectors and the product of their magnitudes. The spectral comparison scores capture the local and global clustering of the nodes in the network. The components are formulated in computing the NDS:

$$NDS = \sqrt{EDS^2 + EWCS^2 + (1 - CRS)^2}$$

An in-house python program is used to calculate the NDS in any pair of networks.

The NDS between the PSNs of the wildtype and mutant chain is generated.

NDS ranges from 0 (indicating absolute congruency/identical networks) to a score of √3 (indicating absolute dissimilarity to the extent of no match between the networks). TM-align tool is employed to generate structure based sequence alignment and structural difference information (Zhang and Skolnick 2005).

Evaluating the Effect of Allostery

In order to study the effect of a perturbation such a disease-causing mutation on the structure of protein, the AlloSigMA server is employed. The server implements a structure-based statistical mechanical model of allostery, abbreviated SBSMMA (Guarnera and Berezovsky, 2016), to quantify the allosteric response that is communicated due to the effect of a perturbation like a molecular binding event or a mutation. The wildtype crystal structure of the protein being analysed is submitted as input to the server and an UP mutation perturbation is introduced. In this case, An UP mutation simulates the effect of mutation to a bulkier residue at the site of the disease-causing mutation. Crystal structures that had missing residues were completed using SWISS-MODEL (Guex and Peitsch, 1997). The AlloSigMA server results in an output of the response free energy of each residue that is accountable for the allosteric signal initiated by the mutation.

RESULTS

The perturbation in the structure of a protein due to disease causing mutations can be studied extensively using their native structural topologies (Ambrus et al., 2015; Ambrus et al., 2016; Szabo et al., 2018). It is understood that the resulting structural change manoeuvres the function or functional levels of the protein that is related to the onset of a disease. Here we study such variations in terms of structural networks of wildtype and disease related mutant. For the analysis, we identified proteins with disease-causing mutational variants from the Humsavar database and their corresponding wildtype and mutant crystal structures from PDB. We identified crystal structure variants corresponding to 74 disease cases and used those structures solved with the best resolution. The effect of mutations on their structure and network is analysed.

Analysis of Protein Structural Network

Protein structure networks are a node-edge representation of the protein structure that efficiently displays the connectivity between different elements of their tertiary structure. Several studies in the past have made use of protein structure networks in studying the connectivity between residues based on features such as their spatial proximity and energy of interaction. We have used an all-atom network model to generate structural network information at the residue level with edges made between residues that are spatially proximal. Two residues are linked with an edge if a pair of their atoms is situated within a distance of 4.5 Å. The strength of the edge depends on the number of such atom pairs between the residues that are forming an edge. We have discussed the criteria for defining an edge in the *Methods* section. We generated the all-atom protein structural networks for all the individual chains of the wildtype and mutant protein structures in our dataset.

The alteration of the connectivity that arises as a result of mutation is studied by comparing the PSNs of the wildtype and the corresponding mutant. The variation in their connectivity is observed by segregating the edges into those that are retained and those that are unique to wildtype or mutant structures (**Supplementary Figure S1**). This means that the edges found to be unique to the wildtype structure are lost in the mutant. Similarly, those edges that are unique to the mutant structure are considered to be gained. The information of edges lost and gained in the wildtype PSN and mutant PSN is presented in **Supplementary Figure S2A**. Every wildtype and mutant structure in the dataset have at least one edge that is unique to it. Of the disease cases that are studied in the dataset, in 28 cases the wildtype has more unique edges than the mutant and in 45 disease cases the mutant has more unique edges. This suggests that in a majority of the disease cases more edges are gained than those that are lost. Only in the case of the cAMP-dependent protein kinase α catalytic subunit that is responsible for primary pigmented nodular adrenocortical disease (by mutation L206R) it is found that the number of edges lost in the wildtype is equal to the number of edges that are gained in the mutant. The wildtype and mutant in this disease case have 1,264 edges, 1,218 of these are retained while the remaining are lost and gained.

The information stored in the protein structure networks are predominantly in their edges and their connectivity. In order to study how well each element of the PSN is connected, we employed the use of a few basic network parameters such as the degree and strength of the nodes in the network. The number of edges that connect to a node constitutes its degree and the sum of all the edge weights connecting to a node spans the strength of each node. It is possible for a node to not form an edge with any other node; such a node is isolated in the network. Alternatively, a node can be well connected with other nodes of the network and form hubs. Hubs are elements in the network that are generally crucial since they are well connected to many other nodes. Perturbations in these nodes can have a more significant effect on the network than those nodes that are not hubs. Nodes from the PSNs in the dataset are found to have a maximum degree ranging from 11 to 18 as shown in the **Supplementary Figure S3**, hence for this analysis we have chosen to consider any node with a degree 11 or higher as a hub node, this ensures that each structure in our dataset is composed of at least one hub.

We observe variability in the number of hubs between the wildtype and mutant crystal structures (**Supplementary Figure S4**). Hubs that are retained in between the conformers are an indication of preserved local networks and retained structure around them. Hubs that are unique to the wildtype and mutant are also identified. Those hubs that are specific to the wildtype structure are lost in the mutant structure and the hubs unique to the mutant are gained. In 37 disease cases the number of hubs lost in the wildtype is greater than the number of hubs that are gained in the mutant and in 28 disease cases the number of hubs gained in the mutant are greater than those lost in the wildtype. In nine other disease cases the number of hubs unique to the wildtype and mutant are equal. There is no loss or gain of hubs in three disease cases. The highest number of hubs lost in wildtype structures is 32 and the highest number of hubs gained in the mutants is 23. The number of hubs unique to the wildtype structure and the number of hubs unique to the mutant are shown as a scatter in the **Supplementary Figure S2B**. The distribution of the number of hubs in the structures of our dataset can be found in the **Supplementary Figure S4, S5**. The functional relevance of the change in number of hubs has been discussed in detail for specific cases in a later section.

Local Site Variation of Structural and Network Parameters

Change in degree of a residue between wildtype and the mutant suggests loss or gain of edges. The strength of an edge (edge weight) that connects two nodes may also change in the mutant. It is expected that a node corresponding to a residue which is buried in the protein structure has high degree and strength since they are in the proximity of several other nodes of the network. We have analysed the variation of network and structure parameters across the topologically equivalent residues and nodes. Since the focus of this work is on the mutation site that brings about the perturbation in the network and structure of the protein that may affect the functional sites, we have focused on studying the variability at these local sites in detail.

The change in degree and strength at the site of mutation reflects the change in local network at the site of perturbation. The change in sidechain atoms of the residue at the site of mutation plays a significant role in its degree that may or may not change in the PSN. For example, the highest gain in degree is in the case of apoptosis inducing factor where a glycine is mutated to a glutamate residue and the degree increases by 5. Likewise, when a phenylalanine is mutated to a serine in the case of Lysine-specific histone demethylase the degree at the site of mutation reduces by 7. The information of the change in degree and solvent accessibility at the site of mutation is shown in **Supplementary Figure S5**. In the dataset we find that at 11 mutation sites the mutated residue undergoes change in solvent accessibility. It is more common to see the mutation site buried in the wildtype whereas in the mutant state they are exposed since at 9 of the 11 sites we observe a buried residue get exposed in the mutant.

Using the information of active site and binding sites available in the Uniport database we identified 151 functional sites in the dataset and analysed the change in network parameters at these sites. The information of the change in degree at the functional site is shown in **Supplementary Figure S6**. No change in degree is observed at majority of the functional sites. The variation of degree at the functional site (ranges from loss of four edges to gain of four edges) is lower as compared to the variation of degree at the mutation sites (ranges from loss of seven edges to gain of five edges). In the dataset, only in the case of Septin-12 protein it is found that a mutation occurs at a site of function, where a threonine that is known to bind to GTP (Castro et al., 2020) is mutated to methionine (T89M) and the degree at the site changes from six in the wildtype to two in the mutant.

Global Structural and Network Variation in the Crystal Conformers

The overall variability in the crystal structures when the protein undergoes a disease-causing mutation has been studied by comparing their structures and networks separately. The structural difference between the conformers is calculated using the root mean square deviation (RMSD) that measures the divergence in the backbone topologies. In order to quantify the variation in the protein structure networks (PSN), a spectral comparison tool that is referred to as the NDS (network dissimilarity score) is used. The spectral comparison method quantifies the extent of dissimilarity between two networks with identical number of nodes. Only those residues that are topologically equivalent are identified by structural alignment and used for the comparison. All the structural and network comparison scores between the wildtype and mutant crystal structures in the dataset is generated using information of their coordinates. **Figure 1** shows the scatter plot between Cα-atom RMSD and all-atom NDS.

The scatter of comparison scores suggests that the variation in the network is not strongly correlated to the variation of their

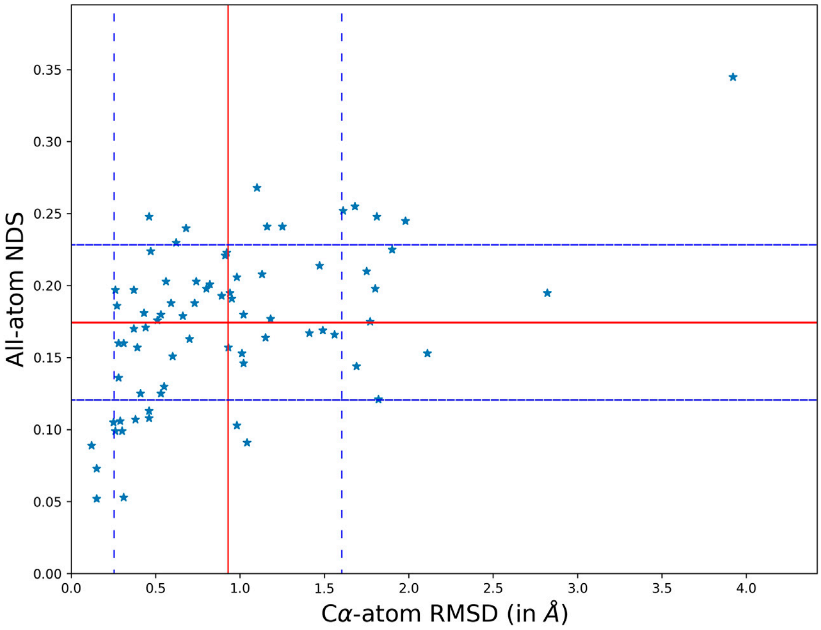


FIGURE 1 | A scatter plot comparing the structural topology (Cα positions) and PSN of the wildtype and mutant using RMSD and NDS respectively. The comparison scores for each disease case are plot on the scatter. It is found that the structural divergence and network dissimilarity do not share strong linear relationship.

structural topologies. The mean and standard deviation in the scores is plot on the scatter using red and blue (dotted) lines respectively. The mean NDS of the disease cases is 0.175 and the mean RMSD is 0.92 Å. A dataset of all pairs of available wildtype structures is used as a control in analysing the significance of the observed variability. In the control dataset the mean NDS is 0.12 and the mean RMSD is 0.57 Å which is relatively lesser than the variability in the disease cases (**Supplementary Figure S7**). It should be noted that RMSD and NDS plotted correspond to Cα positions and all atoms (including sidechains) respectively. Near absence of correlation in **Figure 1** also conveys the message that there are examples with Cα positions well retained between wildtype and the mutant while the sidechain orientations are altered. There are also cases where the sidechain connectivity in networks are highly similar between wildtype and the mutant, but Cα trajectory has undergone a significant change.

Specific Cases of Network and Structure Variability

In the global analysis of protein structure and network variability, we find several cases where the structural topology (Cα positions) is preserved but the all-atom network have changed considerably and the vice versa. In the first type of cases, the network variability is high, NDS is greater than the mean and standard deviation, even though the structures are well superimposed with lower than mean RMSD. In the second type of cases, the networks are not strongly dissimilar i.e. NDS lower than the mean of the dataset, but the structural difference suggests that they might not be as well preserved as their networks with RMSD greater than the mean and standard deviation of the dataset. Three disease cases from the dataset that fall into each of these categories are studied in detail.

Network Variable Cases

Disease Mutation in Medium-Chain Specific Acyl-CoA Dehydrogenase (MCAD) Alters Local Network at the Functional Site

The MCAD mitochondrial protein is known to catalyse the first step of fatty acid beta oxidation in humans. The functional protein is a homo-tetrameric complex with subunits bound to FAD molecules (Lee et al., 1996). The coding gene undergoes a single nucleotide polymorphism (A985G) that results in the protein mutant (K304E) which leads to the disease state (Gregersen et al., 1993). The protein undergoes a significant variation in the all-atom network (NDS 0.248), however the Cα RMSD is quite low (0.46Å). 67 edges and five hubs are lost in the wildtype PSN whereas 83 unique edges and 17 hubs are gained in the mutant (**Supplementary Figure S8**). It is observed that mutational site is far away from the site of function (S142, N191, G377, and R388). The site of function in the protein is shown in **Figure 2A**, the corresponding nodes and their edges in the wildtype PSN and mutant PSN are shown in **Figures 2B,C** respectively. Due to the rearrangement of edges at the nodes corresponding to functional site residues as shown in **Figure 2**, there is change in the local network at the functional site. It is reported that the mutation (K304E) leads to a deficiency of the protein that can result in death at infancy.

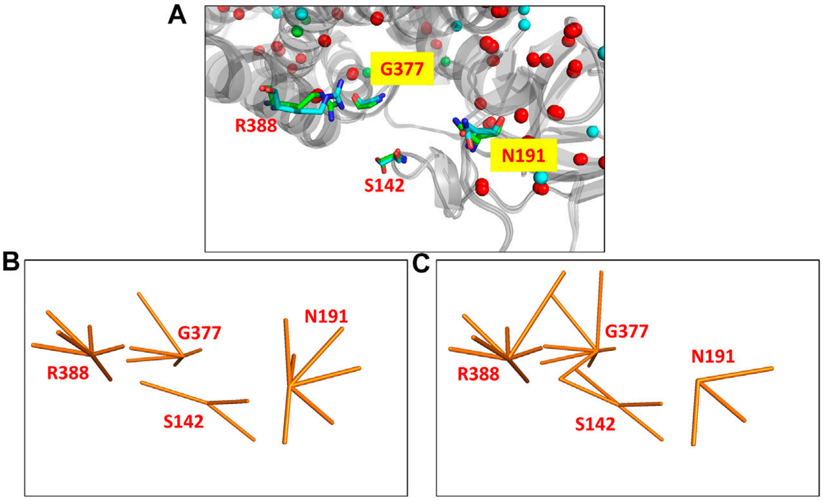


FIGURE 2 | The functional site in the crystal structures of the wildtype (PDB ID: 1EGE) and mutant (PDB ID: 4P13) of the MCAD protein. **(A)** The functional site of the protein consists of four residues (S142, N191, G377, and R388) that are shown (using stick representation) in the superposed structures. The edges corresponding to these residues in the networks are shown in **(B)** the wildtype PSN and **(C)** the mutant PSN (using orange line representation). While N191 looses three edges, S142, G377, and R388 gain 1, 2, and 1 edges respectively.

Porphobilinogen Deaminase Undergoes Disease Mutation That Leads to Loss of Essential Edges and has Reduced Thermostability

Porphobilinogen deaminase is a transferase that catalyses the synthesis of hydroxymethylbilane which is a precursor for heme and porphyrin biosynthesis. The disease mutant has defects of heme biosynthesis, which is mainly due to the enhanced excretion of porphyrins and porphyrin precursors. It is reported that the hydrogen bonding network in the ordered regions of the protein allows for the protein to display higher thermostability (Bustad et al., 2013). It has also been reported that the mutant crystal structure is less thermo stable and has lost its function and hence may be the leading cause for Acute intermittent porphyria (Gill et al., 2009). Although a significant number of edges and hubs are found to be preserved in the PSN, it is observed that 68 edges and 11 hubs that are unique to the wildtype is lost and 46 edges and three hubs unique to the mutant is gained (**Supplementary Figure S9**). Since there is loss of edges around the ordered secondary structures in the wildtype the important network necessary for thermostability is lost.

The Network Around the Functional Site in the Disease Mutant of Glutamine--tRNA Ligase Is Altered

The glutamine tRNA ligase is essential for the biosynthesis of glutamine in humans. The function of this protein is crucial for brain development in infants (Zhang et al., 2014; Ognjenović et al., 2016). The wildtype and mutant structures of the protein are well superposable (RMSD 0.68 Å) although their PSNs are quite dissimilar (NDS 0.24). The mutant node is far from the functional site where minimal variation of edges is observed. However, the significant loss of 176 edges and 32 hubs which are majorly found around the functional site in the wildtype PSN (**Supplementary Figure S10**) can be the cause for reduced aminoacylation activity reported in the mutant to cause microcephaly, progressive, with seizures and cerebral/cerebellar atrophy.

Cases with Backbone Structure Variation

The Mutant Structure of the Major Prion Protein Undergoes a Conformational Switch

The primary physiological function of the major prion protein is unclear. However, the functional state of the protein (**Figure 3A**) forms a well interacting dimer that is known to be involved in several different functions (Knaus et al., 2001). In the disease mutant state (**Figure 3B**), a conformational transition is observed in the C-terminal helix (Non-aligned helix shown in **Figure 3**) that forms a dimer with fewer interaction between the dimeric chains (Lee et al., 2010). The conformational change alters the topology at several other regions of the protein resulting in a high structural difference (RMSD 2.11 Å). However, the network in the topologically equivalent regions of the protein is preserved (NDS 0.153). There is only one hub in the wildtype that is not altered in the mutant and very few edges are rearranged, 19 edges and 23 edges unique to the wildtype and mutant respectively (**Supplementary Figure S11**). The new mutant conformation is found to be associated with Creutzfeldt-Jakob disease where cases are reported of degeneration of neurons and amyloid plaque formation due to protein aggregation.

Calmodulin-1 Mutant Acquires a Closed Conformation With Minimal Change in Network

Calmodulin is a membrane binding calcium transporter protein that transports metal ions across ion channels. A calcium ion binding sequence motif that occurs in pairs is conserved in the structures of this family of proteins (Tsang et al., 2006; Sarhan et al., 2012). There are two pairs of these binding site regions which are far apart in the open conformation of the wildtype structure. In the current case, when one of the calcium binding sites undergoes mutation (N98S), the functional state of the

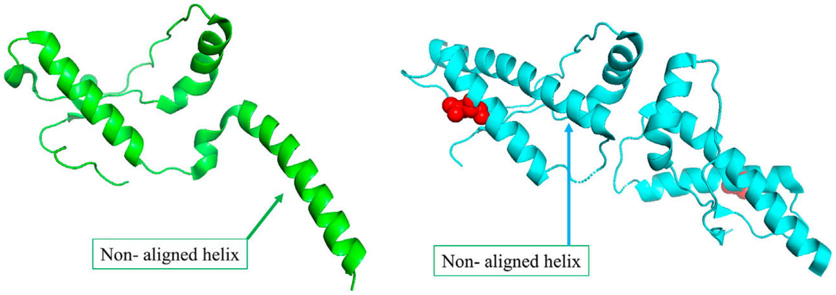


FIGURE 3 | (A) The wildtype conformer (PDB ID: 1I4M) is crystallised as a monomer in the asymmetric unit, although it exists as a dimer functionally. **(B)**The structure of the disease-causing mutant (PDB ID: 3HEQ) shows conformational change in the non-aligned helix. The mutant residue is shown in red spheres.

protein is lost (Wang et al., 2020). The mutant structure has a closed conformation which is reported not to bind to the metal ion at one of the calcium binding sites with the mutation. 21 edges in the wildtype and 18 edges in the mutant are lost and gained respectively. Seven hubs are retained and a single hub in the wildtype is lost in the mutant (**Supplementary Figure S12**). The overall network difference (NDS 0.121) is found to be minimal. However, due to the mutational site region that is found not to align well with the residues in the wildtype results in a large structural difference (RMSD 1.82 Å).

Structural Divergence in Wilms Tumour Protein

The Wilms tumour protein is a transcriptional factor consisting of a DNA binding domain which has four zinc finger repeats that determine sequence specific binding to DNA (Hamilton et al., 1995). While two of the zinc fingers bind to the DNA others are essential for recognising the cognate nucleotide base. One of these zinc fingers that is responsible for recognising the cognate nucleotide base undergoes a mutation (M342R) that enhances the affinity for a different nucleotide base leading to errors in transcription (Wang et al., 2018). The conformation of the wildtype does not superpose well with the mutant (RMSD 1.69 Å). In the PSN, 12 edges are lost in the wildtype and eight edges are gained in the mutant. One new hub is gained in the mutant along with the 1 hub that is retained between the wildtype and mutant PSN (**Supplementary Figure S13**). Hence, the network in the several regions of the protein is still preserved depicting low network dissimilarity (NDS 0.144).

Allosteric Effect due to Disease Causing Mutation

In specific cases where we observe network variation that is far from the site of mutation, we describe the possibility of observing an allosteric signal that repacks the residues resulting in the alteration of PSN. In order to corroborate the exhibition of allostery in these proteins AlloSIgMA (Tan et al., 2020) is employed to quantify the energetics compounding the allosteric effects of a mutation. Crystal structures of the wildtypes of three proteins in our dataset that undergo significant network change upon mutation were studied using AlloSIgMA and UP mutations (A perturbation that simulates the effect of mutation to a bulkier residue) at known disease-causing mutation sites are implemented. The output generated is illustrated in **Figure 4** and discussed in the following section.

DISCUSSION

The protein structure network is an efficient tool in analysing allostery in the protein structure (Süel et al., 2003; Di Paola and Giuliani, 2015). In our study, we have analysed the variation of PSN brought about by disease causing mutations to the native functional protein. We have observed the variability in edges and hubs that are important parameters that make the protein structural network. We have identified edges and hubs that are unique to the wildtype structure that are lost in the mutant where new edges and hubs unique to the mutant structure are gained. The use of such information can be discussed with the help of an example.

The human serum albumin which is found abundantly in blood plasma is known to transport several different molecules including thyroxine (Robbins et al., 1978). In the dataset of disease cases, it is found that the mutant structure of albumin protein undergoes the largest variation in the number of edges and hubs. 294 edges and 25 hubs are lost in the wildtype and 305 edges and 20 hubs are gained in the mutant (**Supplementary Figure S14**). At the site of mutation (R218P) an edge with the residue L238 that is also a hub is found to be lost in the mutant (**Figure 5**). The loss of the edge is indicative of decrease in proximity between the residues suggesting that the thyroxine molecule that binds to K240, hormone binding site (Jacobsen, 1978), can be better accommodated in the mutant. It is reported that the mutation enhances the binding affinity of the protein to thyroxine that causes the elevated serum thyroxine levels associated with familial dysalbuminemic hyperthyroxinemia (FDH) (Petitpas et al., 2003).

We have analysed the variability in the disease cases by comparing their network and structure using the network dissimilarity score and RMSD. A control dataset is employed where the wildtype is compared to all other wildtype structures of the protein that satisfy the criteria for the dataset. The variability in disease cases (mean RMSD 0.92 Å and mean NDS 0.175) is much greater than in the variability in case of only wildtype

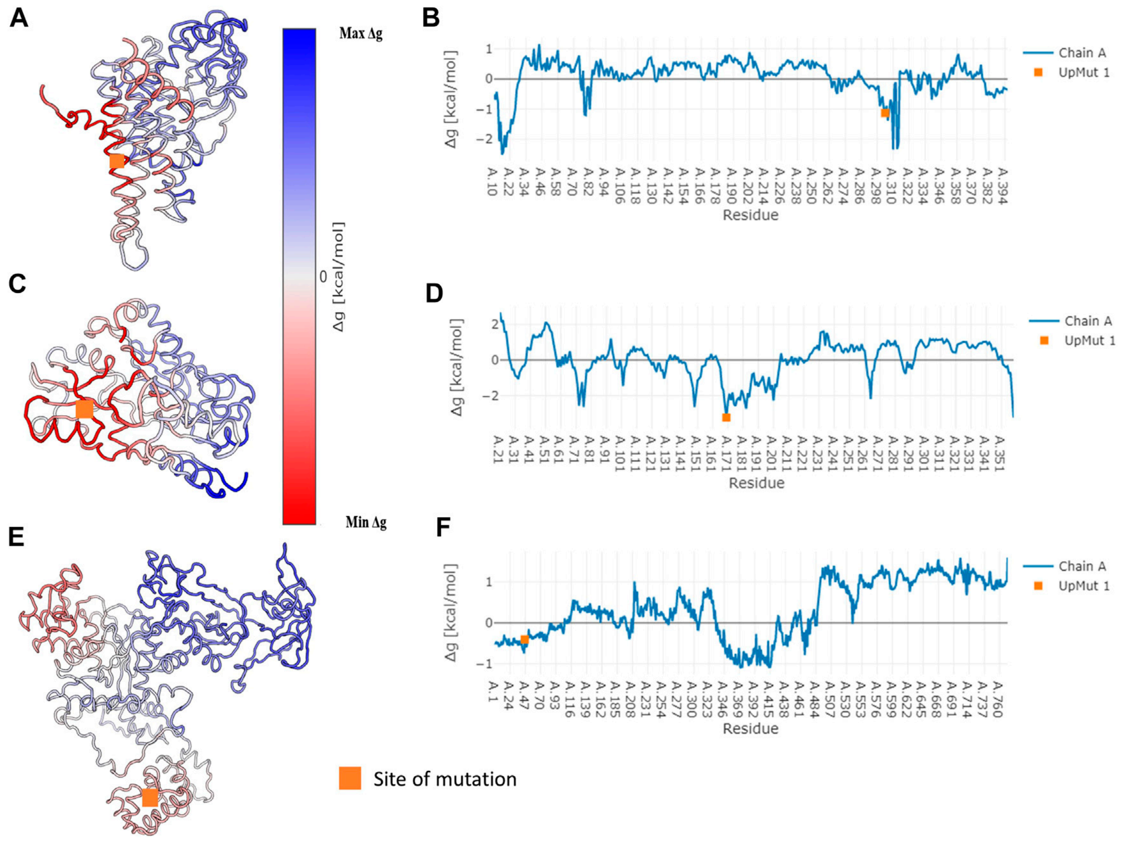


FIGURE 4 | Free energy values obtained for three specific proteins that undergo disease-causing mutation. Specific cases where we observe significant network variability have been subject to the analysis of allosteric effects due to mutation. The AlloSigMA server employs the SBSMMA (Guarnera and Berezovsky, 2016) method to generate the response free energies when perturbations (UP mutation) are introduced at known sites of disease-causing mutations. Cartoon of the wildtype coloured according to their free energy values obtained for the cases of **(A)** Medium-chain specific acyl-CoA dehydrogenase, **(C)** Porphobilinogen deaminase and **(E)** Glutamine--tRNA ligase are shown on the left. Their free energy profiles are illustrated graphically with residue index on the *x*-axis and Δg value on the *y*-axis in **(B)**, **(D)** and **(F)** shown on the right in the same order. The orange square points to the site of mutation.

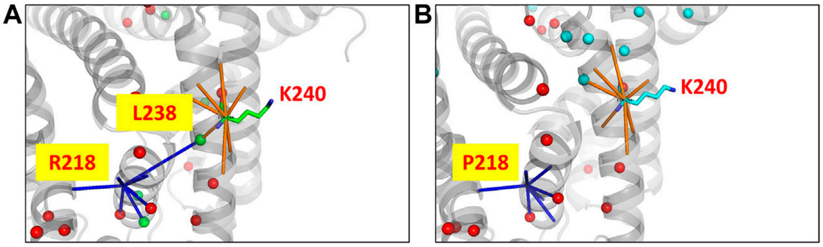


FIGURE 5 | The PSN of human serum albumin protein at the site of mutation and function in the wildtype (PDB ID: 1N5U) and mutant (PDB ID: 1HK3) is shown. The node corresponding to the mutation site makes an edge with a hub node L238 (green sphere) in **(A)** the wildtype PSN which is lost in the case of **(B)** the mutant PSN. It is observed that hubs near to the binding site (K240) are lost, which is indicative of the increase in proximity between the nodes. It has been reported that the mutant structure is able to better accommodate a substrate with greater binding affinity which leads to the FDH disease condintion. Hubs unique to the wildtype and mutant are show in green and cyan sphere representation respectively, those hubs that are retained are shown in red.

structures (mean RMSD 0.57 Å and mean NDS 0.12) which signifies that the mutant structure and network explore diverse conformations with different interconnectivity of residues. The variability observed in protein structural networks is not strongly correlated to the topological structure difference that is used in the traditional analysis of protein structures. It is found that in a few cases, the network variability is relatively higher than the amount of structural difference. The vice versa is also true, where the structural difference is quite large but their networks seem to be well preserved. Such cases have been specifically picked for a detailed analysis of their global and local changes. We have also attempted to provide the functional relevance of the observed variability.

In the disease cases where the site of mutation is not involved with function, allosteric changes brought about in the connectivity of the internal network of the protein seem to affect the function which leads to a disease state. Where the contribution of the mutation may be as minimal as no change in the local network at the site of mutation, a large network alteration can be observed far away from the site of perturbation due to the disturbance in the network of edges connecting each element in the PSN to the other as discussed in the example of glutamine tRNA ligase. A significant improvement in the number of edges and hubs attributing to an improved network stabilises the MCAD protein although the distant mutation site alters the network at the functional site and hence the protein loses its function. Contrarily, a reduction in the number of edges and hubs in the case of the porphobilinogen deaminase protein is attributed to reduced thermostability due to loss of essential edges in the network within the protein. Conformational transition from one state to the other brings structural changes and loss of function in the case of major prion protein. However, their networks are found to be preserved since the aligned regions have retained edges and hubs that are very small in number. Likewise, it is found that there may not be a significant network variation but the structure varies considerably adding to the change in interaction with other molecule due to the mutation that eventually contributes to the alteration of function as observed in the case of Wilms tumour protein.

So as to substantiate the exhibition of allostery due to the mutations, theoretical free energy is computed using the AlloSigMA. The predicted free energy obtained for the specific cases of network variability when an UP mutation (mimicking substitution with a bulkier residue) is implemented at the site of disease-causing mutation are shown in **Figure 4**. A free energy value of zero suggests that the residue may not respond to the perturbation (mutation) whereas a non-zero value suggests that the residue may respond with more or less effect due to the perturbation. In the specific cases with large network variability, it is found that the disease-causing mutations stabilise (negative free energy) the residues around them and communicates the allosteric signal that destabilises (positive free energy) residues elsewhere within the structure. This suggests that the significant change in protein structural network that is observed due to the mutation at a site known to cause a disease is also due to the allosteric mechanism that arises from perturbation of the given site.

In Summary, our work highlights the perturbation of protein structural network as understood from the variability between a wildtype structure and the structure of a disease-causing mutant. Network features such as edges and hubs help to analyse the overall variation of networks while parameters such as degree of each node help to analyse their local network variability. The allostery due to a disease-causing mutation is noticeable from the loss and gain of network elements that result in variation of protein structural networks that is also corroborated using theoretical free energy calculations. We find cases where the network change is confined to the local site of mutation or far away from the site of mutation. We have also noted cases where repacking of sidechains occurs upon mutation and cases where the backbone conformation is altered with preserved sidechain network. From our work, the effect of mutation on the structural network of the wildtype may be used as a learning to extend to the next phase of the project to explore its predictive power of mutant structures and allosteric effects. The major challenge in the future is to translate the learning from the current work to predict the structure of the mutant which is a prerequisite to predict the effect of mutation on the stability and function. Availability of accurate structures of wildtype and reliably modelled mutant structures may be used in the context of thermodynamic cycle towards calculation of free energy difference between the wildtype and the mutant as for example used by Topham et al., (Topham, Srinivasan and Blundell, 1997). The protein structural network approach is an effective tool to understand the structural effects of disease-causing mutation, further we also suggest that the protein structural network approach is a convenient approach to understand the allostery caused by other kinds of structural perturbations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

VP conducted most of the data analysis and wrote the first draft of the manuscript. NN helped in making the dataset and in additional analysis. NS conceived the idea and mentored the project.

FUNDING

Research in our group is generally supported by the FIST grant sponsored by the Department of Science and Technology (DST) and the Department of Biotechnology (DBT), Government of India in the form of IISc-DBT partnership programme. Support from UGC, India – Centre for Advanced Studies and Ministry of Human Resource Development, India are gratefully acknowledged. NS is a J. C. Bose National Fellow. This research is also supported by Bioinformatics and Computational biology centre sponsored by Department of Biotechnology (DBT), Government of India.

ACKNOWLEDGMENTS

We thank Ms. Sohini Chakraborti for providing technical help.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.620554/full#supplementary-material.

REFERENCES

Ambrus, A., Mizsei, R., and Adam-Vizi, V. (2015). Structural alterations by five disease-causing mutations in the low-pH conformation of human dihydrolipoamide dehydrogenase (hLADH) analyzed by molecular dynamics - implications in functional loss and modulation of reactive oxygen species generation by pathogenic hLADH forms. *Biochem. Biophys. Rep.* 2, 50–56. doi:10.1016/j.bbrep.2015.04.006

Ambrus, A., Wang, J., Mizsei, R., Zambo, Z., Torocsik, B., Jordan, F., et al. (2016). Structural alterations induced by ten disease-causing mutations of human dihydrolipoamide dehydrogenase analyzed by hydrogen/deuterium-exchange mass spectrometry: implications for the structural basis of E3 deficiency. *Biochim. Biophys. Acta*. 1862, 2098–2109. doi:10.1016/j.bbadis.2016.08.013

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*. 181, 223–230. doi:10.1126/science.181.4096.223

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., and Kang, H. M. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Bustad, H. J., Vorland, M., Rønneseth, E., Sandberg, S., Martinez, A., and Toska, K. (2013). Conformational stability and activity analysis of two hydroxymethylbilane synthase mutants, K132N and V215E, with different phenotypic association with acute intermittent porphyria. *Biosci. Rep.* 33, e00056. doi:10.1042/BSR20130045

Bateman, A. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.* 58, 899–907. doi:10.1107/S0907444902003451

Brinda, K. V., and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophys. J.* 89, 4159–4170. doi:10.1529/biophysj.105.064485

Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21, 323–325. doi:10.1038/6851

Castro, D. D. V., da Silva, S. M. D. O., Pereira, H. M., Macedo, J. N. A., Leonardo, D. A., Valadares, N. F., et al. (2020). A complete compendium of crystal structures for the human SEPT3 subgroup reveals functional plasticity at a specific septin interface. *IUCrJ*. 7, 462–479. doi:10.1107/S2052252520002973

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238. doi:10.1038/10290

Cohen, E. A., and Barabási, A.-L. (2002). *Linked: the new science of networks*. New York, NY: Perseus Books Group. doi:10.2307/20033300

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1999). Erratum: a DNA polymorphism discovery resource for research on human genetic variation (Genome Research (1998) 8 (1229-1231)). *Genome Res.* 9, 210.

Di Paola, L., and Giuliani, A. (2015). Protein contact network topology: a natural language for allostery. *Curr. Opin. Struct. Biol.* 31, 43–48. doi:10.1016/j.sbi.2015.03.001

Dubay, K. H., Boman, G. R., and Geissler, P. L. (2015). Fluctuations within folded proteins: implications for thermodynamic and allosteric regulation. *Acc. Chem. Res.* 48, 1098–1105. doi:10.1021/ar500351b

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783. doi:10.1093/nar/gkw1121

Gadiyaram, V., Ghosh, S., and Vishveshwara, S. (2017). A graph spectral-based scoring scheme for network comparison. *J. Complex Networks* 5, 219–244. doi:10.1093/comnet/cnw016

Ghosh, S., Gadiyaram, V., and Vishveshwara, S. (2017). Validation of protein structure models using network similarity score. *Proteins* 85, 1759–1776. doi:10.1002/prot.25332

Gill, R., Kolstoe, S. E., Mohammed, F., Al D-Bass, A., Mosely, J. E., Sarwar, M., et al. (2009). Structure of human porphobilinogen deaminase at 2.8 A: the molecular basis of acute intermittent porphyria. *Biochem. J.* 420, 17–25. doi:10.1042/BJ20082077

Gregersen, N., Winter, V., Curtis, D., Deufel, T., Mack, M., Hendrickx, J., et al. (1993). Medium-chain Acyl-CoA dehydrogenase (MCAD) Deficiency: the prevalent mutation G985 (K304E) is subject to a strong founder effect from northwestern Europe. *Hum. Hered.* 43, 342–350. doi:10.1159/000154157

Guarnera, E., and Berezovsky, I. N. (2019a). On the perturbation nature of allostery: sites, mutations, and signal modulation. *Curr. Opin. Struct. Biol.* 56, 18–27. doi:10.1016/j.sbi.2018.10.008

Guarnera, E., Tan, Z. W., Zheng, Z., and Berezovsky, I. N. (2017). AlloSigMA: allosteric signaling and mutation analysis server. *Bioinformatics* 33 (24), 3996–3998. doi:10.1093/bioinformatics/btx430

Guarnera, E., and Berezovsky, I. N. (2016). Structure-based statistical mechanical model accounts for the causality and energetics of allosteric communication. *PLoS Comput. Biol.* 12 (3) e1004678. doi:10.1371/journal.pcbi.1004678

Guarnera, E., and Berezovsky, I. N. (2019b). Toward comprehensive allosteric control over protein activity. *Structure* 27, 866–878.e1. doi:10.1016/j.str.2019.01.014

Guarnera, E., and Berezovsky, I. N. (2020). Allosteric drugs and mutations: chances, challenges, and necessity. *Curr. Opin. Struct. Biol.* 62, 149–157. doi:10.1016/j.sbi.2020.01.010

Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins?. *Proteins* 57, 433–443. doi:10.1002/prot.20232

Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18 (15), 2714–2723. doi:10.1002/elps.1150181505

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., et al. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247. doi:10.1038/10297

Hamilton, T. B., Barilla, K. C., and Romaniuk, P. J. (1995). High affinity binding sites for the Wilms' tumour suppressor protein WT1. *Nucleic Acids Res.* 23, 277–284. doi:10.1093/nar/23.2.277

Jacobsen, C. (1978). Lysine residue 240 of human serum albumin is involved in high-affinity binding of bilirubin. *Biochem. J.* 171, 453–459. doi:10.1042/bj1710453

Knaus, K. J., Morillas, M., Swietnicki, W., Malone, M., Surewicz, W. K., and Yee, V. C. (2001). Crystal structure of the human prion protein reveals a mechanism for oligomerization. *Nat. Struct. Biol.* 8, 770–774. doi:10.1038/nsb0901-770

Lee, H. J. K., Wang, M., Paschke, R., Nandy, A., Ghisla, S., and Kim, J. J. P. (1996). Crystal structures of the wild type and the Glu376Gly/Thr255Glu mutant of human medium-chain acyl-CoA dehydrogenase: influence of the location of the catalytic base on substrate specificity. *Biochemistry* 35, 12412–12420. doi:10.1021/bi9607867

Lorch, M., Mason, J. M., Clarke, A. R., and Parker, M. J. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. *Biochemistry* 38, 1377–1385. doi:10.1021/bi9817820

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi:10.1093/nar/gkx1153

Lee, S., Antony, L., Hartmann, R., Knaus, K. J., Surewicz, K., Surewicz, W. K., et al. (2010). Conformational diversity in prion protein variants influences intermolecular beta-sheet formation. *EMBO J.* 29, 251–262. doi:10.1038/emboj.2009.333

Lorch, M., Mason, J. M., Sessions, R. B., and Clarke, A. R. (2000). Effects of mutations on the thermodynamics of a protein folding reaction: implications for the mechanism of formation of the intermediate and transition states. *Biochemistry* 39, 3480–3485. doi:10.1021/bi9923510

Mitternacht, S., and Berezovsky, I. N. (2011). Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput. Biol.* 7, e1002148. doi:10.1371/journal.pcbi.1002148

Mottaz, A., David, F. P., Veuthey, A. L., and Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 26, 851–852. doi:10.1093/bioinformatics/btq028

Naganathan, A. N. (2019). Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function. *Curr. Opin. Struct. Biol.* 54, 1–9. doi:10.1016/j.sbi.2018.09.004

Ognjenović, J., Wu, J., Matthies, D., Baxa, U., Subramaniam, S., Ling, J., et al. (2016). The crystal structure of human GlnRS provides basis for the development of neurological disorders. *Nucleic Acids Res.* 44, 3420–3431. doi:10.1093/nar/gkw082

Petitpas, I., Petersen, C. E., Ha, C. E., Bhattacharya, A. A., Zunszain, P. A., Ghuman, J., et al. (2003). Structural basis of albumin-thyroxine interactions and familial dysalbuminemic hyperthyroxinemia. *Proc. Natl. Acad. Sci. USA* 100, 6440–6445. doi:10.1073/pnas.1137188100

Robbins, J., Cheng, S. Y., and Gershengorn, M. C. (1978). Thyroxine transport proteins of plasma. Molecular properties and biosynthesis. *Recent Prog. Horm. Res.* 34, 477–519. doi:10.1016/b978-0-12-571134-0.50017-x

Rajasekaran, N., Suresh, S., Gopi, S., Raman, K., and Naganathan, A. N. (2017). A general mechanism for the propagation of mutational effects in proteins. *Biochemistry* 56, 294–305. doi:10.1021/acs.biochem.6b00798

Rignall, T. R., Baker, J. O., McCarter, S. L., Adney, W. S., Vinzant, T. B., Decker, S. R., et al. (2002). Effect of single active-site cleft mutation on product specificity in a thermostable bacterial cellulase. *Appl. Biochem. Biotechnol.* 98-100, 383–394. doi:10.1385/ABAB:98-100:1-9:383

Sarhan, M. F., Tung, C. C., Van Petegem, F., and Ahern, C. A. (2012). Crystallographic basis for calcium regulation of sodium channels. *Proc. Natl. Acad. Sci. USA* 109, 3558–3563. doi:10.1073/pnas.1114748109

Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10, 59–69. doi:10.1038/nsb881

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). DbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi:10.1093/nar/29.1.308

Szabo, E., Mizsei, R., Wilk, P., Zambo, Z., Torocsik, B., Weiss, M. S., et al. (2018). Crystal structures of the disease-causing D444V mutant and the relevant wild type human dihydrolipoamide dehydrogenase. *Free Radic. Biol. Med.* 124, 214–220. doi:10.1016/j.freeradbiomed.2018.06.008

Tee, W. V., Guarnera, E., and Berezovsky, I. N. (2019). On the allosteric effect of nsSNPs and the emerging importance of allosteric polymorphism. *J. Mol. Biol.* 431 (19), 3933–3942. doi:10.1016/j.jmb.2019.07.012

Tiede, S., Cantz, M., Spranger, J., and Braulke, T. (2006). Missense mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTA) in a patient with mucolipidosis II induces changes in the size and cellular distribution of GNPTG. *Hum. Mutat.* 27, 830–831. doi:10.1002/humu.9443

Tokuriki, N., and Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19, 596–604. doi:10.1016/j.sbi.2009.08.003

Topham, C. M., Srinivasan, N., and Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.* 10 (1), 7–21. doi:10.1093/protein/10.1.7

Tsang, W. Y., Spektor, A., Luciano, D. J., Indjeian, V. B., Chen, Z., Salisbury, J. L., et al. (2006). CP110 cooperates with two calcium-binding proteins to regulate cytokinesis and genome stability. *Mol. Biol. Cell.* 17, 3423–3434. doi:10.1091/mbc.E06-04-0371

Taverna, D. M., and Goldstein, R. A. (2002). Why are proteins so robust to site mutations?. *J. Mol. Biol.* 315, 479–484. doi:10.1006/jmbi.2001.5226

Tan, Z. W., Guarnera, E., Tee, W. V., and Berezovsky, I. N. (2020). AlloSigMA 2: paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. *Nucleic Acids Res.* 48 (W1), W116–W124. doi:10.1093/nar/gkaa338

Tan, Z. W., Tee, W.-V., Guarnera, E., Booth, L., and Berezovsky, I. N. (2019). AlloMAPS: allosteric mutation analysis and polymorphism of signaling database. *Nucleic Acids Res.* 47 (D1), D265–D270. doi:10.1093/nar/gky1028

Ung, M. U., Lu, B., and McCammon, J. A. (2006). E230Q mutation of the catalytic subunit of cAMP-dependent protein kinase affects local structure and the binding of peptide inhibitor. *Biopolymers* 81, 428–439. doi:10.1002/bip.20434

Vijayabaskar, M. S., and Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophys. J.* 99, 3704–3715. doi:10.1016/j.bpj.2010.08.079

Wang, D., Horton, J. R., Zheng, Y., Blumenthal, R. M., Zhang, X., and Cheng, X. (2018). Role for first zinc finger of WT1 in DNA sequence specificity: denys-Drash syndrome-associated WT1 mutant in ZF1 enhances affinity for a subset of WT1 binding sites. *Nucleic Acids Res.* 46, 3864–3877. doi:10.1093/nar/gkx1274

Wang, K., Brohus, M., Holt, C., Overgaard, M. T., Wimmer, R., and Van Petegem, F. (2020). Arrhythmia mutations in calmodulin can disrupt cooperativity of Ca2+ binding and cause misfolding. *J. Physiol.* 598, 1169–1186. doi:10.1113/JP279307

Weinkam, P., Chen, Y. C., Pons, J., and Sali, A. (2013). Impact of mutations on the allosteric conformational equilibrium. *J. Mol. Biol.* 425, 647–661. doi:10.1016/j.jmb.2012.11.041

Yang, G., Hong, N., Baier, F., Jackson, C. J., and Tokuriki, N. (2016). Conformational tinkering drives evolution of a promiscuous activity through indirect mutational effects. *Biochemistry* 55, 4583–45933. doi:10.1021/acs.biochem.6b00561

Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 41, D1096–D1103. doi:10.1093/nar/gks966

Yao, X. Q., Momin, M., and Hamelberg, D. (2019). Establishing a framework of using residue-residue interactions in protein difference network analysis. *J. Chem. Inf. Model.* 59, 3222–3228. doi:10.1021/acs.jcim.9b00320

Zhang, X., Ling, J., Barcia, G., Jing, L., Wu, J., Barry, B. J., et al. (2014). Mutations in QARS, encoding glutaminyl-trna synthetase, cause progressive microcephaly, cerebral-cerebellar atrophy, and intractable seizures. *Am. J. Hum. Genet.* 94, 547–558. doi:10.1016/j.ajhg.2014.03.003

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33, 2302–2309. doi:10.1093/nar/gki524

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 March 2021
doi: 10.3389/fmolb.2021.635425

Prediction of Function Determining and Buried Residues Through Analysis of Saturation Mutagenesis Datasets

Munmun Bhasin[1] *and Raghavan Varadarajan*[1,2]*

[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India,* [2]*Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India*

Mutational scanning can be used to probe effects of large numbers of point mutations on protein function. Positions affected by mutation are primarily at either buried or at exposed residues directly involved in function, hereafter designated as active-site residues. In the absence of prior structural information, it has not been easy to distinguish between these two categories of residues. We curated and analyzed a set of twelve published deep mutational scanning datasets. The analysis revealed differential patterns of mutational sensitivity and substitution preferences at buried and exposed positions. Prediction of buried-sites solely from the mutational sensitivity data was facilitated by incorporating predicted sequence-based accessibility values. For active-site residues we observed mean sensitivity, specificity and accuracy of 61, 90 and 88% respectively. For buried residues the corresponding figures were 59, 90 and 84% while for exposed non active-site residues these were 98, 44 and 82% respectively. We also identified positions which did not follow these general trends and might require further experimental re-validation. This analysis highlights the ability of deep mutational scans to provide important structural and functional insights, even in the absence of three-dimensional structures determined using conventional structure determination techniques, and also discuss some limitations of the methodology.

Keywords: deep sequencing, saturation mutagenesis, protein function, activity, stability, phenotype

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom

Reviewed by:
Joseph Marsh,
University of Edinburgh,
United Kingdom
Lipi Thukral,
Institute of Genomics and Integrative
Biology (CSIR), India

Correspondence:
Raghavan Varadarajan
varadar@iisc.ac.in

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *30 November 2020*
Accepted: *25 January 2021*
Published: *11 March 2021*

Citation:
Bhasin M and Varadarajan R (2021)
Prediction of Function Determining and
Buried Residues Through Analysis of
Saturation Mutagenesis Datasets.
Front. Mol. Biosci. 8:635425.
doi: 10.3389/fmolb.2021.635425

INTRODUCTION

Mutagenesis is a tool to learn about proteins, identifying functionally significant protein positions, and understanding determinants of protein folding and stability. Deep mutational scanning involving a combination of saturation mutagenesis, phenotypic screening and next generation sequencing allows high-throughput analysis by measuring the effects of all possible amino acid substitutions on protein function (Fowler and Fields, 2014). Deep mutational scanning reveals the impact of mutations on a specific protein property, for example, interaction with a partner protein or enzymatic activity. A general workflow for a deep mutational scan involves the creation of a library of variants by applying a mutagenesis protocol to the genetic region of interest (Fowler et al., 2010) which can include an entire coding sequence (Adkar et al., 2012). Next, these libraries are subjected to some selection pressure, and this is used to observe the change in the frequency of variants with a particular phenotype. The libraries are sequenced before and after selection to obtain relative occurrences of different mutants in the population and estimate relative enrichment with respect to the wild type sequence (Tripathi and Varadarajan, 2014).

There have been numerous attempts to understand and predict functional consequences of mutations by using computational methods (Bloom et al., 2005; Moretti et al., 2013). The availability of deep mutational scanning data has helped to understand the contribution of every amino acid in a protein to its structure, stability, and function, understand how these mutations regulate protein activity, and to build on this information to predict functional effects of mutations in other contexts. Mutations can affect activity either by altering the specific activity, altering the level of properly folded protein *in vivo*, or by a combination of the above (Tripathi et al., 2016). Identifying which of these is the primary contributor to an observed phenotype is non-trivial.

For understanding the functional role of a protein, it is essential to identify the key catalytic or functionally important residues that we collectively refer to as active-site residues. There are several tools available to predict protein function based on query protein sequence or structural homology with well-characterized proteins (Gherardini and Helmer-Citterich, 2008). One of the common methods used to identify catalytic sites is using sequence conservation (Berezin et al., 2004; Fischer et al., 2008). With the availability of three-dimensional structures of proteins, these methods can be further improved by combining structural and sequence conservation information (Lichtarge et al., 1996; Aloy et al., 2001; Capra et al., 2009). These methods provide cues to design experiments, including site-directed mutagenesis experiments, and help to give an improved prediction of function (George et al., 2005). Such methods are helpful in cases where protein structural information is available. For cases with insufficient structural information, the data from deep mutational scans can be utilized in order to infer functional sites based on the substitution preferences across the protein under study.

In the present study, we have analyzed several deep mutational scanning datasets and observed the mutational sensitivity patterns at buried and exposed positions. Further, the sequence-based predicted accessibility values were incorporated together with the mutational sensitivity scores to predict functional or active-site residues. These residues include residues involved in catalytic activity, substrate binding, as well as protein-protein or protein-ligand interactions. Predicted accessibility scores help in the separation of the exposed from the buried residues. Residues that are sensitive to mutation and predicted to be exposed are likely to constitute the active-site, while the remaining mutationally sensitive residues are likely to be buried.

MATERIALS AND METHODS

Datasets for Large-Scale Mutagenesis

A subset of the published deep mutational scanning datasets was curated. The result was a set of 12 deep mutational scans (**Table 1**). While several other studies have been published, most lack sufficient coverage of single-site mutations over the region of interest, have more than one mutation per read or describe complex phenotypes which preclude easy interpretation of the data. Alternatively, several studies report heatmaps and raw sequencing data without having the underlying numerical values of the processed enrichment scores publicly available.

Data Rescaling

Most of the deep mutational scanning datasets reported mutational effect scores as the log-transformed ratio of mutant frequency before and after selection, divided by wild-type frequency before and after selection. The counts/frequency of the mutational sensitivity scores were considered from the original datasets, and their distribution was plotted. The values were sorted, and the 5th percentile of the value was taken as the minimum value, min*(M)*, for rescaling. The maximum value, max*(M)*, for the rescaling was considered as the value at the peak for the wild type in the histograms. This peak arises because many mutational effect scores are close to that of the WT. The scores were rescaled between 0 and −1 using the formula:

$$M_{rescaled} = (b-a)\frac{M-\min(M)}{\max(M)-\min(M)} + a,$$

where, M is the mutational effect score, a and b are −1 and 0, respectively. With this normalization, the most sensitive positions have mutational effect score ≈−1 and the wild type like mutations have mutational effect score ≈0 (**Supplementary Figure S1** and **Supplementary Table S1**).

Depth and Accessibility Calculations

Both depth and accessibility of each residue were calculated from the available structures deposited in the Protein Data Bank. Amongst the datasets in the study, five of the proteins had high-resolution PDB structures, namely dimeric CcdB structure (PDB ID 3VUB) (Loris et al., 1999), PSD95 pdz3 domain (PDB ID 1BE9), BRCA1 RING domain (PDB ID 1JM7) (Starita et al., 2015), Gal4 (PDB ID 3COQ) (Marmorstein et al., 1992) and TEM1 β-lactamase (PDB ID 1FQG) (Strynadka et al., 1992).

The residue depth calculations were performed using the DEPTH server (http://cospi.iiserpune.ac.in/depth/htdocs/index.html) (Chakravarty and Varadarajan, 1999; Tan et al., 2011). A residue was defined as buried or exposed if the side chain accessibility is ≤5 or >5% respectively, based on the accessibility calculated using the NACCESS program (Adkar et al., 2012).

Prediction of Sequence-Based Surface Accessibility

The sequence-based surface accessibility values were predicted using PROF (Rost and Sander, 1994), a neural network-based method (https://open.predictprotein.org/). These values were compared with the structure-based surface accessibility values, which were calculated using the NACCESS program (Hubbard and Thornton, 1993). NetSurfP was also used for the prediction of sequence-based surface accessibility (Petersen et al., 2009) and compared with the prediction results obtained using PROF. PROF and NetSurfP predictions were also compared with

SPIDER3 (Heffernan et al., 2017), a machine learning method that takes into account the non-local interactions in its predictions.

Prediction of the Active-Site, Buried and Exposed Non Active-Site Residues

The rescaled mutational sensitivity values were averaged across mutations for each position. The averaged mutational sensitivity scores were filtered to include only those positions that had mutational data for a minimum of 10 mutants per position. Also, only those positions were considered for which the predicted sequence-based accessibility values were predicted. Both the scores for averaged mutational sensitivity and PROF accessibility are converted to Z-scores by subtracting the mean value and dividing by the standard deviation. The final score for predicting the active-site residues is obtained by using the following formula:

$$Z_{\text{pred}} = Z_{\text{average}_{\text{mut-sens}}} \pm Z_{\text{prof-acc}},$$

Where, Z represents the z-scores. For the prediction of active-site residues, the two scores are added, whereas the scores are subtracted for the prediction of buried positions. The mean and standard deviation were calculated for the combined score. Residues with scores one standard deviation away from the mean were predicted as active-site or buried.

For the prediction of exposed non-active site residues, the same scores that used the rescaled averaged mutational sensitivity scores along with the sequence-based accessibility scores were considered. The residues that occurred beyond the cut-off for prediction of active-site residues were predicted to be exposed non active-site residues. A similar analysis was performed by incorporating the sequence-based accessibility values obtained using NetSurfP and SPIDER3 to compare the three classes of prediction namely, active-site, buried and exposed non active-site residues.

Evaluation Metrics

We assume the active-site residues to represent the positive samples and non active-site residues to represent the negative samples for the prediction of active-site residues. On the other hand, for the prediction of the buried sites, we consider the buried site residues as the positive samples and the exposed residues as the negative samples. The exposed non active-site prediction considered the positive and negative samples in similar way. To evaluate the performance of prediction, four evaluation metrics are used in this study: sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Matthews Correlation Coefficient} = \frac{(\text{TP}* \text{TN}) - (\text{FP}* \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TP, TN, FP, FN are True Positives, True Negatives, False Positives, and False Negatives, respectively.

RESULTS

Deep mutational scanning involves measurement of large numbers of mutational phenotypes for a given protein using phenotypic screening coupled to deep sequencing (Adkar et al., 2012; Fowler et al., 2010). It can be used to quantify the phenotypic effects of all mutations at each position in a protein. These deep mutational scanning data sets help understand the relationships between amino acid sequence and phenotype. The assay formats used for the deep mutational scans included plate-based activity screens, FACS, phage display and yeast two-hybrid methodologies (Gupta and Vardarajan, 2018). Site-saturation mutagenesis (SSM) has been employed in several studies to probe residue-specific contributions to activity, stability, and binding for whole proteins (Gray et al., 2017). This study analyzed 12 large-scale mutational datasets of 11 proteins from existing deep mutational

TABLE 1 | Large-scale deep mutational scanning datasets used in this study.

Data set	Mutagenized positions	Host	Selection	PDB ID	Citation
Aminoglycoside kinase	264	*E. coli.*	Antibiotic resistance	1ND4	(Melnikov et al. (2014))
BRCA1 RING domain-BARD1 binding	102	*S. cerevisiae*	Binding activity (Y2H)	1JM7	(Starita et al. (2015))
BRCA1 RING domain–E3 ligase activity	102	*S. cerevisiae*	Ubiquitin ligase activity	1JM7	(Starita et al. (2015))
CcdB	100	*E. coli.*	Toxin activity	3VUB	(Adkar et al. (2012))
Gal4 (DBD)	64	*S. cerevisiae*	Transcription factor activity	3COQ	(Kitzman et al. (2015))
G protein (GB1-IgG-Binding domain)	54	*Streptococcus sp. group G*	IgG-Fc binding	1PGA	(Olson et al., (2014))
Hsp90 (ATPase domain)	219	*S. cerevisiae*	Chaperone activity	2CG9	(Mishra et al. (2016))
NUDT15	163	*E. coli*	Abundance and drug sensitivity	5LPG	(Suiter et al. (2020))
Pab1 (RRM domain)	75	*S. cerevisiae*	mRNA binding	1CVJ	(Melamed et al. (2013))
PSD95(pdz3 domain)	83	*E. coli.*	Ligand binding	1BE9	(McLaughlin et al. (2012)
TEM1 β-lactamase	263	*E. coli*	Antibiotic resistance	1FQG	(Stiffler et al. (2015))
Ubiquitin	75	*S. cerevisiae*	Ubiquitin ligase activity	1UBQ	(Roscoe et al. (2013))

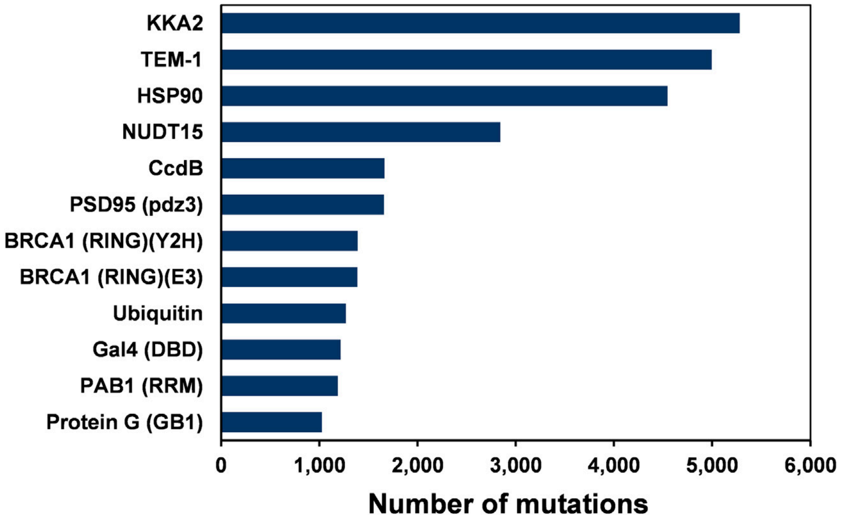


FIGURE 1 | The number of single amino-acid mutations in various deep mutational scanning datasets of 12 proteins.

scan experiments (**Figure 1** and **Table 1**). In the case of BRCA1, there are two independent deep mutational scan experiments, one for BRCA1 BARD1 binding and the other for E3 ligase activity (Starita et al., 2015). In these separate experiments, a multiplexed yeast two-hybrid assay was used to select for the ability of BRCA1 RING domain (2–103) variants to interact with the RING domain of BARD1. The structure is also available for the BRCA1/BARD1 RING-domain heterodimer (1JM7) (Brzovic et al., 2001). Since the present study involves the prediction of the active-site residues based on the mutational effect scores and the sequence-based accessibility predictions, the variants from the same region (2–103) of BRCA1 were used for the E3 ligase function experiment instead of using the E3 ligase scores available for full-length BRCA1 protein.

Some general patterns of mutational sensitivity were observed for the datasets used in the present study. Buried residues have high mutational sensitivity compared to those that are exposed and not part of the active-site. The residues that show high mutational sensitivity at exposed regions are typically involved in an interaction with some other proteins or are part of a catalytic or ligand binding site. As discussed above, these residues are classified as active-site residues. Hence, it is important to examine if these active-site residues can be distinguished from buried residues based on the mutational sensitivity scores, even in the absence of structural data (Tripathi et al., 2016).

Analysis of Mutational Sensitivity Data

The datasets contained effect scores for most mutations at each position. To facilitate comparisons between each data set, the mutational effect scores were rescaled for each protein. To understand the overall trends in mutational sensitivity, the substitution preferences were examined for all the proteins in the dataset. A residue was defined as buried or exposed based on its side-chain accessibility calculated using the NACCESS program. A cut-off of 5% side-chain accessibility was used (Adkar et al., 2012). The interface residues for the proteins in the dataset were determined from the corresponding literature citations of their respective structures.

Most exposed positions have a low mutational sensitivity (**Supplementary Figure S2**). It has been observed that buried residues along with some of the exposed residues have a high mutational sensitivity. Exposed residues that are sensitive to mutations are likely to be a part of the active-site (Wu et al., 2015). We examined if the substitution specific patterns of mutational sensitivity could help to distinguish the active-site residues from the buried ones. The effect of various substitutions was analyzed for different categories, namely aliphatic, aromatic, polar and charged (**Supplementary Figure S2**). In most cases, buried positions tolerated aliphatic substitutions, except when the wild-type residue is an Alanine or Glycine residue. Polar and charged residues are poorly tolerated at buried positions. Exposed

TABLE 2 | Correlation coefficients of surface accessibility predicted using PROF, NetSurfP and SPIDER3 with values calculated from the structure using NACCESS. The oligomeric state of the protein based on the PDB structure is also mentioned.

Protein	Correlation coefficient			Oligomeric state
	PROF	NetSurfP	SPIDER3	
Aminoglycoside kinase	0.66	0.75	0.69	Dimer
BRCA1 RING domain	0.45	0.62	0.66	Monomer
CcdB	0.71	0.75	0.74	Dimer
Gal4 (DBD)	0.73	0.77	0.66	Tetramer
GB1 (IgG-binding domain)	0.67	0.52	0.64	Monomer
Hsp90 (ATPase domain)	0.56	0.64	0.59	Tetramer
NUDT15	0.55	0.63	0.62	Dimer
Pab1 (RRM domain)	0.75	0.81	0.77	Dimer
PSD (pdz3 domain)	0.74	0.81	0.61	Dimer
TEM1 β- lactamase	0.74	0.81	0.79	Monomer
Ubiquitin	0.74	0.84	0.73	Monomer

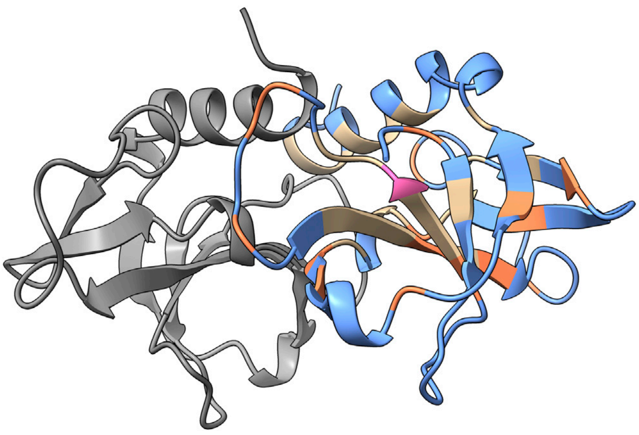


FIGURE 2 | PROF prediction results for CcdB. The sequence-based surface accessibility results obtained from PROF mapped onto the structure of CcdB homodimer (PDB ID: 3VUB). The predictions with respect to the exposed positions are mapped on the structure. One monomer is highlighted in gray and the prediction results are mapped onto the other monomer. The true positives are highlighted in blue, false positives in pink, true negatives in tan and false negatives in orange based on the predictions from PROF and crystal structure accessibilities calculated using NACCESS.

active-site residues showed very high mutational sensitivity including for substitutions to aliphatic residues (**Supplementary Figure S2**). The general trends in mutational sensitivity were similar for most proteins that were considered for the analysis. However, some proteins namely E3 ligase activity of BRCA1 RING domain, NUDT15 and aminoglycoside kinase were exceptionally sensitive to mutation, even at exposed non active-site residues. Even for the same protein, two different activity assays namely BARD1 binding and E3 ligase activity showed very different mutational sensitivity profiles. While this is understandable for active-site residues, it is hard to understand for buried residues where mutations are expected to primarily affect protein levels, rather than specific activity (Bajaj et al., 2008; Tripathi et al., 2016)

Correlation Between Calculated and Predicted Solvent Accessibility

To predict the active-site residues solely from the mutational sensitivity data, the accessibility was predicted based on sequence using PROF (Rost and Sander et al., 1994). Further, the correlation was calculated between the calculated surface accessibility and the predicted accessibility values (**Table 2**). The predicted surface accessibility for the 11 proteins from 12 datasets showed a Pearson's correlation coefficient r ~ 0.6 with the calculated surface accessibility values in most cases. The predicted accessibility information was combined with the mutational sensitivity scores to predict the active-site and buried residues as described in the Methods section.

To illustrate the accuracy of the accessibility predictions, results obtained from PROF and the calculated accessibility from NACCESS are mapped on the structure of CcdB (PDB ID: 3VUB). CcdB is a 101-residue homodimeric toxin found on F-plasmid (**Figure 2**). The true positives, false positives, true negatives and false negatives are highlighted in the figure. Here, true positives are correctly predicted exposed residues while false positives are buried residues that are incorrectly predicted as exposed by PROF. True negatives were correctly predicted buried

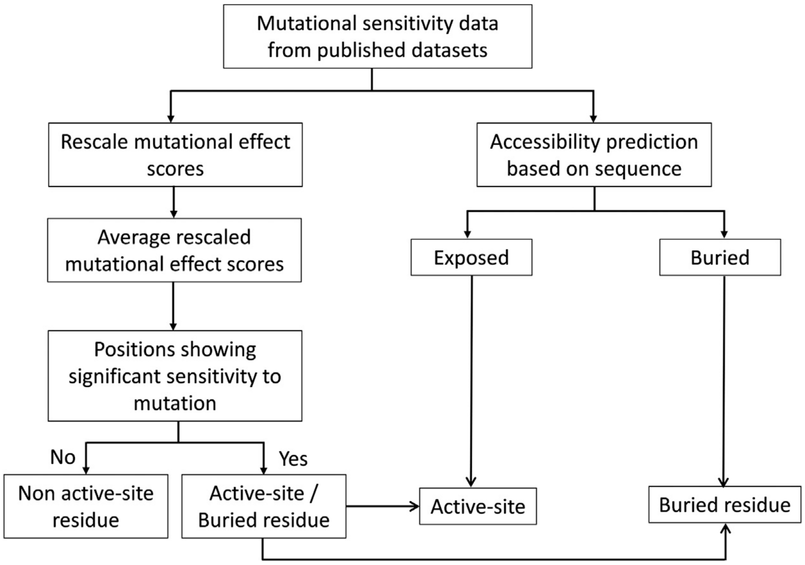

Mutational sensitivity data from published datasets
Rescale mutational effect scores
Accessibility prediction based on sequence
Average rescaled mutational effect scores
Exposed
Buried
Positions showing significant sensitivity to mutation
No
Yes
Non active-site residue
Active-site / Buried residue
Active-site
Buried residue

FIGURE 3 | Flowchart of the methodology for prediction of active-site, buried and exposed non-active site residues. The mutational sensitivity score was determined for each mutant from deep sequencing-based screening. These scores were rescaled and averaged across each position. The sequence-based surface accessibility was predicted using the PROF server. The residues that showed significant sensitivity to mutations and which were predicted to be exposed were further considered to be the active-site residues.

TABLE 3 | Active-site prediction based on the mutational sensitivity data and PROF predicted sequence-based accessibility values.

Dataset	Sensitivity (%)	Specificity (%)	Accuracy (%)	Matthews correlation coefficient
Aminoglycoside kinase	72.7	87.7	87.1	0.34
BRCA1 RING domain-BARD1 binding	45.5	91.4	85.2	0.37
BRCA1 RING domain–E3 ligase activity	50	92.3	86.6	0.42
CcdB	75	98.9	96.9	0.79
Gal4 (DBD)	46.6	86.1	75.9	0.34
GB1 (IgG-binding domain)	85.7	91.5	90.7	0.66
Hsp90 (ATPase domain)	93.3	92	92.1	0.63
NUDT15	50	91.2	85.4	0.41
Pab1 (RRM domain)	62.5	87.9	85.1	0.41
PSD (pdz3 domain)	75	90.6	89.2	0.53
TEM1 β-lactamase	66.6	85.8	85.2	0.26
Ubiquitin	70	96.4	92.3	0.69

TABLE 4 | Prediction of buried sites based on mutational sensitivity data and PROF predicted sequence-based accessibility values.

Dataset	Sensitivity (%)	Specificity (%)	Accuracy (%)	Matthews correlation coefficient
Aminoglycoside kinase	66.6	90.8	85.6	0.57
BRCA1 RING domain-BARD1 binding	38.5	88.2	80.2	0.27
BRCA1 RING domain–E3 ligase activity	38.5	80.6	73.3	0.12
CcdB	68.4	96.1	90.6	0.69
Gal4 (DBD)	50	78.6	77.6	0.13
GB1 (IgG-binding domain)	70	90.9	87	0.58
Hsp90 (ATPase domain)	18.8	84.7	73.5	0.06
NUDT15	69.7	88	84.2	0.55
Pab1 (RRM domain)	80	89.8	87.8	0.65
PSD (pdz3 domain)	55	90.5	81.9	0.48
TEM1 β- lactamase	59.8	91.5	80.9	0.55
Ubiquitin	45.5	83.3	76.9	0.26

residues and false negatives were exposed residues wrongly predicted as buried.

Performance of the Method for Prediction of Active-Site, Buried and Exposed Non-Active Site Residues

Deep mutational scanning plays a crucial role in identifying protein-ligand interfaces and is useful regardless of the structural context. To identify the active-site residues and distinguish them from buried residues, we analyzed the structures of 11 proteins for the dataset used. The dataset comprises proteins that share interfaces with other proteins and includes a protein that binds to DNA. For all the proteins, structure-based solvent accessibilities were calculated to validate the predicted accessibilities (**Figure 3**).

For the prediction of active-site residues, an average sensitivity of ~61% was observed (**Table 3**). This shows that if only the mutational sensitivity scores and sequence-based accessibility values are used, then those residues which are exposed and non-interacting, as well as ones that are buried, are segregated from the active-site residues. There is often a trade-off between specificity and sensitivity. Consistent with this, it was observed that for some of the datasets, there is low sensitivity, i.e., not all active-site residues are identified. In these cases, most of the exposed active-site residues have been incorrectly predicted as buried residues.

