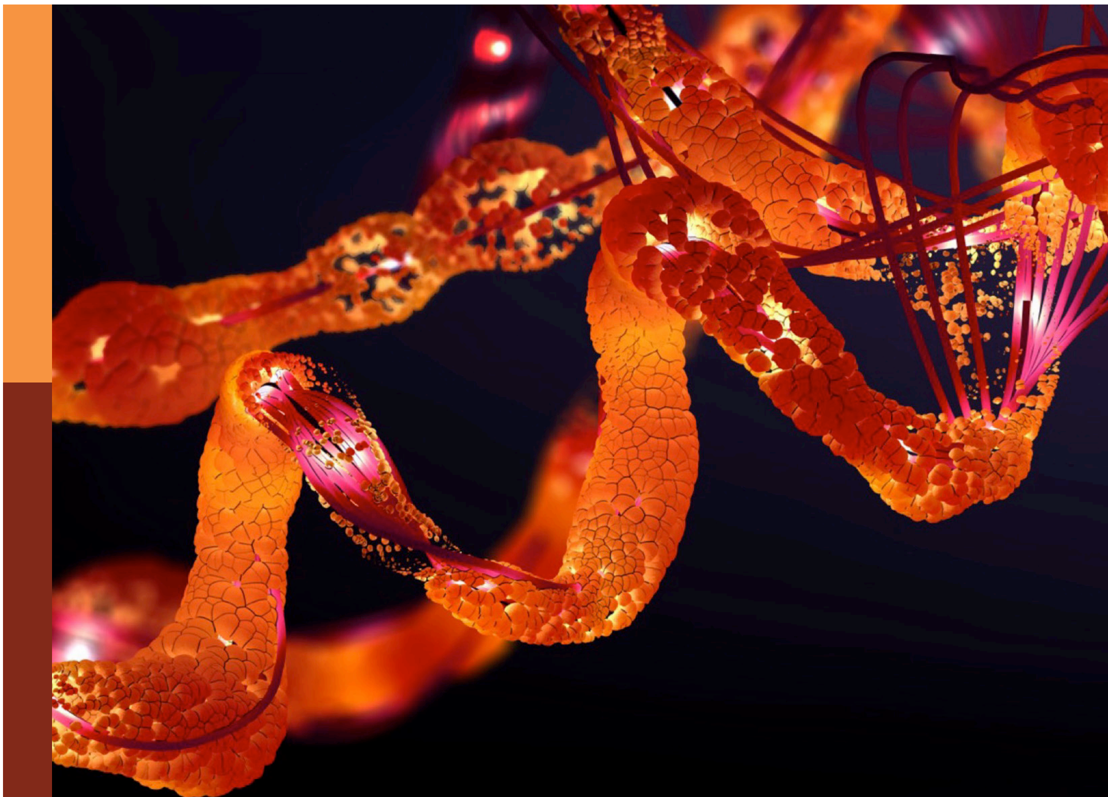
frontiers | Research Topics

Computational and experimental protein variant interpretation in the era of precision medicine

Edited by
Daniele Dell'Orco, Valerio Consalvi, Silvia Morante,
Paola Turina, Tiziana Sanavia, Arthur M. Lesk
and Constantina Bakolitsa

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-8325-4371-9
DOI 10.3389/978-2-8325-4371-9

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 and experimental protein variant interpretation in the era of precision medicine

Topic editors

Daniele Dell'Orco — University of Verona, Italy
Valerio Consalvi — Sapienza University of Rome, Italy
Silvia Morante — University of Rome Tor Vergata, Italy
Paola Turina — University of Bologna, Italy
Tiziana Sanavia — University of Torino, Italy
Arthur M. Lesk — The Pennsylvania State University (PSU), United States
Constantina Bakolitsa — University of California, Berkeley, United States

Citation

Dell'Orco, D., Consalvi, V., Morante, S., Turina, P., Sanavia, T., Lesk, A. M., Bakolitsa, C., eds. (2024). *Computational and experimental protein variant interpretation in the era of precision medicine*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4371-9

Table of contents

TYPE Editorial
PUBLISHED 15 January 2024
DOI 10.3389/fmolb.2024.1363813

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Andrea Mozzarelli,
University of Parma, Italy

*CORRESPONDENCE
Tiziana Sanavia,
✉ tiziana.sanavia@unito.it

RECEIVED 31 December 2023
ACCEPTED 04 January 2024
PUBLISHED 15 January 2024

CITATION
Sanavia T, Turina P, Morante S, Consalvi V, Lesk AM, Bakolitsa C and Dell'Orco D (2024), Editorial: Computational and experimental protein variant interpretation in the era of precision medicine.
Front. Mol. Biosci. 11:1363813.
doi: 10.3389/fmolb.2024.1363813

Editorial: Computational and experimental protein variant interpretation in the era of precision medicine

Tiziana Sanavia[1]*, Paola Turina[2], Silvia Morante[3,4], Valerio Consalvi[5], Arthur M. Lesk[6], Constantina Bakolitsa[7] and Daniele Dell'Orco[8]

[1]Department of Medical Sciences, Computational Biomedicine Unit, University of Torino, Torino, Italy, [2]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [3]Department of Physics, University of Roma Tor Vergata, Roma, Italy, [4]Istituto Nazionale di Fisica Nucleare, University of Roma Tor Vergata, Roma, Italy, [5]Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza University of Roma, Roma, Italy, [6]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, United States, [7]Department of Plant and Microbial Biology and Center for Computational Biology, University of California, Berkeley, Berkeley, CA, United States, [8]Department of Neurosciences, Biomedicine and Movement Sciences, Section of Biological Chemistry, University of Verona, Verona, Italy