It has been observed that active-sites, as well as buried positions, have high mutational sensitivity. Therefore, it is essential that these buried positions are separated from the exposed active-site residues to enhance the accuracy of active-site prediction. To identify buried residues, we employed predicted accessibility values that have been obtained from sequence information. Since sequence-based accessibility Z-scores for buried residues are typically very low, these scores are subtracted from the averaged mutational sensitivity scores to predict them in the absence of structural information. After combining both averaged mutational sensitivity scores and sequence-based accessibility values from PROF, an average specificity of ~90% is observed (**Table 4**). The sensitivity is ~55% as some buried residues are predicted as exposed by the sequence-based accessibility predictor. The overall value of average sensitivity is affected by the low sensitivity of predictions in the case of HSP90 (Mishra et al., 2016). The pattern of mutational sensitivity for the buried positions in this protein is atypical, relative to the overall trend observed in the other large-scale mutagenesis datasets, with many buried positions tolerating charged substitutions. A similarly high degree of tolerance is observed for the BRCA1 RING domain, but only when BARD1 binding, rather than E3 ligase activity is assayed.

TABLE 5 | Prediction of exposed non active-site residues based on mutational sensitivity data and PROF predicted sequence-based accessibility values.

Dataset	Sensitivity (%)	Specificity (%)	Accuracy (%)	Matthews correlation coefficient
Aminoglycoside kinase	94.9	22.1	76.1	0.25
BRCA1 RING domain-BARD1 binding	94.7	25	74.1	0.29
BRCA1 RING domain–E3 ligase activity	92.3	34.8	74.7	0.34
CcdB	97.1	29.6	78.1	0.39
Gal4 (DBD)	92.3	41.2	77.5	0.41
GB1 (IgG-binding domain)	89.2	35.3	72.2	0.29
Hsp90 (ATPase domain)	92.3	36.2	80	0.34
NUDT15	90.3	23.6	67	0.27
Pab1 (RRM domain)	94.2	50	81.1	0.52
PSD (pdz3 domain)	92.7	35.7	73.5	0.36
TEM1 β-lactamase	92.0	31	68.8	0.3
Ubiquitin	97.8	40	80	0.5

The accuracy of prediction results for both active-site and buried residues are ~88% and ~84%, respectively. In the case of prediction of the buried positions, it was observed that the incorporation of sequence-based accessibility values played an important role in improving the results (**Supplementary Figure S3**). This helped to distinguish both the categories of mutationally sensitive positions, namely exposed active-site and buried positions. In contrast, prediction of the exposed non active-site residues prediction did not show significant improvement after incorporating the sequence-based accessibility scores (**Supplementary Figure S3**). Overall, incorporating the sequence-based accessibility values along with the averaged mutational sensitivity scores improves the prediction performance of the method primarily for buried residues and can be useful in identifying key residues in the protein even in the absence of structural information.

Along with the prediction of the active-site and buried residues, the exposed non active-sites can also be distinguished from the other two categories. The high value of sensitivity in these prediction results points to the ability of the method to identify these residues (**Table 5**). In a few cases, it was observed that there are a few exposed positions far from the active-site that show high mutational sensitivity. In the case of TEM1 β-lactamase, it was observed that exposed positions with large side chain show a high mutational sensitivity in comparison to the other exposed non active-site residues. For example, Trp210, Trp229 and Trp290 are exposed residues that are crucial for the structure and activity of β-lactamase (Huang et al., 1996). Mutations at such positions may lead to the instability of the enzyme, thus abrogating its function, though this needs to be confirmed by experiments. In comparison to predictions in the other two categories, prediction specificity was low for exposed non active-site residues, probably resulting from the lower fraction of true negatives in this category.

The Matthew's correlation coefficient (MCC) was computed using either the experimental mutational effect scores or the PROF predicted accessibility values. These values were compared with corresponding values obtained using the combined score for all the three categories of predictions (**Figure 4**). The results show that overall, the combined score yields the best results.

Comparison of the Results Across Other Solvent Accessibility Predictors

In the above analysis, the sequence-based accessibility scores from PROF were considered along with the experimental mutagenesis scores to calculate the prediction sensitivity for the active-site, buried and exposed non active-site residues. The average Pearson's correlation coefficient of predicted accessibility from PROF with calculated surface accessibility from NACCESS is 0.66.

The analysis was also performed with another sequence-based accessibility predictor NetSurfP (Petersen et al., 2009). In this case, the correlation between the predicted accessibility from NetSurfP and calculated accessibility from NACCESS, is improved with an average correlation coefficient of 0.72. The sensitivity, specificity and accuracy of the results were recalculated using NetSurfP rather than PROF for residue accessibility prediction (**Figure 5**, **Supplementary Tables S3–S5**). Results for prediction of buried and exposed non-active site residues are comparable with both accessibility predictors. However for active-site residues the sensitivity was

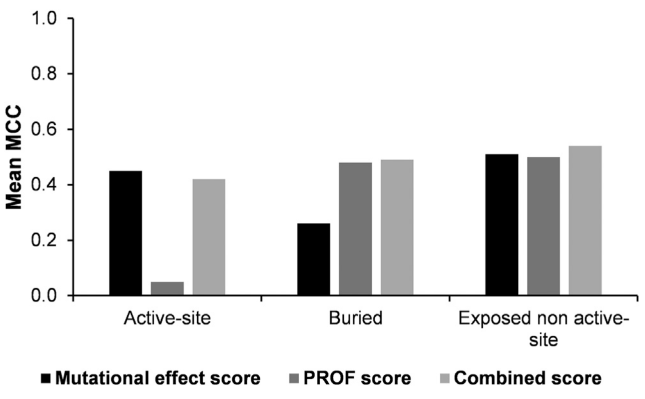


FIGURE 4 | Comparison of Matthew's correlation coefficients of predictions using experimental mutational effect scores alone, PROF predicted accessibility alone and experimental mutational effect scores combined with the PROF predicted accessibility scores (combined score).

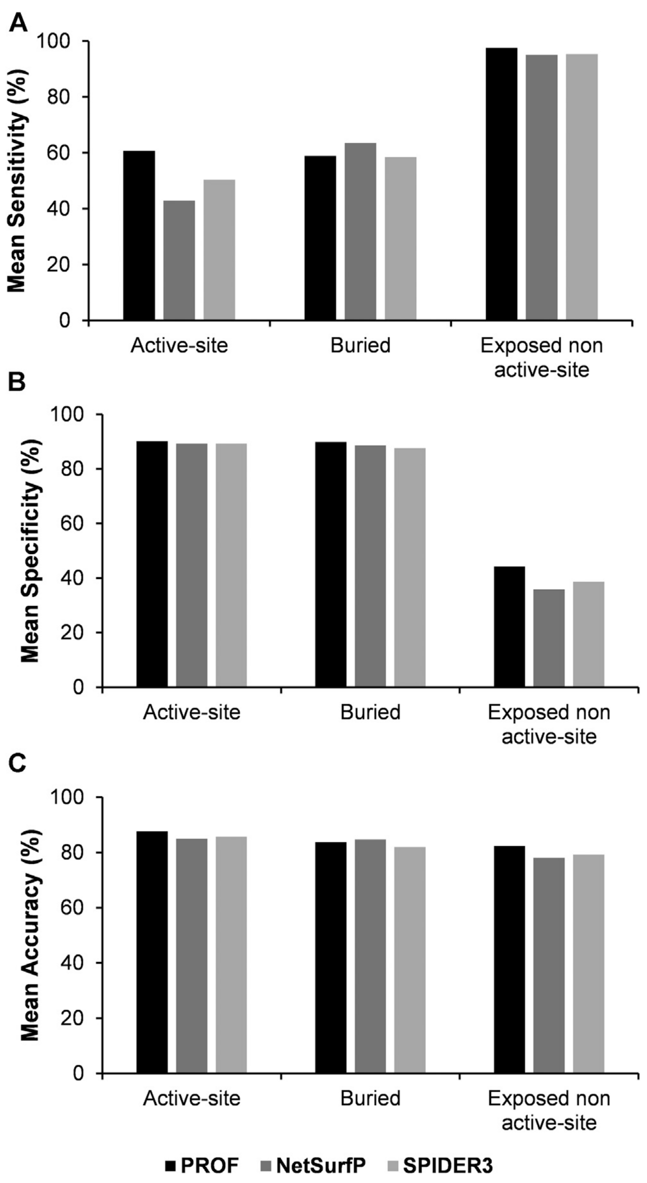


FIGURE 5 | Comparison of mean values of sensitivity, specificity and accuracy of predictions using mutational effect scores combined with the predicted accessibility results from PROF, NetSurfP and SPIDER3 respectively.

lower when the NetSurfP predicted accessibility was used instead of the PROF predicted accessibility.

In order to compare between the various surface accessibility predictors, SPIDER3 (Heffernan et al., 2017) was also used for prediction of active-site, buried and exposed non active-site residues. SPIDER3 is a method that captures long-range, non-local interactions and predicts the protein one-dimensional structural properties. The correlation between the predicted accessibility using SPIDER3 and calculated accessibility using NACCESS was 0.68 which is comparable to the correlation coefficient observed in the case of PROF. After incorporating sequence-based accessibility scores from SPIDER3 with the experimental mutational sensitivity scores, there was a very slight improvement in the prediction sensitivity of the buried positions. However as with NetSurfP, the mean sensitivity values for prediction of active-site residues and mean accuracy values were lower with SPIDER3, relative to PROF (**Figure 5**; **Supplementary Tables S3–S5**).

Comparison of the Results with Mutational Effect Predictors

As there is currently a limited number of complete deep mutational scanning datasets, a similar analysis was carried out by using the predicted mutational effect scores from the computational variant effect predictor SNAP2 (Hecht et al., 2016), which required only the sequence as the input to predict mutational effect scores. An average Pearson's correlation coefficient of 0.5 was see between the experimental and SNAP2 predicted scores. The three categories of residues namely, active-site, buried and exposed non active-site residues were further predicted using the SNAP2 scores by combining them with PROF predicted accessibility (**Figure 6**, **Supplementary Table S6**). The predicted variant effect scores poorly predict active-site residues. However, prediction metrics for buried and exposed non active-site residues are comparable in terms of their sensitivity, specificity and accuracy to those obtained with experimental mutational scores.

DISCUSSION

Deep mutational scanning is a method that is widely used to probe the effects of substitutions on proteins, which helps to identify functionally important residues (Adkar et al., 2012). In this study, we examined if such large-scale mutagenesis datasets, could be used to infer locations of functional sites in proteins and distinguish them from other positions based on their specific mutational sensitivity pattern.

The present analysis reveals that active-site residues are on average more sensitive to mutation than buried residues. Use of sequence-based accessibility predictions further contributes to distinguishing buried positions from the exposed active-site residues. The third category of residues that is largely insensitive to mutation, is exposed non active-site residues There are a few exposed non active-site residues that are mis predicted as active-site residues. One of the reasons for this is their proximity to the active-sites, thus making them sensitive to substitutions. In some cases, these exposed mutationally sensitive residues have accessibility values that are close to the cut-off that is used for classifying them as exposed or buried. Among the datasets considered for prediction of active-site residues, there is one deep mutational scan of the DNA-binding domain (DBD) of Gal4, a yeast transcription factor (Kitzman et al., 2015). Gal4 binds DNA as a homodimer via a Zn_2Cys_6-class domain centered on a pair of Zn^{2+} ions. This helps to maintain the fold of the DNA-binding residues. Substitutions at any of six cysteines completely disrupts the function (Marmorstein et al., 1992). Since these cysteines are both buried, but also involved in the

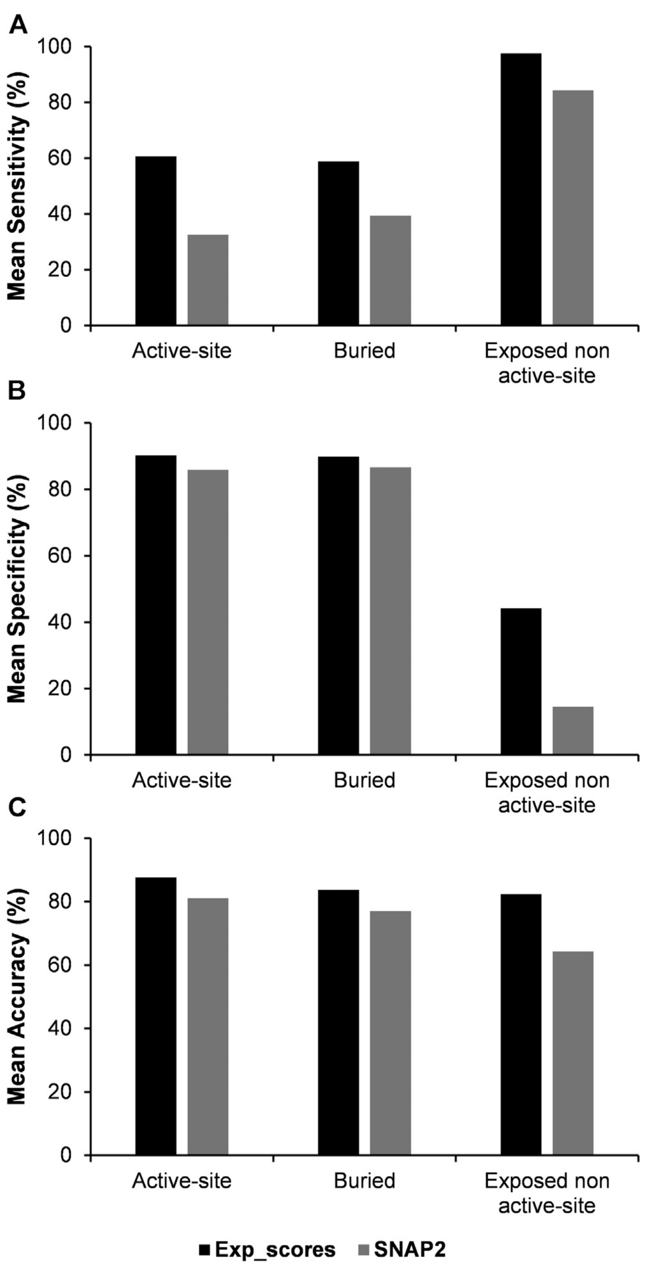


FIGURE 6 | Comparison of mean values of sensitivity, specificity and accuracy of predictions using experimental and SNAP2 predicted mutational effect scores combined with the predicted accessibility results from PROF.

activity of the protein, they are considered as active-site residues for analysis. They have been further excluded in the prediction of the buried residues. It was also observed that sensitivity of predictions decreases for proteins with a large number of interacting partners or with limited mutational sensitivity data. Thus, for the present study only those deep mutational scanning datasets are considered which have an average of at least ten mutants per residue.

For datasets where the relative fitness effects of single amino acid mutations were observed under antibiotic selection, an optimum antibiotic concentration value was selected for prediction. In the case of TEM1 β-lactamase, mutational data foe selection with an ampicillin concentration of 625 μg/ml were used (Stiffler et al., 2015). Higher concentrations of ampicillin result in high mutational sensitivity across the entire protein. This results in inability to separate the key catalytic residues from the non-interacting ones. For aminoglycoside kinase, the relative abundance of mutant vs. wild-type amino acids at each position was examined under kanamycin selection at a range of inhibitory concentrations (Melnikov et al., 2014). At high kanamycin concentration, the mutational sensitivity was again very high, thus data from the lower kanamycin concentration was used for analyzing the pattern of mutational sensitivity. In general, it appears that mutational scanning datasets are most useful when phenotypic screens are carried out under conditions where ~25% of substitutions yield measurable phenotypes.

Amongst the deep mutational scanning datasets analyzed in this study, there are a few cases where there is high mutational sensitivity at non active-site residues. One such example is the deep mutational scan of TEM1 β-lactamase (Stiffler et al., 2015). There are residues that are distal from the active-site but are highly sensitive to substitutions, suggesting possible allostery (Avci et al., 2016). However, it is difficult to know if such mutational sensitivity is because of functional allostery or because of a decreased level of secreted protein, for example because of increased proteolysis. This emphasizes the need to measure both levels of properly folded protein as well as activity. This is not done in most mutational scans.

Since there still relatively few proteins that have been subjected to deep mutational scans, computationally predicted variant effect scores were used in place of experimental data. However, this led to poor predictions for active-site residues. In future, given recent advances in deep learning based structure prediction (Senior et al., 2020), it would be interesting to map computationally predicted variant scores onto structural models to more accurately predict active-site residues.

In addition to identifying buried, active-site and exposed non active-site residues, the present analysis has identified puzzling mutational sensitivity features in some of the proteins in the present dataset, that reflect either our incomplete understanding of determinants of protein stability and function or potential lacunae in the experimental data that need additional validation through repeat experiments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MB carried out the analysis and drafted the manuscript. RV designed the study, supervised the analysis, and corrected the manuscript.

FUNDING

This work was funded by grants to RV from the Department of Biotechnology (grant number-BT/COE/34/SP15219/2015), and Department of Science and Technology, Government of India (EMR/2017/004054). We acknowledge funding for infrastructural support from the following programs of the Government of India: DST FIST, Ministry of Human Resource Development (MHRD), Council of Scientific & Industrial Research (CSIR) and the DBT IISc Partnership Program The funding agencies had no role in study design, results analysis and preparation of manuscript. MB is a CSIR-SRF and RV is JC Bose Fellow of DST.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.635425/full#supplementary-material.

REFERENCES

Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., et al. (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* 20, 371–381. doi:10.1016/j.str.2011.11.021

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. E. (2001). Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* 311, 395–408. doi:10.1006/jmbi.2001.4870

Avci, F. G., Altinisik, F. E., Vardar Ulu, D., Ozkirimli Olmez, E., and Sariyar Akbulut, B. (2016). An evolutionarily conserved allosteric site modulates beta-lactamase activity. *J. Enzyme Inhib. Med. Chem.* 31, 33–40. doi:10.1080/14756366.2016.1201813

Bajaj, K., Dewan, P. C., Chakrabarti, P., Goswami, D., Barua, B., Baliga, C., et al. (2008). Structural correlates of the temperature sensitive phenotype derived from saturation mutagenesis studies of CcdB. *Biochemistry* 47, 12964–12973. doi:10.1021/bi8014345

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., et al. (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20, 1322–1324. doi:10.1093/bioinformatics/bth070

Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., and Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. U.S.A.* 102, 606–611. doi:10.1073/pnas.0406744102

Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M.-C., and Klevit, R. E. (2001). Structure of a BRCA1– BARD1 heterodimeric RING–RING complex. *Nat. Struct. Biol.* 8, 833–837. doi:10.1038/nsb1001-833

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., and Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *Plos Comput. Biol.* 5, e1000585. doi:10.1371/journal.pcbi.1000585

Chakravarty, S., and Varadarajan, R. (1999). Residue depth: A novel parameter for the analysis of protein structure and stability. *Structure* 7, 723–732. doi:10.1016/s0969-2126(99)80097-5

Fischer, J. D., Mayer, C. E., and Söding, J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* 24, 613–620. doi:10.1093/bioinformatics/btm626

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., et al. (2010). High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7, 741–746. doi:10.1038/nmeth.1492

Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nat. Methods* 11, 801–807. doi:10.1038/nmeth.3027

George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., MacArthur, M. W., Porter, C. T., et al. (2005). Effective function annotation through catalytic residue conservation. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12299–12304. doi:10.1073/pnas.0504833102

Gherardini, P. F., and Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. *Brief. Funct. Genomic Proteomic* 7, 291–302. doi:10.1093/bfgp/eln030

Gray, V. E., Hause, R. J., and Fowler, D. M. (2017). Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* 207, 53–61. doi:10.1534/genetics.117.300064

Gupta, K., and Varadarajan, R. (2018). Insights into protein structure, stability and function from saturation mutagenesis. *Curr. Opin. Struct. Biol.* 50, 117–125. doi:10.1016/j.sbi.2018.02.006

Hecht, M., Bromberg, Y., and Rost, B. (2016). Better prediction of functional effects for sequence variants from VarI-SIG 2014: identification and annotation of genetic variants in the context of structure, function and disease. *BMC Genomics* 16, 1–12. doi:10.1186/1471-2164-16-S8-S1

Heffernan, R., Yang, Y., Paliwal, K., and Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* 33, 2842–2849. doi:10.1093/bioinformatics/btx218

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., and Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* 258, 688–703. doi:10.1006/jmbi.1996.0279

Hubbard, S. J., and Thornton, J. M. (1993). *"NACCESS" computer program*. London, United States: Department of Biochemistry and Molecular Biology, University College London..

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nat. Methods* 12, 203–206. doi:10.1038/nmeth.3223

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257, 342–358. doi:10.1006/jmbi.1996.0167

Loris, R., Dao-Thi, M. H., Bahassi, E. M., Van Melderen, L., Poortmans, F., Liddington, R., et al. (1999). Crystal structure of CcdB, a topoisomerase poison from E. coli. *J. Mol. Biol.* 285 (4), 1667–1677. doi:10.1006/jmbi.1998.2395

Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* 356, 408–414. doi:10.1038/356408a0

McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature* 491, 138–142. doi:10.1038/nature11500

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* 19, 1537–1551. doi:10.1261/rna.040709.113

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014). Comprehensive mutational scanning of a kinase *in vivo* reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* 42, e112–e118. doi:10.1093/nar/gku511

Mishra, P., Flynn, J. M., Starr, T. N., and Bolon, D. N. A. (2016). Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep.* 15, 588–598. doi:10.1016/j.celrep.2016.03.046

Moretti, R., Fleishman, S. J., Agius, R., Torchala, M., Bates, P. A., Kastritis, P. L., et al. (2013). Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins* 81, 1980–1987. doi:10.1002/prot.24356

Olson, C. A., Wu, N. C., and Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* 24, 2643–2651. doi:10.1016/j.cub.2014.09.072

Petersen, B., Petersen, T., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* 9, 51. doi:10.1186/1472-6807-9-51

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. A. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* 425, 1363–1377. doi:10.1016/j.jmb.2013.01.032

Rost, B., and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19, 55–72. doi:10.1002/prot.340190108

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., et al. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200, 413–422. doi:10.1534/genetics.115.175802

Stiffler, M. A., Hekstra, D. R., and Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* 160, 882–892. doi:10.1016/j.cell.2015.01.035

Strynadka, N. C., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C., et al. (1992). Molecular structure of the acyl-enzyme intermediate in beta-lactam hydrolysis at 1.7 A resolution. *Nature* 359, 700–705. doi:10.1038/359700a0

Suiter, C. C., Moriyama, T., Matreyek, K. A., Yang, W., Scaletti, E. R., Nishii, R., et al. (2020). Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5394–5401. doi:10.1073/pnas.1915680117

Tan, K. P., Varadarajan, R., and Madhusudhan, M. S. (2011). DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.* 39, W242–W248. doi:10.1093/nar/gkr356

Tripathi, A., Gupta, K., Khare, S., Jain, P. C., Patel, S., Kumar, P., et al. (2016). Molecular determinants of mutant phenotypes, inferred from saturation mutagenesis data. *Mol. Biol. Evol.* 33, 2960–2975. doi:10.1093/molbev/msw182

Tripathi, A., and Varadarajan, R. (2014). Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing. *Curr. Opin. Struct. Biol.* 24, 63–71. doi:10.1016/j.sbi.2013.12.001

Wu, N. C., Olson, C. A., Du, Y., Le, S., Tran, K., Remenyi, R., et al. (2015). Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLOS Genet.* 11, e1005310–27. doi:10.1371/journal.pgen.1005310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Low Diversity of Human Variation Despite Mostly Mild Functional Impact of De Novo Variants

Yannick Mahlich[1]*, *Maximillian Miller*[1], *Zishuo Zeng*[1] *and Yana Bromberg*[1,2]*

[1]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, United States, [2]Department of Genetics, Rutgers University, Piscataway, NJ, United States

Non-synonymous Single Nucleotide Variants (nsSNVs), resulting in single amino acid variants (SAVs), are important drivers of evolutionary adaptation across the tree of life. Humans carry on average over 10,000 SAVs per individual genome, many of which likely have little to no impact on the function of the protein they affect. Experimental evidence for protein function changes as a result of SAVs remain sparse – a situation that can be somewhat alleviated by predicting their impact using computational methods. Here, we used SNAP to examine both *observed* and *in silico* generated human variation in a set of 1,265 proteins that are consistently found across a number of diverse species. The number of SAVs that are predicted to have any functional effect on these proteins is smaller than expected, suggesting sequence/function optimization over evolutionary timescales. Additionally, we find that only a few of the yet-unobserved SAVs could drastically change the function of these proteins, while nearly a quarter would have only a mild functional effect. We observed that variants common in the human population localized to less conserved protein positions and carried mild to moderate functional effects more frequently than rare variants. As expected, rare variants carried severe effects more frequently than common variants. In line with current assumptions, we demonstrated that the change of the human reference sequence amino acid to the reference of another species (a cross-species variant) is unlikely to significantly impact protein function. However, we also observed that many cross-species variants may be weakly non-neutral for the purposes of quick adaptation to environmental changes, but may not be identified as such by current state-of-the-art methodology.

Keywords: variation, adaptation, evolution, nsSNVs, SNAP, cross-species variation, common variation

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology (LMB), Cambridge, United Kingdom

Reviewed by:
Pier Luigi Martelli,
University of Bologna, Italy
Octav Caldararu,
Technical Institute of Denmark, Denmark

****Correspondence:***
Yana Bromberg
yana@bromberglab.org
Yannick Mahlich
ymahlich@bromberglab.org

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *30 November 2020*
Accepted: *01 February 2021*
Published: *18 March 2021*

Citation:
Mahlich Y, Miller M, Zeng Z and Bromberg Y (2021) Low Diversity of Human Variation Despite Mostly Mild Functional Impact of De Novo Variants. Front. Mol. Biosci. 8:635382. doi: 10.3389/fmolb.2021.635382

INTRODUCTION

The vast majority of human genomic variants are single nucleotide variants (SNVs) (Durbin, et al., 2010). Coding region variants trivially make up a much smaller fraction of all variation than do non-coding variants (Lander, et al., 2001). However, the former affect protein structure/function and thus have a disproportionate effect of molecular function of the cellular machinery. For example, each individual genome contains approximately ten thousand of nsSNVs (non-synonymous SNVs, which change the amino acid sequence (Shen, et al., 2013), a combination of which is responsible for a variety of observed phenotypes, including disease (Peterson, et al., 2013; Hassan, et al., 2019). Establishing the effect of any given nsSNV, however, is a difficult task. One gold-standard experimental approach is saturated mutagenesis (SM) (Wells, et al., 1985), which induces variants of interest in a gene and measures the change of resulting protein molecular function. However, SM is too inefficient to thoroughly study the entirety of genomic variation. While the recent development of the deep mutational scanning techniques (Fowler and Fields, 2014) has facilitated high-throughput functional analysis of coding variants, experimental annotation of millions of possible nsSNVs in human genome still remains elusive, Given the inefficiency of large-scale experimental measurements computational methods for variant effect interpretation offer a plausible alternative for the exploration of the human genome.

Genome-wide association study (GWAS) (Visscher, et al., 2017), as well as the *post hoc* polygenetic risk scoring (Torkamani, et al., 2018), has been extensively deployed to establish the associations between complex phenotypes and genetic background. GWAS results, however, are by definition association (not causation) evaluations and are specific to a phenotype. Evaluating variant effect on molecular function requires a different type of techniques. Machine learning models are often used to classify variants into neutral/deleterious (e.g., CADD (Kircher, et al., 2014), DANN (Quang, et al., 2014)), benign/pathogenic (e.g., MutPred2 (Pejaver, et al., 2017), PhD-SNP (Capriotti and Fariselli, 2017)), stable/unstable (e.g., I-Mutant2.0 (Capriotti, et al., 2005)), and effect/no-effect (e.g., Envision (Gray, et al., 2018), SNAP (Bromberg and Rost, 2007), SNAP2 (Hecht, et al., 2015)).

Conservation of residues across homologs is often assumed to indicate structural or functional importance of these residues and their intolerance to substitution (Kumar, et al., 2009). Thus, conservation is used as a proxy for variant effect evaluation, e.g. by tools like SIFT (Ng, 2003) and PROVEAN (Choi and Chan, 2015), and has been widely incorporated as one of the features in many other variant effect predictors (e.g., CADD, DANN, SNAP, PhD-SNP). We previously proposed the concept of cross-species variants (CSV) analysis (Mahlich, et al., 2017), which is similar to but intuitively different from conservation evaluation. Conservation can be directly computed from a multiple sequence alignment (MSA) of homologs built for CSV analysis. However, CSVs specifically describe only the difference between two orthologous reference sequences and do not summarize overall conservation. For example, if the amino acid residue at a specific position of a human protein is glycine, and if the MSA-corresponding position of a mouse ortholog is leucine, then a CSV at this position of this human protein would be glycine > leucine. If this particular glycine > leucine variant also occurs in the human population, the variant is an *observed* CSV. As a rule, these types of human variants, i.e. to residues found in other species, have been presumed to carry no effect on protein function (Ng and Henikoff, 2001; Ng, 2003; Calabrese, et al., 2009; Adzhubei, et al., 2010; Shihab, et al., 2013; Kircher, et al., 2014; Schwarz, et al., 2014; Pejaver, et al., 2020). After all, if an amino acid is observed in a functional protein of an ortholog, its substitution into the human version cannot be expected to drastically affect the function.

Pathogenic amino acid substitutions are, on average, functionally more radical than CSVs (Briscoe, et al., 2004; Miller and Kumar, 2001; Subramanian and Kumar, 2006). A study of the rhodopsin protein, for example, has revealed that variants corresponding to CSVs among vertebrates are less likely to be pathogenic (Briscoe, et al., 2004). Of the 7,293 human-mouse CSVs in 687 human disease genes, only a small fraction (2.2%) corresponds to known human disease variants (Waterston, et al., 2002). Other studies have also estimated that only about 10% of the human-to-other-species amino acid substitutions are involved in disease (Kondrashov, et al., 2002; Subramanian and Kumar, 2006). However, this type of logic may have precipitated a self-fulfilling prophecy, where CSVs that were annotated to be neutral in the development of variant effect-prediction methods (Bromberg and Rost, 2007; Adzhubei, et al., 2010; Kircher, et al., 2014; Pejaver, et al., 2017; Pejaver, et al., 2020) could bias the prediction of previously unseen CSV effects toward neutrality. While unlikely pathogenic, intuitively, a yeast version of the human protein may be less or more functionally efficient, may have unexpected structural effects given the rest of the protein sequence, or may participate in different/additional molecular pathways. Incorporating taxonomic distances between the species included in an alignment improves identification of variant effect (Malhis, et al., 2019). A deeper evaluation of CSVs in terms of their functional effects may thus be warranted.

We previously reported (Mahlich, et al., 2017) that amino acid CSVs have less predicted molecular functional effects on average than human variation recorded by the Exome Aggregation Consortium (Lek, et al., 2016). Here we extend this analysis, by investigating human variation in 1,265 proteins that have orthologs in 20 species spread across the eukaryotic branch of the tree of life. We evaluate the differences in functional impact of the variants that are observed within the human population against those not yet observed, but genetically possible. We show that common variants favor less conserved positions than rare variants, indicating a potential need for flexibility in sequence for the purposes of environment-driven adaptation. We also assessed the differences in predicted impacts on the function of human protein of cross-species variants (CSVs; variant amino acid is found in one of the 20 orthologs) and non-CSVs. We finally suggest that the lack of functional impact of CSVs might be overestimated by the current presumption that evolutionary persistence suggests functional neutrality.

METHODS

Variant Collection

A total of 93,437 human protein-coding transcripts were extracted from GRCh37 p.13 assembly (Church, et al., 2011) in Ensembl BioMart (Kinsella, et al., 2011). From these, we selected 22,346 longest transcripts per gene. We removed transcripts from patches/alternate sequences (http://m.ensembl.org/info/genome/genebuild/haplotypes_patches.html), retaining 19,971 transcripts. For these, we artificially generated all possible non-synonymous single nucleotide variants (73,813,560 nsSNVs). We downloaded the Genome Aggregation Database (gnomAD v2, https://gnomad.broadinstitute.org/downloads) exome data (Karczewski, et al., 2020) and, using SAMtools (Li, et al., 2009), mapped the generated nsSNVs to the corresponding variant allele frequencies where available. We thus collected 2,951,998 variants with gnomAD allele count = 1 and 2,561,015 gnomAD variants with larger allele counts. The remaining 68,300,547 variants were not found in gnomaAD. Note that at the time of data collection gnomAD v2 was the most current version available. The current v3 version of gnomAD is only slightly different in relevant content as its reference genome, GRCh38, recapitulates 99% of GRCh37 (Pan, et al., 2019) and most differences between the two are in the non-coding regions, an area outside this study. We thus expect that results and conclusions reported here would not change with this update.

The allele counts of all nsSNVs causing the same single amino acid substitution (SAV) were further aggregated to represent the frequencies of individual SAVs (**Eq. 1**):

$$\mathrm{freq}\,(\mathrm{SAV}) = \frac{\sum_{i=1}^{k} n_i}{N}, \quad (1)$$

where for any codon, $n_1 \ldots n_k$ are counts of the specific SAV-causing alleles and N is the total numbers of sequenced alleles of that codon. Note that in the process of aggregation some *observed* (allele count >1) SAVs could be derived from the aggregation from multiple single allele nsSNVs. The aggregation of nsSNV frequencies into SAV frequencies, resulted in 2,564,652 *observed* (allele count >1), 2,918,355 *singletons* (allele count =1), and 60,601,329 *synthetic* SAVs in the 19,971 transcripts. Observed variants were further classified as *common* (*freq* (SAV) ≥ 0.01) and *rare* (*freq* (SAV) <0.01).

Collection of Cross-Species Variants

Cross-species variants (CSVs) are the amino acid differences between the human reference protein sequence and the orthologous protein sequence of another species. For example, if the amino acid residue at the third position of the human protein sequence *P* is leucine, and if the amino acid residue at the same position in mouse orthologous protein sequence is glycine, then the CSV at this position in *P* would be L3G. Aiming to span the tree of life with species available in Ensembl BioMart (GRCh37), we considered 20 species for CSV analysis: yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabdiis elegans*), fruitfly (*Drosophila melanogaster*), zebrafish (*Danio rerio*), xenopus (*Xenopus laevis*), anole lizard (*Anolis carolinensis*), chicken (*Gallus gallus*), platypus (*Ornithorhynchus anatinus*), opossum (*Monodelphis domestica*), dog (*Canis familiaris*), pig (*Sus scrofa*), dolphin (*Tursiops truncatus*), mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), tree shrew (*Tupaia belangeri*), tarsier (*Carlito syrichta*), gibbon (*Nomascus leucogenys*), gorilla (*Gorilla gorilla*), bonobo (*Pan paniscus*), and chimpanzee (*Pan troglodytes*). We identified the evolutionary distances of these species from *Homo sapiens* using the TimeTree database (Kumar, et al., 2017). All protein coding DNA sequences (CDS) of these 20 species were downloaded from the Ensembl database (Zerbino, et al., 2018) (release 94, https://uswest.ensembl.org/info/data/ftp/index.html). For every human protein coding transcript *T*, the available orthologous CDS for each of the 20 species was extracted using the Ensembl BioMart (Kinsella, et al., 2011). Each species may have multiple protein coding sequences orthologous to *T*, but only the longest one was selected. We performed multiple sequence alignment (MSA) of *T* and all its orthologs using PRANK (Löytynoja and Goldman, 2005), which translates CDS and aligns protein sequences. Of the 19,971 human transcripts in our set, 1,342 had a full set of the 20 species orthologs in the MSA. In these transcripts (940,328 amino acids) there were 183,540 *observed* (49,541 CSVs/133,999 non-CSVs), 228,774 *singleton* (52,550 CSVs/176,224 non-CSVs), and 5,118,164 *synthetic* SAVs (873,011 CSVs/4, 245,153 non-CSVs).

Cross-Species Variant Effect Predictions

We generated SNAP (Bromberg and Rost, 2007) predictions for all variants in the 1,342 transcripts. SNAP predictions could be made for 1,265 of the proteins; a set of 77 sequences (832,697 variants) did not yield any predictions due to SNAP's sequence length constraints (63 sequences), variant to sequence mapping errors (3 sequences), and unresolvable errors in the SNAP input feature extraction pipeline (11 sequences) as well as an additional 46,840 variants on the remaining proteins. Note that, as in all other proteins in our set, the vast majority (93%) of these variants were *synthetic* (4% *singleton* and 3% *observed*), suggesting that our analyses of effect trends should be largely unaffected by this missing subset. Thus, the final SNAP effect prediction dataset contained 4,650,941 variants in 791,040 positions among 1,265 proteins (**Supplementary Table S1**). Note that for this study we used the original SNAP tool instead of the more recent version SNAP2 (Hecht, et al., 2015). There were two reasons for this choice: 1) SNAP2 used OMIM (Amberger, et al., 2009) disease variants in training, a choice which does not directly reflect variant functional effects, and 2) SNAP effect prediction reliability scores strongly correlate with the functional effect strength (Bromberg, et al., 2013), an observation that has not been explicitly made for SNAP2.

Variant Conservation Scores

For all residues of all proteins in our set we computed two types of conservation scores:

1. We used the PredictProtein pipeline (Yachdav, et al., 2014) to compute ConSurf (Glaser, et al., 2003) conservation scores. ConSurf scores are based on MSAs of up to 150 homologous sequences. Reported scores are normalized so that the average score over all residues of one protein is zero and the standard deviation is one. Lower scores indicate more conserved residues.
2. We extracted from the list of SNAP input features the position-specific independent counts (PSIC) (Sunyaev, et al., 1999). PSIC scores reflect per-residue position-specific weights considering the MSA-based overall level of sequence similarity.

We only retained the conservation scores for those variant positions (104,375) that had both ConSurf and PSIC annotations.

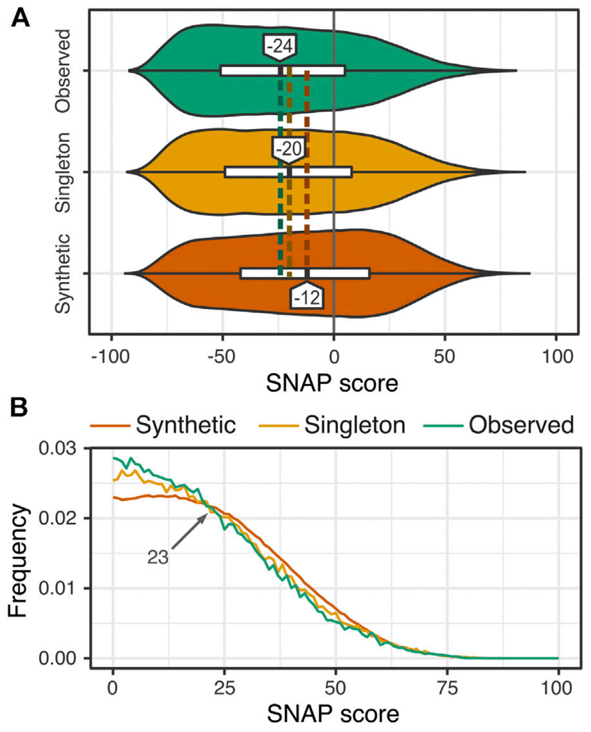


FIGURE 1 | Higher prevalence of effect among the synthetic as compared to observed and singleton variants. **(A)** The distribution of effect predictions for synthetic variants (dark orange; median SNAP = -12) is significantly more right-shifted toward effect (SNAP ≥0; horizontal line) than that of observed variants (green; median SNAP = -24) and singletons (yellow; median SNAP = -20). For all distributions, however, the majority of predictions are neutral (SNAP <0) **(B)** Additionally, synthetic variants show an enrichment of moderate to severe functional effects (SNAP ≥ 23) vs. singletons and observed variants.

Conservation scores across variant subsets were used only once per variant position in the subset. That is, if two rare CSVs were present at one protein position, conservation for this position was only used once toward establishing the distribution of the rare CSV dataset. On the other hand, if a position contained both a common CSV and a rare CSV, the conservation score was included separately into distributions of each subset.

Per-Residue Funtrp Scores

funtrp (Miller, et al., 2019) is a prediction tool that assesses the expected range of functional effects due to the possible variants at a given protein position. funtrp classifies sequence positions as *neutral* (most variants at this position show weak or no effect), *rheostatic* (a full range of variant effects) and *toggle* (most variants have a severe effect). funtrp was trained with deep-mutagenesis data and uses sequence-based features to differentiate between the three residue classes. We used our publicly available webservice (https://services.bromberglab.org/funtrp) to identify funtrp classes for each position of 1,254 of our protein sequences; predictions for the remaining 11 sequences were not returned by the method.

Evaluating Statistical Significance Distribution Differences

For all comparisons of score distributions (e.g. SNAP scores) across variant classes (e.g. rare vs common), we re-sampled said distributions 1,000 times to extract 1,000 observations each time. For each resampling instance, we performed the Kolmogorov-Smirnov test to test the equity of the distributions, reporting the associated *p*-value; the median p-val over 1,000 iterations was reported.

RESULTS AND DISCUSSION

Many Variants Remain to be Sequenced

Single amino acid variant (SAV) effects were determined by SNAP (Bromberg and Rost, 2007) (predicted score range for our variants [−94, + 88]), with negative scores identifying neutral SAVs (no change in function) and positive scores identifying non-neutrals/effect SAVs (activating or deactivating changes in function); score absolute values indicate the reliability of prediction and, for non-neutral variants, the size of the effect (Bromberg, et al., 2013). Note that our definition of effect does not specify whether the effect is detrimental or beneficial to the organism, but rather reports on the change in wild-type functionality of the affected protein.

Overall, more variants were predicted to be neutral than effect, with some difference in fractions of effect variants between *synthetic*, *singleton*, and *observed* variant subsets (**Supplementary Table S1**). The distribution of synthetic variant SNAP scores was significantly different from that of singleton and observed variant scores (Kolmogorov-Smirnov, KS, test *p*-value; synthetic vs. singleton = 8.7*e*−04, synthetic vs. observed = 1.1e−06), while singleton and observed scores were only slightly different (singleton vs. observed *p*-val = 0.14). For *synthetic* variants (median SNAP score = −12; **Figure 1A**), i.e. those that have not been seen in the population, the majority (60%) were predicted to be neutral. These variants are, thus, technically *observable* and may be identified in future sequencing efforts. Those 40% of the synthetic variants predicted to have an effect, had on average more severe impact than the effect variants seen in the human population (combined *observed* and *singleton* sets; 31% effect; **Figure 1B**). Increased predicted effect of synthetic variants is in line with the expectation that these are subject to purifying selection.

Earlier (Bromberg, et al., 2013), we observed a similar trend of more effect variants in the *synthetic* than in the *observed/singleton* set; i.e. 55% effect in *synthetic* SAVs in 100 randomly selected enzymes vs. 46% effect variants in 1000Genomes data (Auton, et al., 2015). However, the fractions of both the *synthetic* and *observed/singleton* of effect variants in our earlier study were significantly higher than the corresponding numbers reported here. Furthermore, the SNAP scores of the *synthetic* variants reported here and those in Bromberg were significantly different (*p*-val 4.0*e*−15); the scores of our combined *observed/singleton* variants also differed from the scores of 1000Genomes variants (*p*-val = 3.0e−12).

While 1000Genomes variants were observed in 85% (1,072 of 1,265) of the transcripts used in this study, our variant set for these proteins was larger, suggesting improved sequencing coverage and accounting for some effect prediction differences. Notably, only 36% of the 1000Genome variants in our proteins had an effect–in line with the 31% effect variants in our *observed/singleton* set and 10% less than in the complete 1000Genomes variant set. Furthermore, of the set of 100 enzymes used in the Bromberg et al. study to generate synthetic variants, only four were present in our protein set. Thus, the difference in effect scores between our earlier study and the current work is most likely due to the specific genes/proteins selected for this study. Genes/proteins in our set have orthologs in each of our selected species, i.e. these are likely ancient and rarely disease-associated (Moreau and Tranchevent, 2012). As the functions of these proteins are important for organism survival, they likely harbor the variants necessary for environment-driven functional adaptation but do not allow for severe disruption upon mutation (Key, et al., 2014; Key, et al., 2014; Ilardo and Nielsen, 2018; Rees, et al., 2020). While the variants in these proteins may still be extremely deleterious, less than three percent in our set were of severe effect (SNAP score ≥50; 130,870 variants; 7.2% of all effect variants) and, as expected, most were *synthetic* (123,962 variants, 3% of all *synthetic*), with few found in the population (6,908 variants, 2% of all *singleton*/*observed*).

Given these fractions of effect variants, we expect at least half a million (neutral *synthetic* CSVs) and possibly over four million (any *synthetic* neutrals and milds/moderates) variants to be possibly observable, i.e. they may be found with more sequencing. As the genes considered here are likely ancient and evolutionarily optimized to resist drastic changes upon mutation, this 12-fold possible increase in the observable variants (vs those already observed) suggests an upper bound of increase in the number of *observed*/*singleton* SAVs that may be collected in the future.

Common Variants May Drive Environmental Adaptations

Despite the fact that common variation is, by definition, widespread in the population, trivially, the vast majority of unique population variants are rare. Variant effect trends are therefore dominated by observations for rare variants, effectively drowning out signal from common variants. We thus aimed to elucidate the difference between common (≥1% SAV frequency) and rare variants. For this part of the analysis we excluded from consideration the *singleton* variants, which are a special case of rare variation and may be disproportionately sequencing errors. We note that common variants are unlikely to be very deleterious/disease-causing as they would not stay common. On the other hand, variants that have no impact on function (*neutrals*) and very weak nonneutrals can be fixed in the population at about the same rate via genetic drift (Kimura and Ohta, 1969).

We also considered the differences between *observed* cross-species variants (CSVs) and non-CSVs (Methods). We expected different evolutionary drivers for the existence of different variant types (e.g. common CSV vs. rare non-CSV) and, in turn, potential differences in their impact on protein function. Note that variants labeled as non-CSV may still be present in the orthologs of species that were not assessed here. However, using more species could also reduce our total protein set if some of the currently used transcripts are absent in the new species transcriptome.

TABLE 1 | Prevalence of human reference amino acids in CSV positions across orthologs.

	Rare		Common	
	CSV (%)	**Non-CSV (%)**	**CSV (%)**	**Non-CSV (%)**
Apes	98	99	59	98
Mammals	83	93	48	86
All	68	83	40	75

Common variants are as frequently CSVs as non-CSVs (691 CSVs vs. 683 non-CSVs, **Supplementary Table S2**). For common CSVs (*reference* substituted by *variant* amino acid), the human reference amino acid is present in a minority (40%) of all 20 species orthologs, but more frequently in mammals (48%) and great apes (59%; **Table 1**). Note that these fractions were computed as the number of shared reference amino acids of all residues aligned, e.g. if for one variant ten of 15 orthologs aligned at the variant position have the human reference amino acid, while for another variant four of the 20 orthologs do, the total fraction of reference amino acid across these variants is 40% (14/35). Given these fairly low fractions, the *variant* amino acids of common CSVs are possibly ancestral, i.e. human *variant* amino acid could have been the *reference* of a potential ancestor. Thus, for humans reinstating the ancestral residue at this position is likely to be detrimental, as it would otherwise remain fixed as reference.

For common non-CSVs the corresponding fractions of reference amino acids across orthologs are 75% (all), 86% (mammals), and 98% (apes; **Table 1**). Thus, *variant* amino acids of common non-CSVs likely represent somewhat newer evolutionary developments and are 1) likely to be beneficial (still effect!) for humans as a whole but may have not been around long enough to become the reference or 2) are non-universal adaptations to persistent environmental conditions, e.g. ethnicity-specific variants (Rees, et al., 2020).

Unlike common variants, rare CSV variants are nearly three-fold less commonplace than non-CSVs. However, just as for common variants, rare non-CSV *reference* amino acids are present in orthologs at a higher frequency than CSV references (83% non-CSV vs. 68% CSV). The preponderance of non-CSV reference amino acids across all species highlights these variants as likely of recent origin, and therefore possibly of any amount (a full range) of effect. Rare CSV variant amino acids, on the other hand, may be ancestral, although the likelihood of this is greatly diminished as compared to common variants (68% rare vs. 40% common reference amino acid across orthologs). If they are ancestral, their extensive elimination from the population would suggest deleterious effects (purifying selection). Independent appearances of the variant in human (as rare variant) and in another species (as reference) is unlikely, but also possible. In this case, the variant amino acid would likely be neutral or slightly deleterious in human.

Further comparing the frequencies of occurrence of reference amino acids across orthologs suggests that rare variants occur at more conserved positions than common variants; reference amino acids of CSVs *vs.* non-CSVs were present across all species for 68% vs. 83% for rare variants and 40% vs. 75% for common ones. Evaluation of

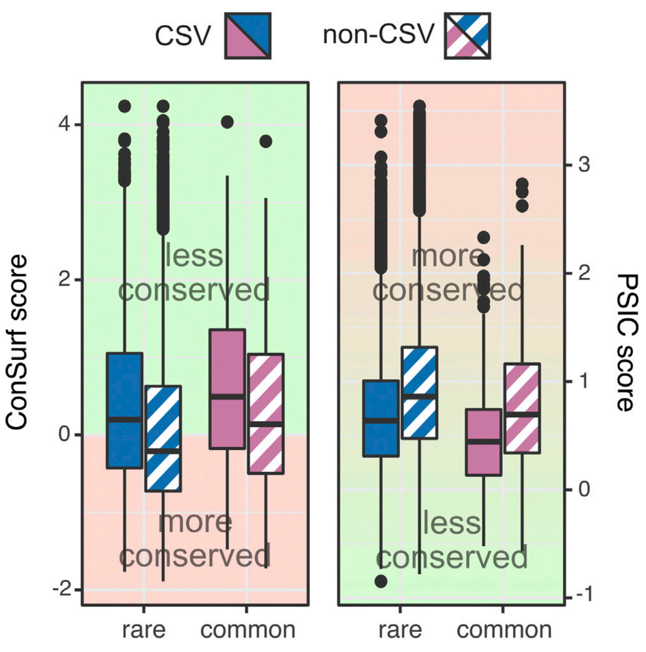


FIGURE 2 | Rare variants more frequently found in conserved protein positions. Rare variants (blue) are more frequently found in conserved positions (ConSurf ≤0) than common variants (purple). Furthermore, non-CSVs (hatched fill) are more frequently present in conserved positions than CSVs (solid fill). Similarly, rare variants carry higher PSIC scores than common variants.

conservation of variant positions using ConSurf (Glaser, et al., 2003) confirmed this observation (**Figure 2**; lower score means more conserved position; KS p-val CSV rare vs. common = 3.2*e*–08, non-CSV rare vs. common = 1.5*e*–09). The protein positions harboring rare variants were on average more conserved (103,609 positions; median ConSurf score = −0.11) than positions with common variants (1,013 positions; median ConSurf score = 0.36). Note that there are only a few 247) positions for which both rare and common variants are present, and these are also only weakly conserved (median ConSurf score = 0.34). A similar trend was observed using PSIC scores (Sunyaev, et al., 1999) of variant positions (**Figure 2**; higher score means more conserved position; median PSIC score of: rare = 0.80, common = 0.55, both = 0.59; KS *p*-val CSV rare vs. common = 4.4*e*–16, non-CSV rare vs. common = 4.2*e*–07).

This is an unexpected result, as variants in conserved positions are often assumed to have an effect, while rare variants, both CSV and non-CSV, are less frequently predicted to have an effect than the corresponding common variants (rare *vs.* common effect variants: 10% *vs.* 20% CSVs and 36% *vs.* 40% non-CSVs; **Supplementary Table S2**). Here we point out that more severe effect (several high score outliers) vs. more frequent effects (many variants have some effect) indicate different score distributions but may result in similar summary statistics (e.g. distribution means). Thus, although common variants have an effect more frequently than rare variants (**Figure 3A**), the former are less frequently severe (SNAP ≥50; 6% rare vs 3.6% common effect variants; **Figure 3B**). Furthermore, rare non-CSVs are enriched in moderate effect variants (SNAP ≥25) vs. common non-CSVs that are mostly mild. Common CSVs, on the other hand, carry more moderate effects than rare CSVs (**Figure 3B**). Note that as CSVs in general score tend to be predicted neutral more often than non-CSVs (**Supplementary Figure S1**, the preponderance of high-scoring common CSVs vs. non-CSVs reinforces the likely adaptational value of common CSVs proposed above. The propensity of rare variants to cause severe effects highlights them as likely culprits of disease. However, rare variants make up nearly three quarters of variation overall and are clearly not restricted to being disease-causing. In fact, they cover a complete range of effect–from strongly effect to reliably neutral (**Figure 3A**).

In an effort to validate our observations of effects of common variants we used funtrp (Miller, et al., 2019) – a method that trained to recognize the range of variant effects possible at a single protein position. It classifies positions into 1) neutrals, where most variants have no effect on protein function, 2) toggles, where

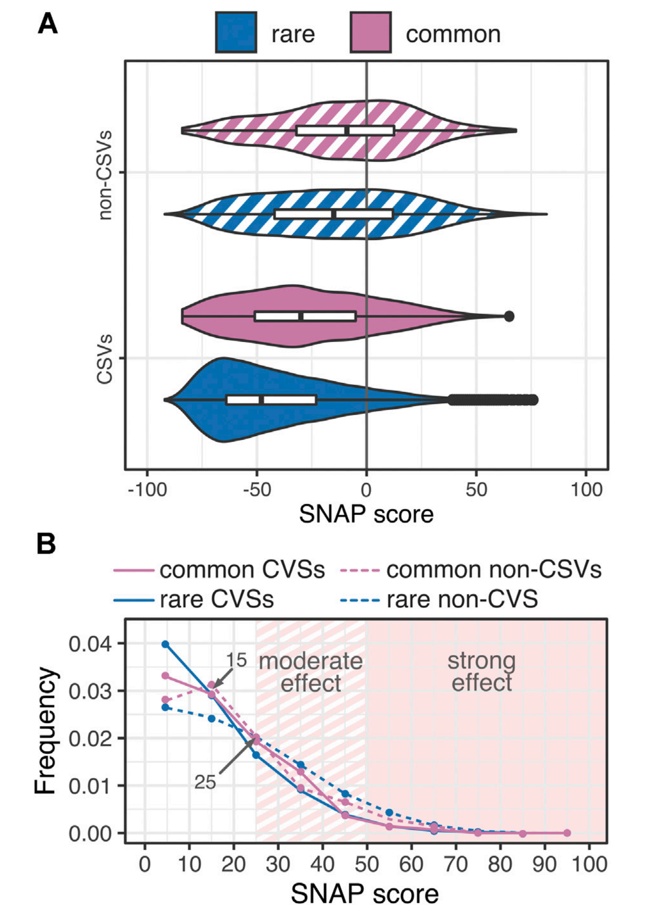


FIGURE 3 | Common variants are more frequently effect than rare variants, but rare variants are more frequently severe. **(A)** The distribution of common CSV and non-CSV predictions (purple) is more right-shifted (more effect) than that of rare variants (blue). Furthermore, **(B)** rare non-CSVs (blue dashed line) are more often of moderate and severe effect than common non-CSVs (purple dashed line). However, common CSVs are more often of mild-moderate effect than rare CSVs. Due to small numbers of variants at each SNAP score (*x*-axis), frequencies are calculated in intervals of 10, e.g. 0 ≤ SNAP <10; points are centered in the interval.

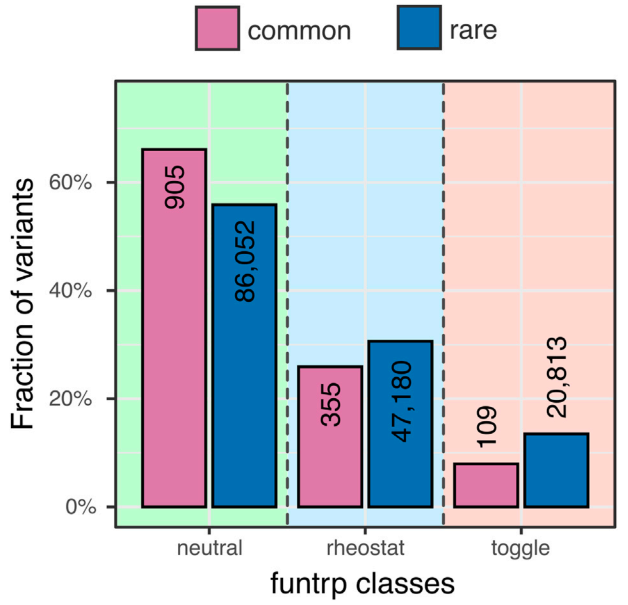


FIGURE 4 | Common variants prefer neutral positions more than rare variants. Neutral positions (green shading) are enriched in common variants (purple) as opposed to rare variants (blue) (66% vs. 56% - actual variant counts shown as numbers in the bars). The fraction of rare variants in rheostatic positions (blue shading) is higher than the corresponding fraction of common variants. However, the ratio of common variants in rheostat positions vs. toggles (pink shading) is higher than that of rare variants.

most variants have severe or knockout effects, and 3) rheostats, where variants cover a range of effect strengths. Overall, funtrp classes reflected SNAP predictions well; median SNAP scores of variants in neutral, rheostat, and toggle positions were −33, −18, and 12, respectively. Common variants were more often found in neutral positions as compared to rare ones (66% vs. 56%, **Figure 4A**). However, of the effect positions (i.e. rheostat and toggle), common variants preferred rheostats (77% common vs. 69% rare variants). As most toggles are conserved (Miller, et al., 2019), this observation is in line with the above finding that rare variants 1) are more likely than common ones to be in conserved positions and 2) that they carry more severe functional effects. Common variants in rheostatic positions, on the other hand, were likely used in evolution to fine-tune functions of affected proteins.

Variant Effect Reflects Evolutionary Time of Reference Amino Acid Origin

We asked whether variant effect is related to the likely evolutionary time of appearance of the human reference. For each species *X*, we collected all effect variants in our dataset where the human and *X* reference amino acids were identical. For mammals, the median effect strengths of the variants affecting these positions were similar. For other species, however, the variant effect was correlated with increasing evolutionary distance between human and the specific species (**Figure 5**). This correlation held true for CSVs and non-CSVs, as well as for *synthetic*, *singleton* or *observed* variants.

Notably for non-CSVs, median effect scores increased more rapidly over evolutionary time than for CSVs. This trend was expected, as variants whose reference amino acids are present in evolutionarily distant species likely disproportionately affect conserved ancestral amino acids. For example, a shared human and yeast reference amino acid is likely present across all or most species in our set. Thus, a CSV at this position (if say, fly amino acid is different) would indicate some flexibility at the position, but a non-CSV would elicit the functional effect associated with the disruption of stringent conservation. However, we found that conservation of the variant position is unlikely the sole contributor to the observed effect gradient. The trend, albeit less pronounced, remained visible if only the variants in positions of low conservation (ConSurf score ≥0.5) were used in the analysis (**Supplementary Table S2**). Importantly, a clear distinction between CSVs and non-CSVs was also still evident, indicating that even in non-conserved positions CSVs and non-CSVs are distinguishable.

Self-Fulfilling Prophecy: Are Cross-Species Variants Really Neutral?

As mentioned previously, CSVs were less often predicted to have an effect than mutations to an amino acid that is not present in other species (non-CSVs); this observation was true for both *synthetic* and observed human variation (**Supplementary Table S1**). The absolute difference in median SNAP scores between CSVs and non-CSVs was 38 (mean =30) for *synthetic* variants and 32 (mean =27) for the *observed*–a full 14–21% of the entire scoring range ([−94, + 88]). CSV scores are most often neutral across all three categories of variation (i.e. *synthetic*, *singleton*, *observed*), while the distribution of non-CSV scores is much more widespread (**Figure 6**). An biological explanation for this observation is that CSVs are indeed more likely to be neutral with respect to protein function, as is expected from their persistence in homologs (Kondrashov, 1995; Sunyaev, et al., 2001). However, another explanation for this stark difference could then be the fact that SNAP was trained using a dataset of cross-species orthologous enzyme variants deemed neutral. Only 30 of these enzymes were in our set of 1,265 proteins and, thus, are not expected to dramatically impact our observations. However, if SNAP learned input feature patterns specific to CSVs, others could be labeled neutral without ever being seen in training. Thus, SNAP could fail to recognize CSVs that have a functional impact without introducing the organism to selection pressures, i.e. functionally non-neutral, but physiologically neutral. In fact, these may be the so called “fuel for evolution” (Bromberg, et al., 2013; Fu, et al., 2013) – the pool of weakly nonneutral variants necessarily present in the population for the purposes of quick adaptation to a changing environment.

In our earlier work we had determined a SNAP threshold of 23 as the upper functional impact limit to the absence of physiological visibility. We have confirmed this threshold for this data set as well, as the score where the fraction of possible/expected variants exceeds those observed (**Figure 6**). Of the *observed* effect CSVs, 76% are in this mild functional effect range, while 58% of all effect non-CSVs are as well. This significantly larger fraction of mild effect CSVs than effect

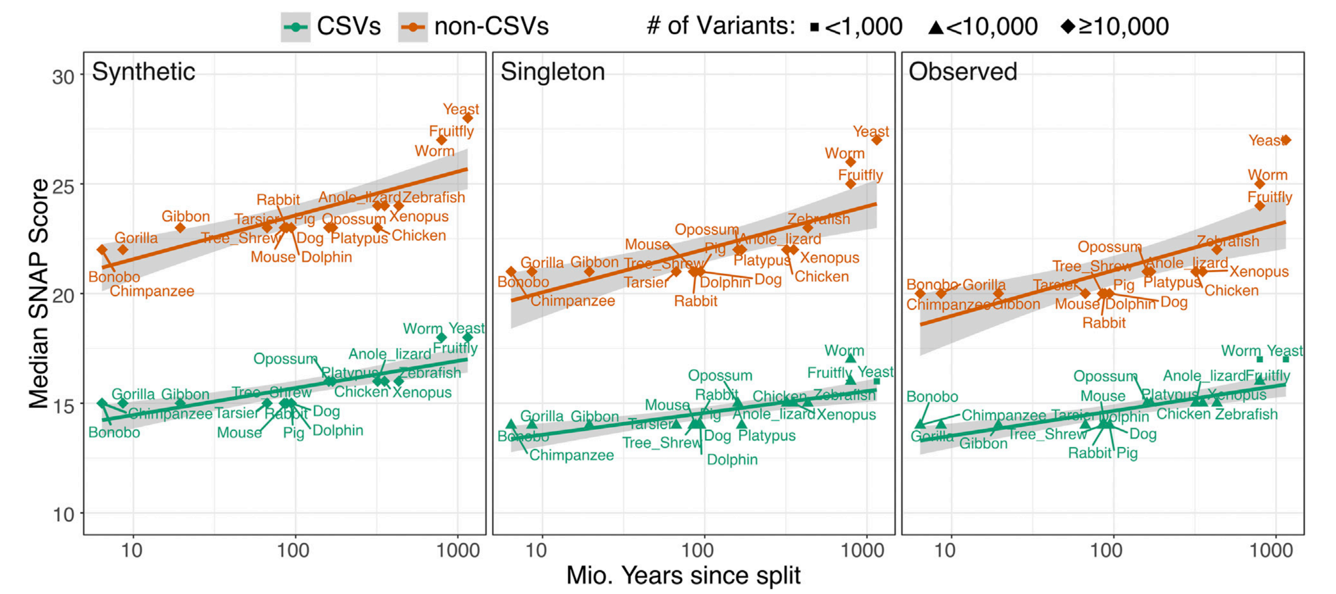


FIGURE 5 | Impact of variants sharing reference amino acids with other species correlates with evolutionary distance. Mean SNAP scores (y-axis) are computed for CSV (green line) and non-CSV (red line) synthetic (left panel), singleton (middle panel), and common (right panel) variants, according to per-species human-shared reference amino acids. Species are placed along the x-axis (logarithmic) according to their distance to ancestor shared with human.

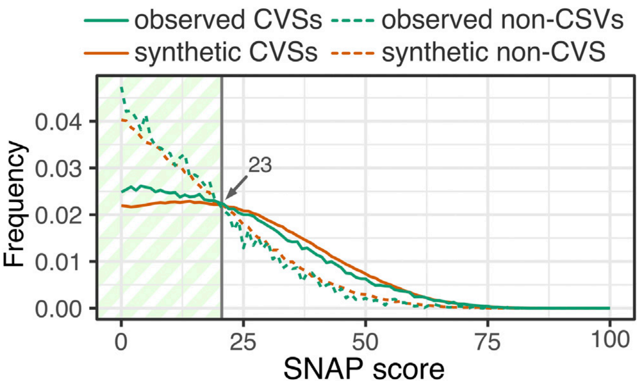


FIGURE 6 | Observed variants enriched in mild effects. Both observed CSVs (green solid line) as well as non-CSVs (green dashed line) are enriched in mild effect variants over their synthetic counterparts (orange CSVs–solid line, non-CSVs–dashed line).

non-CSVs suggests that the former are more likely the functional variants necessary for adaptation.

Although CSVs are more frequently (vs non-CSVs) predicted to be mild in effect, they also vastly outnumber non-CSVs in the neutral score range. Curiously, there is almost no difference between the *synthetic* and *observed* CSV score distributions. However, only 5% of all possible CSVs in our set are observed in the human population–not much more (percentage-wise) than all possible non-CSVs (3%; and fewer in the absolute sense with ~42K *observed* CSVs and ~144K *observed* non-CSVs). It thus remains unclear whether functional constraints are indeed weaker for (often biochemically similar substitutions of amino acids in) CSVs.

Evaluating prediction bias is difficult in the absence of a gold-standard data set and one of neutral CSVs doesn't exist. While funtrp uses site conservation as input, it was not trained to recognize individual variant effect and thus could be used to elucidate our findings. In other words, funtrp forgoes the broad generalization of assigning neutrality to cross-species variants on the basis of the evolution-guided inference (e.g. SNAP and other methods (Ng and Henikoff, 2001; Ng, 2003; Calabrese, et al., 2009; Adzhubei, et al., 2010; Shihab, et al., 2013; Kircher, et al., 2014; Schwarz, et al., 2014; Pejaver, et al., 2020).

In line with our earlier observations, funtrp found that most protein positions in our set are neutral. The distribution of *synthetic*, *singleton*, and *observed* variants across position classes was very similar for CSVs (62/30/8% neutral/rheostat/toggle; **Supplementary Table S3**). Non-CSVs maintained an average 50/33/17% ratio of neutrals/rheostats/toggles, with *observed* non-CSVs more frequently found in neutral and rheostat positions than *singletons* or *synthetic* variants (**Supplementary Table S3**). Thus, both CSVs and non-CSVs were about as likely to localize to rheostatic positions, but non-CSVs were less frequently found in neutrals and twice as often in toggles. Note that while not all variants in neutral positions are necessarily functionally neutral, and non-neutral positions may have some neutral variants, only 62% of observed CSVs are found in neutral positions, while SNAP predicts 90% of observed CSVs to be functionally neutral.

Two conclusions from these results are salient: 1) as expected, CSVs are indeed more frequently neutral than non-CSVs and 2) it appears that SNAP (and likely other predictors) tends to overestimate CSV neutrality. Thus, we suggest that cross-species variants may carry mild to moderate functional effects and should be evaluated accordingly.

CONCLUSION

We investigated a set of single amino acid substitutions (SAVs) in evolutionarily persistent, likely ancient, proteins, i.e. those that we expect to be optimized to tolerate variation. We found that despite the enrichment in severe effects of *synthetic* vs *observed* variants, a large proportion of SAVs might still be found upon broader sequencing of the population. Moreover, we expect that only a small fraction of variants that have yet to be sequenced will have a severe impact and/or be disease causing. We further observed that common variants favor poorly conserved sites. This lower conservation, indicative of more tolerance toward variation, might be providing enough "wiggle" room for environmental adaptations. Rare variants are, on the other hand, are often found in more conserved positions, explaining their enrichment in severe effects in comparison to common SAVs. Curiously, it appears that our ancient proteins have been optimized to the point where disrupting a conserved site does not immediately cause a functional disruption, as seen in the majority of rare variants predicted to be neutral. Finally, we suggest that cross-species variants (CSVs) might indeed be more often neutral than non-CSVs however not as consistently as currently expected. Ultimately, however, this question can only be answered through the development of an effect predictor that is does not make a priori assumptions of CSV neutrality and, which is somewhat harder, does not rely on conservation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YB and YM conceived the presented ideas and designed the study. YM conducted the variant effect analysis. MM produced the effect predictions. ZZ produced the variant and cross-species dataset. YM, MM, and ZZ contributed equally to this work as first authors. All authors contributed to writing the manuscript.

FUNDING

YB, MM, YM, and ZZ were supported by the NIH (National Institutes of Health) grant U01 GM115486 and YB was also supported by the NIH grant R01 MH115958.