KEYWORDS

single-point mutation, protein stability, genotype-phenotype correlation, molecular dynamics simulation, kinetics, pathogenicity, variant interpretation, machine learning

Editorial on the Research Topic
Computational and experimental protein variant interpretation in the era of precision medicine

Most traits of the human phenotype depend on the combination of various genetic factors with environmental influences, and a major challenge is the understanding of the relationship among genetic and phenotype variations (Casadio et al., 2011). In the last years, both advancements in human genome sequencing technologies and the creation of databases collecting information on human variations at the gene and protein levels have hugely enhanced the investigations on the role of these variations in determining health and disease (Austin-Tse et al., 2022). At the same time, the increasing amount of data generated by these resources are requiring new accurate and reliable computer-aided tools to predict phenotype–genotype associations (Brandes et al., 2023; Cheng et al., 2023).

Efficient and powerful analytical methods are necessary for the discovery of unknown etiologies, which is important for rare diseases (Greene et al., 2023). Licata et al. highlighted the most relevant online resources and computational tools for single-nucleotide variant interpretation that can enhance the diagnosis, clinical management, and development of treatments for rare disorders.

A large number of computational methods have been developed for the identification of potentially pathogenic missense mutations. An example provided in this Research Topic is PON-All, a machine learning tool that exploits features of evolutionary conservation, changes in physicochemical properties of amino acids, and biological function annotations from Gene Ontology. The novelty, introduced by Yang et al. with this tool, was to improve the variant interpretation through the inclusion of non-human variants in the learning process, achieving high accuracy on a blind test set.

A complementary approach to understand the effects of missense mutations is by computational predictors of stability. Protein stability perturbations have already been associated with pathogenic missense variants, and they have been shown to significantly contribute to the loss of function in haploinsufficient genes (Birolo et al., 2021). The effects of these variants on protein stability can be measured as the difference in the free energy change of unfolding ($\Delta\Delta G$) between the mutated protein and its wild-type form. Predicting protein stability changes upon genetic variations is still an open challenge (Rollo et al., 2023). Current tools, which can either require the knowledge of the protein tertiary structure or rely on protein sequences only (Pancotti et al., 2021), are less accurate in predicting stabilizing variations than destabilizing ones (Pancotti et al., 2022). Benevenuta et al. investigated possible reasons for such a difference by focusing on the relationship between experimentally measured $\Delta\Delta G$ and some protein properties (protein structural information, different physical properties, and statistical potentials). The results highlighted both the need to design predictive methods able to exploit input features highly correlated with the stabilizing variants and the importance of evaluating these tools on stabilizing, neutral, and destabilizing variants separately. Since this classification is associated with the sign of the protein melting temperature variation, Nobili et al. proposed a full atomistic protein description able to improve the estimation of free energy by modeling its change as a function of the number of hydrogen bonds computed using well-tempered metadynamics and maximal constrained entropy. The authors found a good agreement in the sign of representative values of $\Delta\Delta G$ upon unfolding and the sign of the shift in the melting temperature compared to experimental results.

Comparing experimental characterization and computational predictions, Pacheco-Garcia et al. investigated naturally occurring variants of NAD(P)H:quinone oxidoreductase 1 (NQO1), a multifactorial protein associated with an increased risk of developing cancer and neurological disorders. The authors used computational tools to probe 5,187 variants, and the effects of the clinically relevant missense NQO1 variants were then experimentally characterized in terms of protein levels during bacterial expression, solubility, thermal stability, and coenzyme binding.

Disease-causing variants are supposed to directly affect experimentally measurable features, such as protein function and stability, and the kinetics and thermodynamics of protein–protein recognition, interaction, and binding. Morante et al. used circular dichroism, fluorescence spectroscopy, and melting temperature measurements to investigate key structural aspects of the interaction between wild-type frataxin and some of its variants found in cancer tissues upon Co2+ binding, highlighting the peculiar role of the N-terminal disordered tail in modulating the protein ability to interact with the metal. Dal Cortivo et al. provided a comprehensive biophysical investigation of calmodulin (CaM) by assessing structural, thermodynamic, and kinetic properties of protein–peptide interactions, involving two protein variants associated with congenital arrhythmia (N97I and Q135P) and a highly conserved CaM-binding region in ryanodine receptors RyR1 and RyR2. Specifically, the integration of spectroscopic investigation with molecular dynamics (MD) simulations and protein structure network analysis showed that these disease-associated CaM mutations alter CaM selectivity for the specific RyR channel.

The impact of MD simulations in molecular biology and drug discovery has expanded dramatically in recent years (Hollingsworth and Dror, 2018) since they capture the behavior of proteins and other biomolecules in full atomic detail and at a very fine temporal resolution. Shinwari et al. applied MD to characterize the structural and functional impacts of high-risk non-synonymous single-nucleotide polymorphisms on the TCIRG1 protein, causing congenital neutropenia and osteopetrosis. The analysis identified 15 variants that are likely to be highly deleterious, significantly destabilizing the wild-type protein structure and function. Hashimi et al. systematically investigated the dynamic properties involved in the double-stranded DNA (dsDNA) recognition by the EBNA1 protein, a key nuclear antigen of Epstein–Barr virus (EBV). Stability, flexibility, structural compactness, hydrogen bonding frequency, and binding affinity were altered by disrupting the native protein–DNA contacts, thereby decreasing the binding affinity. Their results revealed hotspot residues (arginine substitutions R521A and R522A), which are likely to become crucial in designing structure-based drugs against EBV infections.

In conclusion, this Research Topic has provided an overview of the current progress in both computational and experimental research and of their interplay in the annotation and interpretation of protein variants to detect pathogenic variations, analyzing their effects at the molecular level. These studies may help to predict the risk of developing specific diseases, the susceptibility to environmental factors, and the personal response to specific drugs.

Author contributions

TS: Writing–original draft, Writing–review and editing. PT: Writing–review and editing. SM: Writing–review and editing. VC: Writing–review and editing. AL: Writing–review and editing. CB: Writing–review and editing. DDO: Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work has received financial support from the Italian Ministry of University and Research (PRIN 2017, 201744NR8S).

Acknowledgments

TS thanks the EU projects GenoMed4All (H2020 RIA n.101017549) and BRAINTEASER (H2020 RIA n.101017598).

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 author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

References

Austin-Tse, C. A., Jobanputra, V., Perry, D. L., Bick, D., Taft, R. J., Venner, E., et al. (2022). Best practices for the interpretation and reporting of clinical whole genome sequencing. *npj Genomic Med.* 7, 27–13. doi:10.1038/s41525-022-00295-z

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

Brandes, N., Goldman, G., Wang, C. H., Ye, C. J., and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* 55, 1512–1522. doi:10.1038/s41588-023-01465-0

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, 1161–1170. doi:10.1002/humu.21555

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381. eadg7492. doi:10.1126/science.adg7492

Greene, D., Pirri, D., Frudd, K., Sackey, E., Al-Owain, M., Giese, A. P. J., et al. (2023). Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat. Med.* 29, 679–688. doi:10.1038/s41591-023-02211-z

Hollingsworth, S. A., and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron* 99, 1129–1143. doi:10.1016/j.neuron.2018.08.011

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., et al. (2022). Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* 23, bbab555. bbab555. doi:10.1093/bib/bbab555

Pancotti, C., Benevenuta, S., Repetto, V., Birolo, G., Capriotti, E., Sanavia, T., et al. (2021). A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes* 12, 911. doi:10.3390/genes12060911

Rollo, C., Pancotti, C., Birolo, G., Rossi, I., Sanavia, T., and Fariselli, P. (2023). Influence of model structures on predictors of protein stability changes from single-point mutations. *Genes* 14, 2228. doi:10.3390/genes14122228

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/fmolb.2021.664436

Structural and Biophysical Investigation of the Key Hotspots on the Surface of Epstein–Barr Nuclear Antigen 1 Essential for DNA Recognition and Pathogenesis

Huma Farooque Hashmi[1]*, *Muhammad Waseem*[2], *Syed Shujait Ali*[3], *Zahid Hussain*[3] *and Kaoshan Chen*[1]

[1]College of Life Sciences, Shandong University, Jinan, China, [2]Faculty of Rehabilitation and Allied Health Science, Riphah International University, Islamabad, Pakistan, [3]Center for Biotechnology and Microbiology, University of Swat, Swat, Pakistan

Epstein-Barr Virus (EBV) is considered the most important human pathogen due to its role in infections and cellular malignancies. It has been reported that this Oncolytic virus infects 90% world's population. EBNA1 is required for DNA binding and survival of the virus and is considered an essential drug target. The biochemical and structural properties of this protein are known, but it is still unclear which residues impart a critical role in the recognition of dsDNA. Intending to disclose only the essential residues in recognition of dsDNA, this study used a computational pipeline to generate an alanine mutant of each interacting residue and determine the impact on the binding. Our analysis revealed that R469A, K514A, Y518A, R521A and R522A are the key hotspots for the recognition of dsDNA by the EBNA1. The dynamics properties, i.e. stability, flexibility, structural compactness, hydrogen bonding frequency, binding affinity, are altered by disrupting the protein-DNA contacts, thereby decreases the binding affinity. In particular, the two arginine substitution, R521A and R522A, significantly affected the total binding energy. Thus, we hypothesize that these residues impart a critical role in the dsDNA recognition and pathogenesis. This study would help to design structure-based drugs against the EBV infections.

Keywords: EBNA1, DNA recognition, MD simulation, in silico mutagenesis, free energy calculations

OPEN ACCESS

Edited by:
Qing-Chuan Zheng,
Jilin University, China

Reviewed by:
Xuanyu Meng,
Soochow University, China
Lin Chen,
Qiqihar University, China

****Correspondence:***
Huma Farooque Hashmi
humahashmi@bs.qau.edu.pk

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

Received: *05 February 2021*
Accepted: *14 June 2021*
Published: *29 June 2021*

Citation:
Hashmi HF, Waseem M, Ali SS, Hussain Z and Chen K (2021) Structural and Biophysical Investigation of the Key Hotspots on the Surface of Epstein–Barr Nuclear Antigen 1 Essential for DNA Recognition and Pathogenesis. Front. Mol. Biosci. 8:664436. doi: 10.3389/fmolb.2021.664436