ACKNOWLEDGMENTS

We would like to thank Burkhard Rost (Technical University Munich) for all discussions of the topics covered here. We are grateful to Nir Ben-Tal (Tel-Aviv University), Daisuke Kihara (Purdue University), and Arun Prasad Pandurangan (Cambridge Biomedical Campus) for the invitation to contribute our thoughts to this collection of articles. Many thanks go to the reviewers, whose comments and suggestions helped improve the paper. Finally, we would also like to express gratitude to all researchers working in genome analysis and to the people who have made their data available to science.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.635382/full#supplementary-material.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7 (4), 248–249. doi:10.1038/nmeth0410-248

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's Online mendelian inheritance in man (OMIM). *Nucleic Acids Res.* 37, D793–D796. doi:10.1093/nar/gkn665

Auton, A., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Briscoe, A. D., Gaur, C., and Kumar, S. (2004). The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. *Gene* 332, 107–118. doi:10.1016/j.gene.2004.02.037

Bromberg, Y., Kahn, P. C., and Rost, B. (2013). Neutral and weakly nonneutral sequence variants may define individuality. *Proc. Natl. Acad. Sci. USA* 110 (35), 14255. doi:10.1073/pnas.1216613110

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35 (11), 3823–3835. doi:10.1093/nar/gkm238

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* 30 (8), 1237–1244. doi:10.1002/humu.21047

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33 (Suppl. l–2), W306–W310. doi:10.1093/nar/gki375

Capriotti, E., and Fariselli, P. (2017). PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res.* 45 (W1), W247–W252. doi:10.1093/nar/gkx369

Choi, Y., and Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31 (16), 2745–2747. doi:10.1093/bioinformatics/btv195

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., et al. (2011). Modernizing Reference Genome Assemblies. *PLoS Biol.* 9 (7), e1001091. doi:10.1371/journal.pbio.1001091

Durbin, R. M., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467 (7319), 1061–1073. doi:10.1038/nature09534

Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* 11 (8), 801–807. doi:10.1038/nmeth.3027

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493 (7431), 216–220. doi:10.1038/nature11690

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., et al. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19 (1), 163–164. doi:10.1093/bioinformatics/19.1.163

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst* 6 (1), 116–e3. doi:10.1016/j.cels.2017.11.003

Hassan, M. S., Shaalan, A. A., Dessouky, M. I., Abdelnaiem, A. E., and ElHefnawi, M. (2019). A review study: computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. *Gene* 680, 20–33. doi:10.1016/j.gene.2018.09.028

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC genomics* 16 Suppl 8 (S8), S1. doi:10.1186/1471-2164-16-S8-S1

Ilardo, M., and Nielsen, R. (2018). Human adaptation to extreme environmental conditions. *Curr. Opin. Genet. Development* 53, 77–82. doi:10.1016/j.gde.2018.07.003

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581 (7809), 434–443. doi:10.1038/s41586-020-2308-7

Key, F. M., Peter, B., Dennis, M. Y., Huerta-Sánchez, E., Tang, W., Prokunina-Olsson, L., et al. (2014) Selection on a variant associated with improved viral clearance drives local, adaptive pseudogenization of interferon lambda 4 (IFNL4). *Plos Genet.* 10 (10), e1004681. doi:10.1371/journal.pgen.1004681

Key, F. M., Teixeira, J. C., de Filippo, C., and Andrés, A. M. (2014) Advantageous diversity maintained by balancing selection in humans. *Curr. Opin. Genet. Dev.* 29:45–51. doi:10.1016/j.gde.2014.08.001

Kimura, M., and Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* ;61 (3), 763-771.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030. doi:10.1093/database/bar030

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46 (3):310. doi:10.1038/ng.2892

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* 99(23), 14878–14883. doi:10.1073/pnas.232565499

Kondrashov, A. S. (1995). Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.* 175 (4), 583–594. doi:10.1006/jtbi.1995.0167

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4 (7), 1073–1081. doi:10.1038/nprot.2009.86

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34 (7), 1812–1819. doi:10.1093/molbev/msx116

Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci.* 102 (30), 10557–10562. doi:10.1073/pnas.0409137102

Lander, E. S., Linton, L. M., Birren, M., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536 (7616), 285–291. doi:10.1038/nature19057

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Mahlich, Y., Reeb, J., Hecht, M., Schelling, M., De Beer, T. A. P., Bromberg, Y., et al. (2017). Common sequence variants affect molecular function more than rare variants? *Sci. Rep.* 7 (1), 1608–1613. doi:10.1038/s41598-017-01054-2

Malhis, N., Jones, S. J. M., and Gsponer, J. (2019). Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat. Commun.* 10 (1), 1556. doi:10.1038/s41467-019-09583-2

Miller, M., Vitale, D., Kahn, P. C., Rost, B., and Bromberg, Y. (2001). Funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.* 47 (21), e142. doi:10.1093/nar/gkz818

Miller, M. P., and Kumar, S. (2001). Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* 10 (21), 2319–2328. doi:10.1093/hmg/10.21.2319

Moreau, Y., and Tranchevent, L. C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* 13 (8), 523–536. doi:10.1038/nrg3253

Ng, P. C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* 11(5):863–874.doi:10.1101/gr.176601

Ng, P. C., and Henikoff, S. (2019). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31 (13):3812–3814. doi:10.1093/nar/gkg509

Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., et al. (2019) Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* 20 (2), 101. doi:10.1186/s12859-019-2620-0

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., et al. 2020. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* 11 (1), 5918. doi:10.1038/s41467-020-19669-x

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., et al. (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. doi:10.1101/134981

Peterson, T. A., Doughty, E., and Kann, M. G. (2013). Towards precision medicine: advances in computational approaches for the analysis of human variants. *J. Mol. Biol.* 425 (21), 4047–4063. doi:10.1016/j.jmb.2013.08.008

Quang, D., Chen, Y., and Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31 (5), 761–763. doi:10.1093/bioinformatics/btu703

Rees, J. S., Castellano, S., and Andrés, A. M. (2020). The genomics of human local adaptation. *Trends Genet.* 36 (6), 415–428. doi:10.1016/j.tig.2020.03.006

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11 (4), 361–362. doi:10.1038/nmeth.2890

Shen, H. (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four caucasians. *PLoS One* 8 (4), e59494. doi:10.1371/journal.pone.0059494

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* ;34 (1), 57–65. doi:10.1002/humu.22225

Subramanian, S., and Kumar, S. (2006). Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC genomics* 7 (1), 306. doi:10.1186/1471-2164-7-306

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10 (6): 591–597. doi:10.1093/hmg/10.6.591

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* 12 (5): 387–394.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19 (9):581–590. doi:10.1038/s41576-018-0018-x

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2018) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101 (1):5–22. doi:10.1016/j.ajhg.2017.06.005

Waterston, R. H., Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520–562. doi:10.1038/nature01262

Wells, J. A., Vasser, M., and Powers, D. B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene* 34 (2–3), 315–323. doi:10.1016/0378-1119(85)90140-4

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., et al. (2014) PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* 42 W337–W343. doi:10.1093/nar/gku366

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, R., Bhai, J., et al. (2018). Ensembl. *Nucleic Acids Res.* 46 (D1), D754–D761. doi:10.1093/nar/gkz966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer PLM declared a past co-authorship with one of the authors YB to the handling editor.

frontiers in Molecular Biosciences

METHODS
published: 25 March 2021
doi: 10.3389/fmolb.2021.620475

ThermoScan: Semi-automatic Identification of Protein Stability Data From PubMed

Paola Turina[1], *Piero Fariselli*[2]* and *Emidio Capriotti*[1]*

[1]Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, Italy, [2]Department of Medical Sciences, University of Torino, Torino, Italy

During the last years, the increasing number of DNA sequencing and protein mutagenesis studies has generated a large amount of variation data published in the biomedical literature. The collection of such data has been essential for the development and assessment of tools predicting the impact of protein variants at functional and structural levels. Nevertheless, the collection of manually curated data from literature is a highly time consuming and costly process that requires domain experts. In particular, the development of methods for predicting the effect of amino acid variants on protein stability relies on the thermodynamic data extracted from literature. In the past, such data were deposited in the ProTherm database, which however is no longer maintained since 2013. For facilitating the collection of protein thermodynamic data from literature, we developed the semi-automatic tool ThermoScan. ThermoScan is a text mining approach for the identification of relevant thermodynamic data on protein stability from full-text articles. The method relies on a regular expression searching for groups of words, including the most common conceptual words appearing in experimental studies on protein stability, several thermodynamic variables, and their units of measure. ThermoScan analyzes full-text articles from the PubMed Central Open Access subset and calculates an empiric score that allows the identification of manuscripts reporting thermodynamic data on protein stability. The method was optimized on a set of publications included in the ProTherm database, and tested on a new curated set of articles, manually selected for presence of thermodynamic data. The results show that ThermoScan returns accurate predictions and outperforms recently developed text-mining algorithms based on the analysis of publication abstracts.

Availability: The ThermoScan server is freely accessible online at https://folding.biofold.org/thermoscan. The ThermoScan python code and the Google Chrome extension for submitting visualized PMC web pages to the ThermoScan server are available at https://github.com/biofold/ThermoScan.

Keywords: protein stability, text mining, document classification, automated literature mining, thermodynamic data

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology (LMB), United Kingdom

Reviewed by:
Elodie Laine,
Université Pierre et Marie Curie, France
Sundeep Chaitanya Vedithi,
University of Cambridge, United Kingdom

***Correspondence:**
Piero Fariselli
piero.fariselli@unito.it
Emidio Capriotti
emidio.capriotti@unibo.it

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *23 October 2020*
Accepted: *18 February 2021*
Published: *25 March 2021*

Citation:
Turina P, Fariselli P and Capriotti E (2021) ThermoScan: Semi-automatic Identification of Protein Stability Data From PubMed. Front. Mol. Biosci. 8:620475. doi: 10.3389/fmolb.2021.620475

Abbreviations: AUC, Area Under the receiving operating characteristic Curve; AUPR, Area Under the Precision-Recall curve; CC, Computational Concepts; F1 score, harmonic mean of PPV and TPR; MCC, Matthews correlation coefficient; NPV, Negative Predicted Values; Q 2, overall accuracy; PMC, PubMed Central; PPV, Positive Predicted Values; TC, Thermodynamic Concepts; TPR, True Positive Rate; TNR, True Negative Rate; TV, Thermodynamic Variables; UM, Units of Measure.

INTRODUCTION

A key aspect for characterizing the relationship between genotype and phenotype is the study of the impact of amino acid variants on protein function and structure (Thusberg and Vihinen, 2009; Compiani and Capriotti, 2013). To address this task, several tools for predicting the effect of variants on protein stability have been developed (Sanavia et al., 2020). The implementation of these methods requires a large and accurate set of experimental data, both for training and benchmarking. Although many protein folding databases were developed in the past (Bava et al., 2004; Fulton et al., 2007; Wagaman et al., 2014; Pancsa et al., 2016; Manavalan et al., 2019) some of them were discontinued or no longer maintained (Bava et al., 2004; Fulton et al., 2007). Among them, ProTherm (Kumar et al., 2006), the most comprehensive resource for thermodynamic data on protein variants, was not updated since 2013, and its maintenance was discontinued. Therefore, the need for curated databases on the thermodynamics and kinetics of protein folding has become urgent for implementation of accurate prediction methods.

In general, the collection of data from scientific literature is an expensive and time-consuming process requiring careful selection of keywords and queries for web searching (Fleuren and Alkema, 2015). As a consequence, during the last decades, several text-mining tools have been developed to speed up the data collection process (Rebholz-Schuhmann et al., 2012). Given the complexity and large variety of biological data, such searching tools were customized to address specific tasks (Huang and Lu, 2016). In particular, different approaches have been developed for identifying protein-protein interactions (Krallinger et al., 2008), drug-drug interactions (Zeng et al., 2019) and drug-phenotype relationships (Garten and Altman, 2009). Other methods identify gene functions (Soldatos et al., 2015) and define the role of molecules involved in biological processes (Wang et al., 2011). Currently, text-mining tools are used in daily life science research activity to improve web search (Ananiadou et al., 2010) and facilitate the database curation process (Yeh et al., 2003; Wei et al., 2012; Karp, 2016).

In this context, we developed ThermoScan, a new method for facilitating the collection and curation of thermodynamic data. Aiming at maximizing the extent of automatic vs. manual curation, ThermoScan is based on a semi-automatic text-mining algorithm for identifying experimental data on protein stability within the publicly accessible literature. ThermoScan reads the Open Access full-text manuscripts, ranks them according to the likelihood of finding the experimental thermodynamic data, and extracts relevant parts of the manuscript from paragraphs and tabular items. In addition, we evaluated the performance of ThermoScan in the detection of thermodynamic data in comparison with two existing web-server tools for documents classification (Fontaine et al., 2009; Simon et al., 2019).

METHODS

ThermoScan is a semi-automatic method for retrieving protein thermodynamic data from literature. The method scans the PubMed Central full-text HTML page and calculates a score for identifying manuscripts reporting experimental protein thermodynamic data in paragraphs and tables.

Datasets

For optimizing and testing the performance of ThermoScan we collected different datasets of articles reporting protein thermodynamic data (positives) or not (negatives). The initial set of positives (Pos-PT) was collected by considering 157 Open Access PMC articles referenced in the ProTherm database. Two negative sets of publications were selected from the PMC Open Access repository using different searching keywords. In detail we considered only the full-text articles available in HTML format and containing the terms "*protein*" and "*stability*" (Neg-PS) or "*protein*" and "*unfolding*" (Neg-PU). For the Neg-PS dataset we restricted the search to the first 2,000 articles. Thus, the Neg-PS and Neg-PU negative sets, obtained by restricting the literature search to the period 2000–2010, were composed of 2,000 and 583 manuscripts respectively.

For testing the performance of ThermoScan, we selected a set of 296 recently published (2011–2019) Open Access PMC articles with a PubMed search of the keywords "*protein*," "*stability*" and "*unfolding*". The manual curation of these articles, based on stringent criteria, allowed the identification of 194 manuscripts reporting experimental protein folding data. The remaining 102 papers, initially retained as negatives, were filtered excluding 37 articles reporting only protein thermodynamic data from binding or *in silico* experiments. With this manual procedure, we generated the New-PSU dataset, composed of 194 positive and 102 negative articles, and the Snew-PSU, composed of the same number of positives and 65 high-quality negatives. The composition of the datasets is summarized in **Supplementary Table S1**. The PMCIDs of the manuscripts collected in all the datasets are available as Supplementary File.

Manuscript Processing and Word Selection

Full-text articles in HTML format are parsed using the BeautifulSoup *Python* library (https://www.crummy.com/software/BeautifulSoup/). BeautifulSoup is used for extracting the text between paragraphs (<*p*>) and tables (<*table*>) tags. After extraction of the text included in the paragraphs and tables of each manuscript, the Natural Language Toolkit (NLTK) platform (https://www.nltk.org/) (Bird et al., 2009) is used for removing stopwords and for the lemmatization process. In particular, we use the *WordNetLemmatizer* function of NLTK for determining the word's lemma. After processing the manuscript with NLTK, the text is analyzed for identifying the words associated with protein thermodynamic concepts. In detail, we compared the frequency of the words in the manuscript of Pos-PT dataset against the Neg-PS dataset using a binomial distribution. The words were ranked on the basis of the *p*-value obtained from the complementary cumulative binomial distribution. Such *p*-value represents the probability of observing, in the Pos-PT dataset, a number of manuscripts with a given word higher than expected from the background probability, as estimated in the Neg-PS dataset. According to the *p*-values, calculated using the binomial survival function of the binomial distribution (**Supplementary Table S2**), the 5 words with lowest score were: unfolding, two-state, denaturant, dichroism and midpoint.

Text Mining and Scoring

ThermoScan processes the full-text article in HTML searching for significant protein thermodynamic words grouped in four classes:

- **Thermodynamic concepts (TC):** Important words frequently appearing in protein thermodynamic studies (unfolding, two-state, denaturant, dichroism, midpoint).
- **Thermodynamic variables (TV)** Words are identified by a regular expression matching the abbreviations of the main thermodynamic variables (ΔG, ΔH, ΔTm, etc.).
- **Units of measure (UM):** Words are identified by a regular expression matching the main units of measure used in thermodynamic experiments (kcal/mol, kJ/mol, etc)
- **Computational concepts (CC):** Words referring to computational studies (simulation, molecular dynamics, force field, predict, etc.).

The text extracted from the manuscript is searched for the 5 words in the first group. If one of the words is found, all the significant terms are extracted using each of the four regular expressions representing the four classes. The codes of the four regular expressions are reported in **Supplementary Materials**.

For each article, ThermoScan calculates an empirical score based on the four classes of words defined above. Our approach returns the total and the single paragraph/table scores. A positive partial score is assigned to the items matching the first three classes (thermodynamic concepts, thermodynamic variables and units of measures), and a negative one to the items matching the fourth class (computational concepts).

The paragraph/table score is calculated by summing the scores of the individual matches without repetitions. The individual scores of the different classes of words are the following:

- two-state = unfolding = denaturant = midpoint = dichroism = 1
- Cp = Tm = 1, ΔX = 2, ΔΔX = 3 (X = Cp, Tm, UG, GU, G, H, T, U)
- °C = 1, E/C = 2 (E = kcal, kJ; C = mol, mole, mole/°C, mol/°C, mol/K, mol/M)
- simulation = molecular dynamics = force field = charmm = gromacs = amber = PBSA = GBSA = predict = −1; md simulation = −2

The total score assigned to the article is obtained by summing all paragraph/table scores. For the classification task, we considered two alternative measures, corresponding to the maximum (Max) or to the average (Mean) paragraph/table score for each paper.

Although not used at this stage for the classification task, ThermoScan additionally searches for thermodynamic data relative to binding processes, considering the following terms: binding, affinity, dissociation, interaction, ppi, protein-protein, kcat/Km.

Method optimization and Testing

For optimizing the performance of ThermoScan we maximized the performance of a binary classifier discriminating between manuscripts reporting protein thermodynamic data and not. In general, this task can have different difficulty levels depending on the selection of the negative set. To select a fair negative set of manuscripts, we considered those collected in the Neg-PS and Neg-PU datasets, which include the terms "*protein*" and "*stability*," or "*protein*" and "*unfolding*," respectively. From Neg-PS and Neg-PU datasets we generated 10 randomly selected sets of 157 negative manuscripts in equal proportion, to be compared with those collected in the Pos-PT dataset. With this procedure we generated 10 training sets that only differ by the subset of negatives. Using the procedure described above, for each manuscript we calculated the maximum (Max) and average (Mean) scores of the extracted paragraphs and tables. In addition, we evaluated the relative contributions of the three main groups of words (thermodynamic concepts, thermodynamic variables and units of measures) to the prediction power of ThermoScan by calculating the performance achieved when using different groups combinations. In particular we evaluated the performance of three alternative methods considering:

- thermodynamic concepts alone (TC);
- thermodynamic variables and units of measures (TV ∪ UM);
- thermodynamic concepts, thermodynamic variables and units of measures (TC ∪ TV ∪ UM).

The results obtained with the three combinations were compared with those obtained by including all four groups of words defined above.

For ThermoScan optimization we selected the classification thresholds that maximized the Matthews Correlation Coefficient (see Methods section in **Supplementary Materials**), and finally we tested the ThermoScan performance on the two testing sets (New-PSU, Snew-PSU) by applying the same classification thresholds.

The performance of ThermoScan was then compared with those achieved by MedlineRanker (Fontaine et al., 2009) and BioReader (Simon et al., 2019). The performances of the two text mining methods (MedlineRanker and BioReader), which are both based on the analysis of the manuscript abstract, were evaluated on the New-PSU, Snew-PSU datasets. All the performance measures are defined in **Supplementary Materials**.

RESULTS

Here we present the results achieved by ThermoScan in the selection of manuscripts reporting experimental protein thermodynamic data from PubMed. We first optimized ThermoScan in a training step, then tested its performance on a blind set of manually curated articles, and finally compared such performance with those achieved by MedlineRanker (Fontaine et al., 2009) and BioReader (Simon et al., 2019).

ThermoScan Optimization

For the optimization of ThermoScan we calculated its performance considering both the maximum (Max) and the average (Mean)

TABLE 1 | Optimized performance of ThermoScan based on the maximum (Max) and average (Mean) scores. The performance measures are defined in **Supplementary Materials**. The standard deviation of all the performance measures are ≤0.01.

Score	TH	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUPR
Max	3.00	0.97	1.00	0.95	0.94	1.00	0.94	0.97	0.99	0.99
Mean	1.36	0.94	0.94	0.95	0.95	0.94	0.89	0.94	0.98	0.99

scores assigned to each part (paragraph/table) of the manuscript. The performance of ThermoScan was calculated using a positive set of 157 manuscripts from Protherm containing protein thermodynamic data (Pos-PT) and a negative set with an equal number of articles not containing any thermodynamic information (randomly selected from Neg-PU and Neg-PS datasets, described in the Methods section). All the performance measures (defined in the **Supplementary Materials**) were averaged over 10 random samplings of the negative subset. The detailed results obtained with both Max and Mean scoring systems are reported in **Supplementary Tables S3, S4**; **Table 1** summarizes the optimal performance measures from **Supplementary Tables S3, S4** for both the Max and Mean scoring systems. In detail, the method based on the maximum score achieved 3% higher accuracy (Q_2) and 5% higher Matthews correlation coefficient (MCC). In **Figure 1**, the Precision (PPV) and Recall (TPR) values from **Supplementary Tables S3, S4** are plotted as a function of Max (**Figure 1A**) and Mean (**Figure 1B**) scoring threshold. The results show that the best performance was achieved with the Max scoring system with threshold ≥3. Alternative scores of the performance are based on the AUC (Area Under the receiving operating characteristic Curve) and on the AUPR (Area Under the Precision-Recall curve) which are shown in **Figure 2**. Also, these results confirm that the Max scoring system achieved the best performance.

In summary, the above analysis shows that the binary classifier results in a higher performance when based on the maximum paragraph/table score rather than on the average score.

ThermoScan Testing and Benchmarking

ThermoScan was tested calculating its performance on two sets (New-PSU and Snew-PSU) obtained by searching in the Open Access PMC articles having the words "*protein*," "*stability*" and "*unfolding*" in their abstracts. The classification was performed using the same threshold values obtained in the optimization steps. The results reported in **Table 2** show that ThermoScan achieved the highest performance on the testing set Snew-PSU, obtained by removing 37 manuscripts of difficult classification, (i.e. reporting protein thermodynamic data from binding or *in silico* experiments only). Indeed, when comparing the performances of both versions of ThermoScan (Max and Mean) on the Snew-PSU and New-PSU datasets, the method results in ~10% better accuracy and 20% better Matthews correlation coefficient on the first one. The version of ThermoScan based on the maximum paragraph/table score achieved an overall accuracy of 91% and a Matthews correlation coefficient of 0.76. These results are the most similar ones to those reached in the optimization step. Furthermore, to estimate the filtering capabilities of ThermoScan, we analyzed a set of ~700,000 manuscripts from the PubMed Central FTP website (https://ftp.ncbi.nlm.nih.gov/pub/pmc/manuscript/), which required on average ~4 s for each article. By using a scoring threshold of 6, ThermoScan selects ~2,200 items (0.3%), which, according to our analysis of the New-PSU testing set, are expected to include less than 4% of false positives. Finally, we compared the performance of ThermoScan with those of MedlineRanker (Fontaine et al., 2009) and BioReader (Simon et al., 2019) which are based on the analysis of the manuscript abstracts. As shown in **Table 3**, ThermoScan, that analyzes the full-text manuscript, results in better performance than MedlineRanker and BioReader on both New-PSU and Snew-PSU datasets. In almost all cases ThermoScan reached ~15% higher overall accuracy and ~30% higher Matthews correlation coefficient with respect to MedlineRanker and BioReader. Given the different amount of information in input, the performance of ThermoScan

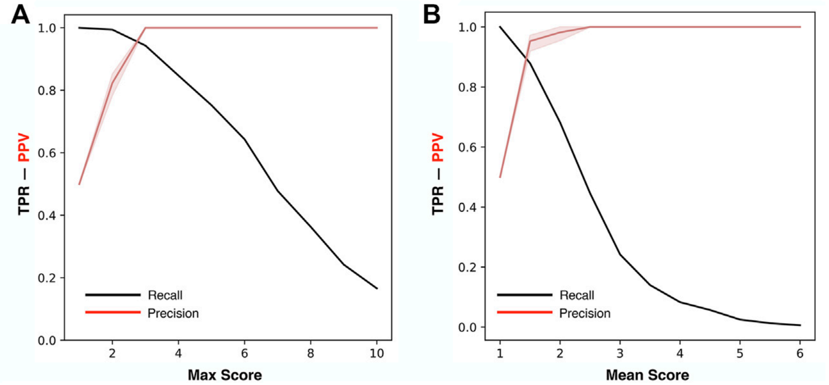


FIGURE 1 | Precision and Recall of ThermoScan at different classification thresholds. The plots show the performance based on the Max **(A)** and Mean **(B)** scores. The performance measures TPR (black) and PPV (red) are defined in **Supplementary Materials**. The shaded area represents the range between the minimum and maximum scoring values.

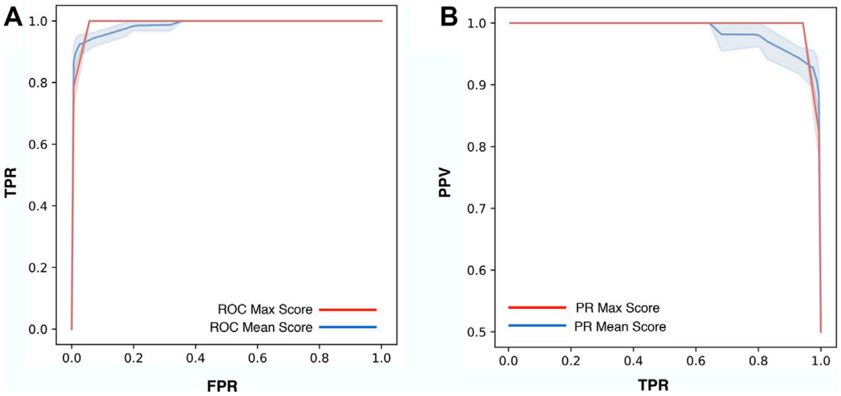


FIGURE 2 | Performance measures of ThermoScan based on the Max (red) and Mean (blue) scores. The plots show the AUC (Area Under the receiving operating characteristic Curve). The shaded area represents the range between the minimum and maximum scoring values **(A)** and the AUPR (Area Under the Precision-Recall curve) **(B)** for the two scoring systems. The TPR, FPR, and PPV performance measures are defined in **Supplementary Materials**.

TABLE 2 | Performance of ThermoScan on the New-PSU and Snew-PSU datasets. The ThermoScan thresholds obtained in the optimization step with maximum and mean paragraph/table scoring methods are 3.00 and 1.36 respectively. The performance measures are defined in **Supplementary Materials**.

Score	Dataset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUPR
Max	New-PSU	0.80	0.49	0.88	0.96	0.78	0.55	0.86	0.86	0.86
	Snew-PSU	0.91	0.75	0.88	0.96	0.92	0.76	0.94	0.96	0.94
Mean	New-PSU	0.80	0.59	0.77	0.91	0.81	0.53	0.85	0.83	0.82
	Snew-PSU	0.89	0.83	0.75	0.91	0.94	0.71	0.92	0.92	0.91

TABLE 3 | Comparison of the performance of ThermoScan (based on maximum paragraph/table score) with BioReader and MedlineRanker on the New-PSU and Snew-PSU datasets. The classification thresholds for BioReader and MedlineRanker and ThermoScan are 0.022, 0.027 and three respectively. The performance measures are defined in **Supplementary Materials**.

Method	Dataset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUPR
BioReader	New-PSU	0.66	0.59	0.50	0.70	0.76	0.28	0.73	0.64	0.72
	Snew-PSU	0.70	0.69	0.43	0.70	0.87	0.34	0.77	0.69	0.75
MedlineRanker	New-PSU	0.63	0.63	0.47	0.63	0.76	0.25	0.69	0.70	0.67
	Snew-PSU	0.70	0.68	0.43	0.70	0.87	0.34	0.78	0.78	0.72
ThermoScan	New-PSU	0.80	0.49	0.88	0.96	0.78	0.55	0.86	0.86	0.86
	Snew-PSU	0.91	0.75	0.88	0.96	0.92	0.76	0.94	0.96	0.94

can not be directly compared with those of MedlineRanker and BioReader. Our analysis shows that full-text classification-based methods do tend to have higher discriminating power than methods based on the analysis of the abstract, even though the latter can deal with larger sets of articles in a shorter amount of time.

Contribution to Performance

To evaluate the contribution to the performance of ThermoScan of each group of words included in the manuscript processing, we assessed the performance of three alternative methods considering a subset of groups (see *Method optimization and testing* paragraph in the Methods section). In particular, we compared the performance of ThermoScan with the three following approaches based on:

i. the thermodynamic concepts alone (TC);
ii. the thermodynamic variables and units of measure (TV ∪ UM);
iii. all previous groups (TC ∪ TV ∪ UM).

On the training sets (Pos-PT, Neg-PS and Neg-PU), the results of the comparison between ThermoScan, which includes four groups of words (TC ∪ TV ∪ UM ∪ CC), and the alternative methods described above are reported in **Supplementary Tables S5, S6**. This analysis shows that the predominant contribution to the classification power is given by the 5 words belonging to the group of the thermodynamic concepts. We also noticed that the combination, which significantly contributes to improve the performance, includes all three groups: both the thermodynamic concepts and variables, together with the units of measure. Indeed, considering the classifier based on the maximum paragraph/table score, the method based on the combination of the three groups of words results in 4% better overall accuracy and 7% better Matthews correlation coefficient with respect to the methods based on thermodynamic concepts alone (**Supplementary Table S5**). Although no significant improvement of the performance is resulting from adding the computational concepts (CC), this negative score, which is included in ThermoScan, is important for penalizing the manuscripts reporting in silico protein stability data. A similar improvement is observed on the testing sets New-PSU and Snew-PSU (**Supplementary Tables S7–S10**). In the testing step we observed an improvement of NPV (negative predicted value) and TNR (true negative rate) of 2 and 4% respectively when comparing ThermoScan with the method based on the three groups of words (TC ∪ TV ∪ UM).

Identification of In-Silico Data and Manuscripts

Identifying in-silico articles, which represented less than 10% of our testing set, remains a critical issue, especially when the article texts include reference to, and description of, experimental data. To penalize articles presenting in-silico data only, we defined a negative score based on the presence of the computational concepts (CC). The maximum penalization score for a paragraph is -2 when the words "md simulation" is found. Although the addition of the CC does not significantly improve the performance of the automatic evaluation, it can help during the manual curation process to detect and discard possible false positives.

ThermoScan Web Server and Code

We developed a web server version of ThermoScan that takes in input a list of manuscript identifiers (PMCID, PMID or DOI) and returns a table with the scores associated with each article. Each identifier in the output is linked to a webpage showing significant paragraphs and tables which include protein thermodynamic terms. Words belonging to the main three classes defined in the Method section (thermodynamic concepts, thermodynamic variables, units of measure) are highlighted in *red*. To facilitate the curation process and avoid the selection of in-silico data, the output of the webserver displays the CC terms in blue and returns a score related to their presence. For better help in identifying the possible presence of thermodynamic data on protein mutants, the potential amino acid variants are highlighted in *green*. For each manuscript, the server calculates the total score and the maximum score for the extracted paragraphs and tables. An example of the ThermoScan server output is available at the page https://shorturl.at/cetwG. To analyze the HTML pages of manuscripts with restricted access, we developed a GoogleChrome app that allows the user to submit the content of a web page, visualized on the user's browser, directly to the ThermoScan server. Furthermore, the ThermoScan python script for the local scanning of the PMC articles is made available through GitHub.

DISCUSSION

In this paper we present ThermoScan, a text-mining algorithm for the selection and fine-grained classification of Open Access PMC articles, aimed at retrieving literature data on the thermodynamic stability of proteins and their variants. Although the direct comparison of the performance of methods with different input features is not straightforward, our results show that ThermoScan, which is based on the analysis of full-text articles, outperforms existing web services based on the analysis of the manuscript abstracts (Fontaine et al., 2009; Simon et al., 2019), thus constituting a new valuable tool to semi-automatically collect protein thermodynamic data. Furthermore, the web interface, which displays relevant parts of the article, makes ThermoScan a valuable complementing tool for refining the search of protein thermodynamic data. In conclusion, our method achieves a high discrimination power by analyzing full-text articles, by fine-tuning the classification thresholds, and by using a tailored subset of specific symbols and words. Given the trend toward an increasing amount of in-silico only studies in the literature repositories, in the future more sophisticated search strategies should be implemented, to avoid the selection of manuscripts reporting in-silico data only, which contribute to increasing the rate of false positives. Nevertheless we expect that ThermoScan will significantly support and accelerate the updating and curation of new databases for collection of protein thermodynamic data. Such data are essential for characterizing the relationship between protein sequence and structure and for the development of more accurate methods for predicting the impact of amino acid variants on protein stability.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

PT analyzed the datasets of manuscripts used for training and testing the method. EC developed the algorithm and the web server. All authors designed the research and contributed to the writing of the manuscript.

FUNDING

This work was supported by the PRIN project, "Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation" of the Ministero Istruzione, Università e Ricerca (201744NR8S).

ACKNOWLEDGMENTS

PF thanks the Ministero Istruzione, Università e Ricerca for the program "Dipartimenti di Eccellenza 20182022D15D18000410001".

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.620475/full#supplementary-material.

REFERENCES

Ananiadou, S., Pyysalo, S., Tsujii, J. I., and Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* 28, 381–390. doi:10.1016/j.tibtech.2010.04.005

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 32, 120D–121D. doi:10.1093/nar/gkh082

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language processing with* Python. Newton, MA: O'Reilly Media Inc.

Compiani, M., and Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry* 52, 8601–8624. doi:10.1021/bi4001529

Fleuren, W. W. M., and Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods* 74, 97–106. doi:10.1016/j.ymeth.2015.01.015

Fontaine, J.-F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M., and Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.* 37, W141–W146. doi:10.1093/nar/gkp353

Fulton, K. F., Bate, M. A., Faux, N. G., Mahmood, K., Betts, C., and Buckle, A. M. (2007). Protein folding database (PFD 2.0): an online environment for the international foldeomics consortium. *Nucleic Acids Res.* 35, D304–D307. doi:10.1093/nar/gkl1007

Garten, Y., and Altman, R. B. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics* 10 (Suppl. 2), S6. doi:10.1186/1471-2105-10-S2-S6

Huang, C.-C., and Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.* 17, 132–144. doi:10.1093/bib/bbv024

Karp, P. D. (2016). Can we replace curation with information extraction software? *Database* 2016, baw150. doi:10.1093/database/baw150

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., and Valencia, A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.* 9 (Suppl. 2), S4. doi:10.1186/gb-2008-9-s2-s4

Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., et al. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34, D204–D206. doi:10.1093/nar/gkj103

Manavalan, B., Kuwajima, K., and Lee, J. (2019). PFDB: a standardized protein folding database with temperature correction. *Sci. Rep.* 9, 1588. doi:10.1038/s41598-018-36992-y

Pancsa, R., Varadi, M., Tompa, P., and Vranken, W. F. (2016). Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res.* 44, D429–D434. doi:10.1093/nar/gkv1185

Rebholz-Schuhmann, D., Oellrich, A., and Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* 13, 829–839. doi:10.1038/nrg3337

Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput. Struct. Biotechnol. J.* 18, 1968–1979. doi:10.1016/j.csbj.2020.07.011

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M. B., and Olsen, L. R. (2019). BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* 19, 57. doi:10.1186/s12859-019-2607-x

Soldatos, T. G., Perdigão, N., Brown, N. P., Sabir, K. S., and O'Donoghue, S. I. (2015). How to learn about gene function: text-mining or ontologies? *Methods* 74, 3–15. doi:10.1016/j.ymeth.2014.07.004

Thusberg, J., and Vihinen, M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.* 30, 703–714. doi:10.1002/humu.20938

Wagaman, A. S., Coburn, A., Brand-Thomas, I., Dash, B., and Jaswal, S. S. (2014). A comprehensive database of verified experimental data on protein folding kinetics. *Protein Sci.* 23, 1808–1812. doi:10.1002/pro.2551

Wang, X., McKendrick, I., Barrett, I., Dix, I., French, T., Tsujii, J., et al. (2011). Automatic extraction of angiogenesis bioprocess from text. *Bioinformatics* 27, 2730–2737. doi:10.1093/bioinformatics/btr460

Wei, C.-H., Harris, B. R., Li, D., Berardini, T. Z., Huala, E., Kao, H.-Y., et al. (2012). Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database* 2012, bas041. doi:10.1093/database/bas041

Yeh, A. S., Hirschman, L., and Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* 19 (Suppl. 1), i331–i339. doi:10.1093/bioinformatics/btg1046

Zeng, Z., Deng, Y., Li, X., Naumann, T., and Luo, Y. (2019). Natural Language processing for EHR-based computational phenotyping. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16, 139–153. doi:10.1109/TCBB.2018.2849968

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 May 2021
doi: 10.3389/fmolb.2021.617016

Mapping OMIM Disease–Related Variations on Protein Domains Reveals an Association Among Variation Type, Pfam Models, and Disease Classes

Castrense Savojardo[1], Giulia Babbi[1], Pier Luigi Martelli[1]* and Rita Casadio[1,2]

[1] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom

Reviewed by:
Yannick Mahlich,
Rutgers, The State University
of New Jersey, United States
Mauno Vihinen,
Lund University, Sweden

***Correspondence:**
Pier Luigi Martelli
pierluigi.martelli@unibo.it

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: 13 October 2020
Accepted: 09 April 2021
Published: 07 May 2021

Citation:
Savojardo C, Babbi G, Martelli PL
and Casadio R (2021) Mapping OMIM
Disease–Related Variations on Protein
Domains Reveals an Association
Among Variation Type, Pfam Models,
and Disease Classes.
Front. Mol. Biosci. 8:617016.
doi: 10.3389/fmolb.2021.617016

Human genome resequencing projects provide an unprecedented amount of data about single-nucleotide variations occurring in protein-coding regions and often leading to observable changes in the covalent structure of gene products. For many of these variations, links to Online Mendelian Inheritance in Man (OMIM) genetic diseases are available and are reported in many databases that are collecting human variation data such as Humsavar. However, the current knowledge on the molecular mechanisms that are leading to diseases is, in many cases, still limited. For understanding the complex mechanisms behind disease insurgence, the identification of putative models, when considering the protein structure and chemico-physical features of the variations, can be useful in many contexts, including early diagnosis and prognosis. In this study, we investigate the occurrence and distribution of human disease–related variations in the context of Pfam domains. The aim of this study is the identification and characterization of Pfam domains that are statistically more likely to be associated with disease-related variations. The study takes into consideration 2,513 human protein sequences with 22,763 disease-related variations. We describe patterns of disease-related variation types in biunivocal relation with Pfam domains, which are likely to be possible markers for linking Pfam domains to OMIM diseases. Furthermore, we take advantage of the specific association between disease-related variation types and Pfam domains for clustering diseases according to the Human Disease Ontology, and we establish a relation among variation types, Pfam domains, and disease classes. We find that Pfam models are specific markers of patterns of variation types and that they can serve to bridge genes, diseases, and disease classes. Data are available as Supplementary Material for 1,670 Pfam models, including 22,763 disease-related variations associated to 3,257 OMIM diseases.

Keywords: protein variations, protein structure, protein domain, variation type, disease-related variations, disease variant databases, Pfam-disease association

INTRODUCTION

In the last decade, several efforts have been devoted to the problem of functional annotation of protein variants with the aim of relating variations to specific diseases (Vihinen, 2017, 2018). A collection of variations of genetic diseases is now available, and this prompted the investigation of molecular mechanisms responsible for protein failure (Schaafsma and Vihinen, 2018). Particularly, variations of non-synonymous proteins can promote the change of the active/binding sites and/or protein instability and can hamper protein–protein and ligand–protein interactions (Kucukkal et al., 2015; Ittisoponpisan et al., 2019; Ofoegbu et al., 2019). Molecular mechanisms can be, therefore, different, and different phenotypes may share common molecular mechanisms, independent of the different genes (Deans et al., 2015; Reeb et al., 2016; Babbi et al., 2019, and references therein). Several studies also focused on determining the most frequent protein variants associated with diseases, with the aim of helping functional annotation, starting from variant sequencing (Niroula and Vihinen, 2017; Zeng and Bromberg, 2019).

Different computational methods are available for the functional annotation of variations, based on different approaches. Routinely, given a specific variation, computational methods return with a computed reliability whether the change of a side chain in a protein is disease-related or not (Niroula and Vihinen, 2016).

An interesting aspect of disease-related protein variants is the protein instability promoted by the variations (Casadio et al., 2011; Savojardo et al., 2019, and references therein). Protein instability may be related to a disease, with this not being the only reason. For functional annotation of disease-related variations, routinely, the chemico-physical properties of the variation and the effect of the variation on the close environment in the protein structure are taken into consideration. It appears that the correlation among the strength of association to disease and the strength of association to the protein structure perturbation is moderate (Savojardo et al., 2019).

The problem of which phenotype is associated with a given variation or a set of variations has been scarcely addressed, and it remains unanswered, given the complexity of the scenario relating phenotypes to variations. Existing databases can relate genes to diseases and/or variations to diseases (MalaCards[1], Rappaport et al., 2017; GeneCards[2], Stelzer et al., 2016; DisGeNet[3], Piñero et al., 2020; eDGAR[4], Babbi et al., 2017; Humsavar[5], UniProt Consortium, 2019; OMIM[6], Amberger et al., 2015).

Protein domains have been adopted to explore associations between genes and human-inherited diseases (Zhang et al., 2011, 2016; Yates and Sternberg, 2013; Wiel et al., 2017, 2019). Models of protein domains are available in the Pfam database[7] (El-Gebali et al., 2019), and they enable the clustering of proteins into protein families, each represented by multiple sequence alignments, mainly based on protein structural alignments and cast into hidden Markov models (HMMs). Initially, similarities of disease phenotypes were exploited within a given domain–domain interaction network, and a Bayesian approach was proposed to prioritize candidate domains for human complex diseases (Zhang et al., 2011). Then, domain–disease associations were inferred from domain–protein, protein–disease, and disease–disease relationships (Zhang et al., 2016). In these studies, the bottom layer of variations in proteins, detected in large-scale sequencing experiments, was not taken into consideration, restraining the analysis only to the already known protein– or gene–disease associations. More recently (Wiel et al., 2017), with the notion of homologous domains in proteins, variants were aggregated to improve their interpretation, and a web server (MetaDome[8], Wiel et al., 2019) was made available for the pathogenicity analysis of genetic variants.

In a previous study (Savojardo et al., 2019), we introduced the notion of variation type, in order to take the physico-chemical properties of the variations into account as well (Casadio et al., 2011). After mapping genetic disease–related variations on a restricted set of human protein three-dimensional (3D) structures, we found that the distribution of disease variation types significantly varies across different structural/functional Pfam models.

In this study, relying on the relationship between genes and phenotypes, we ask the question as to which extent possible patterns of variation types framed into Pfam domains are significant for a reliable association to specific groups of maladies.

MATERIALS AND METHODS

Dataset Construction

The dataset adopted in this study was derived from the Humsavar database[5] release 2020_04 of August 2, 2020, listing all missense variants annotated in human UniProtKB/Swiss-Prot (UniProt Consortium, 2019) entries.

From the initial set of proteins included in the database, we only selected those reporting at least one variant implicated in the disease, excluding proteins reporting only polymorphisms not associated with disease insurgence. Moreover, any variation labeled as "unclassified" (i.e., with uncertain implications in disease) was filtered out. Finally, we only retained disease-related variations associated with a genetic disorder reported in the Online Mendelian Inheritance in Man (OMIM) catalog[9].

The set of neutral variations was extended using data retrieved from the GnomAD database (exome version 2.1.1) (Karczewski et al., 2020). Only variations occurring in our set of proteins, not already included in Humsavar and with clinical significance

[1] https://www.malacards.org/
[2] https://www.genecards.org/
[3] https://www.disgenet.org/
[4] http://edgar.biocomp.unibo.it
[5] https://www.uniprot.org/docs/humsavar
[6] https://www.omim.org/
[7] https://pfam.xfam.org/
[8] https://stuart.radboudumc.nl/metadome/
[9] https://omim.org/

labeled as “Benign/Likely benign” by ClinVar (release 2021-03-23) (Landrum et al., 2020), were retained.

Pfam (El-Gebali et al., 2019) annotations were retrieved from the Pfam-A region annotation file for *Homo sapiens* version 33.1 obtained *via* the Pfam FTP server[10]. From all the annotations available, we only retained those occurring at proteins included in our set of data and covering at least one disease-related variation.

Mapping OMIMs to Disease Ontology

The DO (Human Disease Ontology) OBO (Open Biological and Biomedical Ontology) file release of September 15, 2020, was downloaded[11] and used directly to retrieve annotations for each OMIM disease by means of cross-references. Each retrieved leaf DO term associated to a single OMIM was expanded up to the ontology root term, including all ancestors. Term expansion was computed using an *ad-hoc* script to parse the OBO file.

Computing the Disease Score

For each Pfam domain, we estimated a propensity score for the association to the disease as follows:

$$Score\ (pfam) = \frac{N_d^{pfam} / \left(N_d^{pfam} + N_p^{pfam}\right)}{N_d / \left(N_d + N_p\right)} \tag{1}$$

where N_d^{pfam} and N_p^{pfam} are the number of disease-related and polymorphism variations in the domain *pfam*, while N_d and N_p are the same numbers in the whole dataset. In the dataset, scores range from 1.40 down to 0.03.

Kullback–Leibler Divergence Between Distributions

Differences between probability distributions were evaluated using the Kullback–Leibler divergence:

$$D_{KL} = -\sum_{x \in X} p\,(x) \cdot log_2 \frac{q(x)}{p(x)} \tag{2}$$

where p and q are two discrete probability distributions defined on the same probability space X.

RESULTS

A Dataset of Variations With Annotated Pfam

Overall, our dataset comprises 50,746 variations occurring in 2,959 proteins implicated in 3,884 genetic disorders. Disease-related variations in these proteins are 29,949, accounting for 55% of the total variations. The remaining 20,797 variations are neutral (45%). **Table 1** shows summary statistics about the dataset analyzed in this study.

Restricting the set of proteins to those having Pfam entries covering at least one disease-related variation, we ended up with 2,513 proteins (corresponding to 85% of the initial protein set) implicated in 3,257 distinct genetic diseases. Overall, 1,670 distinct Pfam entries were annotated on these proteins. A subset of 548 out of 1,670 Pfams occurs in two or more proteins in the set. The vast majority (96%) of Pfam entries are of type “Domain” or “Family,” while a very small fraction accounts for “Repeat,” “Coiled-coil,” “Motif,” and “Disordered” types.

TABLE 1 | Summary of the OMIM-related variation dataset of this study.

Number of proteins associated with disease	2959
Number of diseases (OMIM)	3884
Number of variations	54746
Number of disease variations	29949 (55%)[^]
Number of neutral polymorphisms (on the same disease proteins)	24797 (45%)[^]
Number of disease proteins with Pfam covering disease variations	2513 (85%)[#]
Number of Pfams	1670
Number of diseases (OMIM) in proteins with Pfams	3257 (84%)[°]
Number of variations covered by Pfams	31934 (68%)[^]
Number of disease variations covered by Pfams	22763 (71%)[+]
Number of neutral polymorphisms covered by Pfams	9171 (29%)[+]

[^] percentage computed with respect to the total number of variations (54746); [#] percentage computed with respect to the total number of proteins (2959); [°] percentage computed with respect to the total number of diseases (3884); [+] percentage computed with respect to the total number of Pfam-covered variations (31934).

After this reduction, we retained 31,934 variations covered by Pfams, distributed into 22,763 (71%) and 9,171 (29%) disease-related and neutral polymorphic variations, respectively.

Data shown in **Table 1** clearly indicate that the incidence of disease-related variations within Pfam domains is significantly higher than the background (71% against 55%).

Overall Pfam Association With Disease

We were interested in elucidating the overall association between Pfam and OMIM diseases. For each entry in the set of 1,670 Pfam domains in our dataset, we computed the score for the association to disease with the formula reported in Eq. 1. A value greater than 1 for this ratio highlights a higher abundance of disease variations in the Pfams than in the background. The complete result of this analysis is reported in **Supplementary Table 1** for all the 1,670 Pfam entries. About 48% of Pfam entries have a value greater than 1, as a consequence of the overall propensity of disease-related variations to be located within Pfam domains. In general, the distribution of scores is not random and reflects a differential disease association for the different Pfam entries.

In **Table 2**, we list the result for the 20 highest scoring Pfams covering 10 or more proteins. Scores with corrected *p*-values (**Supplementary Table 2**) equal to or lower than 0.1 are highlighted (top scoring Pfams are all significant at 0.1 level). Significance does not hold for some Pfams covering only few variations. In these cases, more data are needed in order to properly evaluate the association to the disease.

Interestingly, Pfam entries reported in **Table 2** can be grouped into few functional classes, including DNA-binding domains

[10] ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/proteomes/9606.tsv.gz

[11] https://disease-ontology.org/

TABLE 2 | The 20 highest scoring Pfam entries mostly associated with diseases.

Pfam ID	Pfam name	Pfam type	No of proteins	No of disease variations	No of neutral polymorphisms	Score[§]
PF00105	zf-C4	Domain	12	60	2	1.36*
PF00250	Forkhead	Domain	10	88	4	1.34*
PF00010	HLH	Domain	14	48	3	1.32*
PF00104	Hormone_recep	Domain	18	195	20	1.27*
PF00307	CH	Domain	11	48	6	1.25*
PF00046	Homeodomain	Domain	42	163	21	1.24*
PF07645	EGF_CA	Domain	17	301	46	1.22*
PF00096	zf-C2H2	Domain	23	80	13	1.21*
PF00029	Connexin	Family	10	319	53	1.20*
PF00017	SH2	Domain	11	72	12	1.20*
PF00520	Ion_trans	Family	48	1020	173	1.20*
PF00004	AAA	Domain	10	70	12	1.20*
PF00400	WD40	Repeat	19	52	9	1.20
PF02770	Acyl-CoA_dh_M	Domain	10	40	7	1.19
PF00169	PH	Domain	11	53	10	1.18
PF00005	ABC_tran	Domain	15	236	49	1.16*
PF07686	V-set	Domain	12	84	18	1.16
PF00271	Helicase_C	Family	17	65	15	1.14
PF00176	SNF2_N	Family	10	63	15	1.13
PF00089	Trypsin	Domain	21	258	87	1.13*

[§] *Score is computed as defined in Eq. 1. Significance of each score was assessed using the Fisher exact test on the corresponding contingency table and correcting for multiple testing using the Benjamini-Hochberg procedure. Individual p-values are listed in* ***Supplementary Table 2****. *Corrected P-values are equal or lower than 0.1.*

(accounting for eight domains/families), transmembrane domains (three), and enzymes (three).

Pfams Have Distinctive Patterns of Disease Variation Types

Going a step further in the analysis, we investigated the composition of disease-related variations occurring in different Pfam domains. In a previous study (Savojardo et al., 2019), the same analysis was performed on a small dataset of highly curated variations covered by 3D structures from Protein Data Bank (PDB). In this study, we extended and complemented the previous results using a larger dataset of Pfam domains and variations. To this aim, we first grouped residues according to their physico-chemical properties, obtaining four major groups, namely, apolar (GAVPLIM), aromatic (FWY), polar (STCNQH), and charged (DEKR) residues. We define a variation type in relation to the conservation or substitution of apolar (a), polar (p), aromatic (r), and charged (c) (**Figure 1**). Then, we computed Pfam-specific distributions of disease-related variations involving substitutions from one group to another (overall, 16 different substitution types are possible). Complete results are reported in **Supplementary Table 3** for all the 1,670 Pfam domains.

In **Figure 1**, we show a heatmap reporting the frequencies of each substitution type for the 20 highest scoring Pfam entries described in the previous section and mostly associated with diseases. For each Pfam entry, we report the Pfam ID, the name, and two numbers in parentheses, indicating the number of proteins and disease-related variations covered by the specific Pfam. For comparison, the last row reports the overall distribution of substitution types computed on the whole set of variation types covered by Pfams.

The results shown in the heatmap of **Figure 1** indicate that the different Pfams are enriched in different variation types and that each Pfam shows a differential pattern with respect to the background. Interestingly, in some cases, the pattern of enriched variation types can be related with the overall function of the Pfam domain and/or the cellular context in which the domain/s are presumably operating.

In **Figure 2**, we report three examples, namely, a selection of DNA-binding domains, growth factors, and transmembrane domains. For DNA-binding domains, we observe a higher concentration of disease-related variations involving a substitution from a charged residue to any different residue type. Contrarily, for growth factor domains, we observe abundant variations involving substitutions from polar to any type of the residue, while transmembrane domains are mostly enriched in substitutions involving apolar wild types. These observations clarify a general trend, pointing to the specificity of the disease variation type per Pfams of functional classes.

From data analysis, we conclude that the distribution of the disease-related variation type patterns observed for the different Pfams is non-random and different from the background distribution (computed considering all the disease-related variation types occurring in Pfams). This observation confirms our previous results obtained with a smaller number of Pfam domains, directly related to human protein structures, and corroborates the notion that distinctive patterns of disease-related variation types are Pfam specific (Savojardo et al., 2019).

Linking the Pfam to Disease Ontology

As a final step of our investigation, we searched for a link between Pfam domains and disease ontology. Disease classification is not a trivial task. Different controlled vocabularies and ontologies

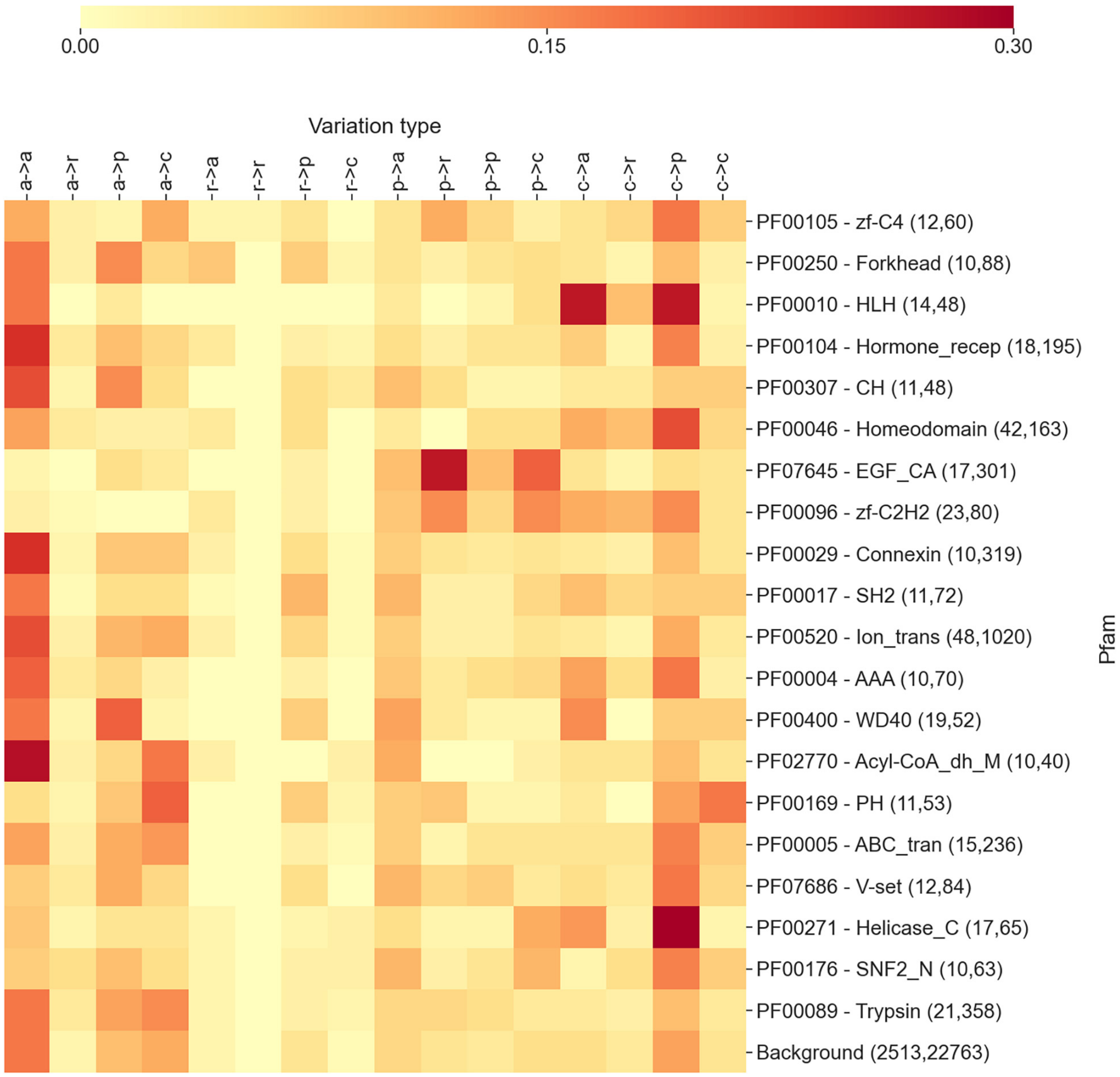


FIGURE 1 | The heatmap reporting the frequency of each variation type as observed within the 20 Pfam entries mostly associated with diseases. For each Pfam, the numbers within parentheses indicate the number of proteins and disease-related variations covered. In variation types, labels are as follows: a, apolar; r, aromatic; p, polar; and c, charged. Mean and median Kullback–Leibler divergences (Eq. 2) between individual Pfam distributions and the background are 2.1 and 2.1 bits, respectively.

such as the Human Phenotype Ontology (HPO)[12] (Köhler et al., 2019) or the DO (Schriml et al., 2019) are available for this purpose. However, none of the ontologies provides a full coverage of the entire space of OMIM diseases, ranging from 82% coverage of HPO to 74% of DO. Moreover, ontologies like HPO are not specifically designed to describe a disease. Instead, they are devised to describe clinically relevant phenotypes. In the current study, we used the DO ontology because, in spite of a slightly lower coverage, it provides a better and less ambiguous classification of diseases.

To obtain a high-level disease classification, we collected all the 3,257 OMIM diseases linked to variations occurring in our 1,670 Pfam domains and mapped them to a set of 17 first-level DO terms. These include 12 terms describing diseases affecting anatomical entities (all child terms of "DOID:7 – disease of anatomical entity" like cardiovascular, endocrine, gastrointestinal, etc.), cellular proliferation diseases (DOID:14566), mental health diseases (DOID:150), metabolic diseases (DOID:0014667), physical disorders (DOID:0080015), and syndromes (DOID:225). We were able to map 2,454 out of 3,257 OMIMs to at least one of the above DO terms. On average, each OMIM was mapped to 1.01 DO,

[12] https://hpo.jax.org/app/

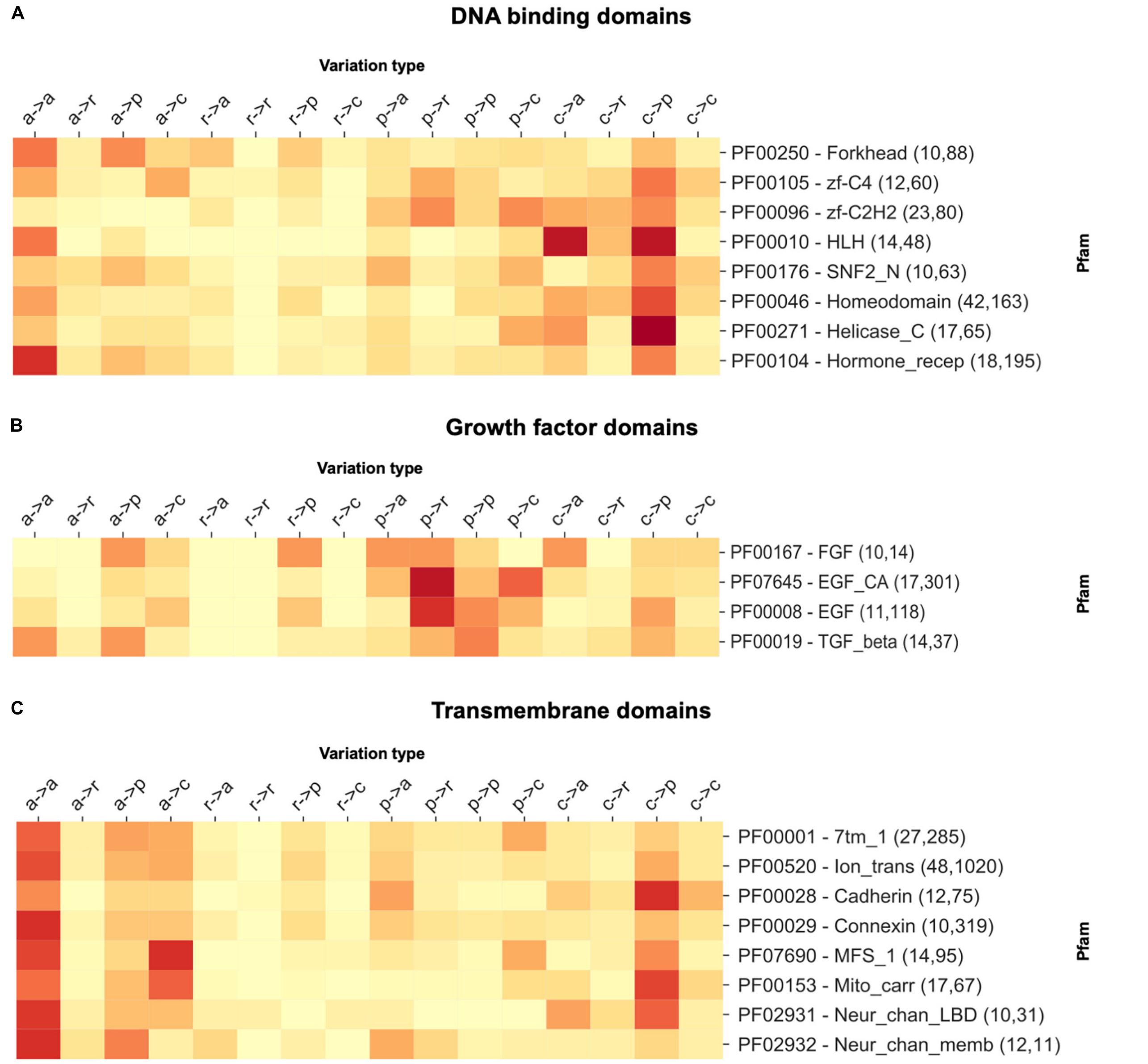


FIGURE 2 | The heatmap reporting the frequency of each variation type as observed within a selection of **(A)** DNA-binding, **(B)** growth factor, and **(C)** transmembrane domains. For each Pfam, the numbers within parentheses indicate the number of proteins and disease-related variations covered. In variation types, labels are as follows: a, apolar; r, aromatic; p, polar; and c, charged.

providing an almost strict classification of each OMIM into a single DO term.

With this mapping, we computed a Pfam-specific distribution of DO-associated disease classes. Complete results are reported in **Supplementary Table 4** for all the 1,670 Pfam entries considered in this study. The data provided in this study indicate that disease classes are not evenly distributed among different Pfam domains, again suggesting a differentiated association between the Pfam and phenotypes.

In **Figure 3**, we show an extract of our analysis, focusing on the 20 highest scoring Pfam domains associated with diseases. The heatmap reports, for each Pfam, the frequency of disease types (in the 17 different classes detailed above) as retrieved from OMIMs associated with substitutions occurring on the specific Pfam. In brackets, close to each Pfam name, we list the number of proteins, disease variations, and OMIMs associated to the Pfam.

Even in this case, the distributions of disease classes appear to be very different from the background (reported in the last row of the heatmap). Remarkably, the aggregation of Pfams into more general functional classes provides an additional level of interpretation. Considering **Figure 3**, we can observe that DNA-binding domains are mostly associated with syndromes, nervous system, and endocrine system disease classes, while enzymes are mostly involved in the metabolic disease

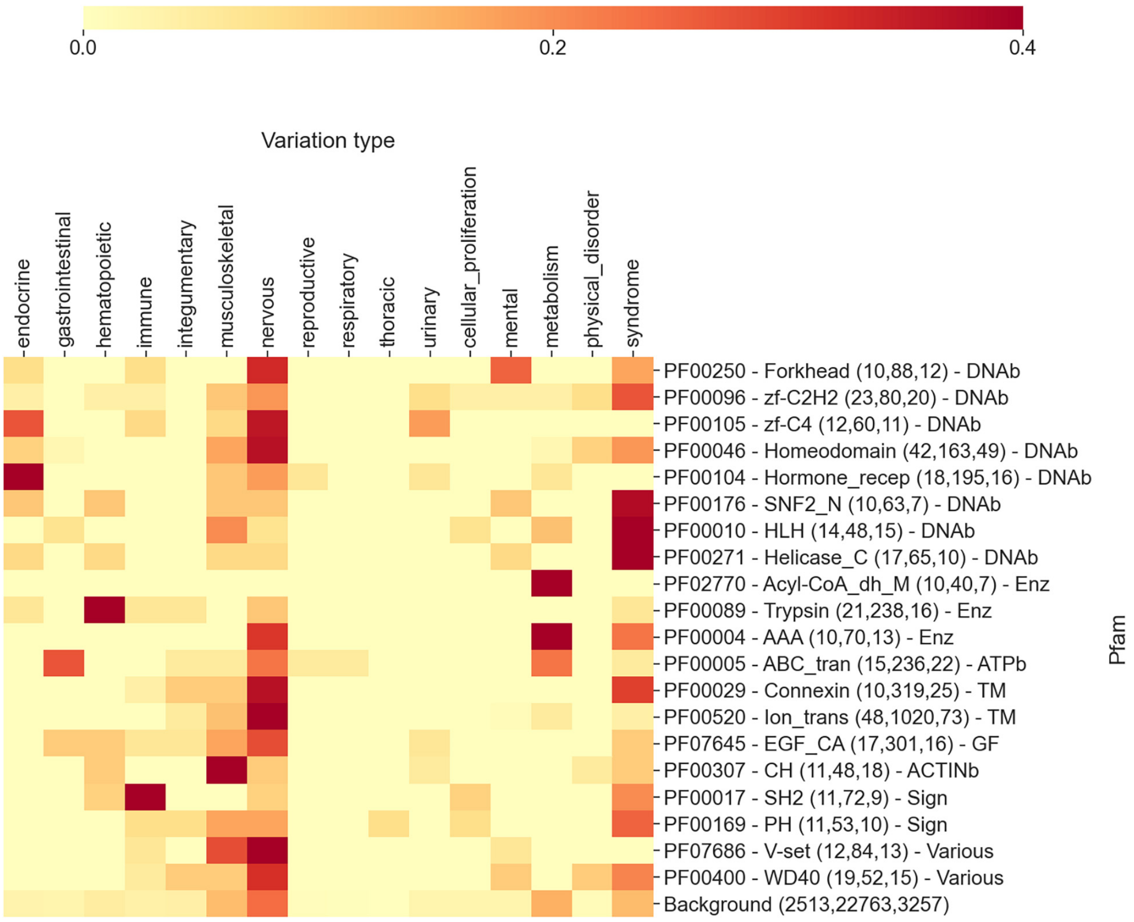


FIGURE 3 | The heatmap reporting, for each Pfam, the frequency of diseases (grouped into 17 different classes extracted from Disease Ontology) as retrieved from OMIMs, after the association *via* the disease type with Pfam. The numbers within parentheses are the number of proteins, the number of disease variations, and the number of Online Mendelian Inheritance in Man (OMIM) diseases associated with the Pfam, respectively. Each Pfam is labeled according to its functional class: DNAb, DNA-binding domain; Enz, enzymatic domain; TM, transmembrane; GF, growth factor; ACTINb, actin-binding domain; Sign, signaling; and Various, various functions associated. Mean and median Kullback–Leibler divergences (Eq. 2) between individual Pfam distributions and the background are 2.5 and 2.7 bits, respectively.

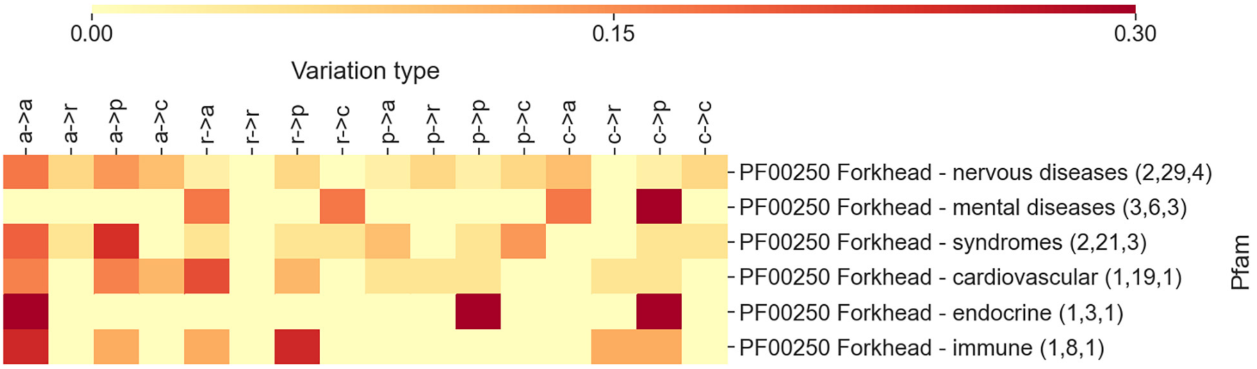


FIGURE 4 | The heatmap reporting, for the Pfam entry PF00250 Forkhead, the frequency of each variation type as observed after separating variations according to disease classes. The numbers within parentheses are the number of proteins, the number of disease variations, and the number of Online Mendelian Inheritance in Man (OMIM) diseases associated with the Pfam (and covered by DOID), respectively.

class. Transmembrane domains show the prevalence of nervous and integumentary disease classes, while growth factors and actin-binding domains are enriched in musculoskeletal diseases. Finally, signaling Pfam domains are prominently associated with immune system diseases. Overall, many of these findings are in line with what we expected. Protein domains have different functions and are involved into different biological processes. Variations occurring in these domains, when disruptive, lead to diseases that are connected to the biological processes in which the proteins are mainly involved. For instance, the fact that variations occurring in transmembrane domains are often linked to neurological diseases is a direct consequence of the involvement of transmembrane proteins (among other functions) in neurotransmission. Similarly, variations in enzymes routinely lead to metabolic diseases.

Some of the Pfams reported in **Figure 3** are associated to more than one disease types. For example, diseases that are associated to the Forkhead domain (PF00250) are distributed into five classes, namely, nervous, mental, endocrine, immune diseases, and syndromes. In **Figure 4**, an additional heatmap is shown trying to link the disease types to the patterns of variation types. Specifically, the patterns of variation types are reported after isolating variations linked to OMIMs in the different disease classes. Interestingly, the patterns show an evident difference among each other. This confirms the level of association that links domains to variation types and diseases.

CONCLUSION AND PERSPECTIVES

In this study, we consider, for the time being, only diseases of genetic origins, with the belief that cancer-related somatic variations are as yet not satisfactorily clustered according to tissue specificity of the plague.

This study, as well as the previous ones (Yates and Sternberg, 2013; Wiel et al., 2017, 2019), aims at establishing a direct mapping among variations, diseases, and phenotypes *via* the protein domains. Our novelty is the introduction of the variation type as a distinguished feature of association to the Pfam domain and to the phenotype. Our findings complement previous ones (Wiel et al., 2017) with the inclusion of the variation type, which adds to the classification of variations and their impact on the protein function, stability, and interaction in the specific context where the gene is active.