INTRODUCTION

Due to the prevalent infections instigated by Herpes viruses, it is considered as an important virus in human pathogens flora. This diverse pathogenic flora Epstein-Barr Virus (EBV) is regarded as the most important human pathogen due to its role in infections and cellular malignancies (Bochkarev et al., 1996; Yasuda et al., 2011) It has been reported that this Oncolytic virus infects 90% world's population. Immortalization of B lymphocytes accompanies the main EBV infection and stimulates them to replicate as lymphoblastic cell lines (Garai-Ibabe et al., 2012). Alongside the B lymphocytes infection, EBV also causes infectious mononucleosis by targeting the epithelial cells (Pope et al., 1968; Haque and Crawford, 1996). Nasopharyngeal carcinoma (NFC), muscle cell sarcoma and gastric carcinomas (GaCa), Hodgkin's lymphoma (HL), Burkitt's lymphoma (BL), extranodal lymphoma of T/NK cell origin and post-transplant lymphoproliferative disease (PT-LPD) are among the EBV associated diseases (Taylor et al., 2015). EBV caused tumors stores the viral genome as a multi-copy episome in the nucleus of infected cells (Tikhmyanova et al., 2014). During the latent infection, progenitor virions are not reproduced, but alternatively, a set of genes essential for survival and proliferation are expressed (Matsuura et al., 2010). Epstein-Barr Nuclear Antigen 1 (EBNA1) acts to preserve the latent viral genome in proliferating cells (Okano, 1998). This protein is expressed in the malignant cell and sustain the proliferation (Leight and Sugden, 2000).

The EBV genome encodes ~100 genes, among which EBNA1 is the key nuclear antigen that works with the other five others (Hiraku et al., 2014). EBNA1 is almost detected in every kind of infection induced by the EBV in both latent and lytic infections (Gianti et al., 2016). This essential antigen is reported to be involved in mitotic segregation of episomes, replication, reactivation, viral transcription, and lytic infection of EBV (Young and Rickinson, 2004). It has also been reported that EBNA1 possess a similar structure to that of human papillomavirus (HPV) E2 protein and the Kaposi's Sarcoma Associated Herpesvirus (KSHV) LANA protein (Garber et al., 2002). In addition to structural similarity, these proteins are reported to have similar function, i.e. DNA binding and episome regulation (Wang et al., 2006). The biochemical and structural properties of this protein are known. EBNA1 works as a dimer with two functional domains (Pope et al., 1968; Haque and Crawford, 1996). The two terminals CTD (carboxy-terminal DNA-binding domain) and ATCTD (amino-terminal chromosome tethering domain) bind the 18ps DNA to initiate the plasmid and viral genome replication (Lindner and Sugden, 2007). Due to the multi-faceted role of EBNA1, it is the primary drug target for the treatment of EBV associated infections.

The crystallographic structure of ENBA1 has been solved, and reported that 459–607 residues at C-terminal are required for DNA binding (Bochkarev et al., 1998). Previous studies determining the dissociation constant (KD) for the EBNA1-DNA association reported that mutating interacting residues, R469A reduced the binding of DNA by 300-fold, Y518A by 80-fold and R522A by 1600-fold (Cruickshank et al., 2000). Additionally, other studies also reported that K514A also reduces the binding affinity significantly. At the same time, others reported that three residues R491E, R491A, and D581E significantly impair the DNA binding (Morgunova et al., 2015). These studies mutated only selected residues, while the impact of others remains a question. To understand each residue's impact and reveal only a few residues that are required explicitly for DNA recognition while others are supplementary interactions, an in-depth investigation is needed. To disclose only the essential residues, this study used a computational pipeline to generate an alanine mutant of each interacting residue and determine the impact on the binding. Highly destabilizing and affinity reducing mutations were subjected to biophysical investigation to reveal their real effect on the binding. Our analysis would help to target the critical hotspots for future rational and structure-based drug designing to curtail the EBV associated lytic and latent infections.

MATERIALS AND METHODS

Structure Retrieval and Preparation

The RCSB protein databank (http://www.rcsb.org/) repository was accessed for structural retrieval. Structural deformities were detected and addressed (Rose et al., 2016). The missing hydrogens were added, and partial charges were assigned. The structure was also analyzed for structural breaks and unknown residues. The structure was minimized and prepared before *in silico* mutagenesis and molecular docking.

Epstein-Barr Nuclear Antigen 1-DNA Docking

For the docking HADDOCK (High Ambiguity Driven protein-protein Docking) (Dominguez et al., 2003) was utilized. It uses biochemical and structural data to drive the docking process. The Guru interface with approximately 500 features considered as the best to predict the docking poses. Using default parameter i.e. lowest intermolecular energies, the best structural complex was extracted. We also used NPDock (Tuszynska et al., 2015), an online web server, which uses the scoring of poses, clustering of the best-scored models, and refinement of the most promising solutions to give the best results. The best scoring complex was retrieved from NPDock and analyzed. To determine the interaction of different residues with the DNA DNAproDB (Sagendorf et al., 2020) was used to extract the interactions from DNA-nucleic acid complexes.

Interface Analysis and Mutants Library Construction

Using the machine learning protocol implemented in MOE (Vilar et al., 2008) the Alanine scanning approach was applied to compute the impact of each residue in the interaction with the DNA. The dAffinity and dStability parameters are essential considerations in the ASM (alanine scanning module) which determine the relative stability and affinity changes upon substitution. The detailed mechanism of this alanine scanning mutagenesis and residue scan approaches has been discussed previously (Junaid et al., 2019). Furthermore, we also used mCSM-NA an online webserver, for the affinity changes prediction upon the alanine substitution uses the graph-based signature model. Residues with high-affinity changes were subjected to molecular dynamics simulation investigation.