The link among the variation type, Pfam domain, and phenotype can greatly reduce the number of possible steps to understand which variations are disease-related or which are not and which phenotype they may promote. In perspective, the association among the variation type, protein domain/s, and phenotype may greatly simplify the problem of genetic variant annotation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RC, PM, and CS: conceptualization, methodology, and writing. CS: software. GB and CS: data curation and visualization. RC and PM: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.617016/full#supplementary-material

REFERENCES

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Babbi, G., Martelli, P. L., and Casadio, R. (2019). PhenPath: a tool for characterizing biological functions underlying different phenotypes. *BMC Genom.* 20, 548–558. doi: 10.1186/s12864-019-5868-x

Babbi, G., Martelli, P. L., Profiti, G., Bovo, S., Savojardo, C., and Casadio, R. (2017). eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genom.* 18, 554–564. doi: 10.1186/s12864-017-3911-3

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Martelli, P. L. (2011). Correlating disease related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* 32, 1161–1170. doi: 10.1002/humu.21555

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., et al. (2015). Finding our way through phenotypes. *PLoS Biol.* 13:e1002033. doi: 10.1371/journal.pbio.1002033

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Ittisoponpisan, S., Islam, S., Khanna, T., Alhuzimi, E., David, A., Sternberg, M., et al. (2019). Can predicted protein 3D-structures provide reliable insights into whether missense variants are disease-associated? *J. Mol. Biol.* 431, 2197–2212. doi: 10.1016/j.jmb.2019.04.009

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J. P., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 47, D1018–D1027. doi: 10.1093/nar/gky1105

Kucukkal, T. G., Petukh, M., Li, L., and Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Biol*. 32, 18–24. doi: 10.1016/j.sbi.2015.01.003

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., et al. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844. doi: 10.1093/nar/gkz972

Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles. Types, Performance, and Choice. *Hum. Mutat.* 37, 579–597. doi: 10.1002/humu.22987

Niroula, A., and Vihinen, M. (2017). Predicting Severity of Disease-Causing Variants. *Hum. Mutat.* 38, 357–364. doi: 10.1002/humu.23173

Ofoegbu, T., David, A., Kelley, L., Mezulis, S., Islam, S., Mersmann, S., et al. (2019). PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants. *J. Mol. Biol.* 431, 2460–2466. doi: 10.1016/j.jmb.2019.04.043

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi: 10.1093/nar/gkz1021

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi: 10.1093/nar/gkw1012

Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y., and Rost, B. (2016). Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLoS Comput. Biol*. 12:e1005047. doi: 10.1371/journal.pcbi.1005047

Savojardo, C., Babbi, G., Martelli, P. L., and Casadio, R. (2019). Functional and Structural Features of Disease-Related Protein Variants. *Int. J. Mol. Sci.* 20, 1530–1544. doi: 10.3390/ijms20071530

Schaafsma, G. C. P., and Vihinen, M. (2018). Representativeness of variation benchmark datasets. *BMC Bioinformatics* 19, 461–479. doi: 10.1186/s12859-018-2478-6

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 47, D955–D962. doi: 10.1093/nar/gky1032

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* 54, 1301–1333. doi: 10.1002/cpbi.5

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vihinen, M. (2017). How to Define Pathogenicity, Health, and Disease? *Hum. Mutat.* 38, 129–136. doi: 10.1002/humu.23144

Vihinen, M. (2018). Systematics for types and effects of DNA variations. *BMC Genomics* 28, 974–989. doi: 10.1186/s12864-018-5262-0

Wiel, L., Baakman, C., Gilissen, D., Veltman, J. A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum. Mutat*. 40, 1030–1038. doi: 10.1002/humu.23798

Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G., and Gilissen, C. (2017). Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum. Mutat.* 38, 1454–1463. doi: 10.1002/humu.23313

Yates, C. M., and Sternberg, M. (2013). Proteins and domains vary in their tolerance of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs). *J. Mol. Biol.* 425, 1274–1286. doi: 10.1016/j.jmb.2013.01.026

Zeng, Z., and Bromberg, Y. (2019). Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front. Genet.* 10:914. doi: 10.3389/fgene.2019.00914

Zhang, W., Chen, Y., Sun, F., and Jiang, R. (2011). Domain RBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Syst. Biol.* 5, 55–75. doi: 10.1186/1752-0509-5-55

Zhang, W., Coba, M. P., and Sun, F. (2016). Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships. *BMC Syst. Biol*. 10, 63–89. doi: 10.1186/s12918-015-0247-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fmolb.2021.633365

Insights Into Mutations Induced Conformational Changes and Rearrangement of Fe^{2+} Ion in *pncA* Gene of *Mycobacterium tuberculosis* to Decipher the Mechanism of Resistance to Pyrazinamide

Asma Sindhoo Nangraj[1†], Abbas Khan[1†], Shaheena Umbreen[2], Sana Sahar[3], Maryam Arshad[4], Saba Younas[5], Sajjad Ahmad[6], Shahid Ali[7], Syed Shujait Ali[7], Liaqat Ali[8] and Dong-Qing Wei[1,9,10]*

[1] Department of Bioinformatics and Biological Statistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] Department of Botany, University of Okara, Okara, Pakistan, [3] The Islamia University of Bahawalpur, Bahawalpur, Pakistan, [4] Government College University Faisalabad, Sahiwal, Pakistan, [5] University of Education, Lahore, Pakistan, [6] Department of Health and Biological Sciences, Abasyn University, Peshawar, Pakistan, [7] Center for Biotechnology and Microbiology, University of Swat, Swat, Pakistan, [8] Department of Biological Sciences, National University of Medical Sciences, Islamabad, Pakistan, [9] Peng Cheng Laboratory, Shenzhen, China, [10] State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Pyrazinamide (PZA) is the first-line drug commonly used in treating *Mycobacterium tuberculosis (Mtb)* infections and reduces treatment time by 33%. This prodrug is activated and converted to an active form, Pyrazinoic acid (POA), by Pyrazinamidase (PZase) enzyme. *Mtb* resistance to PZA is the outcome of mutations frequently reported in *pncA*, *rpsA*, and *panD* genes. Among the mentioned genes, *pncA* mutations contribute to 72–99% of the total resistance to PZA. Thus, considering the vital importance of this gene in PZA resistance, its frequent mutations (D49N, Y64S, W68G, and F94A) were investigated through in-depth computational techniques to put conclusions that might be useful for new scaffolds design or structure optimization to improve the efficacy of the available drugs. Mutants and wild type PZase were used in extensive and long-run molecular dynamics simulations in triplicate to disclose the resistance mechanism induced by the above-mentioned point mutations. Our analysis suggests that these mutations alter the internal dynamics of PZase and hinder the correct orientation of PZA to the enzyme. Consequently, the PZA has a low binding energy score with the mutants compared with the wild type PZase. These mutations were also reported to affect the binding of Fe^{2+} ion and its coordinated residues. Conformational dynamics also revealed that β-strand two is flipped, which is significant in Fe^{2+} binding. MM-GBSA analysis confirmed that these mutations significantly decreased the binding of PZA. In conclusion, these mutations cause conformation alterations and deformities that lead to PZA resistance.

Keywords: PZA, simulation, mutations, PCA, free energy

Edited by:
Daisuke Kihara,
Purdue University, United States

Reviewed by:
Natalia Kulik,
Institute of Microbiology, Academy of Sciences of the Czech Republic (ASCR), Czechia
Pavel Srb,
Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Dong-Qing Wei
dqwei@sjtu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 25 November 2020
Accepted: 07 April 2021
Published: 20 May 2021

Citation:
Nangraj AS, Khan A, Umbreen S, Sahar S, Arshad M, Younas S, Ahmad S, Ali S, Ali SS, Ali L and Wei D-Q (2021) Insights Into Mutations Induced Conformational Changes and Rearrangement of Fe^{2+} Ion in pncA Gene of Mycobacterium tuberculosis to Decipher the Mechanism of Resistance to Pyrazinamide. Front. Mol. Biosci. 8:633365. doi: 10.3389/fmolb.2021.633365

INTRODUCTION

Pyrazinamide (PZA), along with isoniazid (INH) and rifampin (RIF) is a very effective and fast therapy against persistent bacilli (Mitchison, 1985; Aggarwal et al., 2018). Pyrazinamide (PZase) encoded by the pncA gene of *Mycobacterium tuberculosis* (Mtb) transform this prodrug to pyrazinoic acid (POA). POA inhibits the proliferation of latent Mtb at very low pH values (Zhang and Mitchison, 2003; Malik et al., 2019). Studies have shown that resistance is developed against PZA due to mutations in three genes: *pncA, panD*, and *rpsA,* among which *pncA* gene mutations contribute to 72–99% resistance against PZA (Mitchison, 1985; Akhmetova et al., 2015; Miotto et al., 2017).

Mutations in the *pncA* gene have been mapped both in the coding as well as promoter region (Lemaitre et al., 2001; Miotto et al., 2014; Maningi et al., 2015). Recent investigations indicated that Pzase activity is affected due to mutations in D49A, Y64S, W68G and F94A positions (Miotto et al., 2014). The mentioned mutations have been shown to affect enzyme functionality drastically, and together with other reported mutations, influence protein structure integrity, solubility, function stability, and rate of expression (Petrella et al., 2011). More recently, novel pncA mutations are being described as liable to cause PZA resistance (Tan et al., 2014; Junaid et al., 2018).

The crystallographic structure of apo pyrazinamidase has been reported comprising six β-sheets covered by α-helices. This enzyme has metal and substrate binding sites. Iron (Fe^{2+} ion), histidine (His51, His57, and His71), and aspartate (Asp49) residues are part of the metal-binding site, whereas Asp8, Lys96, and Cys138 make the catalytic triads (Chaturvedi and Shrivastava, 2005; Petrella et al., 2011).

Computational approaches are now in routine to decipher mutations mediated biological mechanisms responsible for neutralizing the action of potent drugs. This atomic-level understanding holds great potential in de nova drug design and as such, speeds up novel drug discovery. In particular, advancements in molecular dynamics simulations allows scientist to analyze protein dynamics in environmental milieu replica of real biological cells (Khan et al., 2018a,b, 2019a, 2020a). It has been noticed that binding of PZA to the PZase enzyme altered protein's conformation, which is valuable data-keeping their importance in the quest of novel drug design. Likewise, MD simulations made it possible to study conformational variations in the three-dimensional structure of proteins that may arise following mutation(s) in the sequence (Khan et al., 2019c, 2020b,f). Thus MD simulations decrease time, costs and resources by reducing the number of cases for which experimental evaluation is required (Dolatkhah et al., 2017). We also investigated the molecular mechanism behind the resistance caused by D49N, Y64S, W68G, and F94A mutations (Stoffels et al., 2012; Miotto et al., 2014; Wan et al., 2020). Furthermore, extensive post-simulation analyses were employed to get insights into the atomic level with an ultimate objective to design novel chemical structures that can be effectively used in drug-resistant TB infections—with minimum side effects.

MATERIALS AND METHODS

PZase and PZA Structure Retrieval

The 3D structure of Mtb PZase (accession ID: 3PLI) and PZA (accession ID CID1046) were retrieved from the PDB databank (Rose et al., 2016) and PubChem (Kim et al., 2019) respectively. Water molecules were removed from the protein structure before starting downward analyses. As specific mutant structures of the enzyme were not available, mutations were introduced in the enzyme structure using PYMOL (DeLano, 2002) at particular locations.

Molecular Docking

Energy minimization steps were performed for PZA structure in Open Babel using Universal Force Field (Dallakyan and Olson, 2015). The ligand was optimized with default steepest descent and conjugate gradient algorithms in UCSF Chimera (Goddard et al., 2005). Docking was done in the PatchDock server, where binding conformation clusters were set at RSMD of 4.0 Å (Schneidman-Duhovny et al., 2005). Conformations with the lowest binding score were processed for molecular dynamics simulation using AMBER18 software (Wang et al., 2001; Case et al., 2005).

Impact of Mutations on Protein-Drug Interaction and Stability

Mutations' effect on protein thermodynamic stability was evaluated using mCSM[1] (Pires et al., 2013). The server utilizes graph-based signatures to predicts structural stability impact caused by mutation. mCSM accepts PDB files as input and a list of mutations to predict their effect on protein stability.

Molecular Dynamics Simulation

AMBER18 package was used to perform extensive MD simulations. This was done to investigate the stability of the PZA at the active site of both normal and mutant PZase. Parameters of protein were generated through ff14SB force field, and ligand preparation was done via Amber general force field (GAFF) (Wang et al., 2001; Case et al., 2005). MD simulations were performed for all five systems, including one wild (WT) and four mutants (D49N, Y64S, W68G, and F94A). Each system is solvated in TIP3P water box. Counter ions were added to each system to get charge neutralization. Afterward, two-step energy minimization procedure was adopted; (i) steepest decent minimizations of 6,000 cycles and (ii) conjugate gradient minimization of 3,000 cycles was applied on each system to remove steric clashes and allow system relaxation. Complexes were then heated to 300 K for 0.2 ns, followed by systems equilibration for 2 ns at 300 K. Temperature hold was achieved *via* Langevin thermostat (Zwanzig, 1973). For all systems, MD simulations production run was completed on GPU supported PMEMD code for 100 ns, and each simulation was repeated three times. Long-range electrostatic interactions (Darden et al., 1993; Essmann et al., 1995; Toukmaji et al., 2000) were detected with the particle mesh Ewald method using a cutoff distance of

[1] http:/biosig.unimelb.edu.au/mcsm/

10.0 Å. SHAKE method was applied for covalent bond treatment (Kräutler et al., 2001) (Salomon-Ferrer et al., 2013). CPPTRAJ and PTRAJ (Roe and Cheatham, 2013) packages in AMBER18 were considered for trajectories analysis.

Principal Component Analysis

Principal component analysis was utilized to measure structural fluctuations within the protein of all used complexes (Amadei et al., 1993). CPPTRAJ package calculated the covariance matrix based on Cα coordinates. Eigenvectors and eigenvalues estimation was performed by diagonalizing the covariance matrix, and these values indicate motion direction and fluctuation, respectively. In total, 5000 frames from each system MD trajectories were used to get PCA calculations. The plotting performed on PC1 and PC2 was used for motion monitoring. The lowest energy stable state was determined by the free energy landscape (FEL) and is indicated by deep valleys on the plot, whereas the intermediate state is shown by boundaries between deep valleys (Hoang et al., 2004). In this study, FEL calculations based on PCI and PC2 were obtained by the following equation:

$$\Delta G\,(\mathrm{PC1},\ \mathrm{PC2}) \ = \ -KBT\ln P\,(\mathrm{PC1},\ \mathrm{PC2})$$

Where KB indicates Boltzmann constant, PC1 and PC2 were used to estimate the reaction coordinates, and probability distribution P of the system is shown along PC1 and PC2.

Binding Affinity Estimation

PZA binding free energy with PZase (native and mutants) was estimated through MMPBSA.py script of AMBER over 500 snapshots of simulation trajectories (Miller et al., 2012; Mishra and Koča, 2018). The equation given below is used for binding free energy calculations

$$\Delta G_{\mathrm{bind}} = \ \Delta G_{\mathrm{complex}} - \ [\Delta G_{\mathrm{receptor}} + \ \Delta G_{\mathrm{ligand}}]$$

where ΔG$_{bind}$, ΔGcomplex, ΔGreceptor, and ΔGligand indicate net binding free energy, binding free energy of the complex, protein, and ligand, respectively. The following equation was used to calculate the value of each component:

$$G = G_{\mathrm{bond}} + G_{\mathrm{ele}} + G_{\mathrm{vdW}} + G_{\mathrm{pol}} + G_{\mathrm{npol}}$$

where the energy of bonds, electrostatic, van der Waals interactions, the polar and non-polar contributions are shown by the Gbond, Gele, GvdW, Gpol, and Gnpol, respectively. Whereas Gpol and Gnpol were calculated by the generalized Born (GB) implicit solvent method with SASA.

RESULTS

Mutant PZase Structural Modeling and Docking With PZA

The PZase apo structure (available as crystal structure) with ID: 3PL1 was retrieved from the protein databank and subjected to mutagenesis module in the PyMOL software where D49N, Y64S, W68G, and F94A mutants were created. Before molecular docking, all the structures were minimized by removing bad contacts from newly mutated residues as well as other residues. Following the minimization process, the docking process was completed blindly. Docking results suggested that our docking protocol is reliable, as indicated by the involvement of similar residues in interaction, as reported by a previous study (Junaid et al., 2018, 2020). Two residues such as His137 and Cys138, were reported to be involved in hydrogen bonding interactions with the oxygen of PZA. In the present study, similar results were obtained. The docking score of all complexes, including wild type and mutants, are tabulated in **Table 1**. The more negative binding energy implies better PZase-PZA intermolecular complementarity and higher binding affinity in contrast to the positive binding energy. The complex structure of wild type PZase and the PZA and its interaction pattern are given in **Figure 1A**. The docking score of PZA with both wild type and mutants PZase is in the following order: wild type (−5.21 kcal/mol), D49N (−4.75 kcal/mol), Y64S (−4.1 kcal/mol), W68G (−4.51 kcal/mol) and F94A (−4.18 kcal/mol). This data suggests that the PZA drug has a higher binding affinity for the wild type PZase enzyme in contrast to the mutants. Among the mutants, the lowest binding affinity of the PZA drug was noticed for Y64S and F94A. There is a high possibility that the mutations alter the active pocket conformation and thus not allowing proper PZA binding. The binding interaction pattern of each complex is given in **Figures 1–E**. MD simulations were performed on top scorer conformations to ascertain the effect of the mutation on the PZase structure as well as its binding with PZA.

Dynamics Characterization of Wild Type and Mutant Complexes

Mutations were found to confer instability in enzyme structure as predicted by mCSM web server and also by RMSD plots from a triplicate run of 100 ns MD simulations. mCSM server predicts the impact of each substitution by forecasting the change in conformational energy. As given in **Table 1**, it can be easily pointed that the given mutations induced greater instability compared to the wild type and hence classified as highly destabilizing. Among the four mutants, it was observed that W68G has a profound destabilizing effect on the PZase enzyme with ΔG of −3.14 kcal/mol. This was followed by F94A mutation that contributes to enzyme destabilizing change of −2.94 kcal/mol (Miotto et al., 2017). Among others, the predicted destabilizing energy change for Y64S is −2.2 kcal/mol whereas,

TABLE 1 | Molecular docking scores of the wild and mutant complexes. The mCSM predicted stability changes upon mutation. All the energies are given in kcal/mol.

S. No	Complex	Docking score	Predicted ΔΔG	Outcome
1.	Wild	−5.21	00	–
2.	D49N	−4.75	−2.00	Highly destabilizing
3.	Y64S	−4.1	−2.2	Highly destabilizing
4.	W68G	−4.51	−3.14	Highly destabilizing
5.	F94A	−4.18	−2.94	Highly destabilizing

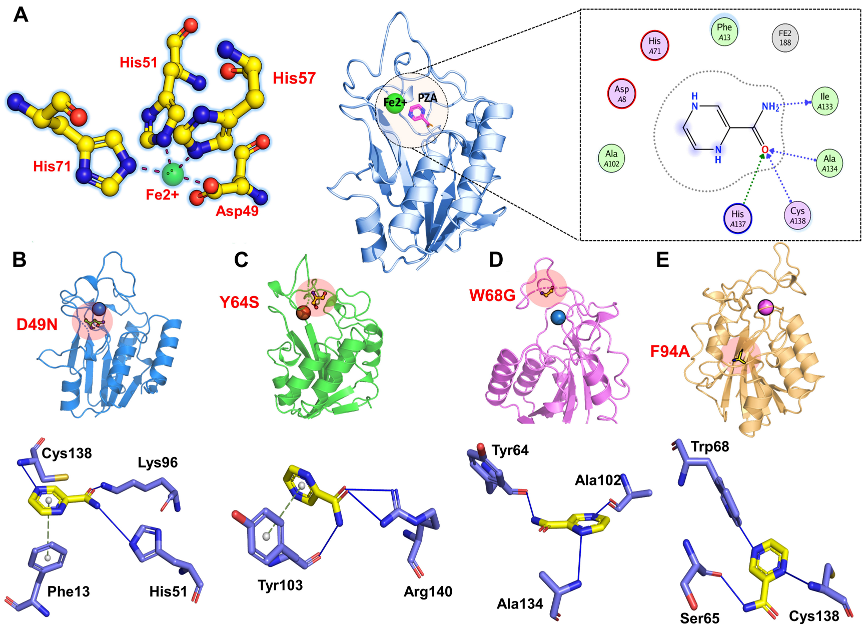


FIGURE 1 | 3D structure of the PZase along with the PZA drug and the Fe^{2+} metal shown in the circle. The figure also shows the binding of Fe^{2+} ion and PZA drug to the wild type **(A)**, D49N **(B)**, Y64S **(C)**, W68G **(D)**, and F94S **(E)**.

for D49N, the energy change is 2.0 kcal/mol. These findings are in line with the docking score of the systems and together, both analysis demonstrated the mutations are responsible for the change of PZase active pocket conformation, thus destabilize the binding network of the PZA drug, as can be seen in **Figure 1**.

For the stability assessment of each system, Cα atoms root-mean-square deviation (RMSD) was calculated based on simulated trajectory. The WT system reached an equilibrium state up to 60 ns, followed by a minor RMSD increased up to a maximum of 1.5 Å. Later, the RMSD continued over 1.5 Å with insignificant fluctuation (**Figure 2**). The D49N mutant system reached an equilibrium state of 2 Å in the first 20 ns, and then the RMSD fluctuated high throughout the simulation time due to system instability compared to WT. The Y64S system, like the WT gained equilibrium in the 50 ns and remained stable with slight fluctuations in the RMSD. The W68G system is in stable conformation till 30 ns with RMSD of 2.2 Å, then retained with RMSD at 1.5 Å and fluctuating slightly from the WT for the rest of time. The F94A system gains equilibrium in the first 10 ns and afterward showing minor fluctuations up to 2 Å. This unstable dynamics behavior of the mutants supports the enzyme conformation changes upon mutations to show resistance against PZA. Further inspection of Cα-RMSD rise for mutants compared to the WT showed that the D49N, Y64S, W68G, and F94A might weaken the active site residues interactions with the PZA. The RMSD of the mutant complexes is comparable with the wild type in terms of RMSD value, but the destability justifies that the different convergences at different intervals faced by the mutant structures but not in the wild type. This explanation of the wild and mutant complexes elucidates that due to small protein, the systems have reached the equilibrium point earlier. Furthermore, it can also be seen that the wild type reached the stability at 1.0Å; however, the other systems gained the equilibrium at ∼1.5Å, which shows the mutations induced structural perturbation in mutant complexes. The RMSD results for the other two replicates are given in **Supplementary Figures 1, 2**.

Local fluctuations due to mutations were examined through Cα, root-mean-square fluctuation (RMSF). Residues fluctuation was noted significantly in the mutant systems compared to the WT. WT system fluctuates at the N-terminus. The D49N mutant system reveals several point fluctuations as compared to WT and other mutations. The RMSF high fluctuation from the WT discloses that the mutations greatly affect the binding of the drug to the active site of the protein. The flexibility of the mutants may justify the binding differences, which can be better revealed by exploring the binding affinity differences. In the case of the mutants, the specific fluctuations at the site of the mutation can be easily distinguished. The RMSF of all the systems is given in **Figure 3**. The RMSF results for the other two replicates are shown in **Supplementary Figures 3, 4**.

Figure 4 presents broader distance distributions in mutants in contrast to WT, indicating more conformation dynamics in the former systems. As three residues: His51, His57, and His71 form a catalytic triad, it is important to understand the effect of these substitutions on the triad dynamics. It can be seen that the wild type, Y64S, and F94A showed a similar pattern of dynamics, while the D49N and W68G possess different triad distance network dynamics. Distinct changes of His57 is due to the loop harboring this residue. Fe^{2+} ion disturbance may reduce PZase activity and may explain the resistance phenomenon of these mutations. This effect was also confirmed by calculating the distance between PZA and the PZase. **Supplementary Figure 5** shows that the

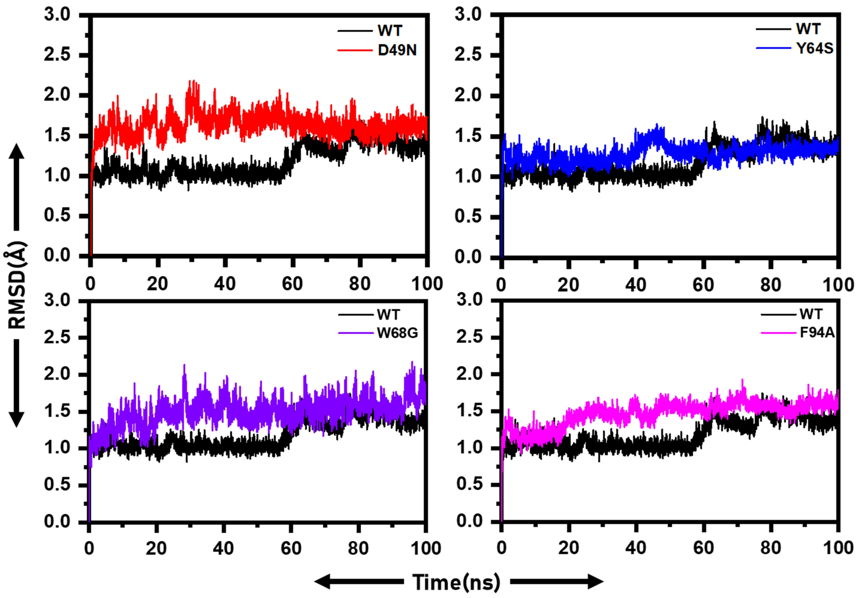


FIGURE 2 | RMSD of wild and mutants' complexes. RMSD of each mutant is superimposed on to the RMSD of the wild type. The X-axis shows the simulation time in nanoseconds, while the y-axis shows the RMSD in Angstrom.

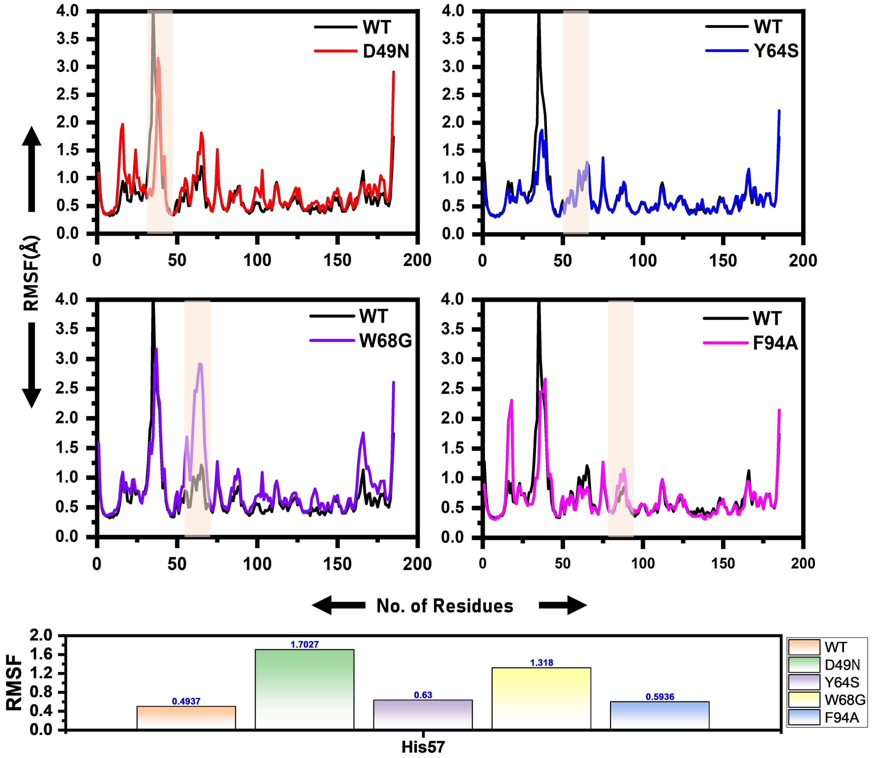


FIGURE 3 | RMSF of the wild and mutants' complexes. The RMSD of each mutant is superimposed over RMSF of the wild type. The x-axis shows residues number, while the y-axis shows RMSF in Angstrom. Shadowed regions depict enzyme amino acids stretch highly affected by the mutation.

distance between the wild type and the PZA is conserved, and the average distance reported was 8Å. However, this distance significantly fluctuates in the case of D49N, W68G, and F94A. While in the case of Y64S, the distance between the PZA and the receptor molecule remained somewhat similar to the wild type. Thus, these results also confirm that mutations have induced

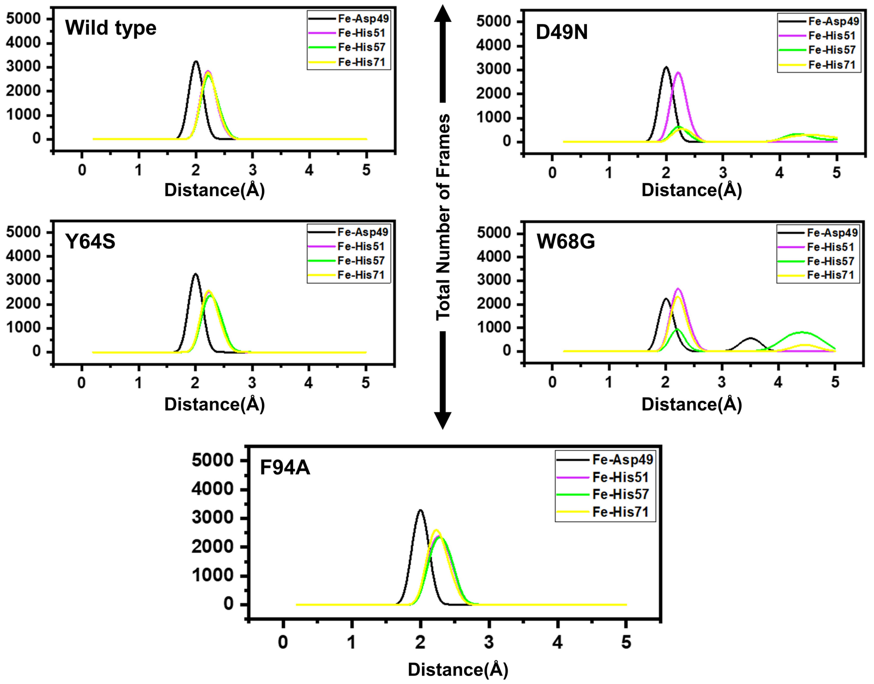


FIGURE 4 | Distances between the Fe^{2+} ion and its coordinating four residues. Furthermore, within each figure inside, there is a legend that shows the distance between Asp49 [O], His51[N], His57[N], His71[N] and Fe^{2+} ion. Each residue from the metal coordinates is differently colored.

structural destabilization and favor PZA unbinding due to their weak attachment.

Furthermore, we also calculated *Rg* to estimate the compactness of each system. The calculated Rg for each complex is given in **Supplementary Figures 6–8**. The results show that D49N is less compact than the wild type. For D49N initially, the higher Rg was observed, which then decreased; however, similar pattern of increasing and decreasing was experienced until 100 ns. In the case of Y64S, the Rg pattern was comparable with the wild type, but at 40–60 ns the Rg converged and a similar pattern was also observed between 95–100 ns.

Similarly, W68G systems were significantly affected. The Rg value significantly increased, and the average Rg was reported to be 15.6Å. The results of F94A and Y64S are comparable. No significant convergence was observed; however, at different intervals, the Rg increased.

Dimensionality Reduction and Clustering the Protein Motions

To understand the protein motion and cluster the related structural frames, PCA was performed. PCA is a mathematical method that transforms several correlated variables into smaller uncorrelated variables called principal components. To comprehensively understand the impact of the substitution on the protein motion initially, the eigenvectors were calculated and presented in **Figure 5**.

As given in **Figure 5**, the first three eigenvectors showed significant variations while rest of the eigenvectors showed localized fluctuations. It was reported that the wild type contributed 41% variance by the first three eigenvectors to the total motion. For D49N, Y64S, W68G, and F94A, variance contribution by the first three eigenvectors is 55, 41, 63, and 32%, respectively. These results, particularly the D49N and W68G mutations, are significantly in uniformity with the RMSD, RMSF, and Rg results because these two mutations significantly affected the overall dynamics of the proteins and PZA binding.

We further plotted the principal components (PC1 and PC2) to cluster the trajectories motion for a perusable understanding. The conformational transition from one to another is represented in different colors (red to blue). Given in **Figure 6**, each dot represents a single frame from the trajectory. The mutant complexes variable phase space as compared to the WT. Together,

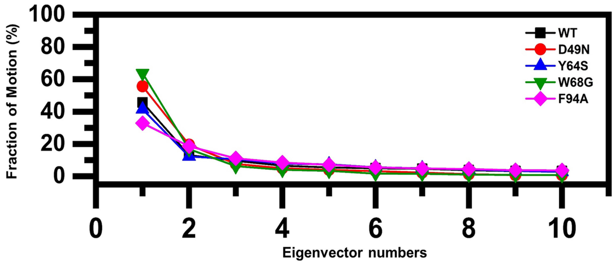


FIGURE 5 | Fractions of the first ten eigenvectors. Using the MD trajectory, the fraction of motions is calculated and given in percentage against the eigenvector numbers.

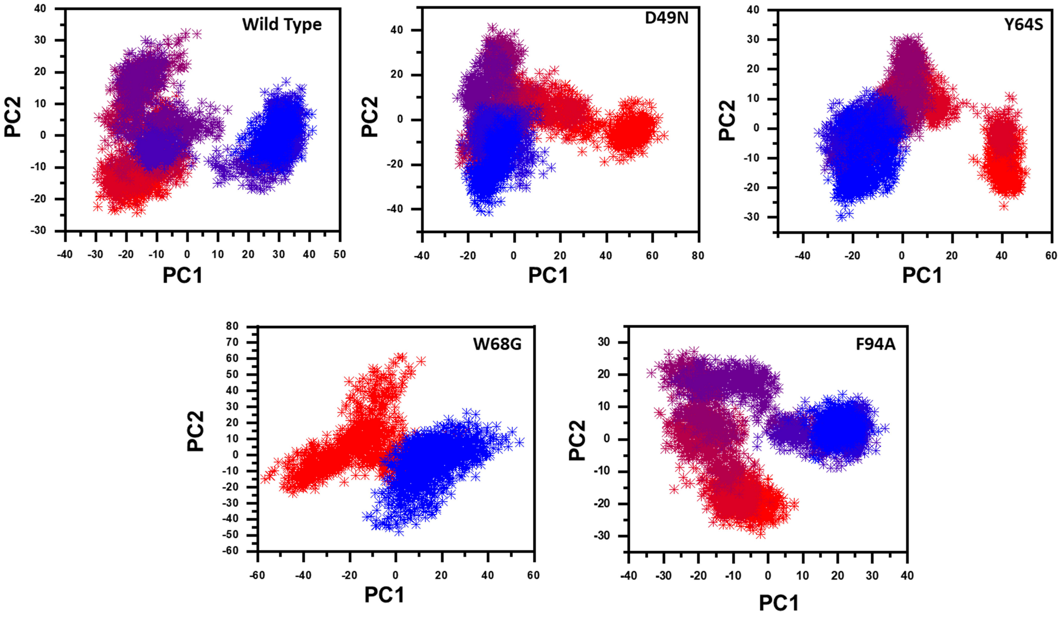


FIGURE 6 | Principal component analysis of all systems, including the Wild type and the four mutants. The first two principal components (PC1 and PC2) are used to project motion in the space phase at 300 K.

all these results indicate that mutations significantly affect the structure that has led to the resistance against PZA drug.

Destabilization of Fe^{2+} Ion by Mutations Induced Conformational Changes

Three histidine residues and one asparagine residue coordinate the Fe^{2+} ion. Li/Merz ion parameters for divalent Fe^{2+} ion was used to generate the topology. Mutations induced by Fe^{2+} destabilization during the simulation were determined by using the free energy landscape. It was found that Fe^{2+} is greatly influenced by the mutations. As given in RMSD and RMSF that the stability of each system is differentially affected, while the residual flexibility also showed variations. As presented in **Figure 7**, in the wild type structure, the Fe^{2+} did not move out significantly, but other regions showed little dynamic differences. The lowest energy conformation was attained at 92 ns. The only metastable state was extracted for wild type PZase is given below, which shows that the protein conformation is not altered during simulation.

On the other hand, as presented in **Figure 7**, the mutant system D49N showed destabilization of the Fe^{2+} ion. The structural coordinates extracted from the simulation trajectory at 70 ns represent the lowest conformation. In the case of D49N, the β-sheet two is significantly affected by transforming conformation.

Y64S has no significant effect on the enzyme and has the lowest energy conformation state attained at 72 ns. As given in RMSD and RMSF, the structural dynamics are not significantly affected by the Y64S mutation. All the analysis performed for Y64S in the manuscript discover consistent results and found Y64S as a comparatively less-lethal mutation than others. On the other hand, as reported above, W68G was significantly involved in structural destabilization and Fe^{2+} rearrangements. Along with the Fe^{2+} replacement and distortion of the coordination, the β-sheet 2 also flipped and thus causes a displacement of Asp49 residues that forms Fe^{2+} coordination along with the three histidine residues. The lowest conformational state of the W68G was extracted (5ns) after attaining the equilibrium (**Figure 8**).

Mutation Diminishes the Binding Affinity of PZA

The MM-GBSA approach was employed to assess the binding affinity of WT and mutated receptors and ligand [1,2]. The last 10 ns trajectory, 500 snapshots, were used as input to estimate dominant forces between the protein and ligand interactions. The total binding free energies ΔGbind of WT and mutants (WT/-8.13, D49N/-5.93, Y64S/-4.88, W68G/-4.02, and F94A/-4.03) were calculated in kcal/mol (**Table 2**). The total energies of mutants compared to the WT indicates that these mutations drop the binding strength of the PZA. The vdW, Elec, and ΔPS energies contribution to the binding energies of the mutants compared to the WT were significantly low. It explores that the mutated proteins have weak binding to PZA. Mutations that are not involved in the direct interaction with the PZA affect orientation coordination of active site residues involved in direct contact with the PZA.

DISCUSSION

Different studies have revealed that the administration of PZA, along with RIF and INH, is efficacious in treating Mtb infections (Gu et al., 2016; Khan et al., 2018c). Mtb resistance to these drugs

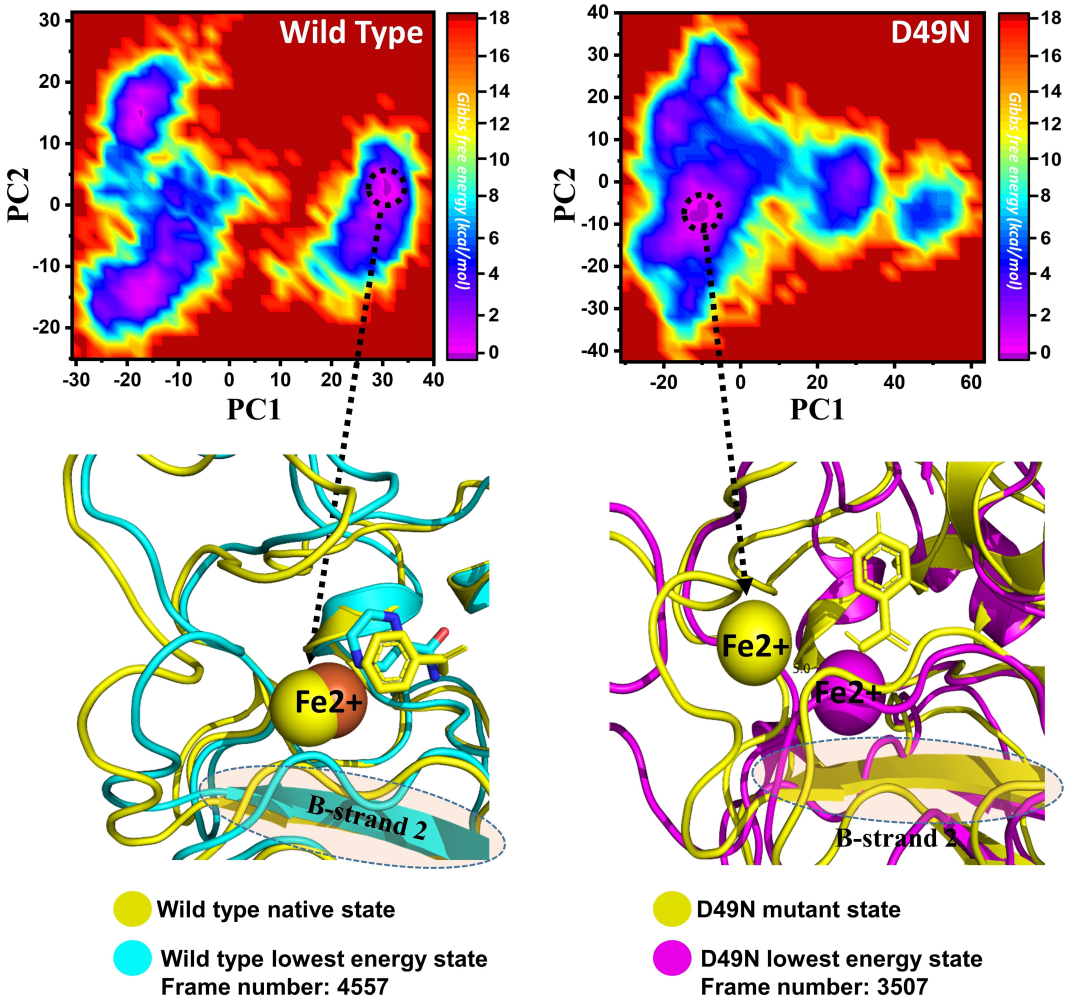


FIGURE 7 | Structural rearrangement of Fe^{2+} and the other regions in the protein given above (WT and D49N mutant). The lowest energy conformation from the wild type (92 ns) for the D49N (70 ns) was extracted and compared with the native state. The circle represents the lowest energy conformation.

renders front-line therapy ineffective, and as a consequence TB patient are exposed to a higher dose of the drugs. This leads to strong side effects on the patients and lower survival chances (Akhmetova et al., 2015; Khan et al., 2018d). Mitchison (1985), Miotto et al. (2014, 2017), Shi et al. (2014), Junaid et al. (2018), and Khan et al. (2020f). Since 1972, PZA was used as an active drug against the Mtb by targeting *panD* gene and had played key role in clearing persistent Mtb. Mutations in pncA, resulting in a loss of function of PZase, represent the primary molecular mechanism for PZA resistance in clinical strains. Pyrazinoic acid (POA) binds to the *pncA* active site and any conformational changes efflux the drug from the active site and this compromise the activity of the drug (Yadon et al., 2017). Additionally, the POA binding pocket is relatively small so any conformational changes result in unviability of the drug (Hewlett et al., 1995). It is clear that how the conformational changes affect the PZA binding, so, in this study, we selected D49N, Y64S, W68G, and F94A mutations at the Fe^{2+} binding site and PZA binding of PZase for their possible role in resistance to PZA (Miotto et al., 2017). We determined how the conformational changes may affect the binding of PZA and hinder the treatment of Mtb and eventually lengthen the eradication of tuberculosis.

TABLE 2 | shows the binding affinity comparison between the wild type and mutant systems.

Complex	vdW	Elec	Δ_{PS}	SASA	MMGBSA	Δ_{TS}	ΔG_{bind}
Wild	−19.25	−23.37	23.48	−3.19	−20.25	−12.12	−8.13
D49N	−16.46	−17.30	19.15	−6.54	−17.21	−11.28	−5.93
Y64S	−18.23	−19.58	17.11	−9.27	−18.05	−13.17	−4.88
W68G	−15.88	−17.21	13.54	−11.10	−15.25	−11.23	−4.02
F94A	−17.14	−15.23	8.22	−10.01	−13.32	−9.73	−4.03

All the energies are given in kcal/mol.

In this regard, *in silico* techniques such as MD simulations were utilized to study the said mutations role in PZase resistance to PZA. These methods are widely used to understand the mechanism of resistance and any binding perturbation

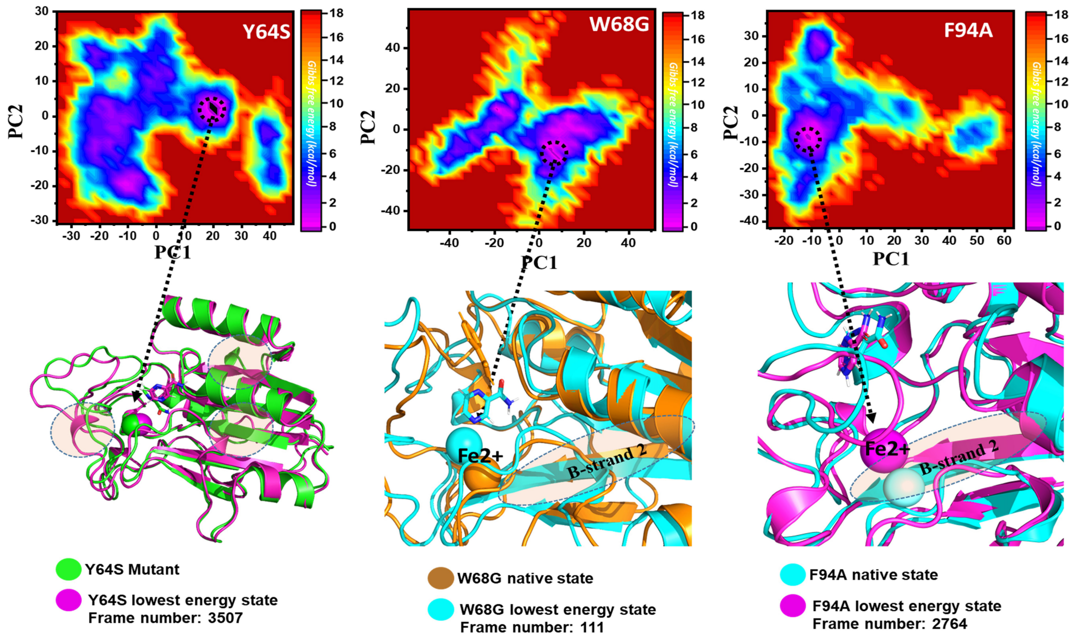


FIGURE 8 | Structural rearrangement of Fe^{2+} and the other regions in the protein given above (Y64S, W68G, and F94A mutants). The lowest energy conformation from each trajectory was extracted and compared with the native state. The circle represents the lowest energy conformation.

caused by mutations (Khan et al., 2020c,e, 2021a,b). It unveils conformational changes of proteins caused by any intrinsic mutations or ligand binding. This information is vital for devising novel strategies to combat drug resistance strains (Khan et al., 2019c, 2020b). Initial investigation of our selected mutations revealed that these mutations had altered the binding affinity of the PZA drug which shows that these mutations have clear role in resistance. Further characterization using biophysical tools revealed that RMSD and RMSF values suggest smaller fluctuations in wild type and higher fluctuations in mutant types. This probably suggest that wild type is highly stable, whereas more fluctuations in mutant type during the course of simulation suggest that the selected mutations are classified as highly destabilizing, and these findings are in line with previous experimental studies conducted on native and mutant (Q10P, D12A, G97D, R123P, T76P, G150A, H71R, W68R, W68G, and K96R). They reported that the mutations causes structural flexibility and thus weaken the drug binding (Khan et al., 2019b). The RMSF high fluctuation in mutant as compared to the WT discloses that the mutations have profound effect on the binding of the drug to the active site of the protein. A previous study carried out by *Muhammad et al.* also concluded that mutations in the PZA enzyme affect the binding orientation of PZA drug by shortening active pocket volume (Junaid et al., 2018, 2020; Khan et al., 2019c, 2020d). Findings of the current study may also suggest that said mutations affect the binding pocket, due to which the binding pocket volume as a whole is disturbed. Any distortion in the functional cavity volume might alter the binding affinity of PZA. This supports the previous study carried out by Vats et al. that mutations at the active pocket decrease the optimum affinity of the drug (Junaid et al., 2018, 2020). The residual flexibility also showed that each mutation displays a different frequency of fluctuations. Conformational dynamics, such as principal component analysis and free energy landscape, which are handy techniques reported by other studies, explored that the binding of Fe^{2+} is significantly affected. The main four residues coordinating the metal ion are disturbed during the simulation. In current study we observed different Rg pattern for the wild type and mutants. In case of wild type, initially Rg value increased and then it remains flat, whereas various patterns of increase/decrease were observed for all the mutants. These patterns suggest that the internal dynamics of each system is impacted by the mutation and eventually contributed to the PZA resistance. This notion is also supported by previous study (Jamal et al., 2020; Karmakar et al., 2020). The lowest energy minima conformation from each trajectory was extracted and compared with the native state that reported significant variations in Fe^{2+} binding and β–stands 2 specifically also confirmed by published literature (Khan et al., 2019c; Junaid et al., 2020). Analogous results have been reported that demonstrate that Fe^{2+} position is affected by catalytic and non-catalytic residues mutation. Furthermore, The Gibbs free energy to estimate the impact of the said substitutions on the binding of PZA. It was witnessed that mutations have significantly reduced the binding affinity of PZA and D49N and W68G being the major which contribute significantly to PZA resistance (Rehman et al., 2019). The study of dynamic behavior provided highly adequate knowledge on the PZase mutation that affected its structure, as well as perspectives into how conformational differences influence protein-ligand interactions which would aid the development of structure-based drug designing against the PZA target of Mtb.

In conclusion, we performed extensive MD simulations in triplets to explore the impact of D49N, Y64S, W68G, and F94A mutations on the PZase resistance to PZA. Our analysis revealed that these mutations affect stability, internal structural dynamics, and the binding energy of PZA. Our study further suggests that the stabilization of Fe^{2+} and β–stand 2 were affected. Hence, there is a dire need to design more potent drugs that would potently inhibit Mtb.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AN, AK, SU, SSA, and SA conceptualized the study. AK, SS, SY, SHA, and MA did the analysis. AK, MA, and SSA draft the manuscript. AN, AK, SHA, and D-QW revised and finalized the manuscript. LA performed the revision and writing improvement. D-QW is an academic supervisor and supervised the study.

FUNDING

D-QW is supported by grants from the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National Science Foundation of China (Grant Nos. 32070662, 61832019, and 32030063), the Science and Technology Commission of Shanghai Municipality (Grant No. 19430750600), the Natural Science Foundation of Henan Province (162300410060), as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14). The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.633365/full#supplementary-material

REFERENCES

Aggarwal, M., Singh, A., Grover, S., Pandey, B., Kumari, A., and Grover, A. (2018). Role of pncA gene mutations W68R and W68G in pyrazinamide resistance. *J. Cell. Biochem.* 119, 2567–2578. doi: 10.1002/jcb.26420

Akhmetova, A., Kozhamkulov, U., Bismilda, V., Chingissova, L., Abildaev, T., Dymova, M., et al. (2015). Mutations in the pncA and rpsA genes among 77 Mycobacterium tuberculosis isolates in Kazakhstan. *Int. J. Tuberc. Lung Dis.* 19, 179–184. doi: 10.5588/ijtld.14.0305

Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins* 17, 412–425. doi: 10.1002/prot.340170408

Case, D. A., Cheatham, T. E. III, Darden, T., Gohlke, H., Luo, R., Merz, K. M. Jr., et al. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* 26, 1668–1688. doi: 10.1002/jcc.20290

Chaturvedi, U. C., and Shrivastava, R. (2005). Interaction of viral proteins with metal ions: role in maintaining the structure and functions of viruses. *FEMS Immunol. Med. Microbiol.* 43, 105–114. doi: 10.1016/j.femsim.2004.11.004

Dallakyan, S., and Olson, A. J. (2015). Small-molecule library screening by docking with PyRx. *Methods Mol. Biol.* 1263, 243–250. doi: 10.1007/978-1-4939-2269-7_19

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N· log (N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

DeLano, W. L. (2002). Pymol: an open-source molecular graphics tool. *CCP4 Newsletter Protein Crystallogr.* 40, 82–92.

Dolatkhah, Z., Javanshir, S., Sadr, A. S., Hosseini, J., and Sardari, S. (2017). Synthesis, molecular docking, molecular dynamics studies, and biological evaluation of 4 h-chromone-1, 2, 3, 4-tetrahydropyrimidine-5-carboxylate derivatives as potential antileukemic agents. *J. Chem. Inf. Model.* 57, 1246–1257. doi: 10.1021/acs.jcim.6b00138

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593. doi: 10.1063/1.470117

Goddard, T. D., Huang, C. C., and Ferrin, T. E. (2005). Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. *Structure* 13, 473–482. doi: 10.1016/j.str.2005.01.006

Gu, Y., Yu, X., Jiang, G., Wang, X., Ma, Y., Li, Y., et al. (2016). Pyrazinamide resistance among multidrug-resistant tuberculosis clinical isolates in a national referral center of China and its correlations with pncA, rpsA, and panD gene mutations. *Diagn. Microbiol. Infect. Dis.* 84, 207–211. doi: 10.1016/j.diagmicrobio.2015.10.017

Hewlett, D., Horn, D. L., and Alfalla, C. (1995). Drug-resistant tuberculosis: inconsistent results of pyrazinamide susceptibility testing. *JAMA* 273, 916–917. doi: 10.1001/jama.1995.03520360030022

Hoang, T. X., Trovato, A., Seno, F., Banavar, J. R., and Maritan, A. (2004). Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7960–7964. doi: 10.1073/pnas.0402525101

Jamal, S., Khubaib, M., Gangwar, R., Grover, S., Grover, A., and Hasnain, S. E. (2020). Artificial intelligence and machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis. *Sci. Rep.* 10:5487. doi: 10.1038/s41598-020-62368-2

Junaid, M., Khan, M. T., Malik, S. I., and Wei, D.-Q. (2018). Insights into the mechanisms of the pyrazinamide resistance of three pyrazinamidase mutants N11K, P69T, and D126N. *J. Chem. Inf. Model.* 59, 498–508. doi: 10.1021/acs.jcim.8b00525

Junaid, M., Li, C.-D., Li, J., Khan, A., Ali, S. S., Jamal, S. B., et al. (2020). Structural insights of catalytic mechanism in mutant pyrazinamidase of Mycobacterium tuberculosis. *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1761879 [Epub ahead of print]

Karmakar, M., Rodrigues, C. H. M., Horan, K., Denholm, J. T., and Ascher, D. B. (2020). Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep.* 10:1875. doi: 10.1038/s41598-020-58635-x

Khan, A., Ali, S. S., Khan, M. T., Saleem, S., Ali, A., Suleman, M., et al. (2020a). Combined drug repurposing and virtual screening strategies with molecular dynamics simulation identified potent inhibitors for SARS-CoV-2 main protease (3CLpro). *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1779128 [Epub ahead of print]

Khan, A., Heng, W., Wang, Y., Qiu, J., Wei, X., Peng, S., et al. (2021a). In silico and in vitro evaluation of kaempferol as a potential inhibitor of the SARS-CoV-2 main protease (3CLpro). *Phytother. Res.* doi: 10.1002/ptr.6998 [Epub ahead of print]

Khan, A., Junaid, M., Kaushik, A. C., Ali, A., Ali, S. S., Mehmood, A., et al. (2018a). Computational identification, characterization and validation of potential antigenic peptide vaccines from hrHPVs E6 proteins using immunoinformatics and computational systems biology approaches. *PLoS One* 13:e0196484. doi: 10.1371/journal.pone.0196484

Khan, A., Junaid, M., Li, C.-D., Saleem, S., Humayun, F., Shamas, S., et al. (2020b). Dynamics insights into the gain of flexibility by Helix-12 in ESR1 as a mechanism of resistance to drugs in breast cancer cell lines. *Front. Mol. Biosci.* 6:159. doi: 10.3389/fmolb.2019.00159

Khan, A., Kaushik, A. C., Ali, S. S., Ahmad, N., and Wei, D.-Q. (2019a). Deep-learning-based target screening and similarity search for the predicted inhibitors of the pathways in Parkinson's disease. *RSC Adv.* 9, 10326–10339. doi: 10.1039/c9ra01007f

Khan, A., Khan, M., Saleem, S., Babar, Z., Ali, A., Khan, A. A., et al. (2020c). Phylogenetic analysis and structural perspectives of RNA-dependent RNA-polymerase inhibition from SARs-CoV-2 with natural products. *Interdiscip. Sci.* 12, 335–348. doi: 10.1007/s12539-020-00381-9

Khan, A., Khan, M. T., Saleem, S., Junaid, M., Ali, A., Ali, S. S., et al. (2020d). Structural insights into the mechanism of RNA recognition by the N-terminal RNA-binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein. *Comput. Struct. Biotechnol. J.* 18, 2174–2184. doi: 10.1016/j.csbj.2020.08.006

Khan, A., Rehman, Z., Hashmi, H. F., Khan, A. A., Junaid, M., Sayaf, A. M., et al. (2020e). An integrated systems biology and network-based approaches to identify novel biomarkers in breast cancer cell lines using gene expression data. *Interdiscip. Sci.* 12, 155–168. doi: 10.1007/s12539-020-00360-0

Khan, A., Saleem, S., Idrees, M., Ali, S. S., Junaid, M., Kaushik, A. C., et al. (2018b). Allosteric ligands for the pharmacologically important Flavivirus target (NS5) from ZINC database based on pharmacophoric points, free energy calculations and dynamics correlation. *J. Mol. Graph. Model.* 82, 37–47. doi: 10.1016/j.jmgm.2018.03.004

Khan, A., Zia, T., Suleman, M., Khan, T., Ali, S. S., Abbasi, A. A., et al. (2021b). Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: an insight from structural data. *J. Cell. Physiol.* doi: 10.1002/jcp.30367 [Epub ahead of print].

Khan, M., Malik, S., Bhatti, A., Ali, S., Khan, A., Zeb, M., et al. (2018c). Pyrazinamide-resistant mycobacterium tuberculosis isolates from Khyber Pakhtunkhwa and rpsA mutations. *J. Biol. Regul. Homeost. Agents* 32, 705–709.

Khan, M. T., Ali, S., Ali, A., Khan, A., Kaushak, A. C., Irfan, M., et al. (2020f). *Insight into the PZA Resistance Whole Genome of Mycobacterium Tuberculosis Isolates from Khyber Pakhtunkhwa, Pakistan*.

Khan, M. T., Junaid, M., Mao, X., Wang, Y., Hussain, A., Malik, S. I., et al. (2019b). Pyrazinamide resistance and mutations L19R, R140H, and E144K in Pyrazinamidase of Mycobacterium tuberculosis. *J. Cell. Biochem.* 120, 7154–7166. doi: 10.1002/jcb.27989

Khan, M. T., Khan, A., Rehman, A. U., Wang, Y., Akhtar, K., Malik, S. I., et al. (2019c). Structural and free energy landscape of novel mutations in ribosomal protein S1 (rpsA) associated with pyrazinamide resistance. *Sci. Rep.* 9:7482.

Khan, M. T., Malik, S. I., Ali, S., Sheed Khan, A., Nadeem, T., Zeb, M. T., et al. (2018d). Prevalence of pyrazinamide resistance in Khyber Pakhtunkhwa, Pakistan. *Microb. Drug Resist.* 24, 1417–1421. doi: 10.1089/mdr.2017.0234

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109.

Kräutler, V., Van Gunsteren, W. F., and Hünenberger, P. H. (2001). A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* 22, 501–508. doi: 10.1002/1096-987x(20010415)22:5<501::aid-jcc1021>3.0.co;2-v

Lemaitre, N., Callebaut, I., Frenois, F., Jarlier, V., and Sougakoff, W. (2001). Study of the structure–activity relationships for the pyrazinamidase (PncA) from Mycobacterium tuberculosis. *Biochem. J.* 353, 453–458. doi: 10.1042/bj3530453

Malik, S. I., Ali, S., Masood, N., Nadeem, T., Khan, A. S., and Afzal, M. T. (2019). Pyrazinamide resistance and mutations in pncA among isolates of Mycobacterium tuberculosis from Khyber Pakhtunkhwa. Pakistan. *BMC Infect. Dis.* 19:116.

Maningi, N. E., Daum, L. T., Rodriguez, J. D., Mphahlele, M., Peters, R. P., Fischer, G. W., et al. (2015). Improved detection by next-generation sequencing of pyrazinamide resistance in Mycobacterium tuberculosis isolates. *J. Clin. Microbiol.* 53, 3779–3783. doi: 10.1128/jcm.01179-15

Miller, B. R. III, McGee, T. D. Jr., Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). MMPBSA. py: an efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* 8, 3314–3321. doi: 10.1021/ct300418h

Miotto, P., Cabibbe, A. M., Feuerriegel, S., Casali, N., Drobniewski, F., Rodionova, Y., et al. (2014). Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *mBio* 5:e01819–14.

Miotto, P., Tessema, B., Tagliani, E., Chindelevitch, L., Starks, A. M., Emerson, C., et al. (2017). A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *Eur. Respir. J.* 50:1701354. doi: 10.1183/13993003.01354-2017

Mishra, S. K., and Koča, J. (2018). Assessing the performance of MM/PBSA, MM/GBSA, and QM–MM/GBSA approaches on protein/carbohydrate complexes: effect of implicit solvent models, QM methods, and entropic contributions. *J. Phys. Chem. B* 122, 8113–8121. doi: 10.1021/acs.jpcb.8b03655

Mitchison, D. (1985). The action of antituberculosis drugs in short-course chemotherapy. *Tubercle* 66, 219–225. doi: 10.1016/0041-3879(85)90040-6

Petrella, S., Gelus-Ziental, N., Maudry, A., Laurans, C., Boudjelloul, R., and Sougakoff, W. (2011). Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PLoS One* 6:e15785. doi: 10.1371/journal.pone.0015785

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2013). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691

Rehman, A. U., Khan, M. T., Liu, H., Wadood, A., Malik, S. I., and Chen, H.-F. (2019). Exploring the pyrazinamide drug resistance mechanism of clinical mutants T370P and W403G in ribosomal protein S1 of Mycobacterium tuberculosis. *J. Chem. Inf. Model.* 59, 1584–1597. doi: 10.1021/acs.jcim.8b00956

Roe, D. R., and Cheatham, T. E. III (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9, 3084–3095. doi: 10.1021/ct400341p

Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., et al. (2016). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* 45, D271–D281.

Salomon-Ferrer, R., Götz, A. W., Poole, D., Grand, S. Le, and Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* 9, 3878–3888. doi: 10.1021/ct400314y

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33(Suppl._2), W363–W367.

Shi, W., Chen, J., Feng, J., Cui, P., Zhang, S., Weng, X., et al. (2014). Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis. *Emerg. Microbes Infect.* 3:e58.

Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R., and Bifani, P. (2012). Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* 56, 5186–5193. doi: 10.1128/aac.05385-11

Tan, Y., Hu, Z., Zhang, T., Cai, X., Kuang, H., Liu, Y., et al. (2014). Role of pncA and rpsA gene sequencing in detection of pyrazinamide resistance in Mycobacterium tuberculosis isolates from southern China. *J. Clin. Microbiol.* 52, 291–297. doi: 10.1128/jcm.01903-13

Toukmaji, A., Sagui, C., Board, J., and Darden, T. (2000). Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* 113, 10913–10927. doi: 10.1063/1.1324708

Wan, L., Liu, H., Li, M., Jiang, Y., Zhao, X., Liu, Z., et al. (2020). Genomic analysis identifies mutations concerning drug-resistance and beijing genotype in multidrug-resistant mycobacterium tuberculosis isolated from China. *Front. Microbiol.* 11:1444. doi: 10.3389/fmicb.2020.01444

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2001). Antechamber: an accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc.* 222:U403.

Yadon, A. N., Maharaj, K., Adamson, J. H., Lai, Y.-P., Sacchettini, J. C., Ioerger, T. R., et al. (2017). A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat. Commun.* 8:588. doi: 10.1038/s41467-017-00721-2

Zhang, Y., and Mitchison, D. (2003). The curious characteristics of pyrazinamide: a review. *Int. J. Tuberc. Lung Dis.* 7, 6–21.

Zwanzig, R. (1973). Nonlinear generalized Langevin equations. *J. Stat. Phys.* 9, 215–220. doi: 10.1007/bf01008729

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 02 June 2021
doi: 10.3389/fmolb.2021.649990

Protein Predictive Modeling and Simulation of Mutations of Presenilin-1 Familial Alzheimer's Disease on the Orthosteric Site

Alejandro Soto-Ospina[1,2], *Pedronel Araque Marín*[3]*, *Gabriel Bedoya*[1], *Diego Sepulveda-Falla*[2,4] *and Andrés Villegas Lanau*[1,2]*

[1]Faculty of Medicine, Group Molecular Genetics, University of Antioquia, Medellín, Colombia, [2]Faculty of Medicine, Group Neuroscience of Antioquia, University of Antioquia, Medellín, Colombia, [3]School of Life Sciences, Research and Innovation in Chemistry Formulations Group, EIA University, Envigado, Colombia, [4]Molecular Neuropathology of Alzheimer's Disease, Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Alzheimer's disease pathology is characterized by β-amyloid plaques and neurofibrillary tangles. Amyloid precursor protein is processed by β and γ secretase, resulting in the production of β-amyloid peptides with a length ranging from 38 to 43 amino acids. Presenilin 1 (PS1) is the catalytic unit of γ-secretase, and more than 200 PS1 pathogenic mutations have been identified as causative for Alzheimer's disease. A complete monocrystal structure of PS1 has not been determined so far due to the presence of two flexible domains. We have developed a complete structural model of PS1 using a computational approach with structure prediction software. Missing fragments Met1-Glut72 and Ser290-Glu375 were modeled and validated by their energetic and stereochemical characteristics. Then, with the complete structure of PS1, we defined that these fragments do not have a direct effect in the structure of the pore. Next, we used our hypothetical model for the analysis of the functional effects of PS1 mutations Ala246GLu, Leu248Pro, Leu248Arg, Leu250Val, Tyr256Ser, Ala260Val, and Val261Phe, localized in the catalytic pore. For this, we used a quantum mechanics/ molecular mechanics (*QM/MM*) hybrid method, evaluating modifications in the topology, potential surface density, and electrostatic potential map of mutated PS1 proteins. We found that each mutation exerts changes resulting in structural modifications of the active site and in the shape of the pore. We suggest this as a valid approach for functional studies of PS1 in view of the possible impact in substrate processing and for the design of targeted therapeutic strategies.

Keywords: presenilin-1, modeling, simulation, quantum mechanics/molecular mechanics, familiar Alzheimer's disease mutations

OPEN ACCESS

Edited by:
Daisuke Kihara,
Purdue University, United States

Reviewed by:
Akio Fukumori,
Osaka University of Pharmaceutical Sciences, Japan
Satoru Funamoto,
Doshisha University, Japan

****Correspondence:***
Pedronel Araque Marín
pedronel.araque@eia.edu.co
Andrés Villegas Lanau
andres.villegas@udea.edu.co

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *06 January 2021*
Accepted: *22 April 2021*
Published: *02 June 2021*

Citation:
Soto-Ospina A, Araque Marín P, Bedoya G, Sepulveda-Falla D and Villegas Lanau A (2021) Protein Predictive Modeling and Simulation of Mutations of Presenilin-1 Familial Alzheimer's Disease on the Orthosteric Site.
Front. Mol. Biosci. 8:649990.
doi: 10.3389/fmolb.2021.649990

INTRODUCTION

Neurodegenerative diseases are characterized by impairment of the central nervous system (Bereczki et al., 2018). Many of these pathologies are produced by deposits of proteins as Huntingtin in the case of Huntington disease (HD), α-synuclein for Lewy body in Parkinson's disease (PD), neurofibrillary tangles by hyperphosphorylation of tau (τ) protein, and senile plaques by accumulation of β-amyloid

(Aβ) peptide in Alzheimer's disease (AD) (Myers, 2004; Paulsen, 2011; Jucker and Walker, 2012; Jucker and Walker, 2013). AD is the most common of neurodegenerative diseases, representing the largest number of reported cases worldwide (Sheikh et al., 2013; Prince et al., 2015).

AD must be divided into familiar AD (FAD) and sporadic AD (SAD). FAD is caused by the inheritance of mutations in the genes for amyloid precursor protein (APP), presenilin-1 (PSEN1), and presenilin-2 (PSEN2), and it can manifest in different ages (Rovelet-Lecrux et al., 2006; Guerreiro and Hardy, 2014; Shao et al., 2017). SAD is thought to be associated with known risk factors for other diseases, for example, high-cholesterol blood levels (dyslipidemia), oxidative stress, inflammation, low cognitive activity, and absence of physical activity (Ballard et al., 2011; Arbor et al., 2016; Yu and Zheng, 2012). The amyloidogenic theory states that AD is the result of the accumulation of Aβ, a fragment of the APP protein. APP metabolism follows two pathways, a non-amyloidogenic and an amyloidogenic pathway. In the non-amyloidogenic pathway, the enzyme α-secretase cleaves APP followed by cleavage by γ-secretase, a transmembrane protein complex, producing a small peptide of 23 amino acids (peptide p3) and an intracellular fragment known as the APP intracellular domain (AICD) (Lichtenthaler, 2012; Eggert et al., 2004). On the other hand, in the amyloidogenic pathway, APP is cleaved by β-secretase first and then by γ-secretase, producing Aβ peptides of diverse length (1–38 to 1–43 amino acids) (Vassar, 2004; Venugopal et al., 2008; Chávez-Gutiérrez et al., 2012). Aβ peptide structure facilitates its oligomerization, resulting in the accumulation of senile plaques in brain parenchyma (Wolfe et al., 1907), (Thal et al., 2008). The γ-secretase enzymatic complex includes four subunits: presenilin-1 (PS1), pharynx-defective 1 (Aph1), nicastrin (NCT), and presenilin enhancer-2 (PEN-2). PS1 is the catalytic unit of the γ-secretase complex, and its orthosteric site is located in aspartic acids 257 and 385, in transmembrane helix 6 and 7, respectively (Chávez-García et al., 2019). PS1 contains low mobility regions including nine α-helixes, and two high mobility regions, so far without a defined structure (Wolfe et al., 1907; Cacquevel et al., 2012; Bai et al., 2015). Only recently, a comprehensive structural analysis of PS1 was possible thanks to protein crystallization and cryogenic electron microscopy (cryo-EM). However, in order to obtain a crystal structure, flexible domains, such as amino acids Met1 to Glu 72 and Ser 290 to Glu 375, were not included in the sequence (Zhou et al., 2019).

The lack of a full crystal structure for PS1, including high-mobility regions, makes difficult to explain the possible role of some mutations, their impact on neuronal pathology, and it hinders the development of effective medical treatments. Moreover, protein loops can have special functions, including domain recognition and regulatory activities. For instance, PS1 loops seem to be responsible of the activation of the catalytic function (Wolfe et al., 1907; Knappenberger et al., 2004; Fukumori et al., 2010).

In order to provide a more precise correlation between PS1 structure and gamma secretase function, it is important to determine the localization and the 3D structure of PS1 missing fragments, using other tools such as bioinformatics and structural modeling. In this work, we have used three different predictive algorithms in order to complete the structural 3D model for PS1.

The dynamic methods that are part of macromolecular systems consist of computational simulations of particles in movement. Molecular dynamics utilize special algorithms to explain motion states and geometrical conformations for systems where several forces are acting simultaneously at various magnitudes of interactions and angles. These are always based on the classic Newtonian physic principles, but under rigid charge distribution (Nosé, 1984; Hospital et al., 2015). There are several structural models proposed in literature for the γ-secretase enzyme (Bai et al., 2015; Zhou et al., 2019; Bai et al., 2015). These are based on the role of protonation and deprotonation of the aspartic acids Asp257 and Asp385, in the substrate immobilization around the pore (Aguayo-Ortiz and Dominguez, 2018; Hitzenberger and Zacharias, 2019; Bhattarai et al., 2020). Additionally, it considers the ability of the enzyme to recognize the extracellular APP which contains the subunit NCT (Bolduc et al., 2015), which is in charge of constraining the substrate by means of hydrogen bonding. All these molecular effects considered in the whole molecular dynamic, along with the configurational arrangements, result in a rigid secondary structure of the enzyme (Hitzenberger and Zacharias, 2019; Aguayo-Ortiz et al., 2017). It is important to consider as well, the effect of all possible mutations of the PS1 that occur far from the active site, with the supporting proteins PEN2 and APH1 which can be modulated according to the protonation degree of the orthosteric site and evidencing that these simulations do not take into account a complete model of the catalytic subunit PS1, to represent the missing fragments of the protein and the correlation of various electronic effects (Chávez-García et al., 2019).

Several hybrid methods in quantum mechanical molecular field have been widely utilized for the study of the macro biological systems. These allow to register and quantify small changes that the enzyme undergoes, considering polarizable electrons as the most susceptible to the measurements of the missense-nonsynonymous variants with quantum methods such as functional density implementing the B3LYP or even the Hartree–Föck method (Murphy et al., 2000; Orlando and Jorgensen, 2010; Náray-Szabó et al., 2013; Roston et al., 2018; Siegbahn and Blomberg, 2018) to calculate the reaction coordinate and the formation or breaking of chemical bonds. These present some disadvantages, principally due to the limited amount of nucleus and electrons considered for a given biological system. Besides, these compute complex matrices that require high computational resources. On the other hand, semi-empirical methods seem to be a promising alternative in the study of biological and polyatomic systems as these intend to solve the Schrödinger equation from an approximated perspective, that is, considering an average between the electron interactions and appropriated theory levels, reducing the computational time. Of particular interest, the semi-empirical Austin model 1 is characterized because its parameters are derived from experimental data in order to solve the Schrödinger equation

(Carvalho et al., 2014; Christensen et al., 2016; Rafique et al., April 2019; Garcia et al., 2020; Grillo et al., 2020). These can be applied efficiently to large molecules to calculate their respective surface potentials (Foresman and Frisch, 1996; Levitt, 2014; Silva et al., 2015; Marín and Soto-ospina, 2020; Cano et al., 2021).

In this work, we have analyzed the structural changes in the active site resulting from seven selected mutations in TM6 and TM7 of PS1, utilizing a hybrid method of quantum mechanics/ molecular mechanics simulation. We have encountered that once a full 3D model for PS1 is achieved. Furthermore, the conformational effects of mutations Ala246GLu, Leu248Pro, Leu248Arg, Leu250Val, Tyr256Ser, Ala260Val, and Val261Phe, localized in the pore, can be explained as polarity changes, torsion angles, distance between helixes, and electronic structures. We proposed this analytical approach as the tool of choice for assessing mutational effects in structurally defined regions within proteins with multiple possible mutations in the nonflexible zones of PS1.