Molecular Dynamics Simulation

To further provide deep insight into the stability and affinity changes upon the alanine substitution, the dynamic features of each complex was determined using the AMBER 20 simulation package. For protein ff14SB, while for DNA, the OLS3 force field was utilized (Salomon-Ferrer et al., 2013). With the TIP3P water model containing 9,784 water molecules, each complex was solvated at 10.0 Å. A total of 29 sodium ions were added to neutralize each system. Multistep energy minimization each 6,000 steps and 3,000 steps of conjugate gradient minimization were completed. Keeping the heating parameters default 300 K for 200 ps, each complex was heated. For density equilibration, using weak restraint for 2 ns at constant pressure was executed. Finally, 200 ns MD using constant pressure was achieved. Langevin thermostat with 1 atm pressure and 300 K for temperature control (Zwanzig, 1973), while Particle Mesh Ewald (PME) algorithm to evaluate long-range interactions respectively (Ryckaert et al., 1977; Roe and Cheatham, 2013) with the cutoff, distances 10 Å. For the covalent bonds involving hydrogen, the SHAKE algorithm was used (Ryckaert et al., 1977). All the simulations were GPU accelerated.

Post-Simulation Analyses

The thermodynamics state function, i.e. RMSD, residual flexibility, i.e. RMSF, structural compactness, i.e. radius of gyration (Rg) and the total number of hydrogen bonds over the simulation were computed by using CPPTRAJ and PTRAJ modules integrated with AMBER (Roe and Cheatham, 2013).

Binding Affinity Calculations

To connect the alanine mutations with the binding affinity changes, the binding free energy of each alanine substituted complex was determined. The free energy scoring function (MMGBSA) is an extensively used approach to evaluate the free energy of a protein-ligand, protein-protein and protein-nucleic acids (Khan et al., 2018; Ali et al., 2019; Khan et al., 2019; Khan et al., 2020a; Khan et al., 2020b; Khan et al., 2020c; Khan et al., 2020d; Khan et al., 2020e; Khan et al., 2020f; Hussain et al., 2020). It used the following equation to calculate the free energy.

$$\Delta G_{bind} = \Delta G_{complex} - \left[\Delta G_{receptor} + \Delta G_{ligand}\right]$$

Each term in the such as electrostatic, van der Waals interactions, polar and nonpolar were predicted using the following equation:

$$G = G_{bond} + G_{ele} + G_{vdW} + G_{pol} + G_{npol}$$

Clustering of MD trajectories using PCA and Free Energy Landscape (FEL).

An unsupervised learning approach known as Principal Component Analysis to describe the motion of MD trajectories (PCA) (Pearson, 1901; Wold et al., 1987) and gain information about the internal motion of the system using CPPTRAJ. For the eigenvector and their atomic coordinates, the spatial covariance matrix was calculated. A diagonal matrix of eigenvalues was generated using the orthogonal co-ordinate transformation. The Principal Components were derived based on the eigenvectors and eigenvalues. The predominant movements during the simulation were plotted using these PCs. (Balsera et al., 1996; Ernst et al., 2015). Furthermore, Free energy landscape (FEL) was constructed to capture the different energy minima at different simulation time.

RESULTS AND DISCUSSION

Structure Retrieval and Epstein-Barr Nuclear Antigen 1-DNA Docking

Using accession number 5T7X the structure of the EBNA1 was retrieved. The structure is a dimer interface of two EBNA1 chains and 18bps DNA. The structural representation of the EBNA1-DNA complex is shown in **Figure 1A**. HADDOCK predicted the correct docking conformation with the binding energy -295.63 kcal/mol. The interactions predicted by the DNAproDB showed that G462, G463, W464, F465, R469, N475, K477, F478, R491, K514, Y518, R521 and R522 are involved in interaction with the DNA. The 3D interaction of these residues with the DNA is given in **Figure 1B**. A different number of hydrogen bonds were formed by each residues ranging from one to five at the interface. The specific hydrogen bonding interactions are shown in **Figure 1C**. These residues contributed to the total binding energy. To potentially determine the impact of each of this residue, alanine scanning revealed its impact on the binding of DNA. Among the 13 residues at the interface G462A and G463A increase the binding affinity while the remaining 11 residues decrease the binding affinity at different folds.

As tabulated in **Table 1**, it can be seen that R469A, K514A, Y518A, R521A and R522A significantly affected the binding of DNA as compared to others. In the case of the R469A, the predicted ΔG was reported to −5.784 kcal/mol, while the dAffinity was also predicted to be reduced (−1,009.21 kcal/mol). The predicted ΔG for K514A was reported to be −3.638 kcal/mol, while the dAffinity was reported to be −1,021.32 kcal/mol, respectively. For the Y518A the predicted ΔG was -3.406 kcal/mol; however, the dAffinity was reported to be −1,020.32 kcal/mol. Intriguingly the dAffinity for R521A and R522A was comparable. The predicted dAffinity for R521A and R522A was reported to be −1,009.60 kcal/mol and −1,009.37 kcal/mol, respectively. Furthermore, the predicted ΔG for R521A was −5.866 kcal/mol, while for R522A, it was −6.008 kcal/mol. In the EBNA-DNA co-crystal structure, the three targeted amino acids are oriented toward the DNA but are too far from the nearest H-bond acceptor in the bases (more than 6 Å) to form H-bonds. Hence, these results also show that R469A, K514A, Y518A, R521A and R522A are required for DNA recognition and are the key hotspots for drug discovery. Thus, these residues were selected for further evaluation and subjected to molecular dynamics simulation to understand its dynamics behaviour and reveal its binding energy differences.