METHODOLOGY

Characterization of Protein PS1 Transmembrane Domains

A preliminary study of the structure of protein PS1 was performed by constructing a plot with the tool TMHMM available in the suite ExPASy (Sonnhammer and Krogh, 1998; Krogh et al., 2001; Möller et al., 2001), based on the primary sequence of the protein. We obtained a plot for the probability distribution through Hidden Markov Models. The position and score of the transmembrane fragments containing the structure inside–outside of cellular membrane were also determined (Artimo et al., 2012; Guex and Peitsch, 1997).

Missing Fragments Structural Prediction and Characterization of Obtained Models

Three different software tools for structure prediction were used to build a hypothetical model of the fragments representing the missing loops of PS1. The crystallized structures for PS1 reported in the protein data bank (PDB) were chose with IDs: 6IYC, 5A63, and 5FN2 subunit B (Presenilin-1) (Bai et al., 2015; Zhou et al., 2019; Bai et al., 2015). The two missing fragments were identified, and the primary sequence was built with a hypothetical model, using an algorithm that was defined to create the models based on homology constitution. The protein's active site was modeled based on the primary sequence of PS1 in FASTA format, and with the templates identified as 5A63 and 5FN2, which have aspartic acid in position 385 (D385) (Bai et al., 2015; Bai et al., 2015). The software tools used for the modeling were I-TASSER from Zhang Lab from the University of Michigan (Yang et al., 2015; Roy et al., 2010; Zhang, 2008) and Phyre2 (Protein Homology/analogY Recognition Engine V 2.0) from the structural bioinformatics group at Imperial College London (Kelly et al., 2015; Kelley and Sternberg, 2009). The models were refined with tools of the suite I-TASSER (Iterative Threading ASSEmbly Refinement), mainly using ModRefiner to optimize the energy from a native structure state and to improve the model for the interaction of backbone with hydrogen bond considering stereochemical optimization of the flexible behavior of the system (Xu and Zhang, 2011). Another refining tool used was the Fragment Guided of Molecular Dynamics (FG-MD). This software begins with classical modeling taking into account the geometrical optimization of angles and removing features that generated an unstable model. These features can improve steric clashes, geometry, and interactions by hydrogen bonding (Zhang et al., 2011). To follow other methodologies of elucidation with *de novo*, the software QUARK was utilized (or ran). This software is a tool for predicting a 3D model from each amino acid. This takes into account folded fragments of peptides and connects them via Monte Carlo simulation considering force fields and without utilizing templates (Xu and Zhang, 2012).

Validation and Minimization of Predicted Models

Each model was validated with an energetic tool from the suite SIB EXpASY, using Z-score values and QMEAN6 for the assessment (Petrey et al., 2003; Soni and Madhusudhan, 2017). Then, the stereochemical distribution was characterized using the EMBL-EBI Procheck software (Laskowski et al., 1993). Ramachandran plots for measuring the dihedral planes between the residues of the peptide bonds in the protein constitution were obtained and a calculation for the angles Phi and Psi in the model was performed. Rampage software from Cambridge University for Ramachandran plot analysis was used for measuring the same angles using another algorithm (Artimo et al., 2012; Petrey et al., 2003; Ramachandran et al., 1963; Crystallography and Bioinformatics Group, 2017). All models were visualized with the software Chimera UCSF version 1.1.1 and aligned through the algorithm 3D by match-maker under the Needleman–Wunsch algorithm. A BLOSUM62 matrix was used for the global alignment of the PS1 protein (Pettersen et al., 2004). Structural minimization was simulated using packages NAMD—Scalable Molecular Dynamics and VMD—Visual Molecular Dynamics, for the elimination of bad initial contacts, to avoid overlapping, and to generate fluidity in the models generated (Phillips et al., 2005; Humphrey et al., 1996). The loops for the best final model for the complete PS1 protein were finally refined with the Modeller software tool. This tool generates a normalized value of discrete and optimized protein statistical potential for the best rotamers in the lateral chain. This is implemented under iterative cycles that consider possible spatial restrictions (Webb and Sali, 2016). Eventually, transversal views of modeled PS1 proteins were obtained using Chimera UCSF v.1.11.

Hydropathicity Index and Phosphorylation Sites Prediction

The analysis of the polarity in the systems was performed with the software ProtScale available in the suite ExPASy, using as measurements the Kyte and Doolittle coefficients. Also, the primary sequences of the wild-type PS1 protein and studied

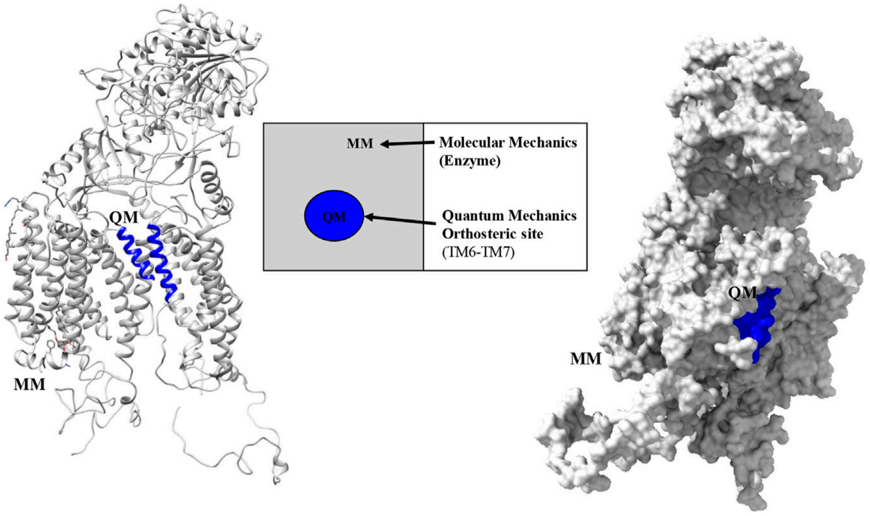


FIGURE 1 | Partition regions observed with the QM-MM hybrid method for PS1 protein in the γ-secretase enzyme.

mutants were assessed, and the hydropathy index was calculated for each amino acid, labeling it as hydrophobic or hydrophilic (Gasteiger et al., 2005). The NetPhos 3.1 Server was used to predict serine, threonine, or tyrosine phosphorylation sites susceptible to phosphorylation by diverse kinases, for instance, PKA, PKC, PKE, RSK, EGFR, or MAPK38 kinases (Artimo et al., 2012; Blom et al., 2004; Blom et al., 1999).

Hybrid Method for Quantum Mechanics and Molecular Mechanics Modeling

The final modeled structure of PS1 protein was used to study the functional behavior of mutations located in TM6: Ala246GLu, Leu248Pro, Leu248Arg, Leu250Val, Tyr256Ser, Ala260Val, and Val261Phe. The hybrid method examines the system based on the z-matrix obtained from the topological consideration for each nucleus, frozen for molecular mechanics purposes. The modeled system used the localization of α-helix TM6 amino acids 243–263 and TM7 383–398 as a representation of the active site. Then, a quantum mechanics calculation was applied to it, with a level of theory that consider the number of atoms and parametrization of the system, using data derived from experimental analysis published in databases for protein structure. However, the selected subsystem for implementing molecular mechanics (MM) observes the whole structure within a classical physics-based description of the remaining PS1 protein. The QM/MM fragment is considered with QM polarization due to the classical MM region for the TM6 and TM7 α-helixes, as shown in (**Figure 1**). The specific method used for this analysis was semi empirical applying the force field Austin model 1 (AM1) to the QM region (Dewar et al., 1993). This method considers the average interaction among electrons to solve the Schrödinger equation in protein macro systems. This method for geometric optimization is very useful given that the study of the canonical function considers 243–261 amino acids for region TM6 and 383–298 amino acids for region TM7. These have, in total, 280 atomic nuclei and 1,560 electrons for electronic description. Considering this, high-level quantum theory methods cannot be easily applied given the required computational resources and the costs of the calculation. The system's total MM region and the classical description of the QM region is carried out with the MMFF(aq) by estimating the system's second solvation sphere. The interface was also saturated with hydrogen atoms (Marques et al., 1995; Halgren, 1996; Halgren, 2000; Mackerell, 2004; Alexeev et al., 2013; Zhou et al., 2019; Soni et al., 2020). The molecules were optimized based on the data of global minimum geometry and energy using the Spartan 18′ software for wave function. This software has a tool for the determination of Spartan surfaces that simulate the optimized structure for each surface, such as density, potential–potential, ionization, orbitals Homo, orbitals Lumo, and electrostatic potential map. All of them resulted from the structural changes in the active site and applied to the electronic structure and their possible interactions.

The total energy in the system is calculated with the equation for mechanical integration:

$$E_{\text{total}} = EQM + EMM_{(\text{total})} - EMM_{(QM)}, \tag{1}$$

with this equation and under a multiscale analysis, the total energy is obtained for wild type and mutations models, tripled for each one, considering the average energy to be included in **Eq. 1**, and with a low standard deviation for each one of the determined systems (Maseras, 1999; Cao and Ryde, 2018).

RESULTS AND DISCUSSION

Modeling of Loop Fragments of Protein PS1

The PS1 structural template used as a baseline for the full structure prediction, was the one published by Zhou et al., Protein Data Bank (PDB) ID: 6IYC, given that it is the most complete structure up to date. Moreover, this structure was

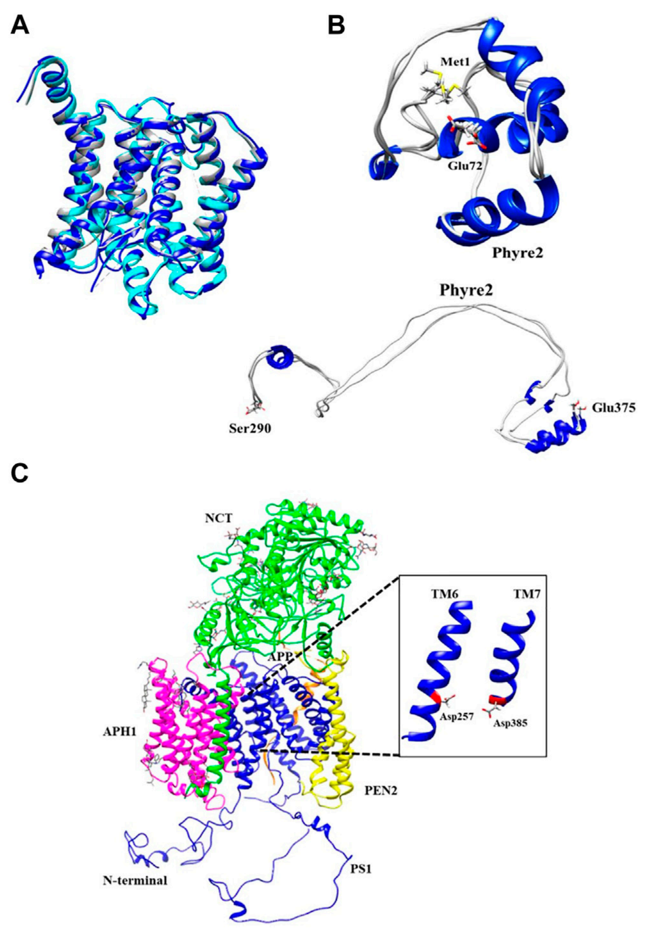


FIGURE 2 | Phyre2 modeling for PS1 structure in the γ-secretase complex. **(A)** Structural alignment of cryo-electron microscopic structure PS1 protein domains as found in PDB ID: 6IYC (Blue), 5FN2 (Cyan), and 5A63 (Gray). **(B)** Refined and unrefined modeling of PS1 N-terminal fragment Met1-Glu72 and Ser290-Glu375, using Phyre2. **(C)** Modeling of the active γ-secretase heteromer. Inset shows PS1 transmembrane domains forming the pocket including TM6 and TM7 together with the active sites Asp257 and Asp385.

aligned with two other structures reported previously PDB ID: 5FN2 and PDB ID: 5A63 (Lu et al., 2014; Bai et al., 2015). Some minimum structural differences were found between the three models (**Figure 2A**). Missing fragments were incorporated as dotted lines. The specific amino acid position of the missing fragments was determined using the tool "sequence" of the Chimera U.C.S.F software as shown in **Supplementary Figure S1A**. The quantitative alignment with a graphical color code for root mean squared deviation (RMSD) showed a high similarity index and identity percentage, as corroborated by the similar topology between models (**Supplementary Figure S1B**).

The Hidden Markov Model software was used to confirm the different transmembrane passes of PS1 based on its primary sequence. Missing fragments are also visible with low probability for a transmembrane pass (**Supplementary Figure S2A**). The primary sequences for each missing fragment (Met1-Glu72 and Ser290-Glu375) were modeled as a tertiary structure. The model of each fragment was generated using structure predictors Phyre2, I-Tasser, and Quark for folding recognition and structural distribution, using homology comparison algorithms that apply forces derived from primary sequences to the model, and compares them with structures identified experimentally. Posteriorly, models were refined using ModRefiner and Fragment Guided Molecular Dynamics (FG-MD). Resulting structures were aligned with Chimera U.C.S.F, and in order to obtain a structural arrangement from the different models for the two missing fragments, Needleman Wunsch and matrix blosum62 approaches were used. Each of the structure predictors applied a different algorithm of assembly together with homology modeling, protein threading for fold recognition from primary sequences, and assembly without a template, using free modeling *ab initio* taking into account force fields in order to produce the spatial distribution of the models. Subsequently, each model was refined with ModRefiner and FG-MD to improve visualization and analysis of not covalent interactions such as hydrogen bonding, disulfide bridges, hydrophobic, and hydrophilic interactions. The final models obtained with the three different predictive software were validated using QMEAN software for energetic calculations and Procheck software for stereochemical analyses. For Met1-Glu72, the best loop model was obtained with the software Phyre2 (**Figure 2B**) with an energetic QMEAN6 value of 0,472 in a range between 0–1. The value for Z-score was set similar to the structural size, −1,867. Ramachandran plot of the favorable region was 71.60%, of the allowed region was 20.90%, of the generously allowed region was 4.50%, and of the forbidden region was 3.00%. The best loop model for the fragment Ser290-Glu375 was obtained also with Phyre2, with an energetic QMEAN6 value of 0.455 in a range between 0–1, Z-score with a value of similarity in structural size −2.331. Ramachandran plots of favorable region was 83.3%, of the allowed region was 12.8%, of the generously allowed region was 1.3%, and of the forbidden region was 2.6% (**Table 1**).

According to the results that were calculated from the QMEAN6 energetic parameter in **Table 1**, Phyre2 was the best structure predicting software for the missing regions (Met1-Glu72 and Ser290-Glu375). This is because it achieved the highest score values in energetic status characterization for the modeled regions compared with the other *ab initio* structure predictors (I-TASSER and QUARK). Lower Z-score value indicated the quality of the models obtained with Phyre2, by assessing the viability of the hypothetical models in relation to structures obtained experimentally that share the same range of values. Finally, the Ramachandran plot showed the distribution of each residue of protein and its dihedral plane with percentages of some residues in the four quadrants of the Cartesian plane i.e., the *x*-axis (Phi) angle and the *y*-axis (Psi) angle. This information was used to validate and assemble the secondary structure of the fragments. We could observe that the Phyre2 models provided high percentage of favorable regions, in this case after refinement. In consequence, the software I-TASSER and Quark *ab initio* generated models (**Supplementary Figure S2B**) were not selected for further assembly with the rest of the PS1 3D model.

Met1-Glu72 and Ser290-Glu375 fragments generated using Phyre2 were integrated into the PS1 6IYC template as obtained

TABLE 1 | Energetic and stereochemical validation of missing fragments of PS1 protein.

Missing region PS1	Model	Software	QMEAN6	Zscore	Ramachandran plot			
					Favorable region	Allowed region	Generously allowed region	Forbidden region
Met1-Glu72 (M1-E72)	Unrefined	I-TASSER	0.263	−3.379	59.70%	28.40%	4.50%	7.50%
	Refined	FG-MD	0.291	−3.176	62.70%	26.90%	4.50%	6.00%
		ModRefiner	0.235	−3.575	73.10%	22.40%	1.50%	3.00%
Met1-Glu72 (M1-E72)	Unrefined	Phyre2	0.472	−1.867	71.60%	20.90%	4.50%	3.00%
	Refined	FG-MD	0.393	−2.438	58.20%	35.80%	6.00%	0.00%
		ModRefiner	0.232	−3.602	88.10%	7.50%	3.00%	1.50%
Met1-Glu72 (M1-E72)	Unrefined	Quark *ab initio*	0.368	−2.62	84.00%	10.70%	5.30%	0.00%
	Refined	FG-MD	0.286	−3.209	65.70%	20.90%	7.50%	6.00%
		ModRefiner	0.309	−3.041	86.60%	10.40%	1.50%	1.50%
Ser290-Glu375 (S290-E375)	Unrefined	I-TASSER	0.421	−2.606	64.10%	34.60%	1.30%	0.00%
	Refined	FG-MD	0.430	−2.529	65.40%	33.30%	1.30%	0.00%
		ModRefiner	0.445	−2.428	82.30%	17.70%	0.00%	0.00%
Ser290-Glu375 (S290-E375)	Unrefined	Phyre2	0.450	−2.374	60.30%	23.10%	11.50%	5.10%
	Refined	FG-MD	0.449	−2.377	56.40%	32.10%	11.50%	0.00%
		ModRefiner	0.455	−2.374	83.30%	12.80%	1.30%	2.60%
Ser290-Glu375 (S290-E375)	Unrefined	Quark *ab initio*	0.404	−2.74	79.10%	13.90%	7.00%	0.00%
	Refined	FG-MD	0.353	−3.151	53.80%	41.00%	2.60%	2.60%
		ModRefiner	0.376	−2.969	76.90%	17.90%	3.80%	1.30%

via Cryo-EM. Consequently, a molecular dynamics approach using the VMD/NAMD package was applied to the assembled structure for minimization of all atoms in order to remove any poor initial contacts, to avoid overlapping and to facilitate the fluidity of the model. The resulting full structural model for PS1 in ribbon (**Supplementary Figure S3A**) and surface density (**Supplementary Figure S3B**) was then assembled into the γ-secretase model in which the 6IYC template was originally included (Zhou et al., 2019) (**Figure 2C**). It can be observed that the full hypothetical model for PS1 was not perturbed by the putative N-terminal and loop fragments (**Supplementary Figures S3A,B**), and that it does not present a structural change in comparison with the original template. In fact, the topology of active sites, Asp257 and Asp385 remain as previously reported (Zhou et al., 2019) (**Figure 2C**, inset). In consequence, the active pore structure modeled can be further used for the analysis of the structural effects induced by the pathogenic mutations in PS1 that directly modify amino acids in transmembranal domains 6 (TM6) and 7 (TM7). Although a molecular dynamics approach could also be used to evaluate the mobility of the different components of the gamma secretase complex and to assess possible effects of flexible fragments in the active pore, our approach using a hybrid quantum mechanics/molecular mechanics (QM/MM) method allows more sensitivity in the measurement of small topological changes and electronic structural modifications (Van Der Kamp and Mulholland, 2013; Omer et al., 2015; Hofer and de Visser, 2018).

Functional Analysis of Mutations in the Orthosteric Site TM6

After determining that the structure of the pore is unaffected by PS1 flexible domains, some mutations were selected from the AD mutations database in Alzforum (Alzforum, 2020). Seven different missense mutations located in TM6 close to the active site (Asp257) were selected: Ala246GLu, Leu248Pro, Leu248Arg, Leu250Val, Tyr256Ser, Ala260Val, and Val261Phe. In order to obtain the most sensitive assessment of topological and electronic structure changes generated by these amino acid substitutions, we applied a hybrid QM/MM approach, including the evaluation of electronic potential, ionization potential, and electrostatic surfaces. For the evaluation of polyatomic systems, we chose the force-field Austin Model 1 (AM1), a semi-empirical method for quantum calculations. In this way, we can obtain a description of the modifications in electronic correlations and changes in atomic nuclei topology when comparing wild-type and mutated PS1. Each PS1 mutation induces specific effects in the protein structure. These effects can be on the topology, the electronic surface, or the electrostatic potential. For each mutation, one of these possible changes generates a stronger impact on the structure of the pore, depending on the distance between the amino acid substituted and the active site Asp257.

Topological Changes Induced by Mutations Ala246Glu, Leu248Pro, and Leu248Arg in PS1

Mutations Ala246Glu, Leu248Pro, and Leu248Arg have effects in the chemical properties of the environment of the pore and a direct effect in the secondary structure of the protein in TM6 and TM7. Mutation Ala246Glu presents a chemical change that increases the polarity due to the high electronegativity conferred by adding two oxygen molecules when substituting alanine by glutamic acid. Increased electronegativity induces the formation of transient dipoles, favoring noncovalent interactions, for instance, hydrogen bonding or acid–base reactions with the adequate distances equal or less 2.7 Å (**Figure 3A**).

In wild-type PS1, position 246 is occupied by alanine, which is not polar and it cannot interact by hydrogen bonding. The lack of polarity in this position brings on London dispersion interactions

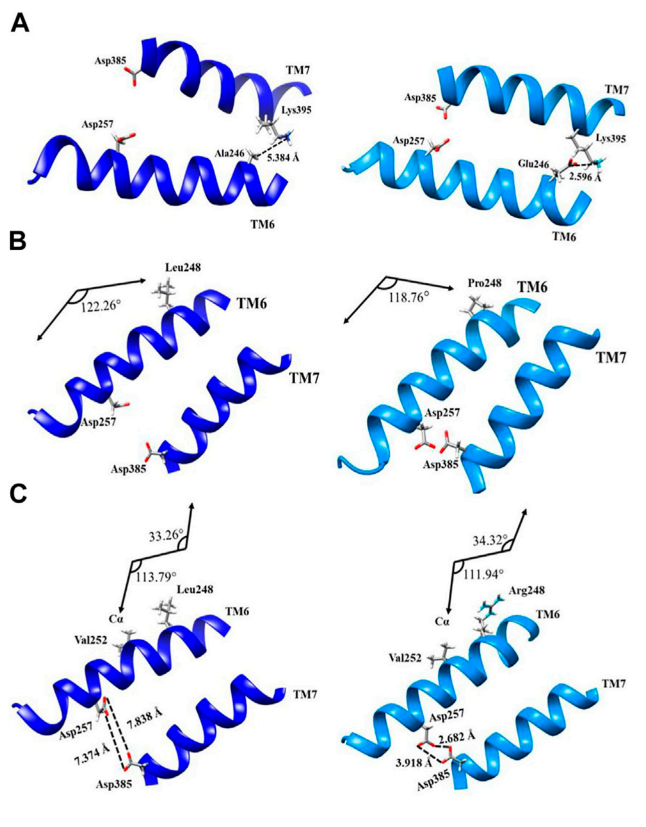


FIGURE 3 | Topological representation of structural changes of PS1 mutations Ala246Glu, Leu248Pro, and Leu248Arg: **(A)** ribbon representation of wild type (left, dark blue) and Ala246Glu mutation (right, light blue) from the interaction with the adjacent α-helix of the transmembrane 7; **(B)** ribbon representation of wild type and Leu248Pro mutation with the kink of α-helix; **(C)** ribbon representation of wild type and Leu248Arg mutation considering the changes in the torsional angles.

inside the helix. With the substitution to glutamic acid in this position, the chemical environment changes and TM6 is brought closer to TM7 on this particular location. PS1 amino acid Lys395 presents a basic behavior, and the Ala246Glu mutation facilitates an interaction via hydrogen bonding, facilitated by the decreased distance between TM6 and TM7. In theory, adequate distances to consider for possible adduct formation should be less than 2.7 Å, with the Ala246Glu mutation, the distance between Glu246 and Lys395 is 2.596 Å, while the same distance between wild-type Ala246 and Lys395 is 5.384 Å. The reduced distance between TM6 and TM7 at this point impairs the interaction between the substrate and the orthosteric site. Previous work has shown that drastic changes in polarity for this mutation can favor interactions different than that of the wild type in the diffusion of the carboxy-terminal (CTF99) fragment. This changes the epsilon-cleavage site (ε) of the enzyme and implies a decrease in the total amount of produced peptide. It might also imply an abnormal substrate processing, which follows the cleavages that occur first at Leu-Val (amino acids 49 and 50) or Thr-Leu (amino acids 48 and 49) for the APP substrate (CTF99), thus blocking production up until amino acids 37 or 38 (Funamoto et al., 2019; Funamoto et al., 2004).

PS1 mutation Leu248Pro does not induce major polarity modifications, but it does induce a topological structural change due to the substitution of a leucine to a proline, which contains a ring of five atoms with a nitrogen inside, facilitating a modification in the torsion angle of the helix. The angle between amino acid 248 and the alpha carbon in the side chain of wild-type PS1 is 122,26°, while with mutation Leu248Pro, this angle measures 118,76° (**Figure 3B**). The effect of these modifications in the torsion of the TM6 helix is similar to the effect observed with mutation Ala246Glu, because it impairs the access of the substrate to the orthosteric site.

PS1 mutation Leu248Arg, on the other hand, modifies polarity in this position. It substitutes leucine, an amino acid with a hydrocarbon side chain, to an arginine, an amino acid with a guanidine group in the extreme of its side chain. The guanidine group contains an electrophilic center, making arginine susceptible to nucleophilic attack by biological systems besides its impact in the amino acid polarity. As with mutations Ala246GLu and Leu248Pro, the distance between TM6 and TM7 decreases. More to the point, the distance between aspartic acids 257 and 385 decreases in the Leu248Arg mutation. The distances of carboxylic groups between aspartic acids 257 and 385 as measured in oxygen atoms sp^2 and sp^3 are 7.838 and Å 7.374 Å, respectively. Meanwhile, mutation Leu248Arg decreases these distances to 2.682 Å and 3.918 Å.

As a result of the decreased distance between them, α-helixes of TM6 and TM7 become susceptible to noncovalent interactions, such as hydrogen bonding or electrostatics bonds, making it difficult to access the pore of the substrate. Furthermore, PS1 mutation Leu248Arg also affects the torsion of the TM6 α-helix, producing a kink in the helix. The alteration of the hydrogen bonding pattern modifies the London dispersion interaction between Val252 and the amino acid in position 248 (in the case of this mutation, arginine) turning the helix closer to TM7. In wild-type PS1, with Leu248, the values for these angles are 33.26° and 113.79°, while with the substitution to Arg248 changes them to 34.32° and 111.94°, respectively. The resulting modification of the torsion in TM6 α-helix represents another argument for a plausible blocking of the active site (**Figure 3C**).

Electronic Surface Changes Induced by Mutations Tyr256Ser and Ala260Val in PS1

PS1 mutations Tyr256Ser and Ala260Val disturb the topological distribution of electrons in the atoms of affected amino acids. These electronic effects can modify the docking with organic ligands, ions, complex peptides, nucleic acids, dendrimers, and others.

Mutation Tyr256Ser occurs adjacent to Asp257, one half of the active site, indicating that it has a direct effect in the structural conformation and processing of the substrate. In wild-type PS1, the phenol functional group in the side chain of Tyr256 has high acidity, which is consistent with a pKa = 10.06. Besides, amino acid deprotonation is oriented from the phenoxide anion stabilized by resonance. When this amino acid is substituted to Ser256, the side chain of serine contains a hydroxyl functional

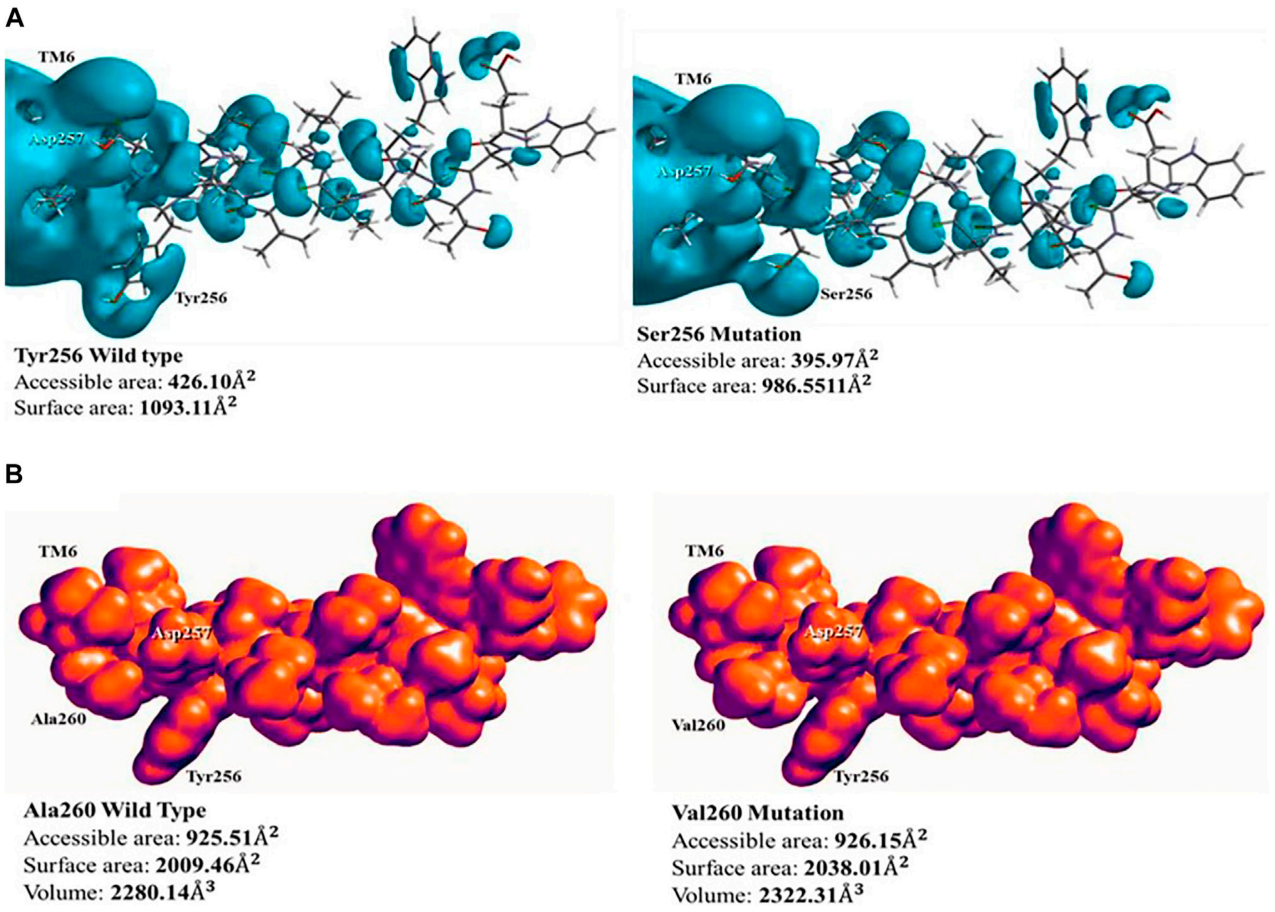


FIGURE 4 | Surface representation of electronic structure of PS1 Tyr256Ser and Ala260Val mutations: **(A)** the potential–potential surface of wild type **(left)** and mutation Tyr256Ser **(right)**. Modification of the surface corresponding to position 256 can be observed; **(B)** density surface of wild type **(left)** and mutation Ala260Val **(right)**, showing the increment in charges volume, blocking potential access to Asp257. The space around Tyr256 is used as a point of reference.

group, this functional group is less acid than phenol, with pKa = 13.60. Therefore, the hydrogen is not released to the reaction medium. Furthermore, the effect of this substitution in the active site was evaluated on the electronic surface structure with a hybrid QM/MM method. With this approach, a potential–potential surface was created for PS1 TM6 domain. This analysis found that the accessible area of interaction in wild-type PS1, containing tyrosine in position 256, was 426.10 Å^2 and the total surface area was 1,093.11 Å^2. With the substitution to serine 256, the accessible area was 395.97 Å^2 and the total surface area was 986.55 Å^2 (**Figure 4A**). Ser256 mutation decreases both areas, and this is an important point to discuss for the possible anchoring of the substrate to the active site. The phenol in the side chain could interact with the substrate by Coulombic interaction with a phenoxide anion or by the effect of the delocalization of electrons in the aromatic ring via stack–stack, stack–cations, or stack–anions interactions. Alternatively, the serine could just interact via hydrogen bonding of the hydroxyl group in the side chain.

PS1 mutation Ala260Val does not present a change of polarity, but it increases the number of carbons in its side chain. The effect of this mutation was measured using a hybrid QM/MM method for the electronic analysis of TM6, by evaluating surface density. In the wild type, with Ala260, the distribution of charges along TM6 has influence on the active site in Asp257 due to its proximity, generating a charges distribution volume of 2,280.14 Å^3 and a total surface area of 2009.46 Å^2. With the substitution to Val260, there is a modification on the distribution of charges with a distribution volume of 2,322.31Å^3 and a total surface area of 2038.01 Å^2 (**Figure 4B**). Taking into account these values, the increase in distribution volume and surface area could be a result of the increased number of carbons, besides the inclusion of a methyl group due to the substitution to valine. Therefore, there is a change in the intrinsic distribution of electronic density in front of the active site Asp257, blocking the access and possible interaction with the substrate in the structural model.

Modifications in the Electrostatic Potential Map in PS1 Mutations Leu250Val and Val261Phe

Aside of topological effects or changes in surface area or density, other possible effects of PS1 mutations could be in the electronic

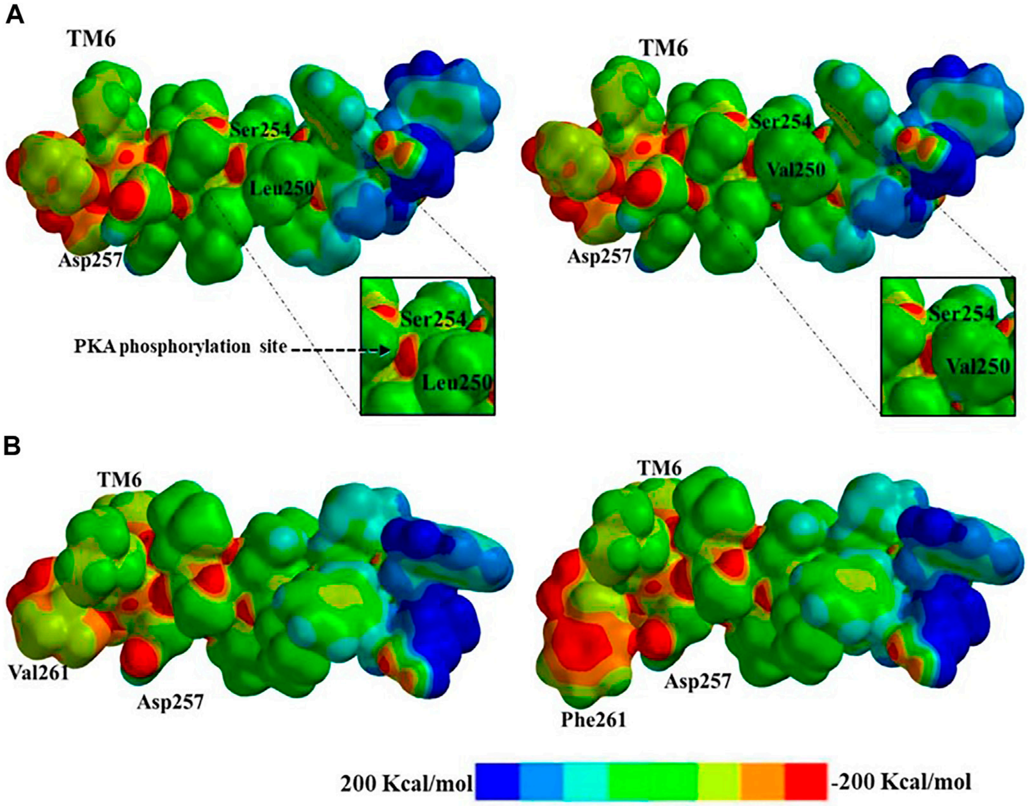


FIGURE 5 | Electrostatic potential map representation: **(A)** electronic distribution and phosphorylation blockage of wild type **(left)** and PS1 mutation Leu250Val **(right)** to putative protein kinase A (PKA) phosphorylation; **(B)** electronic distribution and resulting steric clash of the substitution from wild type **(left)** to PS1 mutation Val261Phe **(right)** due to the aromatic electronic effect in the possible interactions with the substrate. Electronic distribution within the range of −200 Kcal/mol to 200 Kcal/mol is represented by an eight-color scale.

distribution within the electrostatic potential. The PS1 Leu250Val mutation does not present polarity changes given that both amino acids (Leu and Val) do not present any polar feature (London dispersion), the main difference between amino acids is one extra carbon in the structure of valine, and this implies a possible spatial effect, given that the hydropathicity is the same. Using the ProtScale software, we quantified the hydrophobic and hydrophilic forces with the Kyte and Doolittle approach. In this position in PS1, the wild-type Leu250 has a hydropathic index of 2,522, while the substituted Val250 presents a hydropathic index of 2,567, with similar polarity behavior (**Supplementary Figure S4**). Therefore, there are no topological modifications in the α-helix, given that the similarities in hydrophobicity do not affect the London dispersion forces. Likewise, surface electronic density analysis did not detect changes between the wild type and the mutation. However, there was the option to evaluate the electrostatic potential map using Spartan 18.0 software. In effect, there is a modification in the electronic structure in the vicinity of the mutation site, reducing the access to Ser254 in the Val250-mutated PS1 (**Figure 5A**). This modification can have a direct impact in the functionality of the protein. We searched for possible affected interactors using the software XPASY and the tool NetPhosK 2.0. We found that Ser254 is a phosphorylation site for kinase PKA in the wild type situation, with a score 0.50 (**Figure 4A**, insets). In conclusion, when the Val250 substitution takes place, position Ser254 is blocked with the hydrocarbon side chain and its high electronic density site cannot be docked by PKA, resulting in loss of the phosphorylation site.

We did not find experimental reports in the literature that confirm phosphorylation for this position in PS1. However, this protein is highly phosphorylatable, and it has associated functions such as cell signaling, Ser346 being a recognition motif for caspase in apoptosis regulation (Fluhrer et al., 2004). We have also assessed the differential effect of phosphorylation of the A246E mutation in the PS1 transmembrane domain and the N141I mutation in PS2. This has been done considering that these mutations could impact phosphorylation due to its structural localization. However, no differences have been found in the effect of PS1 and PS2 phosphorylation. Likewise, many remaining available phosphorylation sites have been proposed. These remain after the γ-secretase enzyme carries out substrate proteolytic processing. This leads to a structural change in the enzyme that renders amino acids accessible in cases where they were initially inaccessible upon phosphorylation of casein kinase 1 and 2, or of PKA and PKC (Walter and Haass, 2010; Walter

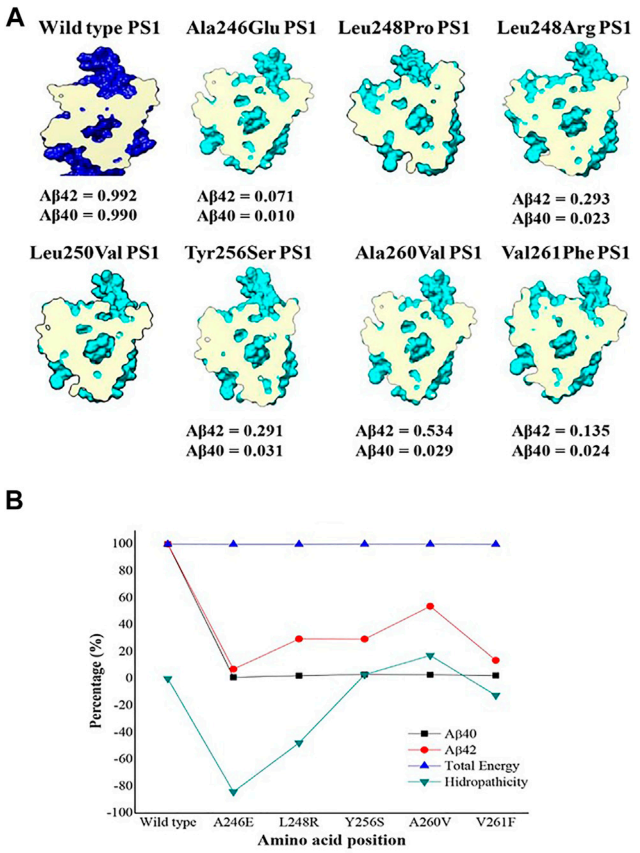


FIGURE 6 | Top view section of wild-type PS1(dark blue) and the seven mutations analyzed in TM6 (pale blue) with a cross section at the same level of the catalytic pocket (magenta). **(A)** β ratio values were obtained from reference (Sun et al., 2016); **(B)** Energy comparison plot, hydrophobicity and total amount of amyloid peptide.

et al., 1997). The position for the phosphorylation site is currently proposed to be serine 367, as it has been found to be closely related to the dynamics of microglia development and has also been found to have a protective function. This encourages autophagosome–lysosome assembly, which increases the degradation of β-CTF99 carboxi-terminal, thereby decreasing amyloid peptide synthesis (Ledo et al., 2020; Bustos et al., 2017).

The last PS1 mutation evaluated is Val261Phe, and here we can observe the change in the aliphatic side chain to an aromatic group with a direct effect on the structure of TM6. Phenylalanine has a high electronic density due to the aromatic ring, and the resulting increased electron density blocks the active site Asp257. The electrostatic map represents the potential and electronic distribution in the range of -200 Kcal/mol to 200 Kcal/mol. In the wild-type PS1, Val261 shows relatively low electronic density, while the Phe261 mutant presents a wider area with higher electronic density with a score of −100 to −150 Kcal/mol (**Figure 5B**). Furthermore, this substitution has an effect in the topology. The aromatic group produces a change in the dihedral angle due to the hybridization of the aliphatic and aromatic carbons in the structure, reducing distances for bonding via steric clash. The angles between the α carbon and the lateral side chain are modified, affecting the structure and the dihedral angle manifest differences in the topological representation with angles of −60.47° for wild type and −132.04° for mutated PS1 Val261Phe. The dihedral changes induce a kink in the α-helix of TM6 and the substrate can be hindered when entering the pore (**Supplementary Figure S5**).

In summary, the consequences of a variety of structural and electronic modifications in the active domains of PS1 as a result of point mutations suggest a possible functional effect in the catalytic activity for the processing of APP as a substrate. This effect could be considered as loss of function given that experimental data from the studied mutations show a decreased production of both Aβ 1–40 and Aβ 1–42 together with increased Aβ 1–42/1–40 ratio (Sun et al., 2016) (**Figure 6A** and **Supplementary Table S1**). Besides, the evaluation of topology, surface area, volume, and electrostatic potential is necessary to understand the structural behavior of PS1. These modifications can be summarized by a top view section in the upper plane of the protein. Due to the combination of the structural effects, the shape of the pore defined by the space and distance between TM6 and TM7 is noticeably modified in PS1 mutants. In some cases, the apparent volume and shape of the pore are radically different, hinting to a possible effect in the accessibility of the pore by the substrate. For instance, in mutation Ala260Val, the structure of the pore is severely modified, and the production of Aβ 1–40 is depleted, while the production of Aβ 1–42 is half of that on the wild type (Sun et al., 2016), hinting to the accessibility effect mentioned above (**Figure 6A**). The energetic calculation is obtained for the entire system with the subtraction formula in quantitative terms and considering the catalytic pocket in the multiscale model, thus obtaining the results reported in **Table 2**.

By analyzing the results in (**Figure 6B**), it is determined from an energetic standpoint that due to several mutations, there is not a significant change at an energetic level. However, there is a decreased size of the catalytic pocket, which leads to the enzymatic function being affected. This explains the decrease in the total amount of peptide for Aβ 1–40 and Aβ 1–42. However, the change of amino acid for mutations is not synonymous if they show an effect on hydrophobicity. Therefore, certain cuts of amyloid-β are favored, which is reflected in the cut ratio at the experimental level with the Aβ42/Aβ40 peptide proportion. As a result, an analysis focused on the lateral chain is validated with changes in topology and electronic structure, as was previously shown. The graph reveals that the increase in hydrophilicity results in the cleavage route that leads to producing Aβ42 peptide instead of Aβ40. This is because it is the most commonly found peptide in the amyloid plaques and the graphical trend of its hydrophobicity is very similar to the data that were experimentally reported on the peptide cuts of 42 amino acids. Likewise, the γ-secretase enzyme with mutations and with these changes in polarity profiles favors certain processing routes that can produce the most frequent amyloid peptide fragments. In addition, due to the mutations, enzyme activity could be modified and could encourage cleaving up to the Aβ38 and Aβ37 amyloid peptide

TABLE 2 | Energetic values calculated in the system with PS1 in the γ-secretase enzyme.

Protein PS1	Energy MM (Full length)	Average MM (Full length)	Energy MM (QM)	Average MM (QM)	Energy QM	Average QM	Total Energy: EQM + EMM-EMM(QM)
Wild type	2962167.1071 2962167.0920 2962167.1261	2962167,108 (+/–) 0,017	−40.6771 −40.6775 −40.6768	−40,67710 (+/–) 0,00040	−5104.5531 −5104.5661 −5104.5621	−5104,5604 (+/–) 0,0067	2957103,225(+/–) 0,023
Ala246Glu	2961014.1399 2961015.0300 2961013.2500	2961014,14 (+/–) 0,89	−161.5506 −161.5481 −161.5523	−161,5503 (+/–) 0,0021	−5534.4918 −5543.0927 −5543.0906	−5540,2 (+/–) 4,9	2955635,5 (+/–) 5,9
Leu248Pro	2964064.3892 2964064.2220 2964064.5480	2964064,39 (+/–) 0,16	772.9771 772.9784 772.9768	772,97740 (+/–) 0,00090	−4998.3374 −4998.3491 −4998.3252	−4998,337 (+/–) 0,012	2958293,07 (+/–) 0,17
Leu248Arg	2959517.4753 2959517.4220 2959517.5312	2959517,48 (+/–) 0,05	−13.3939 −13.3952 −13.3921	−13,3937 (+/–) 0,0016	−4677.5868 −4677.5844 −4660.9302	−4672,0 (+/–) 9–6	2954858,8(+/–) 9,7
Leu250Val	2962020.7393 2962020.7310 2962020.7492	2962020,7398 (+/-) 0,0090	108.0198 108.0192 108.0202	108,01970 (+/–) 0,00050	−4676.2334 −4676.229 −4676.2331	−4676,2318 (+/–) 0,0025	2957236,488 (+/–) 0,011
Tyr256Ser	2961819.2446 2961819.0530 2961819.4348	2961819,24 (+/–) 0,19	30.3606 30.3601 30.3609	30,36050 (+/–) 0,00040	−5235.8072 −5234.7973 −5235.809	−5235,47 (+/–) 0,58	2956553,41 (+/–) 0,77
Ala260Val	2964306.3948 2964306.7205 2964306.0722	2964306,40 (+/–) 0,32	100.4126 100.4121 100.413	100,41260 (+/–) 0,00050	−5168.3129 −5173.4372 −5151.0903	−5164,3 (+/–) 11,7	2959041,7 (+/–) 12,0
Val261Phe	2960691.6238 2960691.5340 2960691.7131	2960691,624 (+/–) 0,090	160.2564 160.2571 160.2583	160,25730 (+/–) 0,0010	−4955.9667 −4955.9701 −4956.8911	−4956,28 (+/–) 0,53	2955575,09 (+/–) 0,62

fragments. These fragments are frequently found in senile plaques but are not as pathogenic as Aβ 40 or Aβ 42, which tend to undergo oligomerization more readily (Murakami et al., 2003; Chen et al., 2017; Morel et al., 2018; Song et al., 2018).

Previously, a molecular dynamics approach by Chávez-García et al. was used to analyze the effects of PS1 mutations in the catalytic domain of the protein (Chávez-García et al., 2019), (Aguayo-Ortiz et al., 2017). Briefly, their approach involved amino acid network and protonation analysis for thirteen PS1 mutations via all-atom molecular dynamics simulation. Among their findings, an increased number of correlations for different mutations were identified. Interestingly, two of the mutations analyzed by this approach are localized in TM6 and were also evaluated in the present work (Ala260Val and Val261Phe). They found that these two mutations presented increased number of correlations and they also suggest that the amino acid substitution might affect the entry gate (Chávez-García et al., 2019). We consider that our approach brings a different view of the problem, and that both approaches (molecular dynamics and hybrid QM/MM) are valid and complementary when analyzing the effect of mutations in structural protein chemistry.

CONCLUSION

Protein functional studies through structural modeling in neurodegenerative diseases is a useful approach for understanding the effect of some genetic variants that translate in specific protein modifications. In the case of PS1, it opens a window to understand how its structure affects its function and those of the γ-secretase complex and its four subunits. Given that PS1 has two flexible domains that have not been resolved satisfactorily via experimental approaches, we have developed a model using structure prediction software. Flexible regions present experimental challenges for protein structure studies, such as their low electron-dense zone with low signal emission that results in low structural resolution for Cryo-EM studies or affecting crystallization for X-ray analysis (Bai et al., 2015; Rossi et al., 2012; Heo et al., 2017). Therefore, an *in silico* approach seems to be the best alternative for resolving the full structure of PS1, until further experimental models are obtained. Homological modeling, using assembly by threading and reconstruction *ab initio*, was used to create a hypothetical construct for the missing fragments with their respective energetic and stereochemical characterization. Our approach included algorithms of molecular dynamics methods that consider force fields and the primary sequences of amino acids in the construction of proteins. The completed model for PS1 was then useful to assess possible effects of the flexible domains in the pore. In our reconstructed model, we observed that the predicted assembly for both flexible fragments did not affect the topology and the connectivity matrix of the most current template for PS1 (Zhou et al., 2019) and did not affect the structure of the pore constituted by TM6 and TM7 (**Supplementary Figure S3**). It is possible that PS1 pathogenic mutations localized in the flexible pores affect pore accessibility by other means different from direct modifications on the active site.

Additionally, we analyzed seven PS1 mutations localized in TM6 and in the proximity of Asp257, in order to assess the direct effect of these mutations in the active site. The structural changes were assessed using a topological approach for distance variations, torsion angles, and dihedral angles and electronic changes with the distribution of charges and surfaces in the system. Macromolecular systems present a problem for structural biology, such as their polyatomic constitution and the high number of possible multiple interactions. In these cases, hybrid QM/MM methods are useful for the study of polyatomic systems given that they assess the interaction of the electronic structure with a stochastic measure of energy and optimization of structural conformers (Murphy et al., 2000; Silva et al., 2015; Zou et al., 2017).

For PS1 mutations Ala246Glu, Leu248Pro, and Leu248Arg, topological changes, such as the modification of distances between TM6 and TM7 as a result of changes in the kink of the TM6 helix, seemed to be the most relevant for their possible effect in the active site. On the other hand, PS1 mutations Leu250Val, Tyr256Ser, Ala260Val, and Val261Phe produce more noticeable modifications in the electronic structure of TM6, affecting the electronic surface, charge distribution volume, and electrostatic potential, finally blocking the access of the substrate. Interestingly, the modification of the electronic distribution for Ser254 elicited by PS1 mutation Leu250Val has a direct effect in the corresponding phosphorylation site with possible functional repercussions. Independently of the specific change, all the studied mutations affected the shape of the pore, possibly affecting the accessibility of the substrates to the active site or affecting the kinetics of its processing. There are experimental data for Aβ processing of five out of the seven PS1 mutations we analyzed. All of them show decreased production of Aβ in comparison with the wild-type enzyme, with some of them increasing the relative production of Aβ 1–42 (Sun et al., 2016), perhaps as an effect of major changes in the pore (**Figure 6A**).

We consider that the use of QM/MM hybrid methods might be an ideal approach for the study of single-point mutation effects in macromolecular systems as complex as that of γ-secretase and PS1. With the development and access to more powerful computational systems, this kind of studies will provide a wide array of possibilities for functional analysis and the development of better targeted drug design.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AS-O, PA, and AV contributed to data modelling and simulation, analysis, and interpretation of the data. ASO, PAM, AV, and DS-F wrote the manuscript. PA, AV, and GB contributed to the guidance of the study. AV, GB, and DS-F critically revised the manuscript. All authors read and approved the manuscript.

FUNDING

AS-O was supported by Colciencias fellowship from Colombia (FP44842-124-2017). This project was sustained by the National Council of Research, Francisco José de Caldas from Colombia, Colciencias by funding number (23411). AV and DS-F was supported by US National Institute of Neurological Disorders and Stroke and National Institute on aging co-funded grant RF1NS110048. AS-O, AV, and PA was supported by EIA University, from Envigado, Colombia, by project CI120180016.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.649990/full#supplementary-material

REFERENCES

Aguayo-Ortiz, R., Chávez-García, C., Straub, J. E., and Dominguez, L. (2017). Characterizing the Structural Ensemble of γ-secretase Using a Multiscale Molecular Dynamics Approach. *Chem. Sci.* 8 (8), 5576–5584. doi:10.1039/c7sc00980a

Aguayo-Ortiz, R., and Dominguez, L. (2018). Simulating the γ-secretase Enzyme: Recent Advances and Future Directions. *Biochimie* 147, 130–135. doi:10.1016/j.biochi.2018.01.007

Alexeev, Y., Mazanetz, M. P., Ichihara, O., and Fedorov, D. G. (2013). GAMESS as a Free Quantum-Mechanical Platform for Drug Research. *Curr. Top. Med. Chem.* 12, 2013–2033. doi:10.2174/1568026611212180008

Alzforum (2020). *ALZFORUM Networking for a Cure*. Osaka, Japan: Biomedical Research Forum. Available at: https://www.alzforum.org/mutations/psen-1 (Accessed May 10, 2021).

Arbor, S. C., Lafontaine, M., and Cumbay, M. (2016). Amyloid-beta Alzheimer Targets - Protein Processing, Lipid Rafts, and Amyloid-Beta Pores. *Yale J. Biol. Med.* 89 (1), 5–21.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., et al. (2012). ExPASy: SIB Bioinformatics Resource Portal. *Nucleic Acids Res.* 40 (W1), 597–603. doi:10.1093/nar/gks400

Bai, X. C., Rajendra, E., Yang, G., Shi, Y., and Scheres, S. H. W. (2015a). Sampling the Conformational Space of the Catalytic Subunit of Human G-Secretase. *Elife* 4, 1–19. doi:10.7554/eLife.11182

Bai, X. C., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., et al. (2015b). An Atomic Structure of Human γ-secretase. *Nature* 525 (7568), 212–217. doi:10.1038/nature14892

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., and Jones, E. (2011). Alzheimer's Disease. *The Lancet* 377 (9770), 1019–1031. doi:10.1016/S0140-6736(10)61349-9

Bereczki, E., Branca, R. M., Francis, P. T., Pereira, J. B., Baek, J.-H., Hortobágyi, T., et al. (2018). Synaptic Markers of Cognitive Decline in Neurodegenerative Diseases: A Proteomic Approach. *Brain* 141 (2), 582–595. doi:10.1093/brain/awx352

Bhattarai, A., Devkota, S., Bhattarai, S., Wolfe, M. S., and Miao, Y. (2020). "Mechanisms of γ-Secretase Activation and Substrate Processing. *ACS Cent. Sci.* doi:10.1021/acscentsci.0c00296

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and Structure-Based Prediction of Eukaryotic Protein Phosphorylation Sites. *J. Mol. Biol.* 294 (5), 1351–1362. doi:10.1006/jmbi.1999.3310

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of Post-translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics* 4 (6), 1633–1649. doi:10.1002/pmic.200300771

Bolduc, D. M., Montagna, D. R., Gu, Y., Selkoe, D. J., and Wolfe, M. S. (2015). "Nicastrin Functions to Sterically Hinder γ -secretase – Substrate Interactions Driven by Substrate Transmembrane Domain. *Proc. Natl. Acad. Sci.*, 1–10. doi:10.1073/pnas.1512952113

Bustos, V., Pulina, M. V., Bispo, A., Lam, A., Flajolet, M., Gorelick, F. S., et al. (2017). Phosphorylated Presenilin 1 Decreases β-amyloid by Facilitating Autophagosome-Lysosome Fusion. *Proc. Natl. Acad. Sci. U. S. A.* 114 (27), 7148–7153. doi:10.1073/pnas.1705240114

Cacquevel, M., Aeschbach, L., Houacine, J., and Fraering, P. C. (2012). "Alzheimer's Disease-Linked Mutations in Presenilin-1 Result in a Drastic Loss of Activity in Purified γ-secretase Complexes. *PLoS One* 7 (4), 1–13. doi:10.1371/journal.pone.0035133

Cano, L., Soto-Ospina, A., Araque, P., and Caro-gomez, M. A. (2021). Diffusion Mechanism Modeling of Metformin in Human Organic Cationic Amino Acid Transporter One and Functional Impact of S189L , R206C , and G401S Mutation. *Front. Pharmacol.* 11, 1–14. doi:10.3389/fphar.2020.587590

Cao, L., and Ryde, U. (2018). On the Difference between Additive and Subtractive QM/MM Calculations. *Front. Chem.* 6 (APR), 1–15. doi:10.3389/fchem.2018.00089

Carvalho, A. T. P., Barrozo, A., Doron, D., Vardi Kilshtain, A., Major, D. T., and Kamerlin, S. C. L. (2014). Challenges in Computational Studies of Enzyme Structure, Function and Dynamics. *J. Mol. Graph. Model.* 54, 62–79. doi:10.1016/j.jmgm.2014.09.003

Chávez-García, C., Aguayo-Ortiz, R., and Dominguez, L. (2019). Quantifying Correlations between Mutational Sites in the Catalytic Subunit of γ-secretase. *J. Mol. Graph. Model.* 88, 221–227. doi:10.1016/j.jmgm.2019.02.002

Chávez-Gutiérrez, L., Bammens, L., Benilova, I., Vandersteen, A., Benurwar, M., Borgers, M., et al. (2012). The Mechanism of γ-Secretase Dysfunction in Familial Alzheimer Disease. *EMBO J.* 31 (10), 2261–2274. doi:10.1038/emboj.2012.79

Chen, G. F., Xu, T. H., Yan, Y., Zhou, Y. R., Jiang, Y., Melcher, K., et al. (2017). Amyloid Beta: Structure, Biology and Structure-Based Therapeutic Development. *Acta Pharmacol. Sin.* 38 (9), 1205–1235. doi:10.1038/aps.2017.28

Christensen, A. S., Kubař, T., Cui, Q., and Elstner, M. (2016). Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* 116 (9), 5301–5337. doi:10.1021/acs.chemrev.5b00584

Crystallography and Bioinformatics Group (2017). Rampage: Ramachandran Plot. [Online]. Available at: http://mordred.bioc.cam.ac.uk/~rapper/rampage.php (Accessed Aug 12, 2017).

Eggert, S., Paliga, K., Soba, P., Evin, G., Masters, C. L., Weidemann, A., et al. (2004). The Proteolytic Processing of the Amyloid Precursor Protein Gene Family Members APLP-1 and APLP-2 Involves α-, β-, γ-, and ϵ-Like Cleavages. *J. Biol. Chem.* 279 (18), 18146–18156. doi:10.1074/jbc.M311601200

Fluhrer, R., Friedlein, A., Haass, C., and Walter, J. (2004). Phosphorylation of Presenilin 1 at the Caspase Recognition Site Regulates its Proteolytic Processing and the Progression of Apoptosis. *J. Biol. Chem.* 279 (3), 1585–1593. doi:10.1074/jbc.M306653200

Foresman, J., and Frisch, E. (1996). *Exploring Chemistry with Electronic Structure Methods*. Second edi. Pittsburgh: Gaussian Inc.

Fukumori, A., Fluhrer, R., Steiner, H., and Haass, C. (2010). Three-amino Acid Spacing of Presenilin Endoproteolysis Suggests a General Stepwise Cleavage of γ-secretase-mediated Intramembrane Proteolysis. *J. Neurosci.* 30 (23), 7853–7862. doi:10.1523/JNEUROSCI.1443-10.2010

Funamoto, S., Morishima-Kawashima, M., Tanimura, Y., Hirotani, N., Saido, T. C., and Ihara, Y. (2004). Truncated Carboxyl-Terminal Fragments of β-amyloid Precursor Protein Are Processed to Amyloid β-proteins 40 and 42. *Biochemistry* 43 (42), 13532–13540. doi:10.1021/bi049399k

Funamoto, S., Tagami, S., Okochi, M., and Morishima-Kawashima, M. (2020). Successive Cleavage of β-amyloid Precursor Protein by γ-secretase. *Semin. Cel Dev. Biol.* 105, 64–74. doi:10.1016/j.semcdb.2020.04.002

Garcia, M. L., de Oliveira, A. A., Bueno, R. V., Nogueira, V. H. R., de Souza, G. E., and Guido, R. V. C. (2020). QSAR Studies on Benzothiophene Derivatives as Plasmodium Falciparum N-Myristoyltransferase Inhibitors: Molecular Insights into Affinity and Selectivity. *Drug Dev. Res.*, 1–21. doi:10.1002/ddr.21646

Gasteiger, B. A. E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., and Appel, R. D. (2005). *The Proteomics Protocols Handbook-Protein Identification and Analysis Tools on the ExPASy Server*.

Grillo, I. B., Urquiza-Carvalho, G. A., Bachega, J. F. R., and Rocha, G. B. (2020). Elucidating Enzymatic Catalysis Using Fast Quantum Chemical Descriptors. *J. Chem. Inf. Model.* 60, 578-591. doi:10.1021/acs.jcim.9b00860

Guerreiro, R., and Hardy, J. (2014). Genetics of Alzheimer's Disease. *Neurotherapeutics* 11 (4), 732–737. doi:10.1007/s13311-014-0295-9

Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An Environment for Comparative Protein Modeling. *Electrophoresis* 18 (15), 2714–2723. doi:10.1002/elps.1150181505

Halgren, T. A. (1996). Merck Molecular Force FieldI-Basis Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* 17, 490–519. doi:10.1002/(sici)1096-987x(199604)17:5/6<490::aid-jcc1>3.0.co;2-p

Halgren, T. A. (2000). MMFF VII-Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular Interaction Energies and Geometries. *J. Comput. Chem.* 20 (7), 730–748.

Heo, S., Lee, J., Lee, J., Joo, K., and Shin, H. C. (2017). Protein Loop Structure Prediction Using Conformational Space Annealing. *J. Chem. Inf. Model.* 57 (5), 1068–1078. doi:10.1021/acs.jcim.6b00742

Hitzenberger, M., and Zacharias, M. (2019). γ-Secretase Studied by Atomistic Molecular Dynamics Simulations: Global Dynamics, Enzyme Activation, Water Distribution and Lipid Binding. *Front. Chem.* 6 (January). doi:10.3389/fchem.2018.00640

Hofer, T. S., and de Visser, S. P. (2018). "Editorial: Quantum Mechanical/Molecular Mechanical Approaches for the Investigation of Chemical Systems – Recent Developments and Advanced Applications. *Front. Chem.* 6 (September), 1–5. doi:10.3389/fchem.2018.00357

Hospital, A., Goñi, J. R., Orozco, M., and Gelpi, J. (2015). Molecular Dynamics Simulations: Advances and Applications. *Adv. Appl. Bioinforma. Chem.* 8, 37–47. doi:10.2147/AABC.S70333

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD- Visual Molecular Dynamics. *J. Mol. Graph.* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5

Dewar, M. J., Zoebisch, E. G., Healy, E. F., and Stewart, J. P. (1993). AM1: A Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* 49 (June), 3903–3909.

Jucker, M., and Walker, L. C. (2012). Pathogenic Protein Seeding in Alzheimer Disease and Other Neurodegenerative Disorders. *Ann. Neurol.* 70 (4), 532–540. doi:10.1002/ana.22615.Pathogenic

Jucker, M., and Walker, L. C. (2013). Self-propagation of Pathogenic Protein Aggregates in Neurodegenerative Diseases. *Nature* 501 (7465), 45–51. doi:10.1038/nature12481

Kelley, L. A., and Sternberg, M. J. E. (2009). Protein Structure Prediction on the Web: a Case Study Using the Phyre Server. *Nat. Protoc.* 4 (3), 363–371. doi:10.1038/nprot.2009.2

Kelly, L. A., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015). The Phyre2 Web Portal for Protein Modelling, Prediction, and Analysis. *Nat. Protoc.* 10 (6), 845–858. doi:10.1038/nprot.2015-053

Knappenberger, K. S., Tian, G., Ye, X., Sobotka-Briner, C., Ghanekar, S. V., Greenberg, B. D., et al. (2004). Mechanism of γ-secretase Cleavage Activation: Is γ-secretase Regulated through Autoinhibition Involving the Presenilin-1 Exon 9 Loop?. *Biochemistry* 43 (20), 6208–6218. doi:10.1021/bi036072v

Krogh, A., Larsson, È., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting Transmembrane Protein Topology with a Hidden Markov Model : Application to Complete Genomes. *J Mol Biol* 305, 567-580. doi:10.1006/jmbi.2000.4315

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* 26 (2), 283–291. doi:10.1107/s0021889892009944

Ledo, J. H., Zhang, R., Mesin, L., Mourão-Sá, D., Azevedo, E. P., Troyanskaya, O. G., et al. (2020). Lack of a Site-specific Phosphorylation of Presenilin 1 Disrupts Microglial Gene Networks and Progenitors during Development. *PLoS One* 15 (8), 1–12. doi:10.1371/journal.pone.0237773

Levitt, M. (2014). Birth and Future of Multiscale Modeling for Macromolecular Systems (Nobel Lecture). *Angew. Chem. - Int. Ed.* 53 (38), 10006–100018. doi:10.1002/anie.201403691

Lichtenthaler, S. (2012). Alpha-secretase Cleavage of the Amyloid Precursor Protein: Proteolysis Regulated by Signaling Pathways and Protein Trafficking. *Curr. Alzheimer Res.* 9 (2), 165–177. doi:10.2174/156720512799361655

Lu, P., Bai, X. C., Ma, D., Xie, T., Yan, C., Sun, L., et al. (2014). Three-dimensional Structure of Human γ-secretase. *Nature* 512 (7513), 166–170. doi:10.1038/nature13567

Mackerell, A. D. (2004). Empirical Force Fields for Biological Macromolecules Overview and Issues. *J. Comput. Chem.* doi:10.1002/jcc.20082

Marín, P. A., and Soto-ospina, A. (2020). Redox Mechanism of Trypanosoma Cruzi Resistance to Nitro Prodrugs Benznidazole and Nifurtimox. *Int. J. Bioinforma. Comput. Biol.* 5 (1), 1–7.

Marques, H. M., Munro, O. Q., Grimmer, N. E., Levendis, D. C., Marsicano, F., Pattrick, G., et al. (1995). A Force Field for Molecular Mechanics Studies of Iron Porphyrinst. *J. Chem. Soc. Faraday Trans.* 1.

Maseras, F. (1999). Hybrid Quantum Mechanics/Molecular Mechanics Methods in Transition Metal Chemistry. *Top. Organomet. Chem.* 4, 165–191. doi:10.1007/3-540-69707-1_5

Möller, S., Croning, M., and Apweiler, R. (2001). Membrane Spanning Regions. *Bioinformatics* 17 (7), 646–653. doi:10.1093/bioinformatics/17.7.646

Morel, B., Carrasco, M. P., Jurado, S., Marco, C., and Conejero-Lara, F. (2018). Dynamic Micellar Oligomers of Amyloid Beta Peptides Play a Crucial Role in Their Aggregation Mechanisms. *Phys. Chem. Chem. Phys.* 20 (31), 20597–20614. doi:10.1039/c8cp02685h

Wolfe, M. S., Xia, W., and Ostaszewski, B. L. (1999). "Two Transmembrane Aspartates in Presenilin-1 Required for Presenilin Endoproteolysis and G -secretase Activity," 117, (1907), 513–517. doi:10.1038/19077

Murakami, K., Irie, K., Morimoto, A., Ohigashi, H., Shindo, M., Nagao, M., et al. (2003). "Neurotoxicity and Physicochemical Properties of Aβ Mutant Peptides from Cerebral Amyloid Angiopathy: Implication for the Pathogenesis of Cerebral Amyloid Angiopathy and Alzheimer's Disease. *J. Biol. Chem.* 278 (46), 46179–46187. doi:10.1074/jbc.M301874200

Murphy, R. B., Philipp, D. M., and Friesner, R. A. (2000). A Mixed Quantum Mechanics/molecular Mechanics (QM/MM) Method for Large-Scale Modeling of Chemistry in Protein Environments. *J. Comput. Chem.* 21 (16), 1442–1457. doi:10.1002/1096-987x(200012)21:16<1442::aid-jcc3>3.0.co;2-o

Myers, R. H. (2004). Huntington's Disease Genetics. *Neurotherapeutics* 1 (2), 255–262. doi:10.1602/neurorx.1.2.255

Náray-Szabó, G., Oláh, J., and Krámos, B. (2013). Quantum Mechanical Modeling: A Tool for the Understanding of Enzyme Reactions. *Biomolecules* 3 (3), 662–702. doi:10.3390/biom3030662

Nosé, S. (1984). A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* 52 (2), 255–268. doi:10.1080/00268978400101201

Omer, A., Suryanarayanan, V., Selvaraj, C., Singh, S. K., and Singh, P. (2015). Explicit Drug Re-positioning: Predicting Novel Drug-Target Interactions of the Shelved Molecules with QM/MM Based Approaches. *Adv. Protein Chem. Struct. Biol.* 100, 89–112. doi:10.1016/bs.apcsb.2015.07.001

Orlando, A., and Jorgensen, W. L. (2010). Advances in Quantum and Molecular Mechanical (QM/MM) Simulations for Organic and Enzymatic Reactions. *Acc. Chem. Res.* 43 (1), 142–151. doi:10.1021/ar900171c

Paulsen, J. S. (2011). Cognitive Impairment in Huntington Disease: Diagnosis and Treatment. *Curr. Neurol. Neurosci. Rep.* 11 (5), 474–483. doi:10.1007/s11910-011-0215-x

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Gimpelev, M., et al. (2003). Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins Struct. Funct. Genet.* 53 (Suppl. 6), 430–435. doi:10.1002/prot.10550

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25 (13), 1605–1612. doi:10.1002/jcc.20084

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* 26 (16), 1781–1802. doi:10.1002/jcc.20289

Prince, M., Wimo, A., Guerche, M., Ali, G.-C., Wu, Y.-T., and Prina, M. (2015). The Global Impact of Dementia. *Alzheimer's Dis. Int.* 13 (4), 1–87. doi:10.1111/j.0963-7214.2004.00293.x

Rafique, R., Khan, K. M., Chigurupati, S., Wadood, A., Rehman, A. U., Karunanidhi, A., et al. (20192020). Synthesis of New Indazole Based Dual Inhibitors of α-glucosidase and α-amylase Enzymes, Their In Vitro, In Silico and Kinetics Studies. *Bioorg. Chem.* 94, 103195. doi:10.1016/j.bioorg.2019.103195

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* 7 (1), 95–99. doi:10.1016/S0022-2836(63)80023-6

Rossi, K. A., Weigelt, C. A., Nayeem, A., and Krystek, S. R. (2012). Loopholes and Missing Links in Protein Modeling. *Protein Sci.* 16 (9), 1999. doi:10.1110/ps.072887807

Roston, D., Lu, X., Fang, D., Demapan, D., and Cui, Q. (2018). Analysis of Phosphoryl-Transfer Enzymes with QM/MM Free Energy Simulations. *Methods Enzymol.* 607, 53–90. doi:10.1016/bs.mie.2018.05.005

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Meur, N. L., Laquerrière, A., Vital, A., et al. (2006). APP Locus Duplication Causes Autosomal Dominant Early-Onset Alzheimer Disease with Cerebral Amyloid Angiopathy. *Nat. Genet.* 38 (1), 24–26. doi:10.1038/ng1718

Roy, A., Kucukural, A., and Zhang, Y. (2010). A Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* 5 (4), 725–738. doi:10.1038/nprot.2010.5

Shao, W., Peng, D., and Wang, X. (2017). Genetics of Alzheimer's Disease: From Pathogenesis to Clinical Usage. *J. Clin. Neurosci.* 45, 1–8. doi:10.1016/j.jocn.2017.06.074

Sheikh, S., Safia, Haque, E., and Mir, S. S. (2013). Neurodegenerative Diseases: Multifactorial Conformational Diseases and Their Therapeutic Interventions. *J. Neurodegenerative Dis.* 2013, 1–8. doi:10.1155/2013/563481

Siegbahn, P. E. M., and Blomberg, M. R. A. (2018). A Systematic DFT Approach for Studying Mechanisms of Redox Active Enzymes. *Front. Chem.* 6 (DEC), 1–9. doi:10.3389/fchem.2018.00644

Silva, J. R., Roitberg, A. E., and Alves, C. N. (2015). A QM/MM Free Energy Study of the Oxidation Mechanism of Dihydroorotate Dehydrogenase (Class 1A) from Lactococcus Lactis. *J. Phys. Chem. B* 119 (4), 1468–1473. doi:10.1021/jp512860r

Song, Y., Li, P., Liu, L., Bortolini, C., and Dong, M. (2018). Nanostructural Differentiation and Toxicity of Amyloid-B25-35 Aggregates Ensue from Distinct Secondary Conformation. *Sci. Rep.* 8 (1), 2–10. doi:10.1038/s41598-017-19106-y

Soni, A., Bhat, R., and Jayaram, B. (2020). "Improving the Binding Affinity Estimations of Protein – Ligand Complexes Using Machine - Learning Facilitated Force Field Method. *J. Comput. Aided. Mol. Des.* 34, 817–830. doi:10.1007/s10822-020-00305-1

Soni, N., and Madhusudhan, M. S. (2017). Computational Modeling of Protein Assemblies. *Curr. Opin. Struct. Biol.* 44, 179–189. doi:10.1016/j.sbi.2017.04.006

Sonnhammer, E. L. L., and Krogh, A. (1998). A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 175–182.