Mutation Stability Correlation (Root Mean Square Fluctuation)

To demonstrate the mutation's stability correlation, the thermodynamics state function Root mean square fluctuation (RMSD) of the wild type and the mutant complexes was calculated as a function of time. A 200 ns simulation trajectory

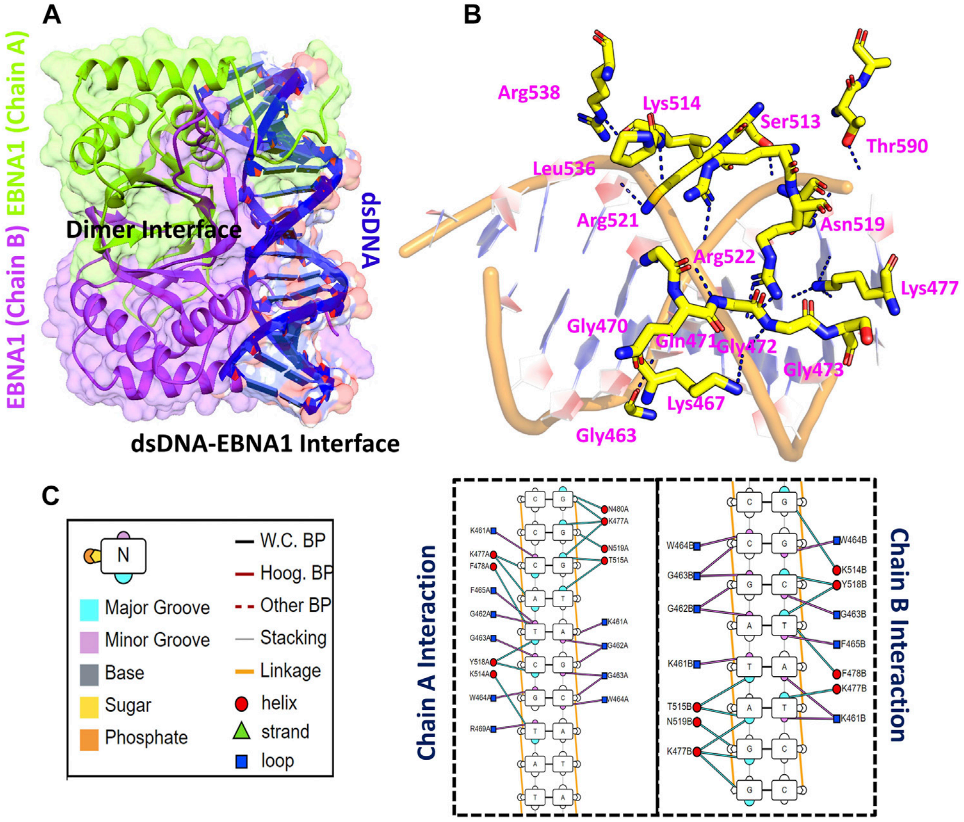


FIGURE 1 | Represent the crystallographic structure of EBNA1-DNA. **(A)** shows the dimer binding to the dsDNA. Both chains are coloured differently. **(B)** shows the 3D interactions of EBNA1 and DNA. **(C)** shows the 2D interactions of and the legend of the interaction pattern. The legend shows the respective interaction between the DNA and EBNA1. The circle in red colour represent the interaction with helix; navy blue colour represent the loop, while the cyan triangle shows the strand. Additionally, the minor groove, major groove and nucleotides are also coloured differently.

TABLE 1 | The table shows the alanine scanning results of the interacting residues. dStability, dAffinity, predicted ΔG and the outcome of each mutation upon substitution is given. Highly affinity reducing mutations are given as bold and were subjected to molecular dynamics simulation-based investigation. All the energies are given in kcal/mol.

Index	Mutant residue	dStability[a]	dAffinity[b]	Predicted ΔG[c]	Outcome
1	G462A	−161.94	−1,297.62	1.482	Increased affinity
2	G463A	−152.55	−1,021.36	1.488	Increased affinity
3	W464A	−153.84	−1,022.16	−1.656	Reduced affinity
4	F465A	−158.65	−1,022.34	−2.878	Reduced affinity
5	**R469A**	**−154.65**	**−1,009.21**	**−5.784**	**Reduced affinity**
6	N475A	−163.71	−1,016.41	−0.592	Reduced affinity
7	K477A	−160.67	−1,020.26	−0.824	Reduced affinity
8	F478A	−161.56	−1,022.69	−2.148	Reduced affinity
9	R491A	−225.26	−1,012.93	−1.986	Reduced affinity
10	**K514A**	**−165.16**	**−1,021.32**	**−3.638**	**Reduced affinity**
11	**Y518A**	**−156.59**	**−1,020.47**	**−3.406**	**Reduced affinity**
12	**R521A**	**−158.75**	**−1,009.60**	**−5.866**	**Reduced affinity**
13	**R522A**	**−156.41**	**−1,009.37**	**−6.008**	**Reduced affinity**

[a] ***dStability*** *= it is the relative stability change upon the mutation. The more the negative the more instable the structure.*

[b] ***dAffinity*** *= it is the relative affinity change between the wild type and mutated complex. Negative dAffinity means the mutation will increase the binding affinity while positive dAffinity mean it will decrease the binding affinity.*

[c] ***Predicted ΔG*** *= it implies a similar formula but different algorithm to calculate the binding differences between the wild type and mutant. ΔG shows change in the binding free energy changes upon the mutation.*

Bold values are Mutations selected for MD simulation.