Sun, L., Zhou, R., Yang, G., and Shi, Y. (2016). Analysis of 138 Pathogenic Mutations in Presenilin-1 on the In Vitro Production of Aβ42 and Aβ40 Peptides by γ-secretase. *Proc. Natl. Acad. Sci.* 114 (4), E476–E485. doi:10.1073/pnas.1618657114

Thal, D. R., Griffin, W. S. T., and Braak, H. (2008). "Parenchymal and Vascular Aβ-Deposition and its Effects on the Degeneration of Neurons and Cognition in Alzheimer's Disease. *J. Cel. Mol. Med.* 12 (5B), 1848–1862. doi:10.1111/j.1582-4934.2008.00411.x

Van Der Kamp, M. W., and Mulholland, A. J. (2013). Combined Quantum Mechanics/molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* 52 (16), 2708–2728. doi:10.1021/bi400215w

Vassar, R. (2004). BACE1: The β-Secretase Enzyme in Alzheimer's Disease. *Jmn* 23 (1–2), 105–114. doi:10.1385/JMN:23:1-2:105

Venugopal, C., Demos, C. M., Rao, K. S., Pappolla, M. A., and Sambamurti, K. (2008). Co-workers, "Beta-Secretase: Structure, Function and Evolution. *CNS Neurol. Disord. Drug Targets* 7, 1–33. doi:10.2174/187152708784936626

Walter, J., Grünberg, J., Capell, A., Pesold, B., Schindzielorz, A., Citron, M., et al. (1997). Proteolytic Processing of the Alzheimer Disease-Associated Presenilin-1 Generates an In Vivo Substrate for Protein Kinase C. *Proc. Natl. Acad. Sci. U. S. A.* 94 (10), 5349–5354. doi:10.1073/pnas.94.10.5349

Walter, J., and Haass, C. (2010). The Phosphorylation of Presenilin Proteins. *Mol. Biol. Alzheimer'S Dis.* 32 (1), 317–331.

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.*, 5.6.1–5.6.3. doi:10.1002/cpbi.3

Xu, D., and Zhang, Y. (2012). Ab Initio protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-Based Force

Field. *Proteins Struct. Funct. Bioinforma.* 80 (7), 1715–1735. doi:10.1002/prot.24065

Xu, D., and Zhang, Y. (2011). Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-step Atomic-Level Energy Minimization. *Biophys. J.* 101 (10), 2525–2534. doi:10.1016/j.bpj.2011.10.024

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* 12 (1), 7–8. doi:10.1038/nmeth.3213

Yu, X., and Zheng, J. (2012). Cholesterol Promotes the Interaction of Alzheimer β-Amyloid Monomer with Lipid Bilayer. *J. Mol. Biol.* 421 (4–5), 561–571. doi:10.1016/j.jmb.2011.11.006

Zhang, J., Liang, Y., and Zhang, Y. (2011). Atomic-level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure* 19 (12), 1784–1795. doi:10.1016/j.str.2011.09.022

Zhang, Y. (2008). I-TASSER Server for Protein 3D Structure Prediction. *BMC Bioinformatics* 9, 40. doi:10.1186/1471-2105-9-40

Zhou, X., Xu, Z., Li, A., Zhang, Z., and Xu, S. (2019). Double-sides Sticking Mechanism of Vinblastine Interacting with α , β -tubulin to Get Activity against Cancer Cells. *J. Biomol. Struct. Dyn.* 37 (15), 4080–4091. doi:10.1080/07391102.2018.1539412

Zhou, R., Yang, G., Guo, X., Zhou, Q., Lei, J., and Shi, Y. (2019).Recognition of the Amyloid Precursor Protein by Human Gamma Secretase. *Science* 0930. 80. doi:10.1126/science.aaw0930

Zou, Y., Wang, F., Wang, Y., Guo, W., Zhang, Y., Xu, Q., et al. (2017). Systematic Study of Imidazoles Inhibiting Ido1 via the Integration of Molecular Mechanics and Quantum Mechanics Calculations. *Eur. J. Med. Chem.* 131, 152–170. doi:10.1016/j.ejmech.2017.03.021

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fmolb.2021.619403

Structural and Genomic Insights Into Pyrazinamide Resistance in *Mycobacterium tuberculosis* Underlie Differences Between Ancient and Modern Lineages

Tanushree Tunstall[1], Jody Phelan[1], Charlotte Eccleston[1], Taane G. Clark[1,2] and Nicholas Furnham[1]*

[1] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom, [2] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Resistance to drugs used to treat tuberculosis disease (TB) continues to remain a public health burden, with missense point mutations in the underlying *Mycobacterium tuberculosis* bacteria described for nearly all anti-TB drugs. The post-genomics era along with advances in computational and structural biology provide opportunities to understand the interrelationships between the genetic basis and the structural consequences of *M. tuberculosis* mutations linked to drug resistance. Pyrazinamide (PZA) is a crucial first line antibiotic currently used in TB treatment regimens. The mutational promiscuity exhibited by the *pncA gene* (target for PZA) necessitates computational approaches to investigate the genetic and structural basis for PZA resistance development. We analysed 424 missense point mutations linked to PZA resistance derived from ∼35K *M. tuberculosis* clinical isolates sourced globally, which comprised the four main *M. tuberculosis* lineages (Lineage 1–4). Mutations were annotated to reflect their association with PZA resistance. Genomic measures (minor allele frequency and odds ratio), structural features (surface area, residue depth and hydrophobicity) and biophysical effects (change in stability and ligand affinity) of point mutations on pncA protein stability and ligand affinity were assessed. Missense point mutations within *pncA* were distributed throughout the gene, with the majority (>80%) of mutations with a destabilising effect on protomer stability and on ligand affinity. Active site residues involved in PZA binding were associated with multiple point mutations highlighting mutational diversity due to selection pressures at these functionally important sites. There were weak associations between genomic measures and biophysical effect of mutations. However, mutations associated with PZA resistance showed statistically significant differences between structural features (surface area and residue depth), but not hydrophobicity score for mutational sites. Most interestingly *M. tuberculosis* lineage 1 (ancient lineage) exhibited a distinct protein stability profile for mutations associated with PZA resistance, compared to modern lineages.

Keywords: *Mycobacterium tuberculosis*, pncA, nsSNPs, non-synonymous Single Nucleotide Polymorphisms, biophysical effects, thermodynamic stability, mCSM, FoldX

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom

Reviewed by:
Wim Vranken,
Vrije University Brussel, Belgium
Raghavan Varadarajan,
Indian Institute of Science (IISc), India

****Correspondence:***
Nicholas Furnham
Nick.Furnham@lshtm.ac.uk

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *20 October 2020*
Accepted: *14 April 2021*
Published: *23 July 2021*

Citation:
Tunstall T, Phelan J, Eccleston C,
Clark TG and Furnham N (2021)
Structural and Genomic Insights Into
Pyrazinamide Resistance
in Mycobacterium tuberculosis
Underlie Differences Between Ancient
and Modern Lineages.
Front. Mol. Biosci. 8:619403.
doi: 10.3389/fmolb.2021.619403

INTRODUCTION

Tuberculosis (TB), is a highly infectious and contagious air-borne disease caused by the bacterium *Mycobacterium tuberculosis*. Despite its ancient origins and the efforts to develop disease control and prevention measures, the disease continues to cause a global public health burden, with increased drug resistance making control difficult. In 2019, WHO reported around 10 million global cases of TB of which 1.4 million result in death (World Health Organization [WHO], 2020). In 2019, 465,000 cases of rifampicin resistant TB (RR-TB), among which 78% cases of multidrug-resistant TB (MDR-TB, defined as having additional resistance to isoniazid) were reported. Among these RR/MDR cases, ∼6% cases were further resistant to one fluoroquinolone and one injectable second line drug, leading to extensively drug resistant TB (XDR-TB) (World Health Organization [WHO], 2020).

The size of the *M. tuberculosis* genome (reference H37Rv strain) is 4.4 Mb, with a high (65%) GC content. The *M. tuberculosis* genome is clonal, and consists of seven main lineages, which vary by their geographical spread (L1: Indo-Oceanic, L2: East Asian, L3: East-Africa-Indian, and L4: Euro-American) (Phelan et al., 2016). The lineages are further classified into ancient (L1, L5–6), modern (L2–4), and intermediate (L7) strains, with L2 being particularly mobile as evidenced by its recent spread to Europe and Africa from Asia (Phelan et al., 2016). The *M. tuberculosis* lineages appear as distinct clades on phylogenetic trees (Coll et al., 2014) and govern disease transmission and dynamics with phenotypic consequences on clinical severity and drug resistance (Ford et al., 2013; Reiling et al., 2013), including recent reports of lineage-specific associations with the latter (Oppong et al., 2019). Drug resistance in *M. tuberculosis* is almost exclusively due to mutations [including non-synonymous Single Nucleotide Polymorphisms (nsSNPs), insertions and deletions (INDELs)] in genes coding for drug-targets or drug-converting enzymes. Changes in efflux pump regulation may also have an impact on the emergence of resistance (Al-Saeedi and Al-Hajoj, 2017) and putative compensatory mechanisms have been described to overcome fitness impairment that arises during the accumulation of resistance conferring mutations (de Vos et al., 2013). Resistance-associated point mutations have been described for all first-line drugs, including rifampicin, isoniazid and pyrazinamide, as well as for several second-line and newer drugs (fluoroquinolones, bedaquiline) (Somoskovi et al., 2001; Boonaiam et al., 2010; Segala et al., 2012), but knowledge is still incomplete.

Pyrazinamide (PZA) is a crucial antibiotic used in WHO recommended combination therapies in the front-line treatment of TB. It is a pro-drug which is activated by the amidase activity of the enzyme pyrazinamidase/nicotinamidase (PZase; MtPncA) encoded by the *pncA* gene, converting PZA to its active form of pyrazinoic acid (POA). Despite its indispensable status in TB treatment, PZA's exact mode of action remains poorly understood. Other genes (*rpsA* and *panD)* have been implicated in PZA resistance (Dookie et al., 2018) with a recent study suggesting that PZA exerts its antibacterial activity by acting as a target degrader of panD, blocking the synthesis of coenzyme A (targeted by POA) (Gopal et al., 2020). Despite this, mutations in the *pncA* gene remain the most common mechanism of PZA resistance (Khan et al., 2019).

Advances in whole genome sequencing (WGS) is assisting the profiling of *M. tuberculosis* for drug resistance, lineage determination and virulence, and presence in a transmission cluster (Phelan et al., 2019a), thereby informing clinical management and control policies. This is reflected in the WHO recommendation for use of rapid molecular testing for detecting TB and drug resistant TB (World Health Organization [WHO], 2020). The use of WGS can uncover new resistance mutations through genome-wide association studies (GWAS) and convergent evolution analysis (Phelan et al., 2016; Coll et al., 2018).

Furthermore, using protein structure, the biophysical effects of point polymorphisms can be investigated allowing a mechanistic understanding of resistance development (Phelan et al., 2016; Kavvas et al., 2018; Portelli et al., 2018). This approach can highlight important functional resistance mutations before they take hold in a population, corroborate drug susceptibility test results, as well as provide insights in highly polymorphic candidate loci (e.g., *pncA*) where many of the putative mutations have low frequency. It has been observed that sites with multiple mutations (>2) are linked to drug resistance (Comas et al., 2011), but such resistance hotspots may not necessarily lie close to the drug binding site. To this effect, sites with 2 mutations are considered as "emerging" or "budding" resistance hotspots (Portelli et al., 2018).

One assessment of the impact of missense mutations is to measure the change in a protein structure's as well as drug-target complex's physical interactions that contribute to its overall stability. Computational approaches (e.g., the *mCSM* suite; Pires et al., 2014a, 2016; Pires and Ascher, 2016, 2017; Rodrigues et al., 2019) have been developed to predict the effects of missense point mutations on overall protein structure stability, as well as the binding affinity/stability of ligand, protein-protein, and protein-nucleic acid interactions within a single framework, based on either an experimentally resolved structure or derived model. Here we apply such approaches to the effects of missense point mutations in the *pncA* gene. In addition, we also analyse biophysical structural features including surface area, residue depth and hydrophobicity for residues and sites associated with missense point mutations.

A crystal structure for pncA from *M. tuberculosis* has been determined as a monomeric enzyme of 186 amino acids (19.6 kDa) (Petrella et al., 2011). The structure comprises a 6-stranded parallel beta sheets, with helices on either side forming a single α/β domain with a metal cofactor (iron, Fe2+) binding site formed of D49, H51, H57, and H71. The substrate binding cavity in MtPncA is small, approximately 10 Å deep and 7 Å wide. It consists of highly conserved residues F13 and W68 that are essential in substrate binding with Y103 and H137 limiting access to this cavity (Petrella et al., 2011). The catalytic triad consisting of C138, D8, K96 is indicative of a cysteine-based catalytic mechanism (Petrella et al., 2011). Leveraging this crystal structure, we developed an *in silico* framework to assess the biophysical impact of *pncA* mutations and their resistance risk as determined by GWAS. In this study, we attempt to understand PZA resistance by exploring the relationship between the genomic features and the biophysical consequences of stability and affinity of nsSNPs, and how this is reflected in differences between *M. tuberculosis* lineages.

MATERIALS AND METHODS

SNP Dataset

The dataset consists of 35,944 *M. tuberculosis* isolates, which has been described recently (Napier et al., 2020). In brief, it encompasses all the main lineages (1, 5, and 6, ancient; 2, 3, and 4, modern; 7 intermediate), and drug susceptibility testing across 8 first-and second-line anti-TB drugs. Across these isolates, mutations in the *pncA* coding region with non-synonymous amino acid changes (nsSNPs) were extracted. These nsSNPs were further annotated for their link with drug resistance as defined by their presence in the TB-Profiler mutation database (Phelan et al., 2019b). Initial analysis aimed at understanding the structure and characterising the active site, followed by *in silico* predictions to quantify the enthalpic and entropic effects of GWAS-identified nsSNPs on the pncA protein structure. Subsequently, additional metadata relating to the clinical isolates were studied in relation to the structural effects of mutations. The general methodology workflow followed in this analysis is similar to the one described previously (Portelli et al., 2018).

Drug and Target: Structural Data

In the absence of a drug (PZA) and target (pncA) complex, respective individual structures were obtained from RSCB PDB database (Berman et al., 2000). The crystal structure of *pncA* in *M. tuberculosis* is available as PDB entry 3PL1 (Petrella et al., 2011), while the structure of PZA was extracted from PDB entry 3R55 (Singh et al., 2011). The molecular motion of pncA was analysed by Normal Mode Analysis using the DynaMut tool (Rodrigues et al., 2018) (**Supplementary Figure 1**).

Protein-Ligand Docking: Autodock Vina

The *pncA*-PZA complex was generated using the software AutoDock Vina, version 1.1.2 (Trott and Olson, 2009). Autodock Vina is an open-source, freely available molecular modelling platform to perform protein-ligand docking. Docking was carried out with default settings and guided by the positioning of the ligand within the active site as descried by Petrella et al. (2011). The complex was generated to facilitate downstream analyses by mCSM-lig (Pires et al., 2016) Autodock Vina returns bound conformations with their respective predicted binding affinity values. The prediction of binding affinity (strength of the ligand interaction with its target) is based on one of several scoring functions, which rank the poses in increasing order of predicted binding affinity. Binding free energy is calculated using a semi-empirical force field, combining experimental and knowledge-based information. The docking poses were visualised and inspected in UCSF Chimera 1.13 (Pettersen et al., 2004) according to the occupation of search space and diversity of pose conformations (**Supplementary Figure 2**). The top two binding poses were closely matched with the conformations generated by Karmakar et al. (2018) and Petrella et al. (2011), respectively (**Supplementary Figure 3**). The best pose was chosen considering the ligand orientation generated by molecular docking performed by Karmakar et al. (2018) and comparing interaction of both poses with active site residues through an Arpeggio (Jubb et al., 2017) analysis (**Supplementary Figure 4**).

Ligand extraction and protonation were carried out using UCSF Chimera, version 1.11 (Pettersen et al., 2004) while identification of rotatable bonds was carried out in Autodock tools (available as part of MGL tools, version 1.5.6) (Morris et al., 2009) where protonation of the ligand is specifically required by Autodock Vina (Trott and Olson, 2009). Similarly, protein extraction and explicit removal of solvent were carried out in UCSF Chimera, version 1.11 (Pettersen et al., 2004), and other steps in the overall protein preparation process were carried out in Autodock tools (part of MGL tools, version 1.5.6) (Morris et al., 2009). All the required parameters to perform docking needed to be included in a configuration file.

In silico Predictions: mCSM DUET, FoldX, mCSM-lig

The computational tools based on mutation cut-off scanning matrix, primarily *mCSM DUET* (Pires et al., 2014a) and *mCSM-lig* (Pires et al., 2016) were used to investigate the structural effects of nsSNPs within the pncA target protein. The effects of nsSNPs within *pncA* were analysed with respect to protein stability (DUET and FoldX (Schymkowitz et al., 2005) and ligand affinity (mCSM-lig). The consequences of these effects were to investigate change in protein fold and function, and effect on mechanism of PZA drug activation, respectively. Results from mCSM-lig (Pires et al., 2016) return both ligand affinity and DUET scores, hence only mCSM-lig was run to obtain both the outputs simultaneously.

A semi-automated pipeline was constructed for mCSM and FoldX to submit and extract results for multiple mutations consecutively using python and shell scripts. Both tools require wild type structure, chain ID and a list of nsSNPs in the X <POS> Y format (X: wild type residue; <POS> : position, Y: mutant residue). The residue symbols (X and Y) are specified as one letter amino acid code. DUET and FoldX estimate mutational impact as a change in Gibbs Free energy ($\Delta\Delta G$) in Kcal/mol. The classification of mutational impact based on $\Delta\Delta G$ from these methods are categorised in opposing ways. For example, $\Delta\Delta G < 0$ of a SNP is classified as a “destabilising” according to DUET, while the same is classified as “stabilising” according to FoldX.

The mutational impact on ligand affinity is calculated as a log fold change between wild type and mutant binding affinities. In addition to SNP identifiers, mCSM-lig requires the ligand affinity of the wild-type protein to be specified in nano Molar (nM) for affinity change calculations. Since the binding affinity returned by AutoDock Vina, version 1.1.2 (Trott and Olson, 2009) is in Kcal/mol, these needed to be converted to nM via Eq. 1 (below). The binding affinity for PZA in nM was 0.9911.

$$\Delta G = -RT\ln K. \quad (1)$$

Equation 1: Calculation of binding free energy, ΔG, where R is the gas constant, 1.987 cal K^{-1} mol^{-1} and T is the absolute temperature, 298 K. Adapted from Morris et al. (1998).

The mCSM suite of tools (Pires et al., 2014a, 2016; Pires and Ascher, 2017; Rodrigues et al., 2019) are based on graph-based measures at an atomic level along with machine learning (ML) tools for predicting enthalpic and entropic effects of stability. mCSM achieves this broadly by generating a signature encompassing the wild-type milieu and change in pharmacophore properties upon mutation (Pires et al., 2014b). Owing to the inter-atomic distance pattern within mCSM describing the wild-type residue environment, novel parameters like residue depth and long-range interactions are implicitly considered. In this manner, mCSM is able to characterise both local and global effects of missense point mutations. The mutational change at the atomic level is considered by using a change in the "pharmacophore count" vector, thus obviating the need to have explicit mutant structure. All mCSM tools (Pires et al., 2014a, 2016; Pires and Ascher, 2016, 2017; Rodrigues et al., 2019) use the atomic changes, while DUET (Pires et al., 2014a) is an ensemble method combining methods of mCSM stability (Pires et al., 2014b) and SDM (Worth et al., 2011; Pandurangan et al., 2017). FoldX, however is an empirical-based prediction tool which summarises the change in stability between mutant and wild type protein structures using a combination of energy terms based on fundamental intramolecular interactions (Schymkowitz et al., 2005).

Other Structural Parameters

Additional structural parameters for wild type structure were also included in the analysis. These were: Accessible (ASA) and Relative Surface Area (RSA), residue depth (RD), hydrophobicity values according to the Kyte-Doolittle scale (KD). The DSSP programme (Kabsch and Sander, 1983; Touw et al., 2015) was run to extract the ASA and RSA values, while RD values calculated as described by Chakravarty and Varadarajan (1999) were calculated using the depth server available at http://cospi.iiserpune.ac.in/depth. The KD values were fetched from the expasy server (Artimo et al., 2012) available at https://web.expasy.org/protscale/.

Data Normalisation: DUET, FoldX, and mCSM-lig

The DUET (Pires et al., 2014a), FoldX (Schymkowitz et al., 2005), and mCSM-lig (Pires et al., 2016) scores associated with each SNP were subsequently normalised between the range of −1 and 1. For mCSM-lig analyses, data was filtered according to distance from interacting site and only residues within a distance of 10 Å of the ligand (PZA) were considered for all ligand affinity analyses.

Minor Allele Frequency and Odds Ratio Calculations: SNP Dataset

Across the *M. tuberculosis* isolates tested for PZA drug susceptibility data, we performed association analysis to estimate the risk of resistance for SNP alleles. For each nsSNP, minor allele frequency (MAF) and odds ratio (OR) were calculated in relation to all samples tested for PZA susceptibility. MAF is the average occurrence of a given nsSNP, and OR is the measure of association of a given nsSNP with PZA resistance. In addition to unadjusted odds ratio (OR), and similar to a GWAS approach, adjusted odds ratio (aOR) were estimated using logistic regression models with a kinship matrix adjusting for a random effect representing the SNP-based relationships between samples (e.g., the lineage-based population structure) (Zhou and Stephens, 2012; Coll et al., 2018). *P*-values were estimated using Fisher and Wald test for unadjusted and adjusted ORs, respectively.

Statistical Analyses

Data was analysed using non-parametric statistical tests. For assessing correlations, Spearman correlation values were calculated. For comparing lineage distributions, the Kolmogorov-Smirnov (KS) test was used. Statistical significance thresholds used are $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

Data Visualisation

All plots were generated using R statistical software, version 4.0.2 (R Core Team, 2014). Protein and ligand structures were generated using UCSF Chimera, version 1.11 (Pettersen et al., 2004).

RESULTS

Analysing the pncA Molecular Motion and pncA-PZA Complex

Molecular motion in pncA was analysed by Normal Mode Analysis (NMA). Regions undergoing the greatest movement were limited to residues in loop regions and mainly concentrated to loop 60–66, followed by loop residues 39–41 and 111–113. Residues at site 165–167 within helix 164–178 showed the least flexibility (**Supplementary Figure 1**). The frequency of mutations in these variable regions was most prominent for sites 62–63 (>2 mutations) while the other sites were limited to at most two mutations (**Figure 1**). Mutations within the most flexible region (residues 60–66) of pncA showed mixed effects in relation to their association with PZA resistance with the single mutation at site 64 related to PZA resistance. Sites 39 and 40 within the other highly flexible region 39–41 were not associated with any mutations in our study, while the two mutations at site 41 were not associated with PZA resistance. The region 111–113 is associated with single mutations at sites 111 and 112 which are not linked to PZA resistance, while site 113 was not associated with any mutations in our study. Sites 165–167, which form part of the helix (164–178), are the most stable according to NMA. Two residues (A165 and D166) within this helix were not associated with any mutations in our study, while a single mutation at site T167 was not associated with PZA drug resistance (**Supplementary Figure 1** and **Supplementary Table 1**). Docking with AutoDock vina (Trott and Olson, 2009) generated nine different conformations as per default settings. In six of these poses, the aromatic ring of PZA was oriented towards the substrate binding residue W68 (**Supplementary Figures 2A,B**). The top two poses (1 and 2) returned by Vina were similar to previous molecular docking studies (Petrella et al., 2011; Karmakar et al., 2018)

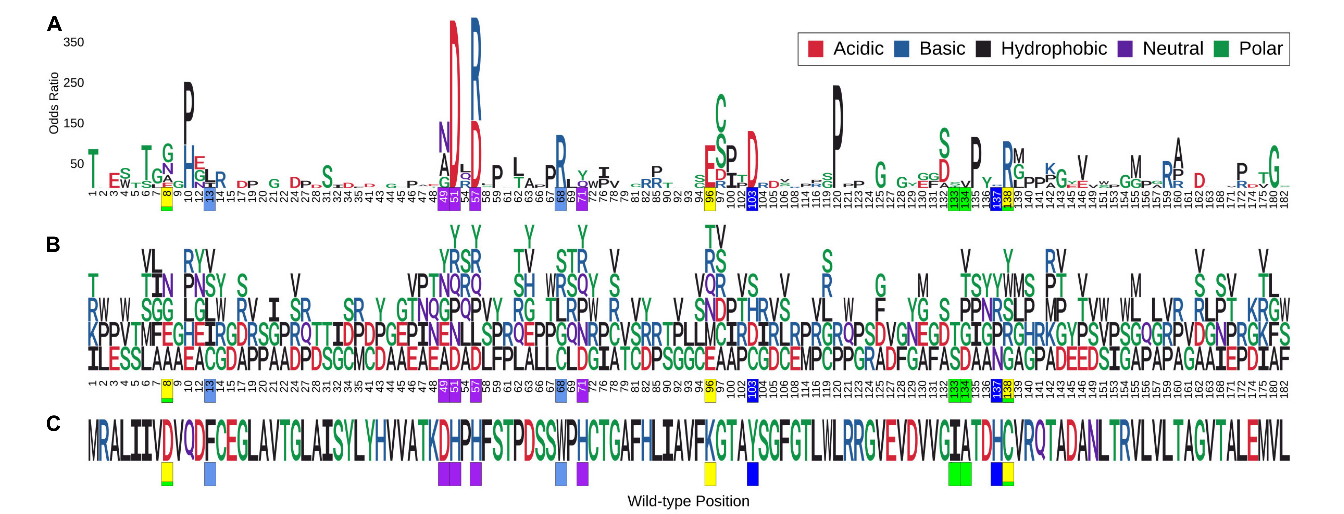


FIGURE 1 | Logo plot showing sites with multiple missense point mutations and association with Odds Ratio. Sites associated with multiple (>2) missense point mutations (i.e., nsSNPs). A total of 386 mutations corresponding to 113 positions on the pncA protein structure were associated with multiple nsSNPs. The horizontal axis in **(A,B)** show the position numbers of sites with multiple nsSNPs, while part **(C)** shows the wild-type residues for each position. The vertical axis in **(A)** represents Odds Ratio (OR) where letters denote mutant residues which are proportional to their corresponding OR highlighting the most resistant mutation at each site and overall. Part **(B)** shows each mutant residue at a given position, highlighting nsSNP diversity by position. The wild-type and mutant residues are coloured according to the amino acid properties as denoted. Positions marked in yellow form the catalytic triad, residues in blue and teal are involved in substrate binding, those in green are involved in hydrogen binding while the ones in purple are involved in the iron centre coordination. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; pncA, pyrazinamidase.

(**Supplementary Figure 3**). A follow-up Arpeggio analysis (Jubb et al., 2017) indicated that pose 1 when compared to pose 2, has more H-bonds (4 vs. 1), fewer aromatic contacts (3 vs. 13), and greater Van der Waals interactions (3 vs. 1) (**Supplementary Figures 4A,B**). Therefore, model with pose 1 was chosen to form the pncA-PZA complex (**Supplementary Figure 5**).

Genomics Data

SNP data from 35,944 *M. tuberculosis* clinical isolates tested for drug susceptibility to a range of first and second line drugs were obtained (Napier et al., 2020). Among these, 39% (n = 13,914) of these isolates were tested for PZA drug susceptibility. The isolates were collected from over 30 different countries and represented the 4 main *M. tuberculosis* lineages (L1, n = 144; L2, n = 1,886; L3, n = 190; L4, n = 2213) (**Supplementary Figure 6**). In order to infer whether the ancestral pncA sequences for each lineage differed, we quantified the number of samples without any mutations in each lineage. The majority of isolates in L1–L4 had an identical *pncA* sequence as the H37Rv reference indicating that the ancestral sequences for these lineages do not differ. The majority were pan susceptible (n = 23,256, 64.7%), with the remainder MDR-TB (n = 6,691, 18.6%), XDR-TB (n = 989, 2.8%), or another type of resistance referred to as DR-TB (n = 5,008, 13.9%) (**Table 1**). From the list, only nsSNPs within the protein coding region of *pncA* (n = 4,731, 13.2%) were considered for our analyses (**Table 1**). The majority of these were MDR-TB (n = 3,290, 69.5%) followed by relatively equal numbers of XDR-TB and DR-TB (n = 625, 13.2% and n = 632, 13.4%, respectively), while only a small percentage were susceptible (n = 184, 3.9%) (**Table 1**). From a total of 13,914 samples tested for PZA drug susceptibility, a minority of those were found to be resistant (n = 2,379, 17.1%) (**Table 1**). However, the burden of PZA resistance among

TABLE 1 | Number of samples analysed.

Item name	Total number (%)
Clinical isolates/samples	35,944
Samples classified Susceptible	23,256 (64.7)
Drug resistant (DR)	5,008 (13.9)
Multi-drug resistant (MDR)	6,691 (18.6)
Extreme drug resistant (XDR)	989 (2.8)
Samples tested for PZA drug susceptibility	13,914
Resistant	2,379 (17.1)
Samples with nsSNPs in the protein coding region of *pncA*	4,731 (13.2)
Susceptible	184 (3.9)
Drug resistant (DR)	632 (13.4)
Multi-drug resistant (MDR)	3,290 (69.5)
Extreme drug resistant (XDR)	625 (13.2)
Samples with *pncA* nsSNPs tested for PZA drug susceptibility	2,289 (48.4)
Samples with *pncA* nsSNPs resistant to PZA	1,677 (73.3)
Unique nsSNPs: No. of sites	424 nsSNPs: 151 sites

Summary of clinical isolates from genome-wide analysis. PZA, pyrazinamide; nsSNPs, non-synonymous Single Nucleotide Polymorphisms.

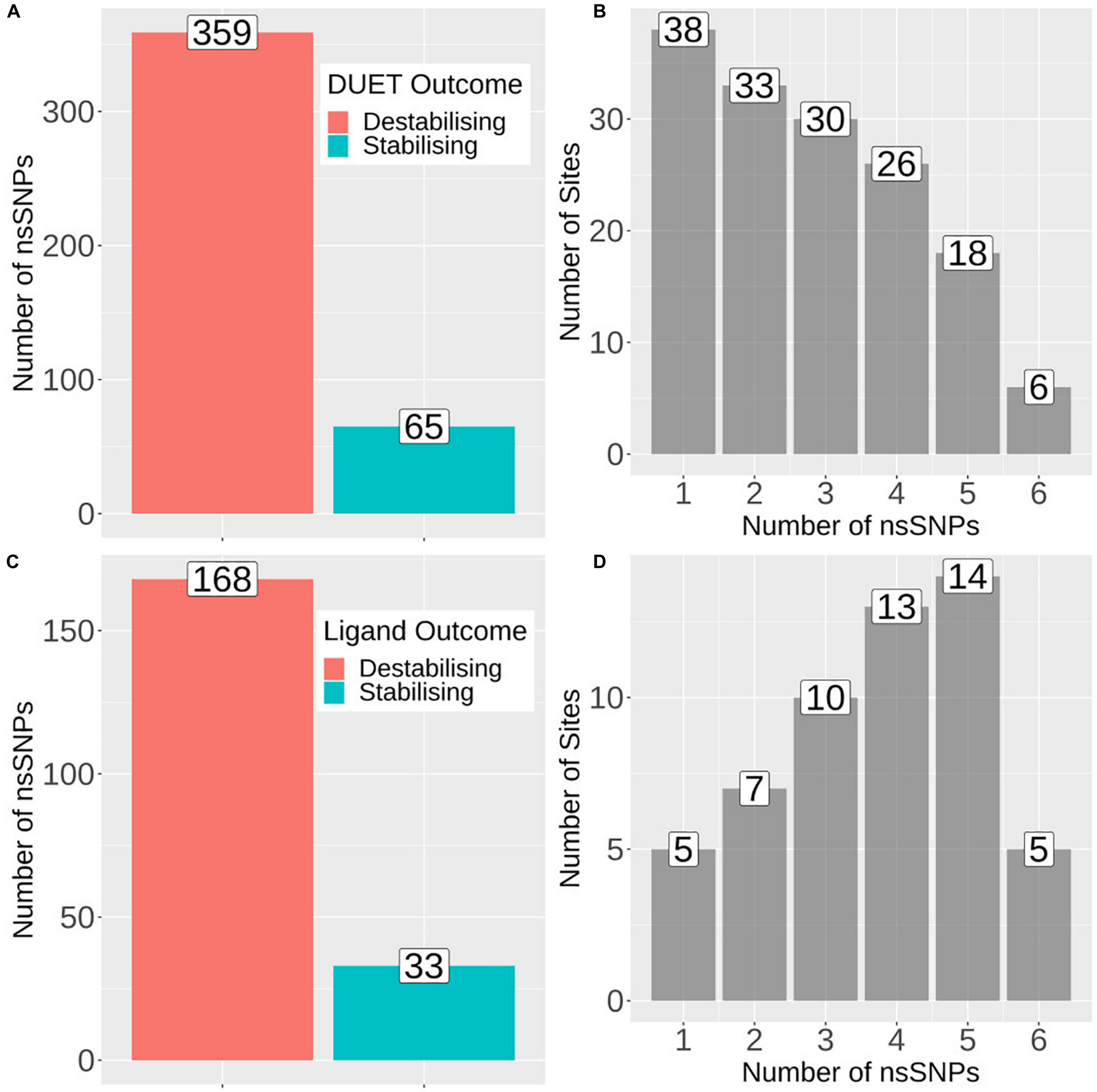


FIGURE 2 | Barplots showing number of mutations and sites associated with protein stability and ligand affinity. **(A)** Number of nsSNPs categorised as destabilising (*n* = 359) and stabilising (*n* = 65) according to DUET protein stability. **(B)** Frequency of sites associated with the number of nsSNPs, where horizontal axis denotes the number of nsSNPs and vertical axis denotes the total number of sites/positions corresponding to the number of nsSNPs. **(C)** Barplot showing the number of nsSNPs categorised as destabilising (*n* = 168) and stabilising (*n* = 33) according to mCSM ligand affinity where sites lie within 10Å of ligand. **(D)** Frequency of sites associated with the number of nsSNPs, where horizontal axis denotes the number of nsSNPs and vertical axis denotes the total number of sites/positions corresponding to the number of nsSNPs. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms.

samples containing nsSNPs in the protein coding region was high (*n* = 1,677, 73.3%) (**Table 1**).

Across the 4,731 isolates, 424 distinct nsSNPs corresponding to 151 distinct amino acid positions on the pncA structure were identified (**Figures 2A,B**). A total of 201 nsSNPs corresponding to 54 amino acid changes were within 10 Å of the ligand binding site (**Figures 2C,D**). The majority of these nsSNP mutations have been annotated as being linked to PZA resistance within the TBProfiler tool (227/424). The majority of these nsSNP mutations have been annotated as being linked to PZA resistance within the TBProfiler tool (227/424; denoted as DM), while the others (197/424; denoted as OM) were assumed to have weak or no links. Genomic measures like minor allele frequency (MAF) and odds ratio (OR) were obtained for a total of 322 nsSNPs, with adjusted OR (aOR) estimated for a total of 163 nsSNPs. Across the majority of these nsSNPs, the MAFs were low (median: 0.02% range: 0.01–2.11%) (**Supplementary Figure 7A**). Similarly, when considering ORs, the majority of the nsSNPs had high ORs (median: 9.70, range: 0.22–414.61) (**Supplementary Figure 7D**). When looking at the distribution of MAF and OR within mutations associated with PZA resistance (DM) and other mutations (OM) (**Supplementary Figures 7B,E**), DM mutations

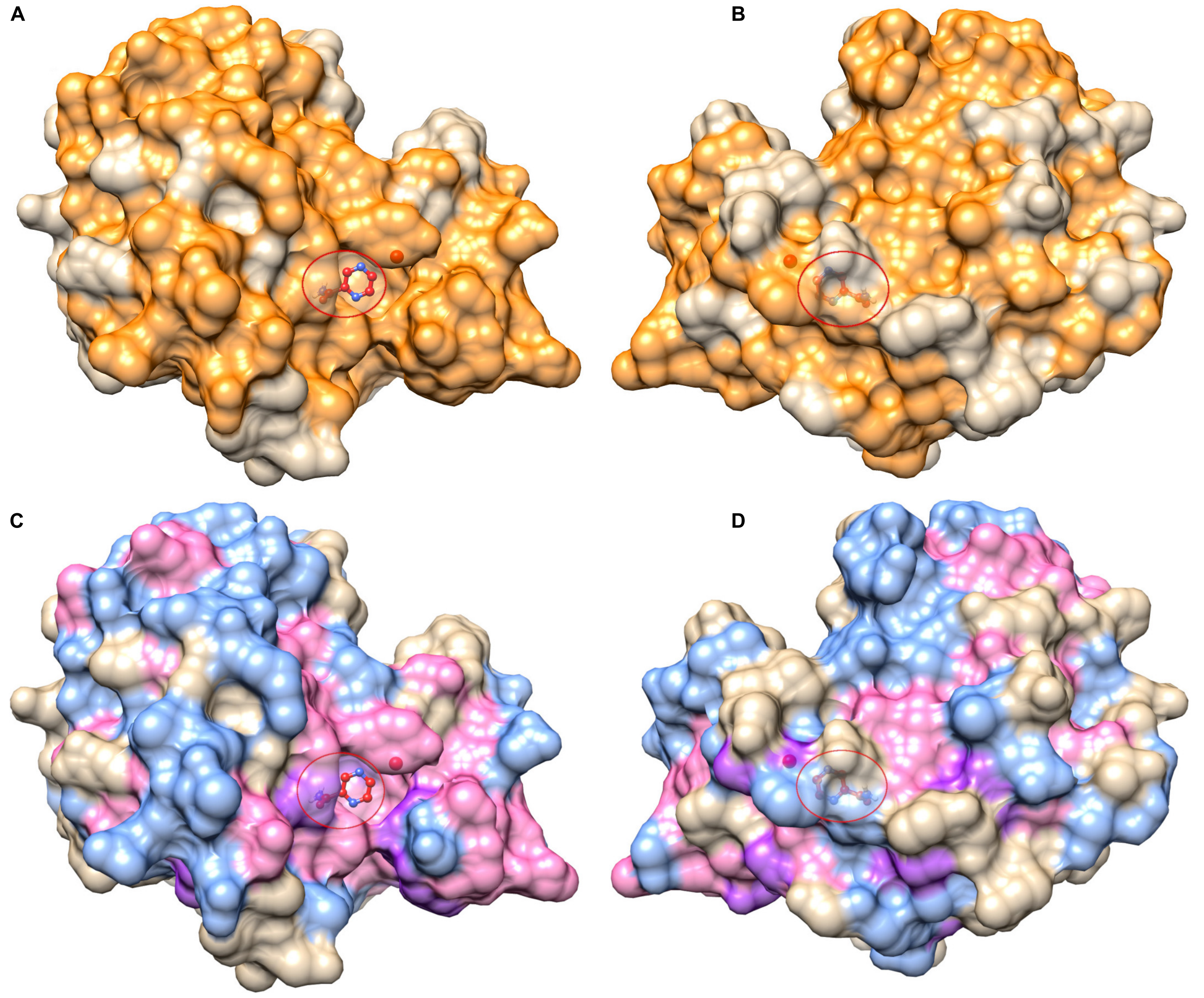


FIGURE 3 | Mutational landscape of pncA structure (3PL1) coloured by positions linked to pyrazinamide drug (PZA) resistance. Panels **(A,B)** show all mutational positions in orange while mutational positions in **(C,D)** are further coloured by mutations classed as either drug resistant mutations (purple) or "other mutations" (blue), while sites linked to mutations belonging to either category are coloured in pink. The right panels **(B,D)** depict the corresponding structure rotated by 180°. The ligand (PZA) is shown as ball and stick within the active site denoted by the red circle. The figure is rendered using UCSF Chimera (version 1.14). pncA, pyrazinamidase.

were associated with significantly higher ($P < 0.0001$) MAF and OR (**Supplementary Figures 7C,F**).

Understanding Mutational Effects on pncA Stability and PZA Binding Affinity

The 424 nsSNPs mapped onto the crystal structure of pncA revealed that mutational landscape of pncA appears distributed (**Figures 3A,B**) throughout the structure. Sites linked to drug resistant mutations were predominant around the PZA binding (active) site, while sites exclusively linked to mutations classed in the "other" category are distal to the active site (**Figures 3C,D**, **4**). Furthermore, active site residues were associated with a multiple point mutation (**Table 2** and **Figures 1B**, **5C**). All active site and hydrogen-bond forming residues with the ligand were associated with multiple mutations (≥2) (**Figure 1B**), thus representing the high diversity of mutations present within pncA. Despite this, there appears to be some degree of clustering around positions 4–14, 46–97, 132–143 involving the active site and metal centre residues (**Figure 5C**).

The biophysical effect of mutations on protomer stability, estimated as $\Delta\Delta G$ (Kcal/mol), was measured using DUET (Pires et al., 2014a) and FoldX (Schymkowitz et al., 2005), while mutational impact on ligand affinity was measured using mCSM-lig (Pires et al., 2016) (see section "Materials and Methods"). Assessing mutational effects on protein stability as measured by

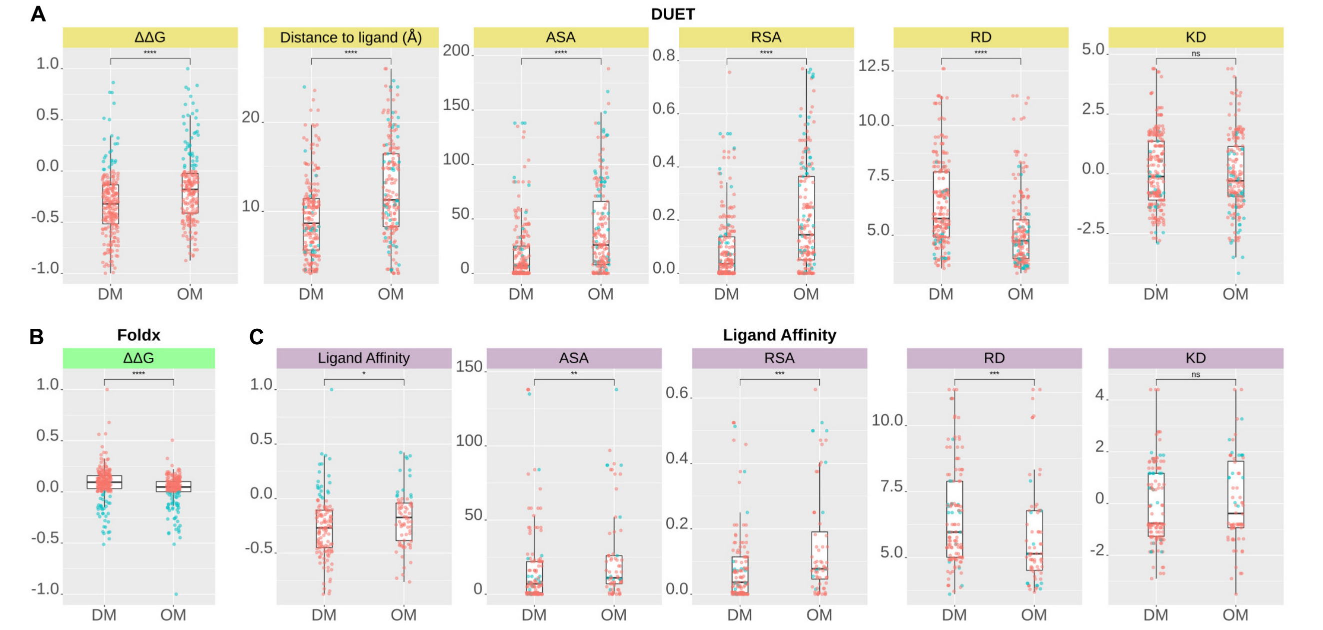


FIGURE 4 | Comparison of structural features between Drug resistance (DM) and other mutations (OM) of pncA gene mutations according to **(A)** DUET protein stability (ΔΔG), **(B)** FoldX stability (ΔΔG), and **(C)** Ligand Affinity. A total of 424 nsSNPs for DUET and FoldX (DM, n = 227, OM, n = 197), while a total of 201 nsSNPs (DM, n = 129 OM, n = 72) lying within 10 Å of PZA for ligand affinity were included in the analysis. DM and OM mutations were compared using Wilcoxon rank-sum (unpaired) and statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). The figure is generated using R statistical software (version 4.0.2). ns, non-synonymous Single Nucleotide Polymorphisms; pnca, pyrazinamidase; PZA, pyrazinamide; Å, Angstroms; ΔΔG, Change in Gibbs free energy in Kcal/mol; ASA, Accessible Surface Area; RSA, Relative surface Area; RD, Residue Depth; KD, Kyte-Doolittle Hydrophobicity values.

DUET, nearly 85% had a destabilising effect (n = 359) compared to nearly 15% mutations with stabilising effects (n = 47) as shown in **Figure 2A**. When assessing ligand affinity, 47.4% (n = 201) SNP mutations were present within 10 Å of the PZA binding site (**Figure 2C**). Similar to DUET stability effects, the majority (84%; n = 168) of nsSNPs were destabilising while 16% (n = 27) were stabilising for ligand binding affinity (**Figure 2C**). More than 50% of the mutational positions were associated with multiple nsSNPs for both protein stability (n = 113) and ligand affinity (n = 49) (**Figures 2B,D**). The average protein stability and ligand affinity effects of all mutations mapped onto the pncA structure (**Figures 5A,B**), highlight mutations with opposing effects for protein stability and ligand affinity. These effects are pronounced for active site residues (I133, A134, H137, C138) (**Figures 5C,D**).

There were 80 sites within *pncA* associated with multiple nsSNPs (>2) (**Figures 1B, 2B**) which included all active residues except I133 which was associated with 2 mutations (**Figure 1B**). Sites with 2 nsSNPs are considered to be budding resistance hotspots (n = 33 for protein stability, n = 7 for ligand affinity). A total of 57 nsSNPs within 5 Å of PZA were considered to be within the first shell of residues lining the active site (**Table 2**). While majority of the mutational sites associated with more than two mutations comprise of destabilising mutations, positions 1, 2, 10, 12, 43, 46, 51, 57, 63, 67, 69, 78, 82, 92, 96, 100, 104, 105, 129, 135–138, 142, 149, 164, 168, and 174 comprised of both stabilising and destabilising mutations (**Figure 5C**). Similarly, for ligand affinity, most mutational sites had destabilising mutational effects, with positions 7, 8, 13, 27, 49, 72, 78, 96, 102, 103, 105, 134, 137, 138, and 162 associated with mutations resulting in mixed stability impact. Position 163 comprised only of mutations with stabilising effects (**Figure 5D**). The budding resistance hotspot active site residue I133 contained both mutations with destabilising effect for protein stability (**Figure 5C**), while stabilising for ligand affinity (**Figure 5D**). Similarly, for budding resistance hotspots, majority of the nsSNPs were associated with destabilising effects. For protein stability, 9/33 sites had mutations with mixed stability (positions 15, 32, 61, 66, 76, 114, 127, 153, and 161) (**Figure 5C**), while only position 20 showed mixed stability effects for ligand affinity (**Figure 5D**).

Mutations With Extreme Effects

Mutations with extreme effects on protein stability and affinity are summarised in **Table 3**. Overall, the most destabilising mutation according to DUET was L4S, where a change from a hydrophobic to a polar residue may contribute to disruption of local conformation (**Table 3**). The closest most destabilising mutational effect on protein stability was from A134D (wild-type residue involved in hydrogen bonding) (**Table 3**), likely resulting in electrostatic and steric clashes due to a change in charge and volume affecting the overall stability negatively. The most stabilising mutation on protomer stability was from active site residue Y103D, while the closest such mutation was C138R (**Table 3**). The stabilising effect of these mutations on the protein stability and ligand affinity is thought to result from the electrostatic interactions working favourably for sites lying within 5 Å of the ligand. The most destabilising mutation according

TABLE 2 | Mutations close to the active site of PZA.

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
1	**A134D**	Others	0.01	2.42	1.00E+00	NA	NA	−2.98	D	3.05	0.58	S	1.03	D	10	0.08	1.87	6.77
2	**A134G**	Others	NA	NA	NA	NA	NA	−1.62	D	3.05	−0.38	D	−1.29	S	10	0.08	1.87	6.77
3	**A134P**	Others	0.01	9.70	1.71E-01	NA	NA	−1.43	D	3.05	0.08	S	−5.20	S	10	0.08	1.87	6.77
4	**A134T**	Others	NA	NA	NA	NA	NA	−1.93	D	3.05	0.88	S	−0.94	S	10	0.08	1.87	6.77
5	**A134V**	Drug associated	0.04	19.43	3.68E-03	1.53	3.07E-05	−0.41	D	3.05	0.12	S	−1.46	S	10	0.08	1.87	6.77
6	**I133S**	Others	0.01	9.70	1.71E-01	NA	NA	−3.22	D	3.05	0.58	S	3.30	D	3	0.02	1.97	7.90
7	**I133T**	Drug associated	0.32	6.44	2.90E-09	0.86	4.86E-03	−2.79	D	3.05	0.70	S	1.58	D	3	0.02	1.97	7.90
8	**D8A**	Drug associated	0.01	19.41	2.92E-02	NA	NA	−0.51	D	3.22	−3.27	D	0.54	D	5	0.03	1.63	9.48
9	**D8G**	Drug associated	0.08	48.69	1.95E-07	1.25	4.42E-02	−0.85	D	3.22	−3.45	D	1.89	D	5	0.03	1.63	9.48
10	**D8E**	Drug associated	0.03	14.56	1.74E-02	1.19	1.46E-01	−0.79	D	3.22	0.01	S	1.90	D	5	0.03	1.63	9.48
11	**D8N**	Drug associated	0.05	29.16	1.49E-04	1.24	7.10E-03	−1.18	D	3.22	−1.66	D	−1.26	S	5	0.03	1.63	9.48
12	**C138G**	Others	NA	NA	NA	NA	NA	−0.02	D	3.28	−0.01	D	1.12	D	12	0.07	1.17	6.70
13	**C138S**	Drug associated	NA	NA	NA	NA	NA	0.00	D	3.28	0.81	S	−0.23	S	12	0.07	1.17	6.70
14	**C138W**	Others	NA	NA	NA	NA	NA	−1.05	D	3.28	0.94	S	−1.72	S	12	0.07	1.17	6.70
15	**C138Y**	Drug associated	NA	NA	NA	NA	NA	−0.52	D	3.28	0.91	S	−0.57	S	12	0.07	1.17	6.70
16	**C138R**	Drug associated	0.09	116.96	6.10E-10	1.74	4.08E-12	0.10	S	3.28	0.35	S	−2.12	S	12	0.07	1.17	6.70
17	**H137N**	Others	0.01	2.42	1.00E+00	NA	NA	0.19	S	3.42	−0.12	D	0.40	D	84	0.38	−1.40	4.60
18	**H137P**	Drug associated	NA	NA	NA	NA	NA	0.37	S	3.42	−0.77	D	2.19	D	84	0.38	−1.40	4.60
19	**H137Y**	Others	0.01	2.42	1.00E+00	NA	NA	0.86	S	3.42	−0.01	D	0.34	D	84	0.38	−1.40	4.60
20	**H137R**	Drug associated	0.03	4.85	1.38E-01	0.56	1.21E-04	−0.27	D	3.42	0.47	S	0.49	D	84	0.38	−1.40	4.60

(Continued)

TABLE 2 | Continued

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
21	**D49G**	Drug associated	0.05	29.16	1.49E-04	1.66	4.38E-08	−1.16	D	3.45	−3.46	D	0.46	D	7	0.04	−1.53	7.89
22	**D49A**	Drug associated	0.04	58.33	2.49E-05	1.67	3.17E-06	−0.45	D	3.45	−3.35	D	−2.07	S	7	0.04	−1.53	7.89
23	**D49N**	Drug associated	0.06	77.84	7.23E-07	1.51	3.14E-04	−1.68	D	3.45	−1.93	D	−0.33	S	7	0.04	−1.53	7.89
24	**D49Y**	Drug associated	0.01	9.70	1.71E-01	NA	NA	−0.74	D	3.45	−1.86	D	−2.67	S	7	0.04	−1.53	7.89
25	**D49E**	Drug associated	0.02	9.70	7.77E-02	NA	NA	−0.47	D	3.45	0.25	S	−0.70	S	7	0.04	−1.53	7.89
26	**A102R**	Others	0.01	2.42	1.00E+00	NA	NA	−0.70	D	3.50	0.17	S	4.13	D	10	0.08	0.03	5.51
27	**A102P**	Others	0.06	14.58	5.08E-04	0.66	5.33E-04	−1.25	D	3.50	−0.23	D	−0.62	S	10	0.08	0.03	5.51
28	**A102V**	Others	0.06	2.43	1.88E-01	0.91	3.00E-01	−0.25	D	3.50	−0.16	D	−1.91	S	10	0.08	0.03	5.51
29	**A102T**	Drug associated	0.01	19.41	2.92E-02	1.75	4.98E-04	−0.72	D	3.50	0.88	S	−2.03	S	10	0.08	0.03	5.51
30	**F13C**	Others	0.01	1.21	1.00E+00	0.64	4.31E-03	−2.32	D	3.55	−0.49	D	2.70	D	24	0.10	0.60	6.93
31	**F13I**	Drug associated	0.03	14.56	1.74E-02	NA	NA	−1.76	D	3.55	−0.45	D	0.89	D	24	0.10	0.60	6.93
32	**F13L**	Drug associated	0.06	34.04	2.89E-05	1.37	2.29E-03	−2.03	D	3.55	−0.43	D	1.10	D	24	0.10	0.60	6.93
33	**F13V**	Others	0.01	1.21	1.00E+00	NA	NA	−2.57	D	3.55	−0.56	D	1.40	D	24	0.10	0.60	6.93
34	**F13S**	Drug associated	0.03	1.62	5.28E-01	0.60	3.07E-04	−3.10	D	3.55	0.22	S	2.59	D	24	0.10	0.60	6.93
35	**K96E**	Drug associated	0.08	107.17	3.58E-09	1.75	2.79E-06	−2.12	D	3.98	−0.67	D	6.92	D	8	0.03	−1.87	5.96
36	**K96Q**	Drug associated	0.03	4.85	1.38E-01	0.64	1.17E-01	−1.32	D	3.98	−0.08	D	1.04	D	8	0.03	−1.87	5.96
37	**K96T**	Drug associated	0.09	58.47	6.68E-09	1.84	2.25E-13	−0.86	D	3.98	−0.57	D	3.54	D	8	0.03	−1.87	5.96
38	**K96M**	Others	0.01	19.41	2.92E-02	NA	NA	0.41	S	3.98	−1.03	D	0.27	D	8	0.03	−1.87	5.96
39	**K96N**	Drug associated	0.01	2.42	1.00E+00	NA	NA	−1.16	D	3.98	0.33	S	2.61	D	8	0.03	−1.87	5.96

(Continued)

TABLE 2 | Continued

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
40	**K96R**	Drug associated	0.11	19.49	1.66E-07	1.43	2.16E-06	−0.17	D	3.98	0.08	S	−0.74	S	8	0.03	−1.87	5.96
41	**H71D**	Drug associated	0.01	9.70	1.71E-01	NA	NA	−2.69	D	4.18	−2.50	D	5.75	D	5	0.02	−0.77	6.25
42	**H71N**	Drug associated	NA	NA	NA	NA	NA	−2.67	D	4.18	−1.34	D	0.64	D	5	0.02	−0.77	6.25
43	**H71P**	Others	0.01	4.85	3.13E-01	NA	NA	−2.36	D	4.18	−2.89	D	3.26	D	5	0.02	−0.77	6.25
44	**H71Q**	Drug associated	0.01	19.41	2.92E-02	1.75	2.12E-04	−2.29	D	4.18	−1.73	D	1.12	D	5	0.02	−0.77	6.25
45	**H71R**	Drug associated	0.05	1.94	3.42E-01	0.88	2.01E-01	−1.93	D	4.18	−0.83	D	−1.52	S	5	0.02	−0.77	6.25
46	**H71Y**	Drug associated	0.18	25.67	4.52E-13	1.48	5.50E-08	−0.46	D	4.18	−1.96	D	−1.78	S	5	0.02	−0.77	6.25
47	**H57D**	Drug associated	0.73	166.91	2.08E-72	1.24	1.05E-01	−1.85	D	4.56	−1.28	D	1.83	D	16	0.07	−1.30	5.63
48	**H57P**	Drug associated	0.03	38.85	8.53E-04	1.55	1.16E-02	−1.23	D	4.56	−2.12	D	0.15	D	16	0.07	−1.30	5.63
49	**H57Q**	Others	NA	NA	NA	NA	NA	−1.29	D	4.56	−0.95	D	0.85	D	16	0.07	−1.30	5.63
50	**H57R**	Drug associated	0.19	254.92	1.02E-20	1.48	9.69E-09	−1.17	D	4.56	−0.28	D	1.25	D	16	0.07	−1.30	5.63
51	**H57L**	Drug associated	NA	NA	NA	NA	NA	−0.06	D	4.56	−1.92	D	−1.11	S	16	0.07	−1.30	5.63
52	**H57Y**	Drug associated	0.02	29.13	4.99E-03	2.08	7.92E-06	0.41	S	4.56	−1.16	D	−0.15	S	16	0.07	−1.30	5.63
53	**W68C**	Drug associated	0.04	24.29	7.49E-04	1.75	1.67E-04	−1.45	D	4.97	−1.58	D	2.68	D	45	0.16	−1.10	5.49
54	**W68G**	Drug associated	0.14	87.93	2.36E-13	1.58	7.39E-11	−2.57	D	4.97	−2.13	D	4.04	D	45	0.16	−1.10	5.49
55	**W68L**	Drug associated	NA	NA	NA	NA	NA	−1.62	D	4.97	−2.24	D	0.19	D	45	0.16	−1.10	5.49
56	**W68R**	Drug associated	0.20	132.41	4.03E-20	1.50	4.26E-09	−1.61	D	4.97	−0.58	D	0.08	D	45	0.16	−1.10	5.49
57	**W68S**	Drug associated	0.01	9.70	1.71E-01	NA	NA	−2.67	D	4.97	−1.04	D	2.65	D	45	0.16	−1.10	5.49

Fifty-seven mutations (nsSNPs) lying within 5 Å of PZA and the corresponding GWAS measures of minor allele frequency (MAF), Odds Ratio (OR), P-values, adjusted OR (aOR), and P-values from Wald test corresponding to aORs, along with structural measures of distance to ligand, DUET, FoldX, ligand affinity values and effect. Wild type residues for mutations highlighted and marked in green are considered to participate in hydrogen bonding, those in yellow form the catalytic triad, residues in teal (and blue) are involved in substrate binding, while the residues in purple are involved in the iron centre. The columns are coloured to highlight the most significant column attribute with deeper colours denoting the greatest effects. The dark colours in MAF, OR, and aOR columns indicate the highest values, while P-values are coloured with the darkest colour showing the most significant values. Values in the DUET, mCSM-lig, and FoldX columns are coloured according to the extent of their respective effects with red indicating destabilising and blue denoting stabilising effects. nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; GWAS, Genome-Wide Association Studies. D, Destabilising; S, Stabilising.

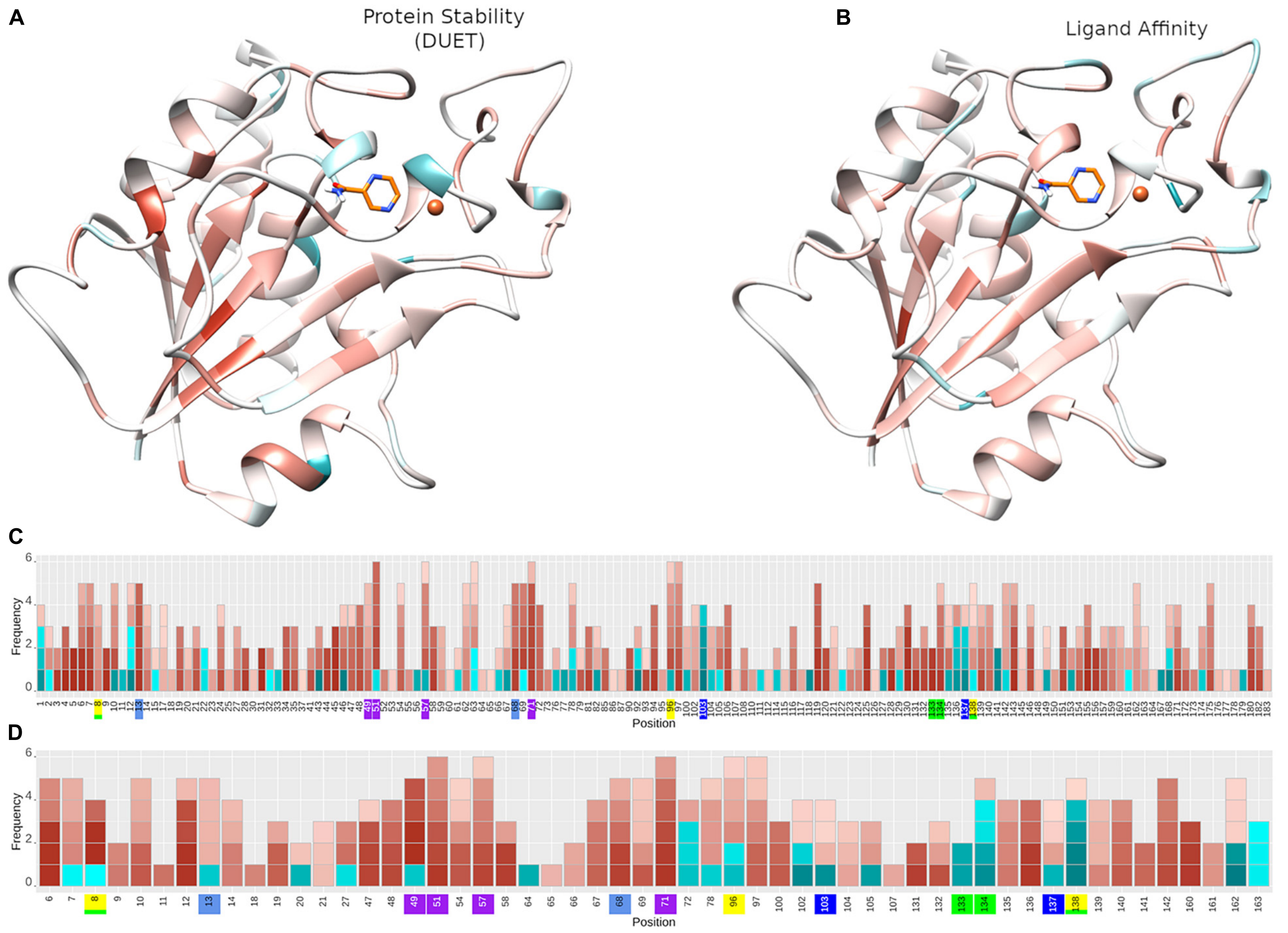


FIGURE 5 | Protein stability and ligand affinity effects of nsSNPs on pncA structure and by position. Mutational impact of nsSNPs on the pncA protein structure coloured by average **(A)** DUET Protein stability (*n* = 424) and **(B)** ligand affinity (*n* = 201). Barplots **(C,D)** showing the frequency of mutations within the pncA gene. The horizontal axis shows the mutational positions within pncA and the vertical axis shows the frequency of mutations. Positions on the horizontal axis are coloured to denote the active site residues: green (residues involved in hydrogen bonding with PZA), yellow (catalytic triad), blue and teal (substrate binding), purple (iron centre). For a given position, each corresponding mutation (nsSNP) is coloured by the level of stability according to **(C)** DUET(*n* = 424) and **(D)** Ligand affinity (*n* = 201) where the horizontal axis denotes amino acid positions in pnca, and is restricted to positions lying within 10 Å of PZA for ligand affinity. Destabilising mutations are depicted in red and stabilising mutations in blue, where colour intensity reflects the extent of effect, ranging from −1 (most destabilising) to + 1 (most stabilising). The structural figures **(A,B)** are rendered using UCSF Chimera (version 1.14). The barplot figure **(C,D)** is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; pncA, pyrazinamidase.

to ligand affinity was D49G located at ∼3.5 Å (**Table 3**). The three subsequent destabilising mutations for ligand affinity were also all within 5 Å of PZA binding site namely D8G (∼3 Å), D49A (∼3.5 Å), and D8A (∼3 Å) (**Supplementary Table 1**), all arising likely due to the loss of charge and volume interfering with ligand interaction. The mutation with the greatest stabilising effect on ligand affinity was G162D, located at ∼8 Å, i.e. outside the first shell of influence (>5 Å) from the ligand. This is possibly due to the resulting electrostatic effects and increase in volume, which may favour hydrogen bond formation with nearby residues and PZA binding, thereby increasing affinity (**Table 3**). The closest most stabilising mutational impact on ligand affinity was due to mutation A134P, though this was a marginal effect (**Table 3**). The most destabilising mutation according to FoldX was C72W, which is located far away from the active site (∼27 Å). Interestingly, mutation A134P was the most stabilising according to FoldX, while the same was estimated to have a destabilising effect according to DUET (**Table 3**). All mutations except A134D and A134P were associated with PZA drug resistance (**Table 3**).

Relating Structural and GWAS Analyses

The minor allele frequencies for the 424 nsSNPs were mapped onto their corresponding amino acid positions of the *pncA* gene (**Supplementary Figure 8**). Position 10 had the highest cumulative minor allele frequency (MAF, ∼2.3%), followed by position 7 (∼1.2%), position 57 (∼1.0%), position 51 (∼0.6%), and position 14 (0.5%). The risk of PZA resistance from the alleles at each SNP was estimated by calculating ORs and *P*-values using a GWAS approach. Additionally, adjusted OR (aOR) which accounted for the confounding effects of lineage were also

TABLE 3 | Mutations with extreme effects.

Mutational effects	Mutation	Mutation class	MAF (%)	OR	*P*-value	Distance to ligand (Å)	Stability ΔΔG	Ligand affinity
Highest OR	H51D	Drug-associated	0.30	414.61	4.49E-33	5.66	−2.2	−1.82
Most frequent mutation	Q10P	Drug-associated	2.11	156.23	1.28E-207	6.02	−0.63	−1.77
Most deStabilising for protein stability (DUET)	L4S	Drug-associated	0.25	28.46	5.63E-18	15.33	−3.87	−1.08
Closest destabilising for protein stability (DUET)	A134D	Others	0.007	2.43	1.00	3.05	−2.98	0.58
Most stabilising for protein stability (DUET)	Y103D	Others	0.22	142.33	1.24E-21	5.42	1.18	0.85
Closest stabilising for protein stability (DUET)	C138R	Drug-associated	0.09	116.96	6.09E-10	3.28	0.10	0.35
Most destabilising for ligand affinity	D49G	Drug-associated	0.05	29.16	0.0001	3.45	−1.16	−3.46
Closest destabilising for ligand affinity	D8G	Drug-associated	0.08	48.69	1.95E-07	3.22	−0.85	−3.45
Most stabilising for ligand affinity	G162D	Drug-associated	0.03	38.85	0.0008	8.32	−1.04	2.23
Closest stabilising for ligand affinity	A134P	Others	0.007	9.70	1.71E-01	3.05	−1.43	0.08
Most destabilising for protein stability (Foldx)	C72W	Drug-associated	0.01	19.41	0.03	7.05	27.46	–
Most stabilising for protein stability (Foldx)	A134P	Others	0.007	9.70	1.71E-01	3.05	−5.2	–

Mutations (nsSNPs) with extreme effects on odds ratio, frequency, thermodynamic stability, and ligand affinity. For ligand affinity, only mutations lying within 10 Å of PZA (pyrazinamide) were considered. nsSNPs, non-synonymous Single Nucleotide Polymorphisms; Å, Angstroms; MAF, minor allele frequency; OR, Odds Ratio; ΔΔG, Change in Gibbs free energy in Kcal/mol.

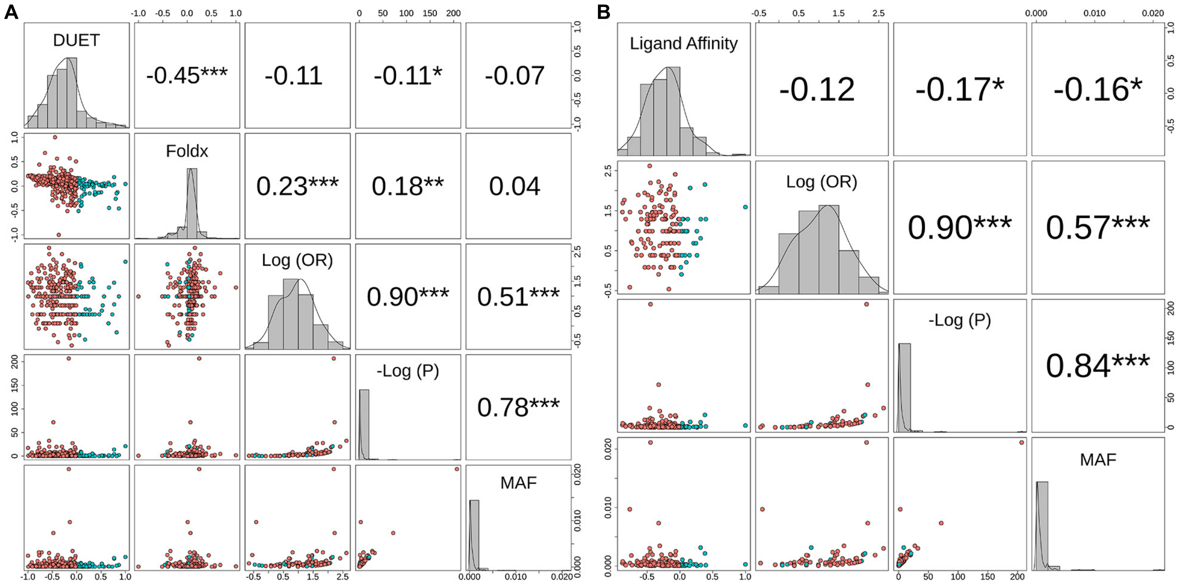


FIGURE 6 | Correlation between biophysical effects and GWAS measures of Odds Ratio (OR), *P*-values (P) and minor allele frequency (MAF). Pairwise correlations between MAF, negative log10 *P*-value [-Log(P)], Log10 (OR) and **(A)** Protein stability (DUET) and FoldX for 424 nsSNPs, **(B)** Ligand affinity of 201 nsSNPs (lying within 10 Å of PZA). The upper panel in both plots include the pairwise Spearman correlation values along with their statistical significance (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). The points in the lower panel represent nsSNPs, coloured according to respective stability effects: **(A)** nsSNPs with destabilising effect for DUET and ligand affinity are coloured red, while for FoldX these appear in blue, **(B)** nsSNPs with stabilising effect for DUET and ligand affinity appear in blue, while for FoldX these appear in red. The diagonal plots display the histogram of the corresponding parameter. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; Units for DUET, FoldX and Ligand Affinity (Kcal/mol).