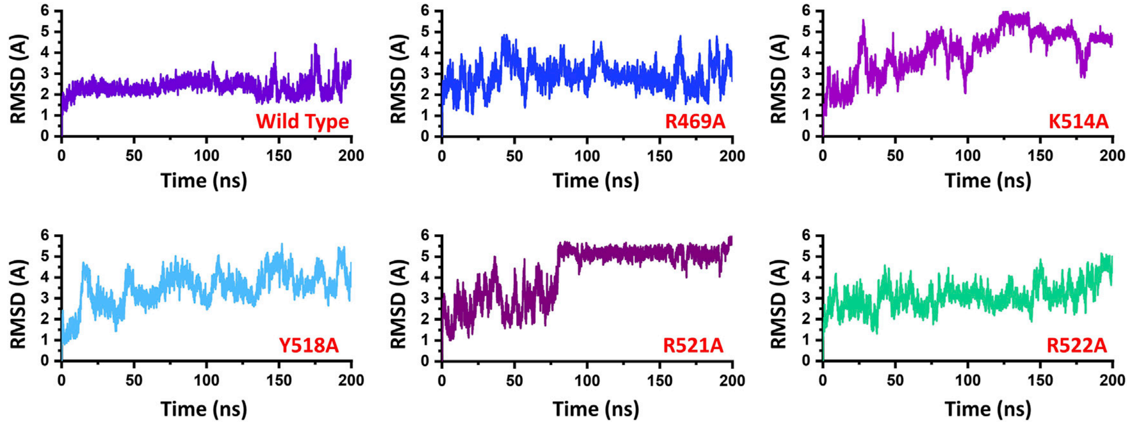


FIGURE 2 | Represent the dynamic stability of EBNA1-DNA bound wild type and mutant complexes. All the complexes are coloured differently and tagged. The *x*-axis shows the time in nanoseconds, while the *y*-axis shows RMSD in Å.

for each complex was analyzed. Results for all the complexes are presented in **Figure 2**. In the case of the wild type, the structure gained stability at 2.0 Å. The structure remained remarkably stable during the simulation. After reaching 150 ns the structure converged, and the RMSD increased which is due to the loop opening and closing surrounded the DNA. It was observed that a loop region between 540–560 deviated from its mean structure significantly and thus the RMSD fluctuated substantially. Furthermore, the terminal of the DNA molecule packs the protein by opening and closing also causes significant structural deviation from its mean position thus causes structural destability. This can be inferred from **Figure 3A** where the loop region in all the complexes fluctuated significantly and **Figures 3B**,C shows the closer look into the loop region which is significantly deviated at different intervals. In the case of the R469A mutation, the complex experiences significant divergence from the initial structure. The equilibrium was never achieved during the 200ns simulation time. During the first 50ns simulation the structure owned significant convergence to the following 100 ns. Between 50 and 150 ns the RMSD remained lower and experienced only one significant convergence at 100 ns. Afterwards, the structure remained unstable until the 200 ns. The average RMSD for R469A was reported to be 3.5 Å. The K514A mutant, which is considered as an important residue for the DNA binding, caused significant perturbation upon the substitution. The complex remained significantly unstable during the 200 ns simulation time. The average RMSD for the first 25 ns remained 3.0 Å. Until the first 25 ns the RMSD remained 2.0 Å; however, a significant convergence was observed abruptly, and this trend continues until 200 ns. On the other hand, the Y518A behaviour was also comparable with the K514A. Significant convergence at different intervals was reported over the simulation and the average RMSD remained 4.0 Å. Furthermore, the two arginine replacements at position R521 and R522 significantly altered the dynamics and interaction of the EBNA1-DNA. These replacements caused significant destabilization of the EBNA1-DNA complex at different interval of the 200 ns. These residues are also reported experimentally to cause significant instability of the complex. The average RMSD remained higher for R521A (5.5 Å), while the average RMSD for R522A remained lower but converged significantly. The RMSD continues to increase during the last 50 ns. Thus these results suggest that the wild type structural topology is required for stable interactions, and the mutation-induced here does not only affect the binding of the complex but also the stability. Hence further study on the impact of the substitution justified the effect of these residues on the binding of dsDNA and its druggability properties.

Residual Flexibility of the Wild Type and Mutant Complexes

Furthermore, to connect the residual flexibility with these substitutions, we estimated RMSF (root mean square fluctuation). The wild and mutant complexes owned comparable flexibility levels. It can be seen that all the structures possess a more similar pattern of flexibility. The residues 460–480, particularly in Y518A and R522A possess more flexible behaviour than the others, which is explained in **Figure 8** that it deviated more than the mean point and the mutations cause an allosteric effect on the flexibility. Significant residual flexibility for region 530–560 can be observed. These results show that complexes possess more rigid structures. The RMSF results for all the complexes are represented in **Figure 3A**. However the loop region which causes structural perturbation and flexibility is shown in **Figures 3B**,C.

Hydrogen Bonding Analysis of the Wild Type and Mutant Complexes

Furthermore, to understand the impact of these substitutions on the total number of hydrogen bonds, we calculated the total number of hydrogen bonds during the 200 ns trajectories and the bonding network between the EBNA1 and DNA. Hydrogen bonding rearrangement was observed during the simulation. Among the key bonding in the wild type R469 residue formed extra two interactions with T11 and A28. Among the others, R522