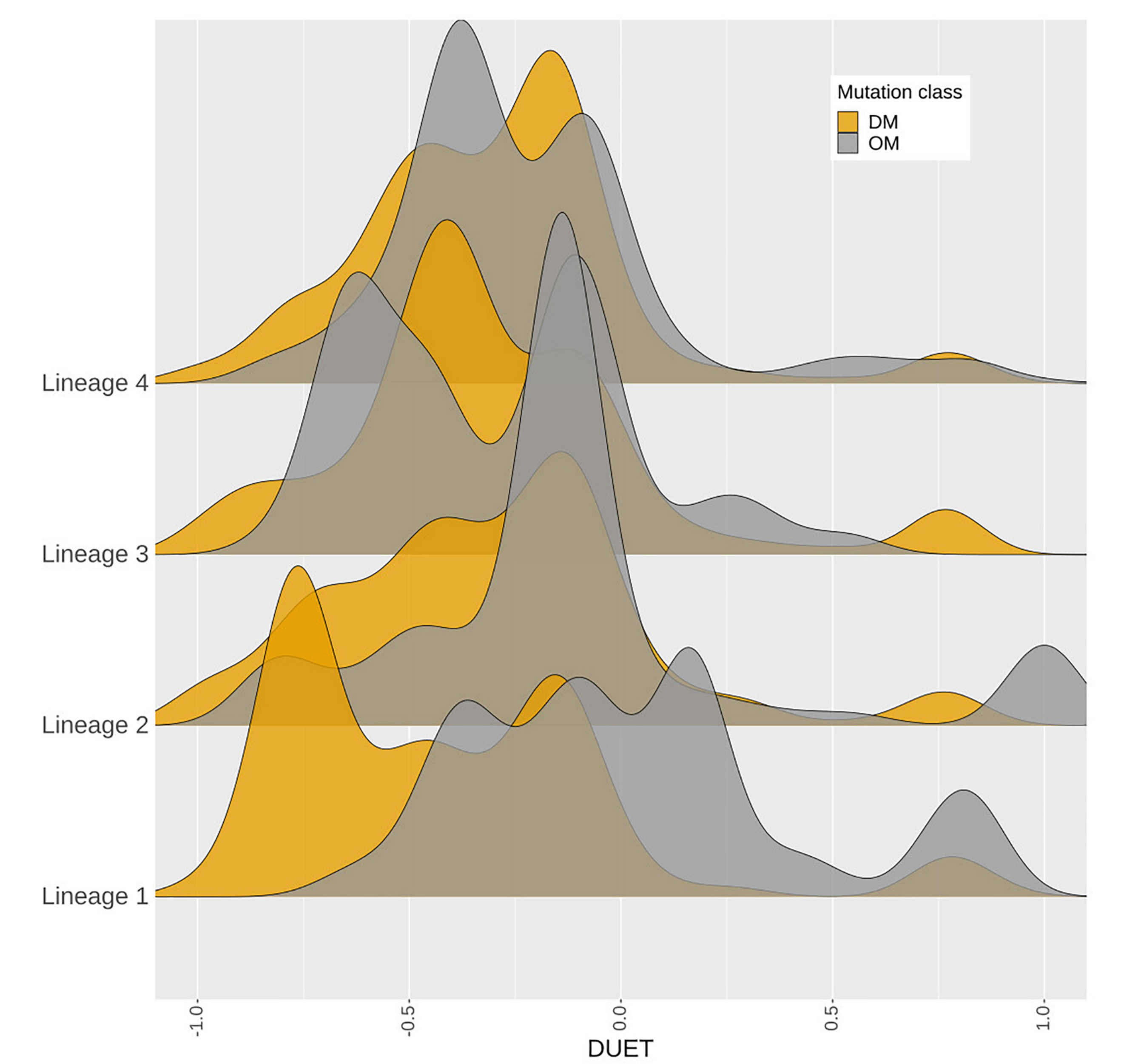


FIGURE 7 | Density distribution of *M. tuberculosis* lineages. A total of 4,433 samples belonging to Lineages 1–4, containing 419'pncA mutations were considered. The horizontal axis shows the DUET stability values (−1, most destabilising) to blue (+1, most stabilising). while the vertical axis shows the density distribution of *M. tuberculosis* lineages coloured by mutation class as either DM (associated with pyrazinamide resistance in orange) or OM (not associated with pyrazinamide drug resistance which appear in grey). DM mutations comprise of a total of 3,565 samples contributing to 226 mutations, while 868 samples contributing to 193 mutations formed part of the OM mutation class. The figure is generated using R statistical software (version 4.0.2). Abbreviations used: nsSNPs: non-synonymous Single Nucleotide Polymorphisms, pncA: pyrazinamidase.

analysed (**Supplementary Figure 9**). The majority of nsSNPs were linked to increased likelihood of being resistant to PZA (OR > 1). For unadjusted ORs, this was 96% (310/322), while for aOR, it was ∼75% (122/163). Wild type position 51 had the highest unadjusted OR (> 350, $P < 10^{-30}$), followed by positions 57, 120 (OR > 250, $P < 10^{-19}$), and subsequently by positions 10, 103, 68, 135, 138, 96, and 180 (OR > 100; $P < 10^{-10}$) (**Figure 1A**, **Supplementary Figure 8**, and **Supplementary Table 1**), with most of these positions being present in the metal binding and active sites.

When assessing sites in relation to mutational diversity, active site residues were among the highest, with residues H51, H57, H71, K96 associated with six distinct mutations, followed by F13, D49, W68, A134, C138 with five mutation each, while residues D8, Y103, H137 were associated with four distinct mutations and residues I133 associated with two distinct mutations (**Figure 1B**). The dominant effect of a highly frequent mutation (Q10P; MAF = 2.1%, *OR* = 156.23) in the population compared to two other mutations observed at the same position namely Q10R (MAF = 0.13%, *OR* = 83.01) and Q10H (MAF = 0.08%, *OR* = 107.17) (**Supplementary Table 1**), makes position 10 prominent in terms of MAF (**Supplementary Figure 8**) while sites involved in the catalytic activity and iron metal centre are more prominent with respect to SNP diversity (**Supplementary Figure 8**). These results suggest that mutations at these structurally and functionally important sites are likely under selective pressure exerted by the drug resulting in this observed mutational diversity.

The relationship between structural measures of stability and OR was visualised as a bubble plot indicating that mutations associated with greater resistance (high OR) tend not to have extreme effects (**Supplementary Figure 10**). Furthermore, this relationship along with MAF, OR, and *P*-values was assessed through Spearman correlations (**Figures 6A,B**). MAF was strongly correlated with *P*-values for all 424 mutations ($\rho = 0.78$, $P < 0.001$) and 201 mutations lying with 10 Å of PZA ($\rho = 0.84$, $P < 0.001$) (**Figures 6A,B**). As expected, OR and *P*-values were strongly correlated ($\rho = 0.9$, $P < 0.001$) for all 424 nsSNPs and 201 nsSNPs close to PZA binding site (**Figures 6A,B**). FoldX stability and DUET stability values showed moderate correlation ($\rho = 0.45$, $P < 0.001$). The negative sign for the DUET and FoldX associations is expected since stability changes measured by these tools have opposite signs (i.e., $\Delta\Delta G < 0$: destabilising in DUET vs. stabilising in FoldX). FoldX $\Delta\Delta G$ values showed weak but significant correlations with OR ($\rho = 0.23$, $P < 0.001$), and *P*-values ($\rho = 0.18$, $P < 0.01$) (**Figure 6A**), while DUET $\Delta\Delta G$ and ligand affinity showed weak and insignificant association with OR ($\rho = -0.1$, $P > 0.05$) (**Figures 1B**, **6A**), including adjusted OR (**Supplementary Figures 9A**, **8B**).

When considering aOR and its relationship with stability and other structural features [i.e., Accessible (ASA), Relative Surface Area (RSA), residue depth (RD), and hydrophobicity values (KD)], there was high correlation ($\rho > 0.6$, $P < 0.05$) with adjusted and unadjusted ORs (**Supplementary Figure 9A**). DUET $\Delta\Delta G$ showed moderate positive correlation between ASA and RSA ($\rho > 0.6$, $P < 0.05$), while moderately negative correlation with RD ($\rho \sim -0.5$, $P < 0.05$), and weak negative correlation with KD values ($\rho \sim -0.2$, $P < 0.05$) (**Supplementary Figure 9A**). The same structural features, however, did not demonstrate correlation with either FoldX $\Delta\Delta G$ (**Supplementary Figure 9A**) or ligand affinity (**Supplementary Figure 9B**).

Structural Differences in Drug Associated Mutations

Comparing stability effect (DUET and FoldX), ligand affinity, ligand distance, and other structural features (ASA, RSA, RD, KD) between mutations associated with PZA drug resistance (DM) and other mutations (OM), revealed statistically significant differences ($P < 0.05$) between all features except hydrophobicity values. The difference in structural features were most prominent when all 424 SNP mutations were considered ($P < 0.0001$) (**Figures 4A,B**) with lesser significance for ligand affinity ($P < 0.05$), ASA ($P < 0.01$), and RSA and RD ($P < 0.001$) values when 201 nsSNPs lying within 10 Å were considered (**Figure 4C**). Mutations associated with PZA resistance have lower DUET (**Figure 4A**, top left) but higher FoldX stability changes (**Figure 4B**, bottom left), and lower binding affinity (**Figure 4C**, second from bottom left) compared to OM. Additionally, it also appears that that while drug mutations need not necessarily occur at the hydrophobic sites (KD values, $P > 0.05$), they tend to lie buried indicated by higher RD values, and consequently lower surface area (ASA and RSA) compared to OM (**Figures 4A,B**).

Distinct Stability Profile for Drug Mutations and Lineage 1

A total of 419 nsSNPs are lineage specific (L1: 74; L2: 277; L3: 104; L4: 311). The greatest diversity of nsSNPs was observed in L3 (54.7%), followed by L1 (51.4%) and Lineage 2 (14.7%) with L4 showing the lowest diversity (14.1%) despite containing the highest number of samples (**Supplementary Figure 6**). Statistical analysis of the DUET $\Delta\Delta G$ distributions revealed significant differences between all lineages except between L3 and L4. Lineage differences for DUET $\Delta\Delta G$ were most prominent between L2 and L4 ($P < 0.0001$), followed by L1 and L4 ($P < 0.001$) (**Supplementary Table 2A**). Within each lineage, mutational distributions were significantly different between DM and OM mutation classes ($P < 0.0001$) except L3 (**Supplementary Table 2B**). Interestingly, a distinct stability profile was observed for DM mutations within L1. Mutations associated with drug resistance showed a marked peak around the extreme end (−0.75 DUET $\Delta\Delta G$) of the destabilising spectrum (**Figure 7**) within L1.

DISCUSSION

Genetic mutations including nsSNPs present within drug-targets and their activating genes are the main drivers of resistance development in TB (Schön et al., 2017). The motivation for investigating the missense mutations within the protein coding region only of the *pncA* gene was to enable understanding of the phenotypic mutational effects in relation to PZA resistance development. While the exact molecular mechanisms of PZA resistance are yet to be fully elucidated, the binding pocket of PZA and its key interactions are well known and characterised (Petrella et al., 2011; Ali et al., 2020; Sheik Amamuddy et al., 2020; Khan et al., 2021). This knowledge was used to guide the molecular docking of PZA to generate the pncA-PZA complex in the absence of an experimentally solved structure of the bound complex in Mtb. While docking generates a variety of ligand conformations (poses), choosing the "best" pose is based on considerations around key molecular interactions formed by the ligand, interaction energy of the docked complex and subject expertise. Using these guides, docking pose 1 was chosen due to its molecular interactions with known key residues and close alignment with previously published studies (Karmakar et al., 2018; Ali et al., 2020; Khan et al., 2021). In addition, we analysed the top two docking poses using the mCSM pipeline (**Supplementary Figure 3**). The resulting mutational effects on pncA stability and ligand affinity did not differ between poses indicating the small differences in pose did not affect downstream analysis. It also suggests that due to the small size of the PZA molecule, the orientation of the aromatic ring within the cavity may have more flexibility in its orientation and interaction with the neighbouring residues, but without drastically impacting the molecular interactions for global protomer stability and ligand binding affinity.

The molecular motion of pncA assessed by NMA was visualised to understand the mutational effects with regard to flexibility (**Supplementary Figure 1**). Sites displaying high mutational frequency or association with drug resistance mutations were not located in regions with high flexibility, with large molecular motions mainly restricted to the loop region 60–66. This suggests the molecular motion in pncA does not interfere with PZA binding as active site residues were not associated with high fluctuations.

Normal mode analysis shows large scale molecular motions. Molecular dynamics (MD) studies offer insights into the finer grained atomic motions and are an excellent way to investigate molecular mechanisms. However, these studies are computationally intensive and are difficult to scale for studying hundreds of mutations. A recent MD study on a subset of mutations found within our dataset analysed seven pncA nsSNPs (F94L, F94S, K96N, K96R, G97C, G97D, and G97S) showed that these destabilising mutations altered the binding pocket, allowing increased PZA flexibility (Khan et al., 2021). All seven mutations were associated with PZA resistance and also showed destabilising effects in our study. A similar study of destabilising mutations R123P, T76P, H7R associated with PZA resistance showed that the mechanism of resistance could be through increasing the flexibility of the region they are located in, thereby changing the binding pocket volume (Ali et al., 2020). Another MD study of mutations P54L and H57P showed that they decrease overall stability along with reduced ligand affinity leading to PZA resistance (Mehmood et al., 2019). All of these observations are concordant with our analysis.

Destabilising effects of nsSNPs are thought to be the main reason for impeding protein function through directly effecting protomer stability or ligand affinity. However, large stabilising effects can have an equally deleterious impact on protein function through rigidification, impeding flexibility and dynamic molecular motions. This has been implicated more generally within a disease context (Gerasimavicius et al., 2020) and more specifically in PZA resistance (Rajendran and Sethumadhavan, 2014). It offers an explanation for the observance of the stabilising mutation site 103. Drug associated mutations at this site (Y103C, Y103H, and Y103S) could result from the rigidification of the binding pocket leading to reduced binding affinity measured as destabilising PZA affinity.

Mutations within *pnca* are scattered along the entire gene length observed in studies (Stoffels et al., 2012; Miotto et al., 2014; Whitfield et al., 2015). While two other genes, *rpsA* and *panD* have also been linked to PZA resistance, a clear link between *rpsA* and PZA resistance is lacking (Shi et al., 2011; Alexander et al., 2012; Simons et al., 2013; Tan et al., 2014) although there is increasing evidence to support *panDs* association with PZA resistance (Pandey et al., 2016; Werngren et al., 2017; Gopal et al., 2020). In our analysis, there were only a few samples with *rpsA* and *panD* mutations, therefore limiting attempts at assessing their synergistic relationship with PZA resistance. Mutations within the *pncA* gene and its promoter remain the most common route to PZA resistance (Dookie et al., 2018) (Khan et al., 2019). Nearly 70% of the MDR isolates and 13% XDR isolates had nsSNPs in the *pncA* coding region. The burden of pncA mutations in the MDR and XDR isolates was lower in our analysis compared to 88.0% and ∼20% observed by Pang et al. (2017). In another study, 70% of the MDR isolates, and significantly higher i.e., 96% of XDR isolates harboured pncA mutations including nsSNPs (Allana et al., 2017). An alternative route to resistance for pncA as a non-essential gene encoding an enzyme that transforms a prodrug to drug would be by INDELs or mutations leading to premature stop codons resulting in the protein being degraded on translation. A recent report analysing the *pncAc.85_86insG* frameshift mutation using structural and biophysical analysis showed the mutation resulted in a truncated and incomplete protein lacking the active site pocket (Karmakar et al., 2018). Despite this obvious route to resistance, only 1% samples in our dataset showed INDELs and stop codons, compared to 13% of samples that showed missense point mutations in pncA. This is consistent with the knowledge that nsSNPs in pncA remain the major route to resistance for PZA (Khan et al., 2019).

Destabilising effects are considered detrimental to the downstream protein function (via disruption of drug affinity, nucleic acid affinity or overall complex stability) and are thus given higher consideration in classifying mutations (Wylie and Shakhnovich, 2011). In our analysis, around 85% of mutations were destabilising for overall protein stability as well as complex affinity. It is thought that the resistant phenotype is imparted either through affecting protein folding, instability of the PZase protein, prevention of coenzyme complex (Gopal et al., 2016) or loss of virulence factor synthesis (Gopal et al., 2016). Further, this is thought to come without a high bacterial fitness cost since pncA is primarily an activator of the PZA drug. This is similar to a recent observation reported in the *katG* gene (target for the anti-TB pro-drug, isoniazid) with a high proportion of destabilising mutations (Portelli et al., 2018). Also, a higher proportion 60% (n = 253) of SNP mutations showed electrostatic changes compared to ∼35% reported by Portelli et al. (2018). This likely due to the larger sample size of our dataset.

All active site residues appear to be under drug selection pressures due to multiple mutations (>2) associated with these with the exception of I133, considered to be an emerging or budding-resistance hotspot. In our analyses, there were 22 such sites while 83 sites within *pncA a*ssociated with > 2 nsSNPs linked to PZA drug resistance (categorised as DM). However mutations were not restricted to the active site, with less than 50% resistant variants lying within 10 Å of the active site of PZA, indicating the possible role of distal residues in resistance development (Portelli et al., 2018). Mutations associated with drug resistance tend to have lower stability, lie buried within the structure with lesser surface area as shown by Karmakar et al. (2020).

Our study compares results from two different computational stability predictors: mCSM and FoldX (Schymkowitz et al., 2005). Unsurprisingly, most mutations were found to have a destabilising effect (**Supplementary Figure 11**). FoldX reported ∼85% (vs. ∼80% estimated by DUET) nsSNPs with destabilising effect. The range for absolute $\Delta\Delta G$ values was greater for FoldX (median: 2.0; range: −5.2, 27.46) compared to DUET (median: −0.1; range: −3.9, 1.2). There was however, 77% agreement between FoldX and DUET outcomes (data not shown). Interestingly, drug associated mutations displayed higher FoldX $\Delta\Delta G$ predictions compared to mCSM-DUET $\Delta\Delta G$ predictions. A possible explanation for this is the differences in the underlying parameters the different methods use. FoldX constructs mutant structures by mutating the target residue and searching for the optimal conformation by iteratively altering the position of the neighbouring side chains. The stability of the mutant structure is estimated using an empirical force field made of several energy terms. This compares to DUET where estimates of the structural effects are based on differences between the wild-type environment and pharmacophore atomic changes resulting from the mutation, without the need to generate mutant structures. With this in mind, it appears that the DM mutations have larger local perturbations in the mutated region considered by FoldX, resulting in higher $\Delta\Delta G$ predictions compared to the lesser effects of surface area considered by DUET. Drug resistance mutations displaying smaller surface area compared to their susceptible counterparts were also observed in recent studies investigating nsSNPs in Mtb genes (Portelli et al., 2018; Karmakar et al., 2020) indicating the role of compensatory mutations, alleviating any fitness penalty in the development of the drug resistance phenotype. The extent of the contribution of surface area in these methods is reflected in the observation of moderate correlations between DUET and structural features, and the weaker associations between FoldX and structural features (**Supplementary Figure 9A**). Structural associations for ligand affinity were also observed to be weak (**Supplementary Figure 9B**) most likely due to the role of factors involved in short-range interactions (like Van der Waal's forces) not considered in our analysis. A similar view emerged in the recent study by Karmakar et al. (2020) where no significant differences were observed for PZA binding affinity.

It has been suggested that frequently occurring mutations may not confer extreme changes in biophysical stability measures, with mild stability effects offering local fitness advantages (Portelli et al., 2018). Our data presented us with the opportunity to test this theory empirically by assessing relationships of stability with GWAS measures of MAF, OR, and *P*-values. At a glance, it appears that mutations with high OR tend be less extreme in their impact on protein stability and ligand affinity (**Supplementary Figure 10**). However, we did not find any significant association with high frequency mutations and extreme changes in stability or affinity parameters (**Figure 6**). One possible explanation is that the fitness landscape is gene and function specific, optimised differently for genes directly coding for drug targets and for non-essential genes like *pncA*. Another major consideration is that resistance is often acquired through a stepwise ordinal accumulation of mutations (Woodford and Ellington, 2007; Ismail et al., 2019). The genetic background can dramatically influence fitness effects associated with mutations (Wong, 2017). Consequently, the mutational impact differs when occurring against a sequence background of extant resistant mutations, a phenomenon known as epistasis (Wong, 2017). Since resistance development is a balanced interplay between fitness effects and cost of resistance, epistasis warrants due consideration in efforts to understand and limit the evolution of multi-drug resistance.

The use of mCSM suite of tools has the advantage of studying global (protein stability) as well as local effects (ligand affinity, protein-protein interaction, and protein nucleic-acid interaction). Additionally, it also provides the methodological consistency for comparing molecular effects and benefits application of machine learning methods (ML) to explore greater mechanistic details. While computationally intensive, ML methods would benefit from using tools such as DynaMut (Rodrigues et al., 2018) which account for protein molecular motions when estimating mutational effect on protein stability. Additionally methods which consider anti-symmetric properties of mutational impact i.e., $\Delta\Delta G\ (A \rightarrow B) = -\Delta\Delta G\ (B \rightarrow A)$ like DeepDDG (Cao et al., 2019) and INPS-MD (Savojardo et al., 2016) have the potential to build robust predictive models and improve the "learning" capability of ML methods in the context of machine learning.

Mtb lineages have been associated with virulence, disease transmission, drug resistance, and clinical outcome (Ford et al., 2013; Reiling et al., 2013; Novais et al., 2017; Correa-Macedo et al., 2019; Oppong et al., 2019; McHenry et al., 2020). Lineage specific differences between lineages 2 and 4 have recently been noted in the development of TB drug resistance, especially related to MDR and XDR strains (Oppong et al., 2019). Our study highlighted the most significant differences between L2 and L4 with respect to protomer stability demonstrating the biophysical phenotypic manifestation of these underlying genotypic changes. The observance of a distinct peak for destabilising mutations related to drug resistance within L1 suggests that the extreme mutational consequences of such mutations in the "ancient" lineage 1 may be rapidly giving way to other "modern" *M. tuberculosis* lineages linked to MDR and XDR-TB and virulence.

Our study is based on a well-characterised clinical dataset sourced globally from over 35 K clinical isolates, and leverages the availability of robust metadata (lineage, geography, DST, etc.) for each isolate. We show that the framework used in our work allows us to investigate the interrelationships between genomic features from GWAS analysis and the biophysical measures of nsSNPs, helping to contextualise the underlying bacterial fitness and mutational landscape. The need to consider multiple stability predictors with different underlying principles to validate these associations has also been highlighted. Lineage associations of drug resistance, and their biophysical consequences, require further investigation and the functional characteristics of mutations should be validated in future experiments. We hope such a framework can be used to understand and inform therapeutic and stewardship efforts.

DATA AVAILABILITY STATEMENT

All pre-generated TB-profiler results were downloaded for all isolates from tbdr.lshtm.ac.uk/sra.

AUTHOR CONTRIBUTIONS

TT was responsible for the molecular docking, integrating genomics and structural data, data analysis, and writing the initial draft. JP made available and generated the genomics data results. CE provided the FoldX pipeline. TC and NF provided the overall supervision and contributed to revising and refining the manuscript. All authors contributed to the manuscript and approved the submitted version.

FUNDING

TT was supported by a BBSRC Ph.D. studentship (Grant No: BB/S507544/1). JP was funded by a Newton Institutional Links Grant (British Council. 261868591). CE was funded by the Medical Research Council UK (Grant No. MR/T000171/1). TC was funded by the Medical Research Council UK (Grant No. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant No. BB/R013063/1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.619403/full#supplementary-material

REFERENCES

Alexander, D. C., Ma, J. H., Guthrie, J. L., Blair, J., Chedore, P., and Jamieson, F. B. (2012). Gene sequencing for routine verification of pyrazinamide resistance in *Mycobacterium tuberculosis*: a role for pncA but Not rpsA. *J. Clin. Microbiol.* 50, 3726–3728. doi: 10.1128/jcm.00620-12

Ali, A., Khan, M. T., Khan, A., Ali, S., Chinnasamy, S., Akhtar, K., et al. (2020). Pyrazinamide resistance of novel mutations in pncA and their dynamic behavior. *RSC Adv.* 10, 35565–35573. doi: 10.1039/d0ra06072k

Allana, S., Shashkina, E., Mathema, B., Bablishvili, N., Tukvadze, N., Shah, N. S., et al. (2017). PncA gene mutations associated with pyrazinamide resistance in drug-resistant Tuberculosis, South Africa and Georgia. *Emerg. Infect. Dis.* 23, 491–495. doi: 10.3201/eid2303.161034

Al-Saeedi, M., and Al-Hajoj, S. (2017). Diversity and evolution of drug resistance mechanisms in *Mycobacterium tuberculosis*. *Infect. Drug Resist.* 10, 333–342. doi: 10.2147/idr.s144446

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Boonaiam, S., Chaiprasert, A., Prammananan, T., and Leechawengwongs, M. (2010). Genotypic analysis of genes associated with isoniazid and ethionamide resistance in MDR-TB isolates from Thailand. *Clin. Microbiol. Infect.* 16, 396–399. doi: 10.1111/j.1469-0691.2009.02838.x

Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J. Z. (2019). DeepDDG: predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model.* 59, 1508–1514. doi: 10.1021/acs.jcim.8b00697

Chakravarty, S., and Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* 7, 723–732. doi: 10.1016/s0969-2126(99)80097-5

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., et al. (2014). A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* 5:4812.

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 50, 307–316.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., et al. (2011). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* 44, 106–110. doi: 10.1038/ng.1038

Correa-Macedo, W., Cambri, G., and Schurr, E. (2019). The interplay of human and *Mycobacterium Tuberculosis* genomic variability. *Front. Genet.* 10:865.

de Vos, M., Müller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., et al. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. *Antimicrob. Agents Chemother.* 57, 827–832. doi: 10.1128/aac.01541-12

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., and Naidoo, K. (2018). Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* 73, 1138–1151. doi: 10.1093/jac/dkx506

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., et al. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* 45, 784–790. doi: 10.1038/ng.2656

Gerasimavicius, L., Liu, X., and Marsh, J. A. (2020). Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep.* 10:15387.

Gopal, P., Sarathy, J. P., Yee, M., Ragunathan, P., Shin, J., Bhushan, S., et al. (2020). Pyrazinamide triggers degradation of its target aspartate decarboxylase. *Nat. Commun.* 11:1661.

Gopal, P., Yee, M., Sarathy, J., Liang Low, J., Sarathy, J. P., Kaya, F., et al. (2016). Pyrazinamide resistance is caused by two distinct mechanisms: prevention of coenzyme a depletion and loss of virulence factor synthesis. *ACS Infect. Dis.* 2, 616–626. doi: 10.1021/acsinfecdis.6b00070

Ismail, N., Ismail, N. A., Omar, S. V., and Peters, R. P. H. (2019). In vitro study of stepwise acquisition of rv0678 and atpE mutations conferring bedaquiline resistance. *Antimicrob. Agents Chemother.* 63:e00292-19.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* 429, 365–371. doi: 10.1016/j.jmb.2016.12.004

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Karmakar, M., Globan, M., Fyfe, J. A. M., Stinear, T. P., Johnson, P. D. R., Holmes, N. E., et al. (2018). Analysis of a Novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.* 198, 541–544. doi: 10.1164/rccm.201712-2572le

Karmakar, M., Rodrigues, C. H. M., Horan, K., Denholm, J. T., and Ascher, D. B. (2020). Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep.* 10:1875.

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., et al. (2018). Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* 9:4306.

Khan, M. T., Malik, S. I., Ali, S., Masood, N., Nadeem, T., Khan, A. S., et al. (2019). Pyrazinamide resistance and mutations in pncA among isolates of *Mycobacterium tuberculosis* from Khyber Pakhtunkhwa, Pakistan. *BMC Infect. Dis.* 19:116.

Khan, T., Khan, A., Ali, S. S., Ali, S., and Wei, D. Q. (2021). A computational perspective on the dynamic behaviour of recurrent drug resistance mutations in the pncA gene from: *Mycobacterium tuberculosis*. *RSC Adv.* 11, 2476–2486. doi: 10.1039/d0ra09326b

McHenry, M. L., Bartlett, J., Igo, R. P., Wampande, E. M., Benchek, P., Mayanja-Kizza, H., et al. (2020). Interaction between host genes and *Mycobacterium tuberculosis* lineage can affect tuberculosis severity: evidence for coevolution? *PLoS Genet.* 16:e1008728. doi: 10.1371/journal.pgen.1008728

Mehmood, A., Khan, M. T., Kaushik, A. C., Khan, A. S., Irfan, M., and Wei, D.-Q. (2019). Structural dynamics behind clinical mutants of PncA-Asp12Ala, Pro54Leu, and His57Pro of *Mycobacterium tuberculosis* associated with Pyrazinamide resistance. *Front. Bioeng. Biotechnol.* 7:404.

Miotto, P., Cabibbe, A. M., Feuerriegel, S., Casali, N., Drobniewski, F., Rodionova, Y., et al. (2014). Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *MBio* 5:e001819-14.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*. 19, 1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Napier, G., Campino, S., Merid, Y., Abebe, M., Woldeamanuel, Y., Aseffa, A., et al. (2020). Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* 12:114.

Novais, E., Bastos, H., Machado, H., Sousa, J., Veiga, M. I., Ramos, A., et al. (2017). Tuberculosis severity and its association with pathogen phylogeny and properties. *Eur. Respir. J.* 50:A3046.

Oppong, Y. E. A., Phelan, J., Perdigão, J., Machado, D., Miranda, A., Portugal, I., et al. (2019). Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics* 20:252.

Pandey, B., Grover, S., Tyagi, C., Goyal, S., Jamal, S., Singh, A., et al. (2016). Molecular principles behind pyrazinamide resistance due to mutations in panD gene in *Mycobacterium tuberculosis*. *Gene* 581, 31–42. doi: 10.1016/j.gene.2016.01.024

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45, W229–W235.

Pang, Y., Zhu, D., Zheng, H., Shen, J., Hu, Y., Liu, J., et al. (2017). Prevalence and molecular characterization of pyrazinamide resistance among multidrug-resistant *Mycobacterium tuberculosis* isolates from Southern China. *BMC Infect. Dis.* 17:711.

Petrella, S., Gelus-Ziental, N., Maudry, A., Laurans, C., Boudjelloul, R., and Sougakoff, W. (2011). 3PL1: crystal structure of the pyrazinamidase of mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PLoS One* 6:e15785. doi: 10.1371/journal.pone.0015785

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera?a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Phelan, J., Coll, F., McNerney, R., Ascher, D. B., Pires, D. E. V., Furnham, N., et al. (2016). Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* 14:31.

Phelan, J., Lim, D. R., Mitarai, S., de Sessions, P. F., Tujan, M. A. A., Reyes, L. T., et al. (2019a). Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci. Rep.* 9:9305.

Phelan, J., O'Sullivan, D. M., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S., et al. (2019b). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 11:41.

Pires, D., and Ascher, D. B. (2016). mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.* 44, W469–W473.

Pires, D., and Ascher, D. B. (2017). mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res.* 45, W241–W246.

Pires, D., Ascher, D. B., and Blundell, T. L. (2014a). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* 42, W314–W319.

Pires, D., Ascher, D. B., and Blundell, T. L. (2014b). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691

Pires, D., Blundell, T. L., and Ascher, D. B. (2016). mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* 6:29575.

Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G., and Furnham, N. (2018). Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. *Sci. Rep.* 8:15356.

R Core Team (2014). *R: a Language and Environment for Statistical Computing*. Vienna: R Development Core Team.

Rajendran, V., and Sethumadhavan, R. (2014). Drug resistance mechanism of PncA in *Mycobacterium tuberculosis*. *J. Biomol. Struct. Dyn.* 32, 209–221. doi: 10.1080/07391102.2012.759885

Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., Ernst, M., et al. (2013). Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio* 4:e00250-13.

Rodrigues, C. H. M., Myung, Y., Pires, D. E. V., and Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* 47, W338–W344.

Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46, W350–W355.

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 32, 2542–2544. doi: 10.1093/bioinformatics/btw192

Schön, T., Miotto, P., Köser, C. U., Viveiros, M., Böttger, E., and Cambau, E. (2017). *Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin. Microbiol. Infect.* 23, 154–160. doi: 10.1016/j.cmi.2016.10.022

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388.

Segala, E., Sougakoff, W., Nevejans-Chauffour, A., Jarlier, V., and Petrella, S. (2012). New mutations in the Mycobacterial ATP synthase: new insights into the binding of the diarylquinoline TMC207 to the ATP $ynthase C-Ring $tructure. *Antimicrob. Agents Chemother.* 56, 2326–2334. doi: 10.1128/aac.06154-11

Sheik Amamuddy, O., Musyoka, T. M., Boateng, R. A., Zabo, S., and Tastan Bishop, Ö (2020). Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of *Mycobacterium tuberculosis* pyrazinamidase. *Comput. Struct. Biotechnol. J.* 18, 1103–1120. doi: 10.1016/j.csbj.2020.05.009

Shi, W., Zhang, X., Jiang, X., Yuan, H., Lee, J. S., Barry, C. E., et al. (2011). Pyrazinamide inhibits trans-translation in *Mycobacterium tuberculosis*. *Science* 333, 1630–1632. doi: 10.1126/science.1208813

Simons, S. O., Mulder, A., van Ingen, J., Boeree, M. J., and van Soolingen, D. (2013). Role of rpsA gene sequencing in diagnosis of pyrazinamide resistance: table 1. *J. Clin. Microbiol.* 51, 382–382. doi: 10.1128/jcm.02739-12

Singh, R. P., Pandey, N., Singh, A. K., Sinha, M., Kaur, P., Sharma, S., et al. (2011). Crystal structure of the complex of goat lactoperoxidase with Pyrazinamide at 2.1 A resolution. doi: 10.2210/pdb3R55/pdb

Somoskovi, A., Parsons, L. M., and Salfinger, M. (2001). The molecular basis of resistance to isoniazid, rifampin, and pyrazinamide in *Mycobacterium tuberculosis*. *Respir. Res.* 2, 164–168.

Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R., and Bifani, P. (2012). Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 56, 5186–5193. doi: 10.1128/aac.05385-11

Tan, Y., Hu, Z., Zhang, T., Cai, X., Kuang, H., Liu, Y., et al. (2014). Role of pncA and rpsA gene sequencing in detection of pyrazinamide resistance in *Mycobacterium tuberculosis* Isolates from Southern China. *J. Clin. Microbiol.* 52, 291–297. doi: 10.1128/jcm.01903-13

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 43, D364–D368.

Trott, O., and Olson, A. J. (2009). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461.

Werngren, J., Alm, E., and Mansjö, M. (2017). Non- pncA gene-mutated but pyrazinamide-resistant *Mycobacterium tuberculosis*: why is that? *J. Clin. Microbiol.* 55, 1920–1927. doi: 10.1128/jcm.02532-16

Whitfield, M. G., Soeters, H. M., Warren, R. M., York, T., Sampson, S. L., Streicher, E. M., et al. (2015). A global perspective on pyrazinamide resistance: systematic review and meta-analysis. *PLoS One* 10:e0133869. doi: 10.1371/journal.pone.0133869

Wong, A. (2017). Epistasis and the evolution of antimicrobial resistance. *Front. Microbiol.* 8:246.

Woodford, N., and Ellington, M. J. J. (2007). The emergence of antibiotic resistance by mutation. *Clin. Microbiol. Infect.* 13, 5–18. doi: 10.1111/j.1469-0691.2006.01492.x

World Health Organization [WHO] (2020). *WHO Report on TB 2020*. Geneva: WHO.

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). SDM-a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* 39, W215–W222.

Wylie, C. S., and Shakhnovich, E. I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses (in press). *Proc. Natl. Acad. Sci. U. S. A.* 108, 9916–9921. doi: 10.1073/pnas.1017572108

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 20 August 2021
doi: 10.3389/fmolb.2021.646288

Packpred: Predicting the Functional Effect of Missense Mutations

Kuan Pern Tan[1,2†], *Tejashree Rajaram Kanitkar*[3†], *Chee Keong Kwoh*[2] *and Mallur Srivatsan Madhusudhan*[3]*

[1] *Bioinformatics Institute, Singapore, Singapore,* [2] *School of Computer Engineering, Nanyang Technological University, Singapore, Singapore,* [3] *Indian Institute of Science Education and Research, Pune, India*

Predicting the functional consequences of single point mutations has relevance to protein function annotation and to clinical analysis/diagnosis. We developed and tested Packpred that makes use of a multi-body clique statistical potential in combination with a depth-dependent amino acid substitution matrix (FADHM) and positional Shannon entropy to predict the functional consequences of point mutations in proteins. Parameters were trained over a saturation mutagenesis data set of T4-lysozyme (1,966 mutations). The method was tested over another saturation mutagenesis data set (CcdB; 1,534 mutations) and the Missense3D data set (4,099 mutations). The performance of Packpred was compared against those of six other contemporary methods. With MCC values of 0.42, 0.47, and 0.36 on the training and testing data sets, respectively, Packpred outperforms all methods in all data sets, with the exception of marginally underperforming in comparison to FADHM in the CcdB data set. A meta server analysis was performed that chose best performing methods of wild-type amino acids and for wild-type mutant amino acid pairs. This led to an increase in the MCC value of 0.40 and 0.51 for the two meta predictors, respectively, on the Missense3D data set. We conjecture that it is possible to improve accuracy with better meta predictors as among the seven methods compared, at least one method or another is able to correctly predict ~99% of the data.

Keywords: missense mutation effect prediction, amino acid depth, local environment/clique, statistical potential, meta predictor

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom

Reviewed by:
Wim Vranken,
Vrije University Brussel, Belgium
Thiyagarajan S,
Institute of Bioinformatics and Applied
Biotechnology, India

****Correspondence:***
Mallur Srivatsan Madhusudhan
madhusudhan@iiserpune.ac.in

†These authors have contributed equally to this work

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *25 December 2020*
Accepted: *19 July 2021*
Published: *20 August 2021*

Citation:
Tan KP, Kanitkar TR, Kwoh CK and Madhusudhan MS (2021) Packpred: Predicting the Functional Effect of Missense Mutations.
Front. Mol. Biosci. 8:646288.
doi: 10.3389/fmolb.2021.646288

INTRODUCTION

Amino acid substitutions could affect protein stability, alter/impair its function, and possibly lead to disease conditions (Zhang et al., 2012). Several such single amino acid substitutions in proteins, also called missense mutations, are implicated in diseases such as cystic fibrosis, diabetes, cancer etc. (Roach et al., 2010; Stranger et al., 2011). Data from clinical studies and from large-scale projects such as the Human Genome Project (Craig Venter et al., 2001), the HapMap Project (Frazer et al., 2007), the Exome Sequencing Project, and the 1,000 Genomes Project (Altshuler et al., 2012) have unearthed such single amino acid mutations. It would be instrumental to have a fast and automated computational method to accurately predict the functional effect of these mutations. Such an exercise could also provide valuable insights into the development of personalized medicine.

Several computational methods predict the effect of missense mutations. The methods use sequence or structure information or a combination of the two. The sequence-based methods rely on previously known protein sequences and their characterizations deposited in databases. For example, in the SIFT method (Ng and Henikoff, 2003), mutational effect prediction is made based on a customized position-specific substitution matrix (PSSM), constructed using PSI-BLAST (Altschul et al., 1997) and MOTIF finder (Smith et al., 1990) to identify conserved local sequence regions. A majority of structure-based methods are based on machine learning algorithms. These methods employ different feature sets and machine learning architectures. For example, I-mutant2.0 (Capriotti et al., 2005) is trained on features such as pH, temperature, and mutation type using a support vector machine. AUTO-MUTE 2.0 (Masso and Vaisman, 2014) constructs a statistical contact potential with Delaunay tessellation and trains their models with additional attributes such as ordered identities of amino acids, pH, and temperature. PoPMuSiC-2.0 (Dehouck et al., 2009) uses a linear combination of 26 different statistical energy functions in an artificial neural network architecture. mCSM (Pires et al., 2014b) utilizes a graph metric to summarize physicochemical interactions within a cut-off distance as pattern signatures and trains them using a Gaussian process regression model. SDM (Pandurangan et al., 2017), which does not rely on machine learning, constructs an environment-specific amino acid substitution matrix based on observed substitutions in evolutionary time. DUET (Pires et al., 2014a) is a meta-algorithm that consolidates the methods of mCSM and SDM (Worth et al., 2011). Missense3D (Ittisoponpisan et al., 2019) is another structure-based method that uses 17 structural properties to predict the effect of the mutation. Dynamut2.0 (Rodrigues et al., 2020) uses normal mode analysis and graph-based signatures. Polyphen (Adzhubei et al., 2010) is a hybrid method that combines sequence and structural features to predict the effect of a mutation. It uses an improved version of PSSM, information from the Pfam database, and structural features such as accessible surface area and volume of an amino acid to make a prediction. SuSPect (Yates et al., 2014) is another hybrid-based method that uses PSSMs and Pfam domain profiles (Finn et al., 2014). It also includes information from protein–protein interaction networks and searches in the database for known functional annotations of a mutated position. Despite these various efforts and algorithms, the functional fate of point mutations remains a challenging problem.

A missense mutation could lead to functional instability by either disrupting its structure or by affecting its interaction interface and/or active sites without necessarily impacting its structure. A mutational effect predictor should hence take into account the effect of mutation on both overall structural stability and its functional relevance. In this study, we describe Packpred, which addresses both these aspects. For structural features, Packpred uses an environment-dependent multi-body statistical potential and a depth-dependent substitution matrix, FADHM. We had previously established that FADHM scores are useful in predicting the effects of point mutations (Farheen et al., 2017). The multi-body statistical potential considers the observed/expected ratio of cliques of residues. The greater the value of the ratio, the more energetically stable is the packing of amino acids in the residue clique. We further categorized these residue cliques based on their residue depths. Residue depth (Chakravarty and Varadarajan, 1999; Tan et al., 2011, Tan et al., 2013) measures the degree of burial and hence the solvation effect on amino acids. Depth has been shown to correlate well with the structural stability and free energy change of cavity-creating mutations in globular proteins (Chakravarty and Varadarajan, 1999; Tan et al., 2011). Our depth-based statistical potential hence assesses the effect of mutation on local packing stability. To capture the functional relevance of amino acids, we used residue position Shannon entropy from a multiple sequence alignment of homologs of the query sequence. By this, we exploit evolutionary information to quantify the degree of observed variation at the position of mutation. Usually, the lesser the variation, the greater is the functional importance of the residue.

MATERIALS AND METHODS

Data Sets

Statistical Potential Data Set

A set of 3,753 protein structures (**Supplementary Table S1**) obtained from the Protein Data Bank (PDB) (Berman et al., 2000) was used to construct the clique statistical potential. The structures in this set have a resolution of 2.5 Å or better, an R-free of 0.25 or better, and are nonredundant at 30% sequence identity. To account for atomic position fluctuations (protein dynamics) while considering amino acid cliques, 10 homology models were built using Modeller9.11 (Šali and Blundell, 1993) with the native protein serving as both target and template in a self-alignment. The "refine very slow" option was used to relax the molecular structures with the aim of maximizing atomic position flexibility. These homology models along with the native structure (i.e., 11 structures for each protein) were then used to build the statistical potential.

Saturation Mutagenesis Data Sets

Saturation mutagenesis data sets of two proteins, T4-lysozyme (Rennell et al., 1991) and controller of cell division or death B (CcdB) (Adkar et al., 2012), were used in this study. T4-lysozyme is a 164 amino acid residue protein with our reference structure being PDB: 2LZM, which was solved at a resolution of 1.7 Å (Weaver and Matthews, 1987). Each position except the first was mutated to 13 other amino acids (A, C, E, F, G, H, K, L, P, Q, R, S, and T). After excluding key catalytic site residues (D10, E11, R145, and R148P), the data set consists of 1,966 mutations. CcdB is a cytotoxin (an inhibitor of DNA gyrase) with 101 amino acids. Its native structure was solved at a resolution of 1.4 Å [PDB: 3VUB (Loris et al., 1999)]. Full saturation mutagenesis (mutating each position to all other 19 amino acids) was performed at all positions of the protein. After removal of active site residues (I24, I25, N95, F98, W99, G100, and I101), a final set of 1,534 mutations was obtained. In both saturation mutagenesis experiments, an assessment was made on the phenotypic effect for each mutation. For T4-lysozyme, the phenotypic effect was gauged based on the plaque-forming ability of the mutant. Subject to the same experimental condition, a mutant is assigned to one of the four levels of sensitivity if the size of the plaque is (1) similar to native control, (2) significantly smaller, (3) with hazy morphology or difficulty in discerning plaques, and (4) no plaque formation (Rennell et al., 1991). For CcdB, the mutational sensitivity score

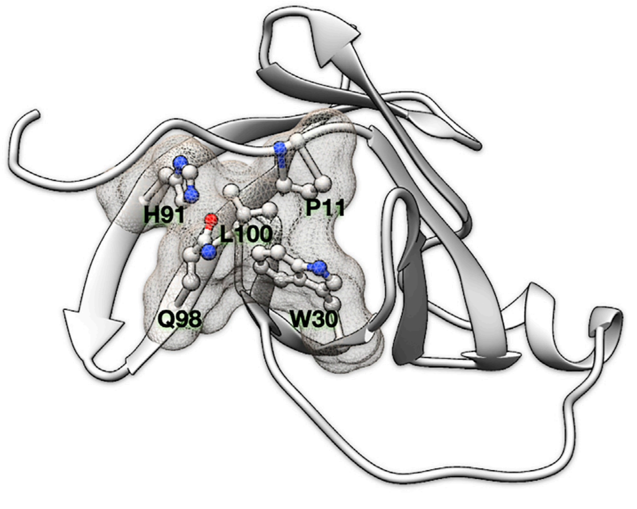


FIGURE 1 | Residue clique of amino acids. A 5-residue clique (P11, W30, H91, Q98, and L100) of cut-off 7.5 Å shown in ball and stick representation and enveloped with a meshed molecular surface from human recombinant MTCP-1 protein (PDB: 1A1X).

was quantitatively defined as the titer number at which the protein activity (in this case, inducing cell death) decreases by 5-fold or becomes more relative to its previous dilution. Values of mutational sensitivity range from 2 to 9 in CcdB, and we scaled the T4-lysozyme values to range from 2 to 5. For both data sets, a mutation is regarded as neutral if there is no perceptible phenotypic difference as compared to its native sequence (MS score = 2 in CcdB and T4-lysozyme) and is regarded as destabilizing otherwise.

Missense3D Data Set

The Missense3D data set consists of 4,099 mutations from 606 proteins extracted from Humsavar (Bateman et al., 2017), ClinVar (Landrum et al., 2014), and ExAC (Karczewski et al., 2017) (Ittisoponpisan et al., 2019). Humsavar lists all the annotated missense variants from humans reported in UniProt and SwissProtKB. ClinVar catalogs variations in humans and their associated phenotypes. ExAC is an exome aggregation consortium that describes the aggregation and analysis of human exomes. The analysis includes quantification of the pathogenecity of variants. The data set of 4,099 mutations consists of 1,965 disease-associated variants and 2,134 neutral variants (not associated with any known disease yet). Packpred parameters were trained on the T4-lysozyme data set and tested on the CcdB and Missesense3D data sets.

Structural and Sequential Features

Residue Depth

Depth is defined as the distance of a protein atom to the nearest bulk water molecule (Chakravarty and Varadarajan, 1999). The quantity measures the degree of burial of the atom. Depth has been shown to be capable of concisely describing the protein environment, as substantiated by its utilities in protein design and function predictions (Tan et al., 2011, Tan et al., 2013; Farheen et al., 2017). Atom depth values were computed using default parameters. The depth of a residue clique is defined as the average depths of its constituent atoms.

Cliques of Amino Acid Residues

A clique is defined as a sub-graph in which all possible pairs of vertices are linked. We define a (N, d_{cut}) "residue clique" to be a clique of N amino acids within a linkage distance of d_{cut}. We consider two amino acids as linked when at least four or more than half of the side chain non-hydrogen atoms (whichever are smaller) are within d_{cut} from atoms of another amino acid (**Figure 1**). For glycine, the C^{α} atom is used in lieu of the side chain. Residue cliques defined with different combinations of N and d_{cut} (N ranges from 2 to 4 and d_{cut} ranges from 7.0 to 10.5 Å in step of 0.5 Å) have been computed and investigated in this study.

Statistical Potential and Residue Clique Score

A residue clique statistical potential is constructed by adopting the formulation of Sippl's potential of mean force (Sippl, 1990),

$$E^{c} = -kT \log\left(\frac{\left(P^{c}_{obs} + \alpha P^{c}_{\exp}\right)}{\left(P^{c}_{\exp} + \alpha P^{c}_{\exp}\right)}\right), \quad (1)$$

where E^c is the pseudo potential energy and c is a residue clique of type $\{r_1, r_2, \ldots\}$, where the r_i's are the amino acid types; P^{c}_{obs} is the observed number of residue clique c; P^{c}_{exp} is its expected number in a hypothetical reference state without energetic interactions; α is the ratio of pseudo-count introduced to account for sparse statistics and is taken as 0.00 in our study. $-kT$ is a constant and is assumed as one in this study.

For each (N, d_{cut}) clique, the statistical potential is built at five different levels of depth (2.80 – 5.25 Å, 4.25 – 6.25 Å, 5.25 – 7.25 Å, 6.25 – 8.25 Å, and 7.25 Å–∞). To calculate the score of a residue clique (S), the mean μ and standard deviation σ of its depth are first computed. A Gaussian probability density function $N(x\,|\,\mu,\sigma)$ is then accordingly built. The clique score is computed as the weighted sum of integrands at every depth level as follows:

$$S^{c}_{\mu,\sigma} = \sum_{d\in D} \frac{1}{d_f - d_i} \int_{x=d_i}^{x=d_f} E^{c}_{d} \cdot N\left(x \vee \mu,\sigma\right) dx, \quad (2)$$

where d is one depth level, and d_i and d_f are the lower and upper bounds of the level.

Most residue cliques in a protein are overlapping with one another, and an amino acid residue can participate in multiple cliques. The score of a residue is taken as the average of all such cliques (refer to **Supplementary Text S1** for example). The score of a protein is further taken as the average of all its residue scores.

Shannon Entropy

Shannon entropy (H) is a measure of variation observed at a given position. It is calculated from a multiple sequence alignment obtained by a PSI-BLAST search against the uniref50 database (Altschul et al., 1997). H for a given position is then calculated as follows:

$$H = -\sum_{\mathrm{i}=1}^{20} \mathrm{P_i} \log_2 \mathrm{P_i}, \tag{3}$$

where Pi is the fraction of amino acid i observed at a given position.

FADHM Scores

FADHM scores are depth-dependent pairwise amino acid substitution likelihood scores extracted from the FADHM matrices. The FADHM matrices quantify the substitution frequencies at different depths obtained by performing protein–protein structural alignments. A detailed account of the FADHM score can be found elsewhere (Farheen et al., 2017)

The Packpred Score for Mutations

The Packpred score is given as follows:

$$PS = 1.5\,(S) + 1.75\,(H) + 0.5\,(FADHM), \tag{4}$$

where PS is the Packpred score, S is the residue clique score obtained from the statistical potential, H is the Shannon entropy, and FADHM is the depth-based amino acid substitution likelihood score. The weights were obtained by training on the T4 saturation mutagenesis data set (**Supplementary Table S2**). The coefficients for S, H, and FADHM (weights) were systematically sampled in the range 0–3 with a step size of 0.25. The cut-off score threshold that best discriminates neutral mutations from destabilizing ones was 1.6 in the training data (see *Training and Testing Packpred Score*). Mutation with a score greater than 1.6 is neutral and is destabilizing otherwise. To score a mutant, we modify the clique composition without explicitly modeling the mutant protein structure, with the mutant amino acid inheriting all the properties of the wild-type residue.

Packpred is implemented as a web server at http://cospi.iiserpune.ac.in/packpred/. A standalone version is also available for download.

Matthews's Correlation Coefficient

We gauge the binary classification performance of Packpred using Matthews's correlation coefficient (MCC) (Matthews, 1975), which is given as follows:

$$MCC = \frac{TP \cdot TN\text{–}\mathrm{FP} \cdot \mathrm{FN}}{\sqrt{[TP + FP]\,[TP + FN]\,[TN + FP]\,[TN + FN]}}, \tag{5}$$

where TP, TN, FP, and FN represent true-positive, true-negative, false-positive, and false-negative predictions.

RESULTS

Training and Testing Packpred Score

Packpred uses a linear combination of sequence position Shannon entropy, a residue clique statistical potential, and a depth-dependent substitution matrix (FADHM) to predict the functional effect of missense mutations. The Shannon entropy part of the score estimates the functional importance of residues based on evolutionary information. The clique statistical potential and the substitution matrix gauge the effect of the mutation on the local environment/structure. The statistical potential computes the observed and expected probabilities to calculate a score for a clique. The FADHM scores are taken from substitution matrices that are derived from structural alignments of proteins. The substitution likelihood scores are calculated by categorizing a protein in three regions based on residue depths (exposed, intermediate, and buried). The substitution scores indicate the likelihood of a residue getting replaced by another at a given depth.

We performed a grid search in the range of 0–3 with a step size of 0.25 for S, H, and FADHM to optimize the coefficients (weights) of each component of the linear combination Packpred score. The optimization was to maximize Matthews's correlation coefficient (MCC) (see below) of the T4 lysozyme saturation mutagenesis training data set. The weights that gave the highest MCC on the training set were 1.5, 1.75, and 0.5 for the clique statistical potential, Shannon entropy, and FADHM, respectively. We also obtained a cut-off threshold that distinguishes the destabilizing from the neutral ones from this training exercise. The cut-off was sampled in the range of 0–2 with a step size of 0.1. Mutations with scores greater than 1.6 are classified as neutral, and scores below 1.6 are classified as destabilizing. The T4-lysozyme training set consists of 1,362 (~69%) neutral and 604 (31%) destabilizing mutations, of which Packpred correctly identifies 1,049 (~77%) neutral mutations and 406 (~67%) destabilizing mutations (**Supplementary Table S3**). In the T4 training exercise, we observe similar MCC values for different combinations of weights of the grid search. Although the MCCs are similar, the underlying predictions and the linear combination scores are different (refer to **Supplementary Text S2** for an example).

The weights and threshold obtained from the training set were applied to two testing sets, the CcdB saturation mutagenesis data set (**Supplementary Table S4**) and the Missense3D data set (**Supplementary Table S5**). The CcdB data set has 1,258 (~80%) neutral mutations and 276 (~20%) destabilizing, while the Missense3D data set has 2,134 (~52%) neutral and 1965 (~48%) disease mutations, respectively. We used the PDB structures 2LZM and 3VUB to obtain Packpred scores of T4-lysozyme and CcdB, respectively. The biological unit of CcdB is a dimer, and we did all the calculations using this dimeric state structure for CcdB. Packpred correctly predicts 864/1,258 (~68%) neutral and 253/276 (~92%) destabilizing mutations from the CcdB testing set and 1,670/2,134 (~78%) neutral and 1,123/1965 (~57%) disease-causing mutations from the Missense3D data set.

We compared Packpred's binary classification with several popular methods such as i-mutant2 (Capriotti et al., 2005), mCSM(Pires et al., 2014b), SDM(Pandurangan et al., 2017), dynamut2 (Rodrigues et al., 2020), FADHM(Farheen et al., 2017), and Missense3D (Ittisoponpisan et al., 2019) (**Table 1**). All the predictions were made using default parameters. Packpred was the best performing method on the T4-lysozyme training set and the Missense3D testing set, with MCC values of 0.42 and 0.36, respectively. The next best method is Missense3D with MCC values of 0.40 and 0.33 for the T4 and Missense3D data sets,

TABLE 1 | Performance of some methods on T4, CcdB saturation mutagenesis, and Missense3D data sets.

Method	MCC for T4-lysozyme saturation mutagenesis data set	MCC for CcdB saturation mutagenesis data set	MCC for Missense3D data set
i-mutant 2.0	0.30[a]	0.36[a]	0.06
mCSM	0.22[a]	0.39[a]	0.05
SDM2	0.24[a]	0.33[a]	0.14
Dynamut2	0.09	0.15	0.06
Missense3D	0.40	0.39	0.33
FADHM	0.38[a]	**0.48**[a]	0.27
Packpred	**0.42**	0.47	**0.36**

[a]Values taken from FADHM article.
The best MCC values are in bold.

TABLE 2 | Prediction performance of seven methods on the Missense3D data set. The best score in each assessment metric is shown in bold font. Values of Class1 are used to describe the results in the manuscript.

Metric	Packpred	FADHM	Missense3D	Dynamut2.0	mCSM	i-mutant	SDM
MCC	**0.36**	0.27	0.33	0.06	0.05	0.06	0.14
Sensitivity (Class 0)	0.57	0.39	0.40	0.84	**0.92**	**0.92**	0.80
Specificity (Class 0)	0.78	0.85	**0.89**	0.20	0.10	0.12	0.34
Precision (Class 0)	0.71	0.71	**0.76**	0.49	0.49	0.49	0.52
F1 (Class 0)	0.63	0.50	0.53	0.62	**0.64**	**0.64**	0.63
Sensitivity (Class 1)	0.78	0.85	**0.89**	0.20	0.10	0.12	0.34
Specificity (Class 1)	0.57	0.39	0.40	0.84	**0.92**	**0.92**	0.80
Precision (Class 1)	**0.66**	0.60	0.62	0.59	0.59	0.60	0.63
F1 (Class 1)	0.72	0.70	**0.73**	0.31	0.18	0.20	0.44
Accuracy	**0.68**	0.62	0.65	0.51	0.50	0.50	0.55

respectively. The MCC of Packpred on the CcdB data set is 0.47 and is marginally outperformed by the best performing method, FADHM, which has an MCC of 0.48 (**Table 1**).

The clique potential and FADHM were earlier trained on 3,753 and 2,384 PDB entries, respectively. 89 of these PDBs are common to the 606 PDB entries that comprise the Missense3D testing set (**Supplementary Table S6**). These 89 overlapping entries include not just those that are identical but also those that are homologs (with sequence identities of 30% or greater). The overlapping PDBs account for 463 of 4,099 mutations in the Missense3D data set. Omitting these 463 mutations and using the other 3,636 mutations resulted in an MCC of ~0.37, comparable to the value of 0.36 obtained over the entire Missense3D data set of 4,099 mutations.

Analysis of the Predictions on the Missense3D Data Set

The Missense3D data set has a balanced representation of ~48% disease-associated mutations and ~52% neutral mutations. The data set, however, is skewed in terms of amino acid abundance when compared to natural abundance (**Supplementary Figure S1**). For instance, arginine has the highest representation and accounts for ~16% (664/4,099) of the Missense3D data set, while its natural abundance is ~5%. The next most abundant amino acid in the Missense3D data set is glycine, which accounts for ~9% (372/4,099) of the data (natural abundance is ~7%). The most frequent mutant is also arginine (347/4,099), followed by serine (343/4,099). There are 2,233 mutations in the exposed environment (depth less than 5 Å), 1,258 in the intermediate environment (depth between 5 and 8 Å), and 608 in the buried environment (depth greater than 8 Å).

We assessed the performance of various methods on the Missense3D data set using metrics including sensitivity, specificity, precision, accuracy, and F1 (**Table 2**). Packpred outperforms all other methods in MCC, precision, and accuracy. Missense3D has the highest sensitivity and F1. Packpred has less sensitivity than FADHM and Missense3D, indicating potential for improvement. Packpred has a specificity of 0.57, indicating a higher number of false-positive predictions. mCSM and i-mutant outperform all other methods in specificity. However, mCSM, i-mutant, SDM, and dynamute predict a large number of false negatives (**Table 3**) that affect their MCC. Hence, we compare Packpred with FADHM and Missense3D in the next sections unless otherwise stated. Packpred has fewer false positives among FADHM and Missense3D and has the highest number of false negatives. The high false-positive rate contributes to its lower specificity.

We analyzed the results structurewise (**Supplementary Table S7**). Packpred correctly predicted all mutations from 264 (out of 606) structures and at least 50% mutations correctly from 507 structures. It could not correctly predict any mutation from 56 structures. In these 56 PDBs, the maximum mutations in any one protein were four, while the average number of mutations per PDB in the whole set was ~6. These 56 structures did not follow any particular discernible pattern or trait.

Packpred has limitations in several areas. One of which is its high number of false-positive predictions that also affects its

TABLE 3 | Confusion matrix values for the different prediction methods. The values in bold font show the best in each category. TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

Metric	Packpred	FADHM	Missense3D	Dynamut2.0	mCSM	i-mutant	SDM
TP	1,670	1816	**1890**	440	229	251	713
FP	842	1,203	1,177	312	**158**	164	420
TN	1,123	762	788	1,650	**1804**	1798	1,542
FN	464	318	**244**	1,685	1896	1874	1,412

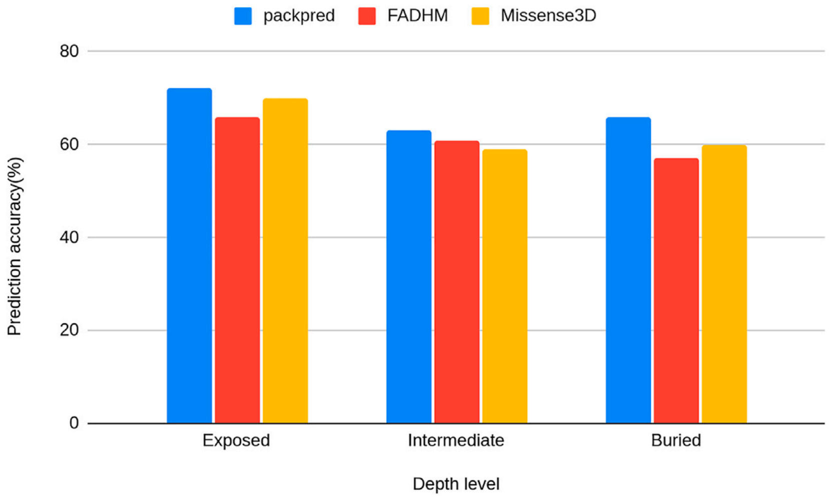


FIGURE 2 | Histograms of the prediction accuracy of Packpred, FADHM, and Missense3D at different depth levels (exposed to the solvent, intermediate, and buried).

specificity. Other methods have a higher specificity but underperform in their sensitivity by overpredicting true negatives. Packpred has fewer true positives than FADHM and Missense3D, indicating another potential area for improvement. With more true positives, it is likely that Packpred's F1 value would also improve, which is currently bested by Missense3D. Packpred has scores higher than 0.65 in all other metrics (accuracy, precision, sensitivity, and F1), indicating its overall balanced performance. We also calculated MCC (**Supplementary Table S8**) for each native amino acid type from the Missense3D data set. We found that of all 20 types of amino acids, Packpred has the highest MCC of 0.40 for Ile, Leu, and Val amino acids and the lowest MCC for Cys with an MCC of 0.17. Similar to Packpred, FADHM also has the lowest MCC of 0.04 for Cys amongst all the amino acid types. FADHM has the best MCC of 0.47 for I, which also happens to be the single best MCC for an amino acid among other methods. Missense3D, in contrast to Packpred and FADHM, has the best prediction for Cys with an MCC of 0.43 and has the lowest MCC of 0.02 for Trp among other amino acid types. These results show us amino acid–wise prediction performances and possibly contain useful hints on where one could improve the method.

We stratified the Missense3D data to particular depth zones to assess the performance of these methods at particular depths. Packpred has 597/2,233 (~72%) correct predictions from the exposed environment, 796/1,258 (~63%) from the intermediate, and 400/608 (~66%) from the buried environment. Packpred is the least accurate in predicting the effect of mutations in the intermediate environment. Interestingly, Missense3D is also the least accurate in this intermediate zone (**Figure 2**).

Meta Predictions

Of the 4,099 mutants, at least one of the seven methods we tested made an accurate prediction in 4,036 cases. This motivated us to make two different meta predictions by combining the different methods.

The first meta prediction makes use of the method that performs the best for particular amino acids. We studied the wild-type (native) amino acid–wise trends of all seven methods. For instance, native amino acids N, K, Q, R, and T are best predicted by Missense3D, FADHM outperforms other methods in the prediction of I and M amino acids, and Packpred is the best at predicting A, D, E, G, L, P, V, and Y. In fact, all seven methods feature as the best method for at least one amino acid (**Supplementary Table S9**). Interestingly, we found that Packpred has the highest percentage (68%) of correct predictions when averaged over the 20 amino acids and with the lowest standard deviation (4%). In contrast, FADHM and Missense3D have averages of 62 and 64% with standard

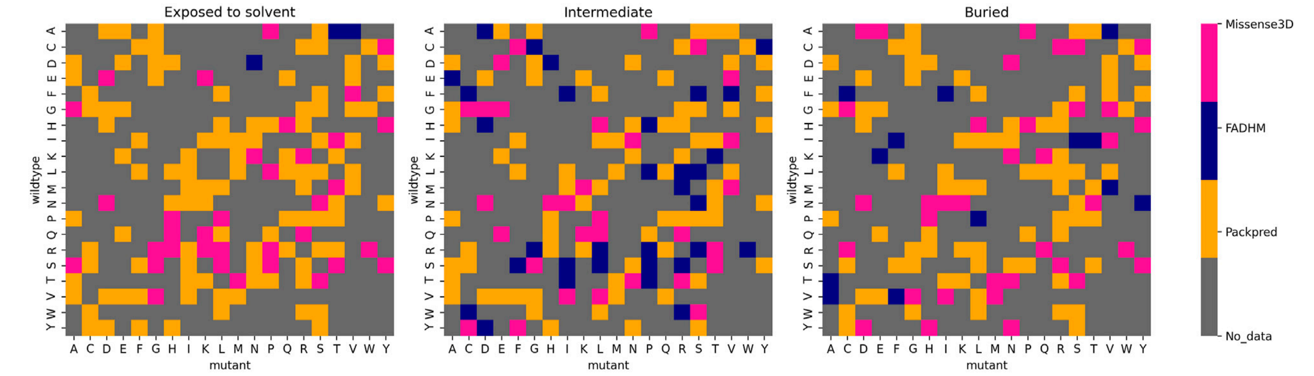


FIGURE 3 | Best performing methods for each wild-type mutant amino acid pair at different depth levels.

deviations of 7 and 10%, respectively. The other methods all have averages less than 60% with standard deviations between 11 and 14% (**Supplementary Table S9**). Packpred predictions are consistently well performing across the different native amino acid types. We then used these prediction strengths of each of the methods to get a hypothetical hybrid/meta prediction scheme (**Supplementary Table S10**) that combines predictions from all of the methods and has an MCC of 0.40 over the Missense3D data set, easily outperforming all the individual methods.

The second hypothetical meta prediction only involves Packpred, FADHM, and Missense3D as these were the methods that did consistently well over all different data sets and amino acids. Here, we considered the method that best predicted wild-type mutant pairs. Furthermore, we segregated these amino acid pairs into different depth categories—exposed to the solvent (depth <5 Å), intermediate (depth between 5 and 8 Å), and buried (depth >8 Å). Our meta prediction then chose the best performing method for a particular pair at a particular depth level. For instance, the wild-type mutant pair A→D, Packpred has the best predictions in an exposed environment, FADHM in the intermediate environment, and Missense3D in the buried environment (**Figure 3**). In case of a tie between methods, the one with the better MCC was chosen. By thus combining the strengths of the three methods, the MCC of the predictions rises to 0.51 for the Missense3D data set (**Supplementary Table S11**). An analysis to rationalize/explain why certain methods are best for certain pairs/environments did not yield any illuminating results. It is clear, however, that there is some degree of complementarity in these different methods, and perhaps a more rigorous treatment of the results from the individual methods could further improve prediction accuracy.

We would like to emphasize here that the purpose of exploring these meta predictions was to simply test the extent to which we could possibly improve results with such an approach. In a more rigorous implementation of this method, we would have to train and test the meta-predictor separately, something that is beyond the scope of this study. Choosing the best results from our testing set, as we have done here, merely represents the possible limit up to which we could improve on predictions.

Rank Ordering the Degree of Phenotypic Change by Mutations

We wanted to investigate if the Packpred scores are indicative of the degree of change/disruption caused by a mutation. The degree of change is measured experimentally using the mutational sensitivity score, which categorizes each mutation into one of four and eight levels in T4-lysozyme and CcdB data sets, respectively. We chose to use Spearman's rank correlation coefficient (SCC) to measure the performance of rank-ordering, as it makes no assumption on a linear relationship between the scores and the phenotypical change. SCC is calculated as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{6}$$

where d is the difference between the actual and the predicted ranks of a mutation, and n is the number of levels. The SCC for T4 and CcdB data sets is −0.48 and −0.54, respectively. At best, this correlation is weak and indicates that these scores could be further improved.

Assessing Robustness of Packpred

Last, we assessed the robustness of Packpred. For this, we changed the training set to include only 149 point mutations that result from a single nucleotide change in codons. The Missense3D data set is made of only these 149 different mutations. We created three additional training sets that all contain instances of only these 149 mutations. The first contains mutations from only the T4 lysozyme data set, the second set contains mutations from T4 in a 50:50 ratio of neutral-to-deleterious mutations, and the third set has mutations from the T4 and CcdB data sets in a 50:50 ratio of neutral-to-deleterious (**Supplementary Table S12**). The ratio was chosen based on the neutral-to-deleterious ratio of the Missense3D test set. For every combination of the training set, we obtained different optimal weights for the features of the linear combination (**Supplementary Table S12**). Interestingly, the accuracy of the method as gauged by the MCC value over the Missense 3D data set was consistently between 0.34 and 0.35 (**Supplementary Tables S13–15**).