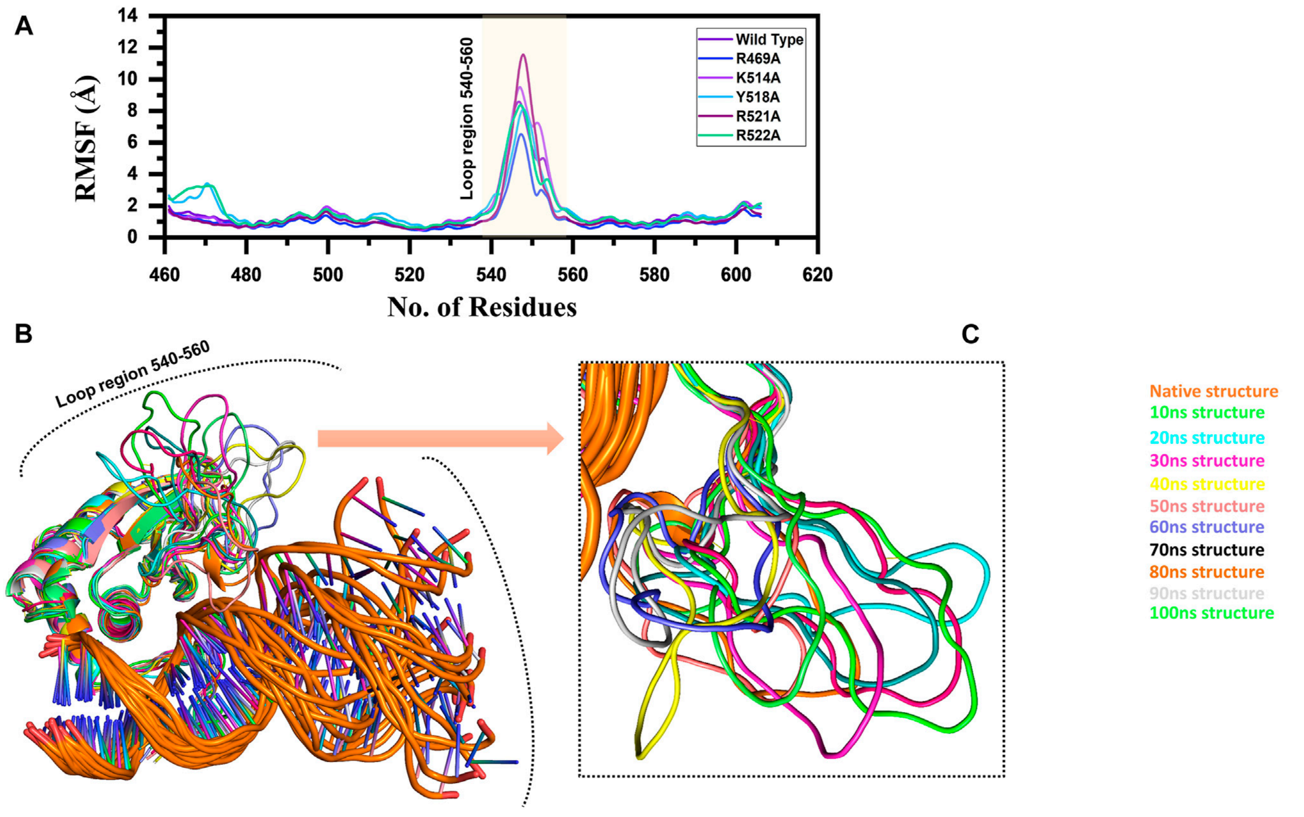


FIGURE 3 | (A) Represent the residual flexibility of EBNA1-DNA bound wild type and mutant complexes. All the complexes are coloured differently and tagged. The *x*-axis shows the total number of residues, while the *y*-axis shows RMSF in Å. **(B)** and **(C)** shows the highly fluctuated regions on the protein's structure.

formed one additional hydrogen bonds with 1.83 Å. Initially, a total of 15 bonds were observed, while after simulation with these three extra interactions formed and the total bonds were reported to be 18 in total in wild type. On the other hand, the R469A lost multiple interactions during the simulation, particularly those formed with R469 residue, consequently remaining 13 hydrogen bonds between EBNA1 and DNA. Among these, Lys461, Arg521 and Arg522 created multiple interactions while the other residues were involved in single interaction only. This shows that in the wild type complex, R469 formed three interactions while those are lost here signifies its role in recognition.

Similarly, only 11 hydrogen bonds were observed in K514A complex. Among the hydrogen bonding interaction, K514 lost its interaction while R521 and R522 also lost three interactions during the simulation. This shows that the mutation has allosterically affected the other residues and destabilized the interaction with the DNA molecule. Moreover, with 12 hydrogen bonds between EBNA1 (Y518A) and DNA complex R469, R521 and R522 lost their multiple interactions which were reported to be sustained in wild type complex. Furthermore, the two essential residues R521A and R522A which significantly contributed to the total binding energy reported in substantial hydrogen bonds reduction between EBNA1 and DNA. In the case of R521A only 10 hydrogen bonds were reported, while only nine bonds were reported between R522A and DNA. In R522A five bonds formed by R521 and R522 are lost. This shows that the two arginine moieties play a significant role in recognition of DNA. After all, the average number of hydrogen bonds were calculated for each complex. As given in **Figure 4**, a significant drop can be observed in the mutant complexes, particularly in the K514A, R521A and R522A complexes. The average number of hydrogen bonds in the wild type was reported to be 98, while in the mutant complexes (R469A), the H-bonds were reported to be 94, (K514A) H-bonds were observed to be 90, Y518A reported 94, while the significant drop was observed in R521A (86) and R522A the total H-bonds were 88. Thus it can be seen that upon substitution, H-bonds count was decreased and thus, these residues potentially act as druggable hotspots. The hydrogen bonding results for all the complexes are represented in **Figure 4**.

Next, to connect the protein conformation changes with the compactness of each complex, we calculated Rg (radius of gyration) as a function of total frames in a trajectory. In the case of the wild, the complex remained more compact than the others. The average Rg for the wild type was reported to be 21.0 Å. In the case of R469A, the same pattern was observed. The results of R469A and wild type was comparable. On the other hand, the K514A the Rg remained higher during the simulation time period. During the first 100 ns. the Rg was observed to be higher, which significantly increased between 100 and 125 ns. Afterwards, the Rg remained uniform. In the case of Y518A, the