DISCUSSIONS

In this study, we have developed a method to predict the effect of missense mutations on the structure and function of a protein. We believe that such predictions could be tested by assaying the protein for its function. Our method, Packpred, is constructed in a way that it is sensitive to structural changes effected by the mutation and any functional changes it may effect without perturbing the structures. To assess the impact of the mutation on the structure (and hence the function) of the protein, we devised a multi-body clique statistical potential. This statistical potential evaluates the strength of the interaction in a local neighborhood (amino acid clique). To assess the impact of mutation, we consider the same residue neighborhood environment while replacing the wild-type amino acid with the mutant. The score of the clique with the wild-type residue and with the mutant are then computed. An inferior score for the mutant in comparison to the wild type would be indicative of a destabilizing mutation. The structural stability of introducing the mutant residue is also gauged using a depth-dependent substitution matrix, FADHM, whose efficacy at detecting the fate of mutations we had previously benchmarked and tested. To account for functional changes that could happen even when the structure is not affected by the mutation, we invoke evolutionary information from a multiple sequence alignment using Shannon entropy. The more conserved the position, the more likely that it is going to affect function. These different scores are taken together in a linear combination, whose coefficients were optimized using the T4-lysozyme saturation mutagenesis data set of ~2,000 mutations. Packpred was tested on two different data sets, another saturation mutagenesis data set (CcdB) and the Missense3D data set. Its performance on these data sets was also compared to those of six other methods including FADHM, Missense3D, Dynamut2.0, mCSM, i-mutant2.0, and SDM. With the exception of the CcdB data set, where it marginally underperforms FADHM, Packpred clearly outperformed all other methods on all data sets. Among the methods, Packpred balances well between predicting true positives and true negatives (neutral and disease-causing mutations) and hence has the best MCC values. Packpred has the best accuracy and is close to the best specificity, precision, and F1. It loses out to the best methods in these measures and on sensitivity as methods such as mCSM predict a disproportionately large number of negatives. When the performance of the different methods is compared on an (wild-type) amino acid by amino acid basis, Packpred performs consistently well, with prediction accuracies never falling below 60%, while maintaining an average of 68%, which is easily the best among the methods tested. Qualitatively, a similar picture also emerges when the results are broken down into wild-type mutant amino acid pairs.

We also investigated whether Packpred (and other methods) preferred certain types of structures over others. No clear deduction could be made from these analyses. However, there was one trend that could be considered for further improvements—Packpred, similar to Missense3D and FADHM, performed the worst in the intermediate amino acid depth environment. Mutational effects in exposed and buried (according to residue depth) environments were better predicted. Perhaps, the intermediate depth levels need to be further stratified, which in the case of Packpred would be reflected in the FADHM matrix values and in the clique statistical potential. Improvements could also be thought of by examining the reasons for why Packpred was unable to accurately predict the fate of 72 mutants that were all accurately called by the other six methods. We could also dissect the 23 correct predictions that Packpred made that were missed by all other methods to determine the relative strength of Packpred in comparison to the other methods.

Packpred relies on the sequence and structure of a given protein to predict the effect of a mutation. It is likely that these predictions could be impacted by the accuracy/resolution of the protein structure. The two structural features that Packpred extracts from structures are amino acid depth and structural neighbors. To whatever extent these two features get affected by the quality/accuracy/resolution of the structure would predicate the impact it would have on the final predictions. For the structures in the Missense3D data set, they all have resolutions of 2 Å or better. For this set, there appears to be no correlation between the accuracy of the prediction and the resolution of the structure (**Supplementary Figure S2**). In an independent study, we are exploring the use of homology models along with low-resolution structures from the PDB to quantify the impact of structural accuracy on Packpred predictions.

The clique statistical potential that has many tunable parameters such as the number of amino acids in the clique, cut-off distance, and definitions of what constitutes a “contact” between residues. Packpred could improve by investigating these aspects too, and this would form an independent study in itself. Similarly, further tweaks to the FADHM matrix, as briefly discussed above, could also possibly improve overall prediction accuracy. Shannon entropy accounts for the degree of variation at a given site/position and does not change depending on the type of mutation. In our method, we use Shannon entropy in conjunction with the clique potential and FADHM to get a wholesome picture of sequence and structure conservation. However, it is likely that a more nuanced version of the entropy measure and/or other scores for conservation may help get more accurate predictions. In its current implementation, Packpred categorizes mutations as being neutral or destabilizing. When we tried to correlate the score with a discretized value of the function, the correlations were around −0.5. Perhaps, with some of the improvements discussed above, this correlation would also improve.

One important observation from our findings is that of the 4,099 mutations, 4,036 were correctly called by at least one of the methods. There exists great complementarity between the methods tested here. We were tempted to then use two simple meta prediction methods. We designated the predictions involving a particular wild-type amino acid or a wild-type mutant amino acid pair to the method that best predicted this type. Such a simple-minded approach gave us MCCs of 0.40 and 0.51 for the amino acid and the amino acid pair type predictions, respectively, where the best predicting method, Packpred, had an MCC of 0.36 (Missense3D data set). It is conceivable that a different method of combining the results from these different methods could vastly increase the accuracy of predicting the functional fate of single amino acid changes.

We assessed the robustness of Packpred by training it on the T4 set and a combination of the T4 and CcdB saturation mutagenesis data sets. Each of the training sets gave us different optimal values of feature weights. These different weights did not, however, affect the overall performance of the method on the Missense3D testing set. In earlier results too, we had observed that different weight combinations gave rise to similar performances on the training set. We believe that one of the primary reasons for the different optimal weights is the fact that the three features in Packpred do not all affect predictions at the same level of granularity. The statistical potential and the substitution matrices (FADHM) give a score for particular mutations, whereas the Shannon entropy score gives a single value for a position, regardless of the type of mutation. Given the myriad of different environments and levels of conservation in different positions of the protein, the contribution due to each of these features is not uniformly the same across a protein. The positive aspect of these predictions is that despite the lack of consensus of optimal values of the different features, the overall prediction accuracy does not appear to suffer. This is probably indicative of the fact that the features of the algorithm are important, and perhaps a different way of combining these features may yield consistently better results.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

KT and TK contributed equally to this study. MM, KT, and TK conceptualized and planned the study. CK advised. KT and TK carried out all the computations.

FUNDING

This work was supported by a Wellcome Trust-DBT India Alliance Senior Fellowship.

ACKNOWLEDGMENTS

MM. would like to acknowledge the DBT-Wellcome India Alliance for a Senior Fellowship. We would like to thank Chandra Verma, Prof. Raghavan Varadarajan, and members of the COSPI laboratory for discussions and constructive criticism. We thank Gulzar Singh and Parichit Sharma for their contribution to the Packpred web server. We also thank Swastik Mishra and Neelesh Soni for their technical assistance with the Packpred software.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.646288/full#supplementary-material

REFERENCES

Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., et al. (2012). Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure* 20, 371–381. doi:10.1016/j.str.2011.11.021

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., et al. (2012). An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature* 491, 56–65. doi:10.1038/nature11632

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., et al. (2017). UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi:10.1093/nar/gkw1099

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* 33, W306. doi:10.1093/nar/gki375

Chakravarty, S., and Varadarajan, R. (1999). Residue Depth: A Novel Parameter for the Analysis of Protein Structure and Stability. *Structure* 7, 723–732. doi:10.1016/S0969-2126(99)80097-5

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The Sequence of the Human Genome. *Science* 291, 1304–1351. doi:10.1126/science.1058040

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009). Fast and Accurate Predictions of Protein Stability Changes upon Mutations Using Statistical Potentials and Neural Networks: PoPMuSiC-2.0. *Bioinformatics* 25, 2537–2543. doi:10.1093/bioinformatics/btp445

Farheen, N., Sen, N., Nair, S., Tan, K. P., and Madhusudhan, M. S. (2017). Depth Dependent Amino Acid Substitution Matrices and Their Use in Predicting Deleterious Mutations. *Prog. Biophys. Mol. Biol.* 128, 14–23. doi:10.1016/j.pbiomolbio.2017.02.004

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: The Protein Families Database. *Nucleic Acids Res.* 42, D222. doi:10.1093/nar/gkt1223

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., et al. (2007). A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* 449, 851–861. doi:10.1038/nature06258

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., and Sternberg, M. J. E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated?. *J. Mol. Biol.* 431, 2197–2212. doi:10.1016/j.jmb.2019.04.009

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes. *Nucleic Acids Res.* 45, D840–D845. doi:10.1093/nar/gkw971

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: Public Archive of Relationships Among Sequence Variation and Human Phenotype. *Nucleic Acids Res.* 42, D980. doi:10.1093/nar/gkt1113

Loris, R., Dao-Thi, M. H., Bahassi, E. M., Van Melderen, L., Poortmans, F., Liddington, R., et al. (1999). Crystal Structure of CcdB, a Topoisomerase Poison from E. coli. *J. Mol. Biol.* 285, 1667–1677. doi:10.1006/jmbi.1998.2395

Masso, M., and Vaisman, I. I. (2014). AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. *Adv. Bioinformatics* 2014, 278385. doi:10.1155/2014/278385

Matthews, B. W. (1975). Comparison Of The Predicted And Observed Secondary Structure Of T4 Phage Lysozyme. *Biochim. Biophys. Acta.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Res.* 31, 3812–3814. doi:10.1093/nar/gkg509

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: A Server for Predicting Effects of Mutations on Protein Stability. *Nucleic Acids Res.* 45, W229–W235. doi:10.1093/nar/gkx439

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014a). DUET: A Server for Predicting Effects of Mutations on Protein Stability Using an Integrated Computational Approach. *Nucleic Acids Res.* 42. doi:10.1093/nar/gku411

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014b). MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* 30, 335–342. doi:10.1093/bioinformatics/btt691

Rennell, D., Bouvier, S. E., Hardy, L. W., and Poteete, A. R. (1991). Systematic Mutation of Bacteriophage T4 Lysozyme. *J. Mol. Biol.* 222, 67. doi:10.1016/0022-2836(91)90738-R

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., et al. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* 328, 636–639. doi:10.1126/science.1186802

Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2020). DynaMut2: Assessing Changes in Stability and Flexibility upon Single and Multiple point Missense Mutations. *Protein Sci.* 30, 60. doi:10.1002/pro.3942

Šali, A., and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 234, 779–815. doi:10.1006/jmbi.1993.1626

Sippl, M. J. (1990). Calculation of Conformational Ensembles from Potentials of mena Force. An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins. *J. Mol. Biol.* 213, 859–883. doi:10.1016/S0022-2836(05)80269-4

Smith, H. O., Annau, T. M., and Chandrasegaran, S. (1990). Finding Sequence Motifs in Groups of Functionally Related Proteins. *Proc. Natl. Acad. Sci. U. S. A.* 87, 826–830. doi:10.1073/pnas.87.2.826

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and Promise of Genome-wide Association Studies for Human Complex Trait Genetics. *Genetics* 187, 367–383. doi:10.1534/genetics.110.120907

Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R., and Madhusudhan, M. S. (2013). Depth: A Web Server to Compute Depth, Cavity Sizes, Detect Potential Small-Molecule Ligand-Binding Cavities and Predict the pKa of Ionizable Residues in Proteins. *Nucleic Acids Res.* 41, W314. doi:10.1093/nar/gkt503

Tan, K. P., Varadarajan, R., and Madhusudhan, M. S. (2011). DEPTH: A Web Server to Compute Depth and Predict Small-Molecule Binding Cavities in Proteins. *Nucleic Acids Res.* 39, W242–W248. doi:10.1093/nar/gkr356

Weaver, L. H., and Matthews, B. W. (1987). Structure of Bacteriophage T4 Lysozyme Refined at 1.7 Å Resolution. *J. Mol. Biol.* 193, 189–199. doi:10.1016/0022-2836(87)90636-X

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). SDM – A Server for Predicting Effects of Mutations on Protein Stability and Malfunction. *Nucleic Acids Res.* 39, W215. doi:10.1093/nar/gkr363

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. E. (2014). SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *J. Mol. Biol.* 426, 2692–2701. doi:10.1016/j.jmb.2014.04.026

Zhang, Z., Miteva, M. A., Wang, L., and Alexov, E. (2012). Analyzing Effects of Naturally Occurring Missense Mutations. *Comput. Math. Methods Med.* 2012, 1–15. doi:10.1155/2012/805827

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

REVIEW
published: 28 September 2021
doi: 10.3389/fmolb.2021.643849

Application of Computational Methods in Understanding Mutations in *Mycobacterium tuberculosis* Drug Resistance

Grace Mugumbate[1], Brilliant Nyathi[2], Albert Zindoga[2] and Gadzikano Munyuki[2]*

[1]Department of Chemical Sciences, Midlands State University, Gweru, Zimbabwe, [2]Department of Chemistry, Chinhoyi University of Technology, Chinhoyi, Zimbabwe

The emergence of drug-resistant strains of *Mycobacterium tuberculosis* (*Mtb*) impedes the End TB Strategy by the World Health Organization aiming for zero deaths, disease, and suffering at the hands of tuberculosis (TB). Mutations within anti-TB drug targets play a major role in conferring drug resistance within *Mtb*; hence, computational methods and tools are being used to understand the mechanisms by which they facilitate drug resistance. In this article, computational techniques such as molecular docking and molecular dynamics are applied to explore point mutations and their roles in affecting binding affinities for anti-TB drugs, often times lowering the protein's affinity for the drug. Advances and adoption of computational techniques, chemoinformatics, and bioinformatics in molecular biosciences and resources supporting machine learning techniques are in abundance, and this has seen a spike in its use to predict mutations in *Mtb*. This article highlights the importance of molecular modeling in deducing how point mutations in proteins confer resistance through destabilizing binding sites of drugs and effectively inhibiting the drug action.

Keywords: mutations, drug resistance, computational tools, *Mycobacterium tuberculosis*, molecular modeling

OPEN ACCESS

Edited by:
Daisuke Kihara,
Purdue University, United States

Reviewed by:
Zhili Zuo,
Kunming Institute of Botany (CAS), China
Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology (LMB), United Kingdom

****Correspondence:***
Grace Mugumbate
mugumbateg@msu.ac.zw

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *18 December 2020*
Accepted: *16 August 2021*
Published: *28 September 2021*

Citation:
Mugumbate G, Nyathi B, Zindoga A and Munyuki G (2021) Application of Computational Methods in Understanding Mutations in Mycobacterium tuberculosis *Drug Resistance.*
Front. Mol. Biosci. 8:643849.
doi: 10.3389/fmolb.2021.643849

INTRODUCTION

Drug resistance in tuberculosis chemotherapy is fast becoming a health crisis on a global scale. The emergence of multidrug-resistant (MDR), extensively drug-resistant (XDR), and totally drug-resistant (TDR) strains of *Mycobacterium tuberculosis* (*Mtb*) has been observed as a result of ineffective directly observed treatment short-course (DOTS) (Bihari et al., 2008; Whalen, 2006) among a myriad of other factors. MDR is due to resistance to at least one first-line drug (**Figure 1**) including isoniazid (INH) which inhibits mycolic acid synthesis (Bollela et al., 2016) and rifampicin (RIF) that inhibits RNA synthesis (Zhang et al., 2019). Other TB drugs facing resistance include ethambutol (EMB) that targets the arabinogalactan synthesis (Zhang and Yew, 2009), streptomycin (STR) that inhibits protein synthesis (Ruiz et al., 2002), and pyrazinamide (PZA) that inhibits pantothenate and CoA synthesis, disrupting plasma membrane and energy metabolism (Zhang et al., 2014).

Resistance to first-line drugs leads to the implementation of treatment regiments belonging to the second-line drugs which are fluoroquinolones, kanamycin/amikacin and capreomycin/viomycin, and ethionamide whose mechanisms of action involve introducing negative supercoils in DNA molecules, inhibiting protein synthesis, and disrupting cell wall biosynthesis by inhibiting mycolic acid synthesis, respectively (**Table 1**). XDR and TDR are,

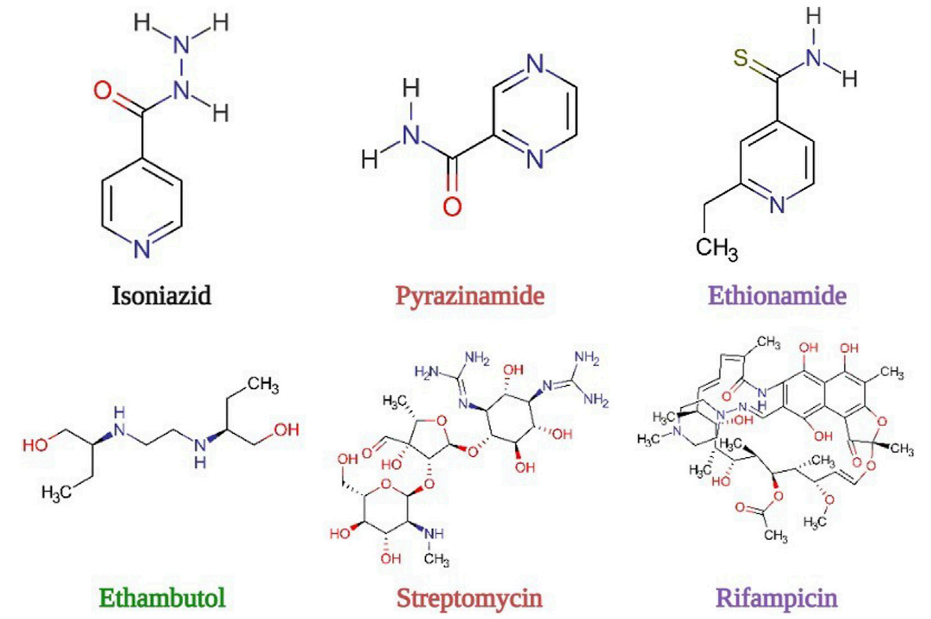


FIGURE 1 | Structures of first-line drugs and ethionamide, a second-line drug.

TABLE 1 | Drug targets and the mode of action (Louw et al., 2009; Zhang and Yew, 2009).

Drug	Target	Gene	Drug mode of action
Ethambutol	Arabinosyl transferase	*embCAB*	Inhibits arabinogalactan synthesis
Streptomycin	Ribosomal protein S12 16S rRNA 7-Methylguanosine methyltransferase	*rpsL* *rrs* *gidB*	Inhibits protein synthesis
Pyrazinamide	Pyrazinamidase	*pncA*	Disrupts plasma membrane and energy metabolism (inhibits pantothenate and CoA synthesis)
Rifampicin	β subunit of RNA polymerase	*rpoB*	Inhibits RNA synthesis
Isoniazid	Fatty acid enoyl acyl carrier protein reductase A Catalase peroxidase β-Ketoacyl-ACP synthase NADH dehydrogenase Alkyl hydroperoxidase reductase	*InhA* *katG* *kasA* *ndh* *ahpC*	Inhibits mycolic acid synthesis
Ethionamide	Flavin monooxygenase Fatty acid enoyl acyl carrier protein reductase A Transcriptional repressor	*ethA* *InhA* *ethR*	Disrupts cell wall biosynthesis by inhibition of mycolic acid synthesis
Kanamycin/Amikacin	16S rRNA	*rrs*	Inhibits protein synthesis
Capreomycin/ Viomycin	rRNA methyltransferase 16S rRNA	*tlyA* *rrs*	
Fluoroquinolones	DNA gyrase	*gyrA* *gyrB*	Introduces negative supercoils in DNA molecules

therefore, due to resistance to several second-line drugs including fluoroquinolones in conjunction with MDR. For better management of drug resistance and rapid detection of resistance, knowledge of the mechanism of resistance at the molecular level is extremely important for an effective treatment regimen to be prescribed.

More often, drug resistance in *Mtb* is associated with mutations within the drug targets; however, not all mutations within the organism are associated with resistance. Drug resistance mechanisms are driven mainly by single-nucleotide polymorphisms or other polymorphisms resulting in the modification of drug targets (Palomino and Martin, 2014). Therefore, understanding the mechanism of action and resistance of the drugs is of paramount importance. Of the first-line drugs, ethambutol, which is active against fast-multiplying bacteria, disrupts the synthesis of arabinogalactan in the cell wall by targeting the *mycobacterial* arabinosyl transferase enzyme encoded by the gene *embB*, encapsulated in the embCAB operon, and mutations in the *embB*306 gene confers ethambutol resistance (Zhang and Yew, 2009). On the other hand, streptomycin, a drug active against slow-growing bacteria, irreversibly binds to the 30S ribosome subunit, blocking translation thereby inhibiting protein synthesis. Chromosomally acquired streptomycin resistance is associated with mutations in the *rpsL*, *rrs*, *and gidB* encoding for ribosomal protein S12, 16S rRNA, and 7-methylguanosine methyl transferase, respectively (Zhang and Yew, 2009). Similarly, resistance to rifampicin, a key component in the first-line treatment of TB that binds to the β subunit of RNA polymerase, has been linked to mutations in a region of the 81 bp region of the *rpoB* gene. Whilst the gene encodes for the β subunit of RNA polymerase, rifampicin resistance is mostly due to mutations at positions 516, 526, and 531 (Goldstein, 2014; Uddin et al., 2020). This is achieved by inhibition of elongation of the messenger RNA, which interferes with transcription (Uddin et al., 2020).

Pyrazinamide is also a key antituberculosis (TB) drug that substantially enhances the activity of novel agents bedaquiline (BDQ) and pretomanid (PA50) in murine models of TB. A vital attribute of this prodrug is its ability to inhibit semidormant bacteria in acidic environments. In its activity, the prodrug is converted by pyrazinamidase/nicotinamidase to its active form, pyrazinoic acid which inhibits membrane transport by disrupting the bacterial membrane energetics. Resistance to pyrazinamide is mainly characterized by mutations clustered at positions 3–17, 61–85, and 132–142 in the *pncA* gene that codes for *mycobacterial* enzyme pyrazinamidase (PZase) (Zhang et al., 2014). The association of multiple mutations throughout the *pncA* gene with PZA resistance makes it difficult to develop a test for detecting PZA resistance (Piersimoni et al., 2013). In most instances, molecular methods are applied to investigate PZA resistance by screening mutations in *pncA* genes in distinct epidemiological regions offering a much more rapid alternative method compared to that of conventional bacteriology (Khan et al., 2019). Miotto identified 280 mutations in 1950 clinical strains (Miotto et al., 2014), which were categorized into four groups: very high–confidence resistance mutations, high-confidence resistance mutations, mutations with an unclear role, and mutations not associated with phenotypic resistance based on the confidence level.

Isoniazid and ethionamide are effective drugs for the treatment of TB; however, several clinical MDR-TB strains have shown high levels of resistance (Machado et al., 2012). Structurally, INH and ETH are highly similar, both containing the pyridine ring; however, ETH is a second-line drug primarily used to treat MDR-TB, and just like INH, it is a prodrug that requires metabolic activation (DeBarber et al., 2000). Although the active metabolites of both drugs inhibit an NADH-enoyl acyl protein reductase, InhA, the drugs have independent activation pathways. The validated drug target InhA is an enzyme involved in fatty acid biosynthesis II, which is important in the bio-production of mycolic acids. These long-chain fatty acids are responsible for the unique impermeable nature of the *Mycobacterium tuberculosis* cell wall (Dover et al., 2004; Timmins and Voja, 2006).

INH is activated by the catalase-peroxidase KatG to INH-NAD and INH-NADP adducts that effectively inhibit InhA (Timmins and Voja, 2006). Resistance to INH has been attributed to mutations or deletion in the active site of the *katG* gene, which encodes the enzyme, KatG (Hameed et al., 2018), at position S315 and position 15 in the InhA promoter region. Also, mutations in *ahpC, kasA,* and *ndh* encoding for alkyl hydroperoxidase reductase, β-ketoacyl ACP synthase, and NADH dehydrogenase, respectively, are associated with INH resistance (Nayak et al., 2017). Cross-resistance occurs between INH and its structural analog, and ETH has been attributed to mutations in the InhA promoter.

On the contrary, ETH is activated by the enzyme EthA encoded by the gene Rv3854c to the toxic S-oxide then to 2-ethyl-4-aminopyrimidine (DeBarber et al., 2000; Baulard et al., 2000). The transcription of the FAD-containing monooxygenase, EthA, is controlled by another gene *ethR* that encodes the protein, EthR. Earlier studies of the resistance mechanism of ethionamide revealed that an increase in the amount of EthR, a member of the TetR repressors, reduces the amount of EthA and results in ethionamide resistance by *mycobacterium tuberculosis* (DeBarber et al., 2000; Baulard, 2000). Mutation studies on MDR-TB isolates revealed the presence of EthR F110L mutants implicated in resistance to ETH. The residue F110 occupies a central position in the long cylindrical and hydrophobic ligand-binding site of EthR.

Similar to INH, ethionamide (ETH) is a second-line prodrug activated by the monooxygenase encoded by the *ethA* gene. Once activated, it forms an adduct with NAD, which inhibits the enzyme enoyl-ACP reductase, thus disrupting mycolic acid synthesis. Transcription of the monooxygenase, ethA is negatively regulated by ethR; hence, allosteric inhibition of ethR would enhance activation of ETH and computer some of the mutation processes.

The advances in computational techniques and expansions in bioinformatics and chemoinformatics have brought a sigh of relief in the study of mutations and provided a rapid drug susceptibility testing important in the detection and control of MDR/XDR TB (Shinnick et al., 2005). Therefore, in this article, we analyze the effective application of computational techniques and tools in the study and understanding of molecular target mutations in conferring drug resistance to first-line drugs and also analyze how we are applying these methods to identify inhibitors that would circumvent resistance in ethR, a gene implicated in the resistance of ethionamide as well as highlight prospects in fast and cost-effective advances to understand drug resistance of antituberculosis drugs.

METHOD

To give a detailed account of how the computational techniques have been applied in the study of the contributions of mutations to the emergence of drug-resistant *Mycobacterium tuberculosis*,

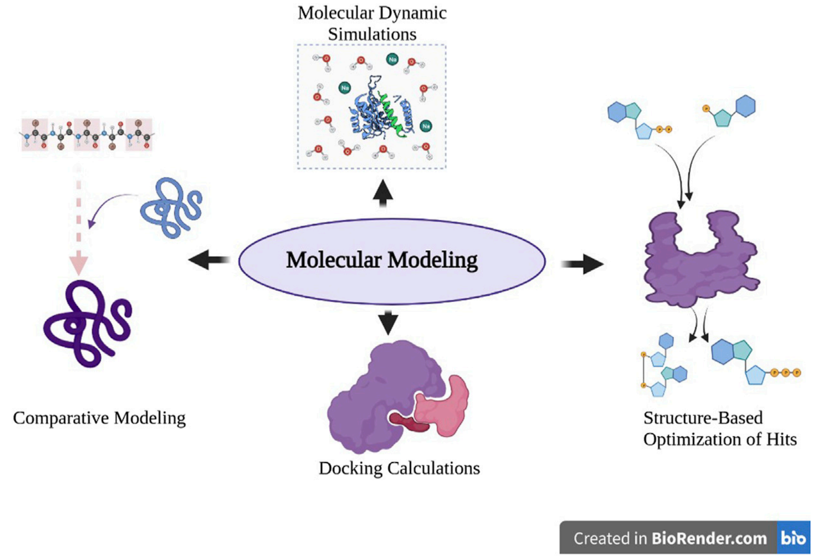

Molecular Dynamic Simulations
Molecular Modeling
Comparative Modeling
Docking Calculations
Structure-Based Optimization of Hits
Created in BioRender.com bio

FIGURE 2 | Common computational approaches applied to study the effect of mutations on protein–ligand interactions.

an extensive literature search was performed. A description of the mechanisms of action of the first-line drugs rifampicin and isoniazid as well as ethionamide, a second-line drug is given. An analysis of the common computational methods used to study the mutations in relevant genes for each drug was performed. Lastly, a detailed account of the importance of F110 in ethR, a transcription regulator implicated in the resistance of ethionamide, is presented. Modeling of the proteome for mycobacteria, and identification of the hotspots and druggability of the proteins are given.

Computational Approaches

A variety of computational techniques that include comparative (homology) modeling, molecular dynamics, protein–ligand docking, and structure-based optimization of ligands (**Figure 2**) have been successfully used to study the impact of mutations at atomic levels on protein–ligand binding and interactions and how they negatively affect ligand affinity by the mutant proteins (Phelan et al., 2016; Zhang et al., 2019; Jamal et al., 2020). Advanced approaches that include machine learning alongside artificial intelligence, bioinformatics, and cheminformatics databases have also been successfully used to build models and tools that can predict mutation and determine their capabilities in conferring resistance (Jamal et al., 2020; Ghosh et al., 2020; Sandgren et al., 2009).

Effect of Mutation in rpoB on Protein–RIF Interactions

Pang and co-workers approached RIF mutations with a computational approach. They used homology modeling to generate a three-dimensional structure of the wild type *rpoB* based on the crystal structure of *Thermus aquaticus* (Taq) core RNAP complexed with RIF. Discovery Studio 3.1 was used for this structural analysis exercise. The protein was modeled using a Build Homology module within the Protein Data Bank; a loop refinement module from Modeller was used to perform structural refinements, and energy minimizations were performed with the Smart Minimizer algorithm. The Build Mutants module was used for building mutants Ser531Leu, His526Asp, His526Gly, His526Leu, His526Arg, and Leu533Pro, and the Align and Superimpose Proteins module was used to compare the wild-type and mutant structures. Their study sought to evaluate the effects of mutating specific amino acid residues involved in the binding of RIF on protein–ligand interactions. The mutated protein–ligand interactions are evaluated subsequently using the Analyze Ligand Interactions and Structure Monitor module. They discovered that the mutated target protein had some level of resistance for RIF as it showed a decrease in its binding affinity. Mutations in His526Asp and Ser531Leu significantly reduced the affinity of *rpoB* for RIF by introducing charge repulsion and conformational changes in *rpoB*, respectively. The other strains with mutations His526Gly, His526Leu, His526Arg, and Leu533Pro exhibited low-level resistance (Pang et al., 2013). On the other hand, Zhang approached this challenge in exploring resistance mechanisms by combining the molecular dynamics simulation, molecular mechanics generalized-Born surface area calculation, dynamic network analysis, and residue interaction network analysis. Molecular dynamics simulations were all performed with the Amber14 package, and it was observed that the binding free energies of RIF with the three mutants H451D/Y/R decreased with molecular mechanics generalized-Born surface area calculations. Dynamic network analysis and residue interaction network analysis indicated increased flexibility within the binding pocket due to mutation of residue 451 which in turn weakened Q438, F439, M440, D441, and S447 residue interactions within the binding pocket. Such flexibility allowed for residues meant that a hydrogen bond to RIF was lost, thus accounting for decreased RIF binding in the mutant RNA polymerase. Changes within the binding pocket in the H451R mutant are extensive, giving too much freedom for RIF to move within the pocket (Zhang et al., 2019). Therefore, H451D/Y/R mutations increased the flexibility of the active pocket which in turn weakened the binding ability of *Mtb* RNA polymerase with RIF. Thus, the H451D/Y/R mutations weaken the interaction of the mutated residue with its adjacent residues. In similar work involving homology modeling of *rpoB* and docking calculations of RIF, Kumar and Jena have shown that two mutants S450L and H445Y exhibit low binding affinity toward the wild type *rpoB*, which has high affinity for the RIF molecule (Kumar and Jena, 2014).

Singh and co-workers investigated mutations of H451. The *Mtb rpoB* sequence was obtained from UniProt, the structure was built through comparative modeling with Modeller, and it was mutated computationally at position 451 using PyMol. GROMACS version 5.0 molecular dynamics simulation was performed on all the structures to obtain stable structures at 40ns. On the stable structures, RIF was docked onto them with AutoDock 4.2, and ligand–RIF complexes were subjected to molecular dynamics and molecular mechanics for estimation of free binding energies in wild-type and mutant systems. Resistance in the mutants arises due to changes within the binding pocket when polar and hydrophobic amino acids were replaced, which affected packing and folding in the vicinity, and relocation of the binding site itself rendering the RNA exit channel inaccessible to the drug (Singh et al., 2017).

The aforementioned studies on RIF resistance all have a consensus on the conference of resistance by the mutations in the target protein. They showed that mutations in *rpoB* cause structural changes within the binding pocket and its vicinity. They also indicated that interactions between the binding pocket residues are changed as a result of a mutation within the binding pocket and its vicinity greatly affecting the location and structure of the binding pocket. Most of these studies concluded that mutations that cause extensive structural changes will affect the way RIF sits in the binding pocket and increase freedom for the ligand in the pocket, which greatly decreases its affinity. Mutations that specifically occur within the binding pocket starve the RIF ligands of residues that contribute to a better binding affinity.

Effect of Mutation in InhA, and katG on INH Binding

Computational studies of INH resistance in *Mtb* have been extensively studied (Jena et al., 2014). INH is activated by *katG* and converted to an active intermediate displaying antimycobacterial properties; in the presence of NADH, an INH-NAD adduct is formed. It is the adduct that inhibits *InhA* (2-trans-enoyl-acyl carrier protein reductase), blocking the synthesis of mycolic acid (Dookie et al., 2018). In one study, homology modeling was employed to predict the 3D structure of *Mtb* UDP-galactopyranose mutase (Glf) and NADH Dehydrogenase (Ndh) with Modeller9v14, and the sequence in the FASTA format was obtained from the NCBI database. The NAD binder server was used in the identification of the binding site; docking studies and visualization were performed with AutoDock Vina Tool 1.5.4 and Pymol, respectively. FADH2 and NADH were both found to have a high affinity for Glf; thus, overexpression of Glf utilizes more NADH reducing its concentration. This results in decreased INH-NAD adduct formation thereby causing INH resistance (Nayak et al., 2017)

In another study, on the influence of mutation in INH, *katG* mutations S315T/S315N were modeled with Modeller9v10 and compared with the wild-type *katG*. It was observed that INH was forming a hydrogen bond with the mutant *katG* which hindered radical formation. AutoDock Tool 1.5.4 docking calculation indicated INH-NAD is more effective at inhibiting *InhA* compared to INH (Jena et al., 2014). The *katG* mutation S315T was computationally observed to decrease the flexibility of binding site residues, and *katG* mutants at His276Met, Gln295His, and Ser315Thr decreased the stability and flexibility of the mutant protein associated with INH resistance. Mutation of the arylamine N-acetyltransferase (NAT) enzyme increases the stability and catalytic activity of the enzyme making the NAT-INH interaction ineffective. Mutations in the *ahpC* result in overexpression of the protein, which is a compensatory mechanism for loss of activity due to the *katG* mutation; thus, the ability to defend against oxidative stress is maintained within the system (Jena and Wankhade, 2016; Waghmare, and Harinath, 2016). Just as in the RIF studies, conformational changes and pocket flexibility changes greatly affect the atomic-level interactions between the target protein and the drug compound, and the trend shows a decreased affinity for the drug by mutant protein targets.

Other Studies on the Effect of Mutation Mtb Drug Resistance

Deedler applied machine learning approaches to *Mtb* isolates that had undergone whole-genome sequencing. Nonparametric classification tree and gradient-boosted tree models were used to predict drug resistance alongside uncovering any associated new mutations. Resistance markers to drugs other than the drug of interest was used in fitting separate drug models for each drug based on the presence and absence of the co-occurrent resistance markers. Predictive performance testing was performed alongside laboratory drug-susceptibility testing. The performance was highest for resistance to first-line drugs, amikacin, kanamycin, ciprofloxacin, moxifloxacin, and multidrug-resistant tuberculosis. The inclusion of resistance markers led to improved results (Deelder et al., 2019).

In a bid to understand the molecular consequences of polymorphisms within loci associated with antituberculosis drugs, Portelli and co-workers employed computational methods to quantify point mutations in conferring resistance. Homology models of target proteins were built with UCSF Chimera 1.1, and protein–ligand docking and protein–ligand interactions were carried out with GLIDE and Arpeggio, respectively. Portelli et al., 2018 concluded that mutational effects are mostly imparted *via* steric or electrostatic changes within the protein, leading to functional changes and affecting target–drug interactions. They also noted that most phenotypically resistant mutations act allosterically, and the introduction of variants affects the drug–protein complex stability, leading to resistance. Frequently occurring mutations do not confer extreme changes in parameters; the protein retains its functionality, but the drug–protein complex is weakened. Mildly stabilizing mutations may confer local fitness advantages. Drug-resistant mutations within the protein are enhanced while maintaining stability within the protein function. It was also concluded that concurrent mutations in close topological proximity enable localized effects of the mutation, and their combination with external mutations ensures different mechanisms that lead to drug resistance.

Nakatani and Helen, 2017 predicted that the *alr* M319T mutation observed in an XDR strain of *Mtb* would likely confer resistance to D-cycloserine (DSC) as it had been noted that the acquisition of this mutation occurred with treatment of DSC suggesting that the mutation is sufficient and necessary to confer resistance. Molecular modeling of the C-8T, M319T, Y364D, and R373L mutations provided insights into how resistance is conferred upon treatment with DSC. DSC covalently binds to an *alr* cofactor pyridoxal 5′-phosphate (PLP); this act irreversibly inhibits *alr* through disruption of the *alr*-PLP covalent bond (Fenn et al., 2003). A generated model of *Mtb, alr*, and DSC highlighted the residues 319 and 364 located directly in the active site. A mutation to aspartic acid at residue 364 introduced a shorter negatively charged side chain. Such a change affects the positioning of the phosphate moiety in PLP, potentially affecting PLP orientation in the active site. The location of the residue 319 mutation could alter the interactions with 364, likely affecting DSC inhibition. *Alr* functions as a homodimer, and the R373L mutation is not located directly in the active site; however, it is close to M319 and D320 and the dimer interface. Such a mutation is most likely going to disrupt molecular interactions at the dimer interface and greatly destabilizing the DSC binding site. This study was strategic to the pharmaceutical sector in 2015 amidst a Global Drug Facility declaration of a price reduction of the DSC drug. Understanding the resistance mechanisms was important for facilitating phenotypic and genotypic drug susceptibility testing (Stop, 2015).

Malik et al. (2012) provided insights into fluoroquinolone resistance through functional genetic analyses and structural modeling techniques. Crystal structures of the N-terminal and C-terminal domains for *gyrA* and *gyrB* were superimposed on the crystal structure of the complex of *Streptococcus pneumoniae* gyrase with a DNA substrate and levofloxacin, all obtained from the Protein Data Bank using the tool Coot (Emsley and Kevin, 2004). This study highlights that *gyrB* mutations M330I, V340L, R485C, D500A, D533A, A543T, A543V, and T546M are not sufficient to confer drug resistance. N538D, E540V, and R485C + T539N mutations did confer resistance to all fluoroquinolones whilst N538K and E540D conferred resistance to moxifloxacin only, and D500H and D500N mutations conferred resistance only to levofloxacin and ofloxacin. The importance of this study was in explaining minimum inhibitory concentrations as observed in experimental work; molecular modeling explained how resistance came about to be through a 3D spatial orientation of substitute residues in the mutant proteins.

Effect of Mutation on ethR-Ligand Binding Affinity

Resistance to ETH has been linked to mutations in the *ethR, ethA,* and *inhA* genes (Hameed et al., 2018) that collectively play crucial roles in the activity of the drug. As a regulator, the N-terminus helix-turn-helix (HTH) domains of the dimeric EthR bind DNA sequences responsible for the transcription of EthA and suppress its expression (Wolff and Nguyen, 2012). This process is controlled by small–molecular weight ligands that bind to the allosteric binding pocket of EthR located in the C-terminal end (Mugumbate et al., 2015). Binding of the ligands induces molecular conformational changes that increase the distances between DNA binding domains of the enzyme, inhibit DNA binding, and hence increase the transcription of EthA. For this reason, EthR has been validated as a suitable drug target for a new collection of antituberculosis compounds that would boost the activity of ETH. Targeting the resistance pathway of antituberculosis drugs has long been proposed (Wolff and Nguyen, 2012); therefore, independent research groups have deposited the apo and bound structures of EthR into the Protein Data Bank (PDB, https://www.rcsb.org/). These structures reveal that the protein is characterized by a long hydrophobic and promiscuous pocket that binds to structurally diverse small molecules like dioxane and long molecular chains with more than 30 atoms. The residue F110 is centrally positioned in the binding pocket with its aromatic side chain strategically positioned to participate in protein–ligand interactions (**Figure 3**).

In a previous study (Bishi, et al.), we carried out docking calculations of a Maybridge dataset containing more than 200 drug-like compounds to investigate binding modes and protein–ligand interactions. The results indicated that F110 played a crucial role in ligand binding, supporting the observation that F110L drastically reduces ligand affinity (Brossier et al., 2011). Most ligands were stabilized by a cascade of pi–pi interactions, where F110 played a central role by linking pi–pi interactions from the ligand to Phe114 (**Figure 3**) in a way that will stabilize the bound ligand and increase ligand affinity. This implies that the F110L mutation disrupts the pi–pi cascade and reduces the ligand–binding affinity.

Application of Machine Learning and Artificial Intelligence Approaches

Bioinformatics was employed for studies concerned with mutations focusing on *Mtb*. Ghosh and co-workers developed a Drug Resistance–Associated Genes database (DRAGdb) which is a repository of mutational data of drug resistance–associated genes (DRAGs) across ESKAPE (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter*

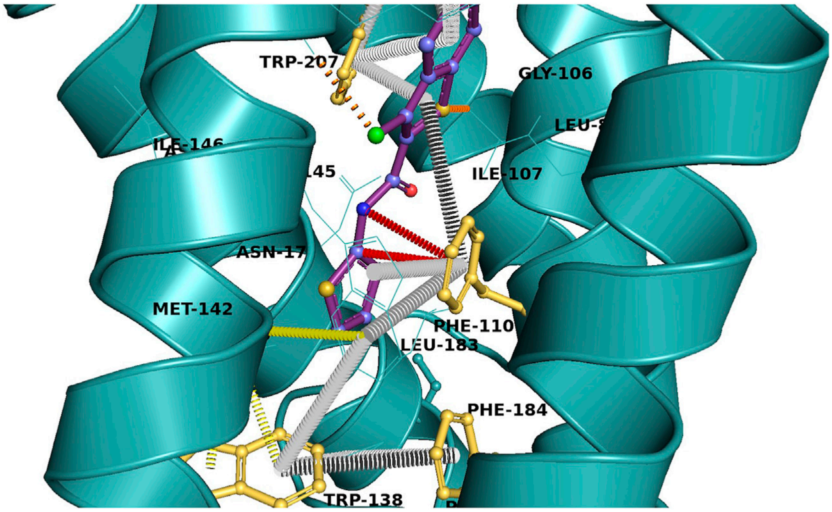


FIGURE 3 | The residue F110 facilitates the pi–pi cascade (grey rings) between aromatic residues in the binding pocket of EthR (yellow) and the ligand (purple), which later translates into a structural modification of the HTH motif and inhibition of DNA binding. Analysis of the interactions was performed using Aperggio (http://bleoberis.bioc.cam.ac.uk/arpeggioweb/) and viewed using PyMol.

baumannii, Pseudomonas aeruginosa, and Enterobacter spp.). Homoplasy is observed in six genes namely *gidB, gyrA, gyrB, rpoB, rpsL,* and *rrs* with mutations related to drug resistance being observed at the codon level. A single-nucleotide mutation that was associated with resistance to amikacin, gentamicin, rifampicin, and vancomycin in *Staphylococcus aureus* was an indication of pleiotropy. The database compiles *Mtb* drug-resistance genes across bacterial species allowing for homoplasy and pleiotropy identification in genes (Ghosh et al., 2020).

In their recent efforts, Jamal and co-workers developed machine learning algorithms alongside artificial intelligence to study and predict resistance in the genes *rpoB, inhA, katG, pncA, gyrA,* and *gyrB* for the drugs rifampicin, isoniazid, pyrazinamide, and fluoroquinolones. Machine learning algorithms naïve Bayes, k nearest neighbor, support vector machine, and artificial neural network were used to build the prediction models. Further molecular docking and molecular dynamics simulations were carried out on predicted resistance causing mutant proteins and their wild-type counterparts. This study evaluated protein conformation and its impact to confirm the observed phenotype (Jamal et al., 2020).

DISCUSSION

The application of machine learning and artificial intelligence in mutation studies is a fast-growing trend in computational research. At the center of it all, bioinformatics and cheminformatics databases are contributing a lot of information that is needed by machine learning algorithms to predict drug resistance–conferring mutations. Information was gathered across species in the experimental work, where the previously mentioned mutations and their influence on drug resistance were observed in detail, and the lack of hereafter is used to train machine learning algorithms in identifying possible novel mutations that might occur and probe their potential in conferring resistance. Usage of multiple layers or algorithms (deep learning) and artificial intelligence has greatly improved the accuracy of drug resistance and mutation prediction tools that have been made available to researchers (Deelder et al., 2019; Jamal et al., 2020).

Structure-guided drug discovery has lately become paramount to combat the emergence of *Mtb* drug-resistant strains which pose a concern to global public health. The rapid expansion of genome sequencing and pathway annotations has shown a positive impact on the progress of drug discovery. Computational tools have been developed to address the effect of mutations on the structure and function of proteins. The mutation cutoff scanning matrix (mCSM) is a machine learning approach which predicts the structural and functional effects of mutations on the target proteins. Its variants are capable of predicting the effects of mutations on protein stability, protein–protein interaction, and protein–ligand interactions (Pandurangan and Blundell 2020). EnCOM and FoldX are tools that are capable of predicting the effects of mutations on flexible protein conformations (Schymkowitz et al., 2005; Frappier et al., 2015). Rapid assessment of many mutations that are difficult to access with experimental methods has been made possible through predictive learning with machine learning algorithms (Waman et al., 2019).

Machine learning techniques have also been developed to address the need to improve TB resistance prediction in less-studied drugs. Rapid detection of antimicrobial resistance is vital in the prevention of existing drug resistance amplification, given that resistance markers are known; machine learning techniques are capable of timely prediction of resistance for a given *Mtb* drug. Machine learning methods are capable of ranking mutations

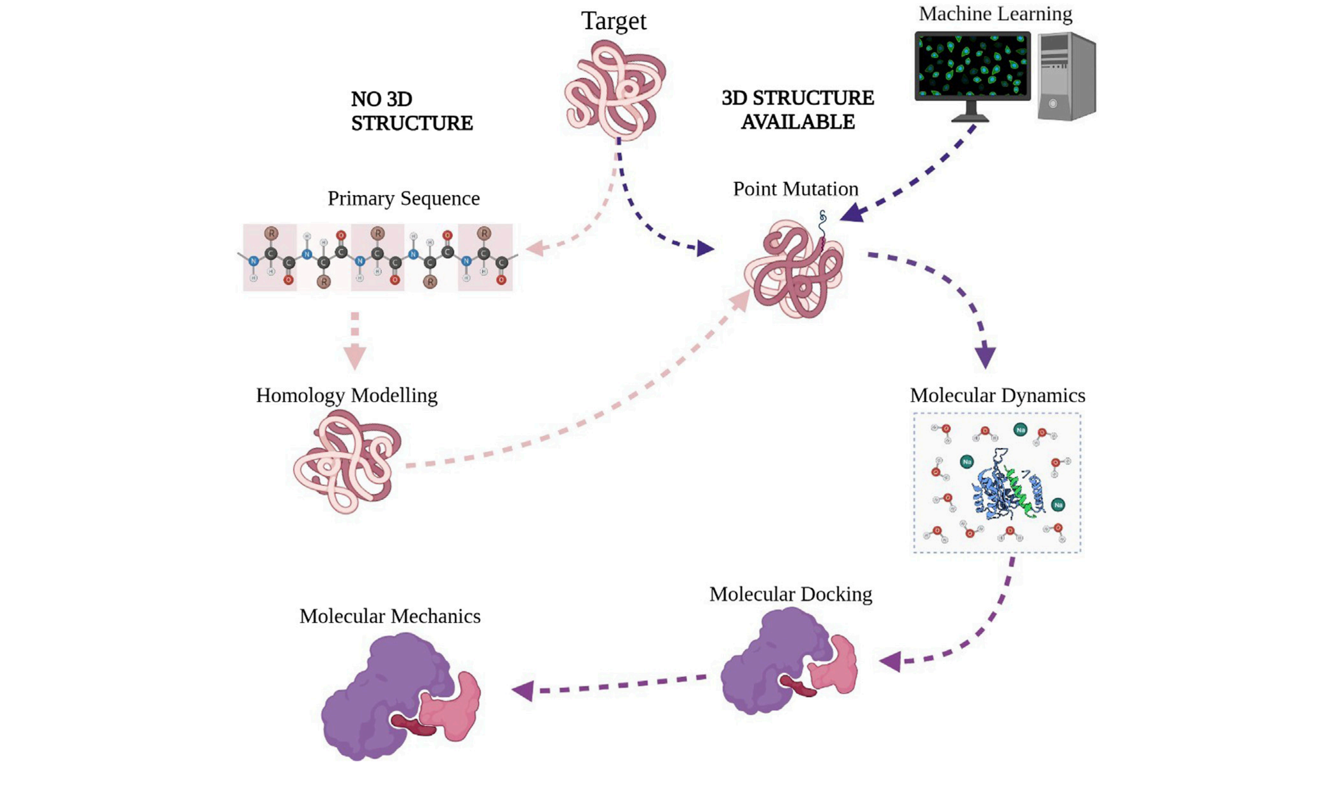


FIGURE 4 | Proposed protocol for studying the effect of mutations on protein–ligand interaction.

regarded as important and mutations from other genes associated with resistance to other drugs; this reflects on multidrug resistance from taking second-line drugs after taking first-line drugs, which is a huge advantage over experimental methods (Kouchaki et al., 2019).

DeepAMR has been developed with the task of identifying co-occurrent resistance within anti-TB drugs. This machine learning technique had a high performance with mean AUROC (Area Under the Receiver Operating Characteristics) from 94.4 to 98.7% for predicting resistance to four first-line drugs, RIF, EMB, INH, and PZA multi-drug resistant TB (MDR-TB) and pan-susceptible TB (PANS-TB: MTB that is susceptible to all four first-line anti-TB drugs). DeepAMR achieved its best mean sensitivity of 94.3, 91.5, 87.3, and 96.3% for INH, EMB, PZA, and MDR-TB, respectively. High-performance machine learning models have made the predictions of co-occurrent drug resistance to be performed timely and prevented amplification of existing resistance (Yang et al., 2019).

The use of machine learning and artificial intelligence makes them possible to identify novel resistance markers which are very difficult and costly to investigate with experimental methods. The timely and rapid prediction of drug resistance has made it possible for drugs to be returned to the discovery pipeline for optimization in a structure-guided drug design approach. To this end, the application of these techniques makes it possible for scientists to comprehensively study the protein–drug interactions at very little cost and shorter time frames.

Proposed Computational Protocol

The Application of computational tools (**Table 2**) in understanding mutations that confer drug resistance in *Mycobacterium tuberculosis* still require a canonization of the process for a standard result output. Initially, 3D structures of drug targets are obtained from the Protein Data Bank followed by point mutations which may be performed by changing an amino acid in a protein sequence with Pymol (**Figure 4**). In the absence of a 3D structure, the primary sequence of the protein is obtained from UniProt (a freely accessible database of protein sequences and functional information). 3D structures are modeled through a process known as homology/comparative modeling of proteins with a standalone program such as Modeller or, alternatively, an online server such as SWISS-MODEL (expasy.org). Molecular dynamics simulations are performed for energy minimizations of the wild-type and mutant drug targets obtaining the most stable protein structures; standalone programs such as GROMACS and Amber are used for performing the task (Singh et al., 2017; Zhang et al., 2019). In computational chemistry, energy minimizations which may also be referred to as geometry optimization entail the exploration of the conformational space for a collection of

TABLE 2 | Computational tools used in the study of mutations.

Computational tool	Use	References
AutoDock	Molecular docking and visualization	Morris et al. (2009)
Glide	Molecular docking and visualization	Friesner and Mainz (2006)
Pymol	Molecular visualization	DeLano (2002)
Gromacs	Molecular dynamics simulations	Van Der Spoel et al. (2005)
Amber	Molecular dynamics simulations	Case et al. (2005)
Modeller	Homology or comparative modeling of protein 3D structures	Webb and Sali (2016)
Discovery Studio	Molecular visualization	Studio (2008)
Arpeggio	A web server for calculating and visualizing interatomic interactions in protein structures	Jubb et al. (2017)
UCSF Chimera	Interactive visualization and analysis of molecular structures and related data	Pettersen et al. (2004)
DeepAMR	Predicting co-occurrent resistance of *Mycobacterium tuberculosis*	Yang et al. (2019)
EnCom	Predicting the effects of mutations on flexible protein conformations	Frappier et al. (2015)
FoldX		Schymkowitz et al. (2005)

TABLE 3 | Databases used alongside computational packages.

Database	Information contained	References
UniProt	Protein sequence and functional information	Consortium (2015)
Protein Databank	Protein 3D Structures	Berman et al. (2000)
DRAGdb	Mutational data of drug resistance–associated genes	Ghosh et al. (2020)
NCBI	Biological data and small-molecule database	Wheeler et al. (2006)
ChEMBL	Binding, functional, and ADMET information for a large number of drug-like bioactive compounds	Gaulton and LouisaBellis (2012)

atoms to find a proper molecular arrangement in space which is energy favorable and stable; it is also referred to as the global energy minimum (Jabeen et al., 2019). The resultant structures are then subjected to molecular docking, where the position of the ligand when bound to a protein receptor is predicted for the drug's wild type and mutated targets. AutoDockTools and Glide among other standalone software packages may be used for this task (Kumar and Jena 2014; Jamal et al., 2020). Protein–ligand complex structures may also undergo energy minimization with molecular dynamics (Kumar and Jena, 2014; Portelli et al., 2018). In the presence of a 3D structure complexed with the preferred drug, molecular mechanics is employed to probe free binding energies and compare protein–ligand complexes for wild-type and mutated drug targets (Zhang et al., 2019). Recent trends that are being explored in the field of computational work include the usage of machine learning algorithms to build prediction tools (Lee et al., 2020). Studies that make use of mathematical models alongside bioinformatics for drug resistance mutations have also been reported (Fonseca et al., 2015). There has also been an exploration of artificial intelligence alongside machine learning algorithms for drug resistance mutation predictive tests (Deelder et al., 2019).

CONCLUDING REMARKS

With the increase in the number of drug-resistant and multidrug-resistant strains of *Mtb*, a need has arisen for techniques that are rapid for extensive studies of the previously mentioned mutations. Computational methods (**Figure 4**) present us with the opportunity to rapidly carry out these studies in silico with outputs comparable with experimental work at high confidence at even lower costs. Such methods have been extensively employed in exploring drug resistance in rifampicin, isoniazid, and ethionamide with the findings correlating to what is observed in experimental work; structural changes within the mutant protein drastically reduce protein–ligand binding affinity.

Machine learning and artificial intelligence have brought about massive changes and advancements in studying mutations and drug resistance in *Mtb* and other diseases. These techniques have made it possible to identify resistance markers within the whole genome Muzondiwa et al., 2020, to predict drug resistance for a given molecule, and to predict co-occurrent drug resistance between two or more drugs. The techniques are driven by big data (**Table 3**), and to that effect, smaller specific repositories/ databases (Drug Resistance–Associated Genes database) have been created for the sole purpose of helping researchers who are studying mutations. Computational tools have also been created for the identification of resistance markers and prediction of drug resistance (DeepAMR). Predictive learning makes it possible for scientists to identify potentially unwanted drug characteristics that may not be picked up with experimental methods, greatly reducing the risk for drug failure and saving time and money in the process.

AUTHOR CONTRIBUTIONS

BN and AM wrote the manuscript, and GrM and GaM were involved in concept development and editing the manuscript.

REFERENCES

Baulard, A. R., Joanna, C. B., Jean, E.-N., Selwyn, Q., Ruth, A. M., Patrick, J. B., et al. (2000). Activation of the pro-drug ethionamide is regulated in mycobacteria. *J. Biol. Chem.* 275 (36), 28326–28331. doi:10.1074/jbc.M003744200

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Bihari, S., Arun, S., Rawat, T., and Katiyar, S. (2008). An Analysis of Failure of Category II DOTS Therapy. *Indian J. Community Med.* 33 (2), 129. doi:10.4103/0970-0218.40886

Brossier, F., Veziris, N., Truffot-Pernot, C., Jarlier, V., and Sougakoff, W. (2011). Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* 55 (1), 355. doi:10.1128/AAC.01030-10

Bollela, V. R., Namburete, E. I., Feliciano, C. S., Macheque, D., Harrison, L. H., and Caminero, J. A. (2016). Detection of KatG and InhA Mutations to Guide Isoniazid and Ethionamide Use for Drug-Resistant Tuberculosis. *Int J Tuberc. Lung Dis.* 20 (8), 1099–1104. doi:10.5588/ijtld.15.0864

Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26 (16), 1668–1688. doi:10.1002/jcc.20290

Consortium, U. P. (2015). UniProt: a Hub for Protein Information. *Nucleic Acids Res.* 43, D204–D212. doi:10.1093/nar/gku989

DeBarber, A. E., Mdluli, K., Bosman, M., Bekker, L.-G., and Barry, C. E. (2000). Ethionamide Activation and Sensitivity in Multidrug-Resistant Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci.* 97 (17), 9677–9682. doi:10.1073/pnas.97.17.9677

Deelder, W., Christakoudi, S., Phelan, J., Benavente, E. D., Campino, S., McNerney, R., et al. (2019). Machine Learning Predicts Accurately Mycobacterium Tuberculosis Drug Resistance from Whole Genome Sequencing Data. *Front. Genet.* 10 (SEP), 1–9. doi:10.3389/fgene.2019.00922

DeLano, W. L. (2002). Pymol: An Open-Source Molecular Graphics Tool. *CCP4 Newsl. Protein Crystallogr.* 40 (1), 82–92.

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., and Naidoo, K. (2018). Evolution of Drug Resistance in Mycobacterium Tuberculosis: A Review on the Molecular Determinants of Resistance and Implications for Personalized Care. *J. Antimicrob. Chemother.* 73 (5), 1138–1151. doi:10.1093/jac/dkx506

Dover, L. G., Cerdeño-Tárraga, A. M., Pallen, M. J., Parkhill, J., and Besra, G. S. (2004). Comparative Cell wall Core Biosynthesis in the Mycolated Pathogens,Mycobacterium tuberculosisandCorynebacterium Diphtheriae. *FEMS Microbiol. Rev.* 28 (2), 225–250. doi:10.1016/j.femsre.2003.10.001

Emsley, P., and Cowtan, K. (2004). Coot: Model-Building Tools for Molecular Graphics. *Acta Crystallogr. D Biol. Cryst.* 60 (12), 2126–2132. doi:10.1107/S0907444904019158

Fenn, T. D., Stamper, G. F., Morollo, A. A., and Ringe., D. (2003). A Side Reaction of Alanine Racemase: Transamination of Cycloserine. *Biochemistry* 42 (19), 5775–5783. doi:10.1021/bi027022d

Fonseca, J. D., Knight, G. M., and McHugh, T. D. (2015). The Complex Evolution of Antibiotic Resistance in Mycobacterium Tuberculosis. *Int. J. Infect. Dis.* 32, 94–100. doi:10.1016/j.ijid.2015.01.014

Frappier, V., Chartier, M., and Najmanovich, R. J. (2015). ENCoM Server: Exploring Protein Conformational Space and the Effect of Mutations on Protein Function and Stability. *Nucleic Acids Res.* 43, W395–W400. doi:10.1093/nar/gkv343

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *J. Med. Chem.* 49 (21), 6177–6196. doi:10.1021/jm051256o

Gaulton, A., BellisBellis, L. J. A., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi:10.1093/nar/gkr777

Ghosh, A., N., S., and Saha., S. (2020). Survey of Drug Resistance Associated Gene Mutations in Mycobacterium Tuberculosis, ESKAPE and Other Bacterial Species. *Sci. Rep.* 10 (1), 1–11. doi:10.1038/s41598-020-65766-8

Goldstein, B. P. (2014). Resistance to Rifampicin: A Review. *J. Antibiot.* 67, 625–630. Nature Publishing Group. doi:10.1038/ja.2014.107

Hameed, H. M. A., Islam, M. M., Chhotaray, C., Wang, C., Liu, Y., Tan, Y., et al. (2018). Molecular Targets Related Drug Resistance Mechanisms in MDR-, XDR-, and TDR-Mycobacterium Tuberculosis Strains. *Front. Cel. Infect. Microbiol.* 8, 114. doi:10.3389/fcimb.2018.00114

Jabeen, A., Mohamedali, A., and Ranganathan, S. (2019). "Protocol for Protein Structure Modelling," in *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic, 252–272. doi:10.1016/B978-0-12-809633-8.20477-9

Jamal, S., Khubaib, M., Gangwar, R., Grover, S., Grover, A., Hasnain, S. E., et al. (2020). Artificial Intelligence and Machine Learning Based Prediction of Resistant and Susceptible Mutations in Mycobacterium Tuberculosis. *Sci. Rep.* 10 (1), 1–16. doi:10.1038/s41598-020-62368-2

Jena, L., Waghmare, P., Kashikar, S., Kumar, S., and Harinath, B. C. (2014). Computational Approach to Understanding the Mechanism of Action of Isoniazid, an Anti-TB Drug. *Int. J. Mycobacteriology* 3 (4), 276–282. doi:10.1016/j.ijmyco.2014.08.003

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: a Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* 429 (3), 365–371. doi:10.1016/j.jmb.2016.12.004

Khan, M. T., Malik, S. I., Ali, S., Masood, N., Nadeem, T., Khan, A. S., et al. (2019). Pyrazinamide Resistance and Mutations in PncA Among Isolates of Mycobacterium Tuberculosis from Khyber Pakhtunkhwa, Pakistan. *BMC Infect. Dis.* 19 (1), 1–7. doi:10.1186/s12879-019-3764-2

Kouchaki, S., Yang, Y., Walker, T. M., Sarah Walker, A., Wilson, D. J., Peto, T. E. A., et al. (2019). Application of Machine Learning Techniques to Tuberculosis Drug Resistance Analysis. *Bioinformatics* 35 (13), 2276–2282. doi:10.1093/bioinformatics/bty949

Kumar, S., and Jena, L. (2014). Understanding Rifampicin Resistance in Tuberculosis through a Computational Approach. *Genomics Inform.* 12 (4), 276. doi:10.5808/gi.2014.12.4.276

L, J., and G, W. (2016). Computational Approach in Understanding Mechanism of Action of Isoniazid and Drug Resistance. *Mycobact Dis.* 06 (01), 1–3. doi:10.4172/2161-1068.1000202

Lee, B. M., Harold, L. K., Almeida, D. V., Aung, H. L., Forde, B. M., Hards, K., et al. (2020). Predicting Nitroimidazole Antibiotic Resistance Mutations in Mycobacterium Tuberculosis with Protein Engineering. *Plos Pathog.* 16 (2), e1008287. doi:10.1371/journal.ppat.1008287

Louw, G. E., Warren, R. M., Gey Van Pittius, N. C., McEvoy, C. R. E., Van Helden, P. D., and Victor, T. C. (2009). A Balancing Act: Efflux/Influx in Mycobacterial Drug Resistance. *Antimicrob. Agents Chemother.* 53 (8), 3181–3189. doi:10.1128/AAC.01577-08

Machado, D., Isabel, C. B., João, P., Liliana, P., Pedro, B., Isabel, J. B., Bruno, V., et al. (2012). Contribution of efflux to the emergence of isoniazid and multidrug resistance in Mycobacterium tuberculosis. *PLoS One* 7 (4), e34538. doi:10.1371/journal.pone.0034538

Malik, S., Willby, M., Sikes, D., Tsodikov, O. V., and Posey, J. E. (2012). New Insights into Fluoroquinolone Resistance in *Mycobacterium tuberculosis*: Functional Genetic Analysis of gyrA and gyrB Mutations. *PloS one* 7 (6), e39754. doi:10.1371/journal.pone.0039754

Miotto, P., Cabibbe, A. M., Feuerriegel, S., Casali, N., Drobniewski, F., Rodionova, Y., et al. (2014). *Mycobacterium tuberculosis* Pyrazinamide Resistance Determinants: a Multicenter Study. *MBio* 5 (5), e01819. doi:10.1128/mBio.01819-14

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30 (16), 2785–2791. doi:10.1002/jcc.21256

Mugumbate, G., Abrahams, K. A., Cox, J. A. G., Papadatos, G., van Westen, G., Lelièvre, J., et al. (2015). Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and *In Vitro* Validation. *PloS one* 10, 3e0121492. doi:10.1371/journal.pone.0121492

Muzondiwa, D., Mutshembele, A., Pierneef, R. E., and Reva, O. N. (2020). Resistance Sniffer: an Online Tool for Prediction of Drug Resistance Patterns of *Mycobacterium tuberculosis* Isolates Using Next Generation Sequencing Data. *Int. J. Med. Microbiol.* 310 (2), 151399. doi:10.1016/j.ijmm.2020.151399

Nakatani, Y., Opel-Reading, H. K., Merker, M., Machado, D., Andres, S., Kumar, S. S., et al. (2017). Role of Alanine Racemase Mutations in *Mycobacterium* *tuberculosis* D -Cycloserine Resistance. *Antimicrob. Agents Chemother.* 61, e01575–17. doi:10.1128/AAC.01575-17

Nayak, T., Jena, L., and Bc, H. (2017). "Austin Tuberculosis : Research & Treatment Isoniazid Drug Resistance : Computational Study to Understand the Mechanism of over Expressed UDP- Galactopyranose Mutase Enzyme in Causing Drug Resistance in Tuberculosis. *Austin Tuberculosis: Res. Treat.* 2 (1), 2–7. Available at: https://austinpublishinggroup.com/tuberculosis/fulltext/atrt-v2-id1006.php. doi:10.4103/ijmy.ijmy_174_17

Palomino, J. C., and Martin, A. (2014). Drug Resistance Mechanisms in Mycobacterium Tuberculosis. *Antibiotics* 3 (3), 317–340. doi:10.3390/antibiotics3030317

Pandurangan, A. P., and Blundell, T. L. (2020). Prediction of Impacts of Mutations on Protein Structure and Interactions: SDM, a Statistical Approach, and mCSM, Using Machine Learning. *Protein Sci.* 29 (1), 247–257. doi:10.1002/pro.3774

Pang, Y., Lu, J., Wang, Y., Song, Y., Wang, S., and Zhao, Y. (2013). Study of the Rifampin Monoresistance Mechanism in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 57 (2), 893–900. doi:10.1128/AAC.01024-12

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera? A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25 (13), 1605–1612. doi:10.1002/jcc.20084

Phelan, J., Coll, F., McNerney, R., Ascher, D. B., Pires, D. E. V., Furnham, N., et al. (2016). *Mycobacterium tuberculosis* Whole Genome Sequencing and Protein Structure Modelling Provides Insights into Anti-tuberculosis Drug Resistance. *BMC Med.* 14 (1), 1–13. doi:10.1186/s12916-016-0575-9

Piersimoni, C., Mustazzolu, A., Giannoni, F., Bornigia, S., Gherardi, G., and Fattorini, L. (2013). Prevention of False Resistance Results Obtained in Testing the Susceptibility of Mycobacterium Tuberculosis to Pyrazinamide with the Bactec MGIT 960 System Using a Reduced Inoculum. *J. Clin. Microbiol.* 51 (1), 291–294. doi:10.1128/JCM.01838-12

Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G., and Furnham, N. (2018). Understanding Molecular Consequences of Putative Drug Resistant Mutations in Mycobacterium Tuberculosis. *Sci. Rep.* 8 (1), 1–12. doi:10.1038/s41598-018-33370-6

Ruiz, P., Rodríguez-Cano, F., Zerolo, F. J., and Casal, M. (2002). Investigation of the In Vitro Activity of Streptomycin Against Mycobacterium Tuberculosis. *Microb. Drug Resist.* 8 (2), 147–149. doi:10.1089/107662902760190707

Sandgren, A., Michael, S., Preetika, M., Brian, K. W., George, M. C., Megan, B. M., et al. (2009). Tuberculosis drug resistance mutation database. *PLoS Med.* 6 (2), e1000002. doi:10.1371/journal.pmed.1000002

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX Web Server: an Online Force Field. *Nucleic Acids Res.* 33, W382–W388. doi:10.1093/nar/gki387

Shinnick, T. M., Iademarco, M. F., and Ridderhof, J. C. (2005). National Plan for Reliable Tuberculosis Laboratory Services Using a Systems Approach. Recommendations from CDC and the Association of Public Health Laboratories Task Force on Tuberculosis Laboratory Services. *MMWR. Recomm. Rep: Morbidity Mortality Weekly Rep.* 54 (RR-6), 1–12. Recommendations and Reports/Centers for Disease Control. Available at: http://europepmc.org/abstract/MED/15829862.

Singh, A., Grover, S., Sinha, S., Das, M., Somvanshi, P., and Grover, A. (2017). Mechanistic Principles behind Molecular Mechanism of Rifampicin Resistance in Mutant RNA Polymerase Beta Subunit of Mycobacterium Tuberculosis. *J. Cel. Biochem.* 118 (12), 4594–4606. doi:10.1002/jcb.26124

Stop, T. B. (2015). *Stop TB Partnership's Global Drug Facility (GDF) Achieves Historic price Reduction for MDR-TB Drug Cycloserine*. Geneva, Switzerland: Stop TB Partnership.

Studio, Discovery (2008). Dassault systemes BIOVIA, Discovery studio modelling environment, Release 4.5. Accelrys Softw Inc, 98–104. doi:10.4016/8372.01

Timmins, G. S., and Deretic, V. (2006). Mechanisms of Action of Isoniazid. *Mol. Microbiol.* 62 (5), 1220–1227. doi:10.1111/j.1365-2958.2006.05467.x

Uddin, M. K. M., Rahman, A., Ather, M. F., Ahmed, T., Rahman, S. M. M., Ahmed, S., et al. (2020). Distribution and Frequency of rpoB Mutations Detected by Xpert MTB/RIF Assay Among Beijing and Non-Beijing Rifampicin Resistant *Mycobacterium tuberculosis* Isolates in Bangladesh. *Idr* 13, 789–797. doi:10.2147/IDR.S240408

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* 26 (16), 1701–1718. doi:10.1002/jcc.20291

Waman, V. P., Vedithi, S. C., Thomas, S. E., Bannerman, B. P., Munir, A., Skwark, M. J., et al. (2019). Mycobacterial Genomics and Structural Bioinformatics: Opportunities and Challenges in Drug Discovery. *Emerg. Microbes.Infect.* 8, 109–118. doi:10.1080/22221751.2018.1561158

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics* 54 (1), 5–6. doi:10.1002/cpbi.3

Whalen, C. C. (2006). Failure of Directly Observed Treatment for Tuberculosis in Africa: A Call for New Approaches. *Clin. Infect. Dis.* 42 (7), 1048–1050. doi:10.1086/501022

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2006). Database Resources of the National center for Biotechnology Information. *Nucleic Acids Res.* 34, D173–D180. doi:10.1093/nar/gkj158

Wolff, K. A., and Nguyen., L. (2012). Strategies for Potentiation of Ethionamide and Folate Antagonists against Mycobacterium Tuberculosis. *Expert Rev. anti-infective Ther.* 10, 971–981. doi:10.1586/eri.12.87

Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W., et al. (2019). DeepAMR for Predicting Co-occurrent Resistance of *Mycobacterium tuberculosis*. *Bioinformatics* 35, 3240–32493249. doi:10.1093/bioinformatics/btz067

Zhang, Q., An, X., Liu, H., Wang, S., Xiao, T., and Liu, H. (2019). Uncovering the Resistance Mechanism of Mycobacterium Tuberculosis to Rifampicin Due to RNA Polymerase H451D/Y/R Mutations from Computational Perspective. *Front. Chem.* 7 (December), 1–13. doi:10.3389/fchem.2019.00819

Zhang, Y., Shi, W., Zhang, W., and Mitchison, D. (2014). Mechanisms of Pyrazinamide Action and Resistance. *Microbiol. Spectr.* 2 (4), 1–12. doi:10.1128/microbiolspec.mgm2-0023-2013

Zhang, Y., and Yew, W.-W. (2015). Mechanisms of Drug Resistance in *Mycobacterium tuberculosis*: Update 2015. *Int. J. Tuberculosis Lung Dis.* 19, 1276–1289. doi:10.5588/ijtld.15.0389

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership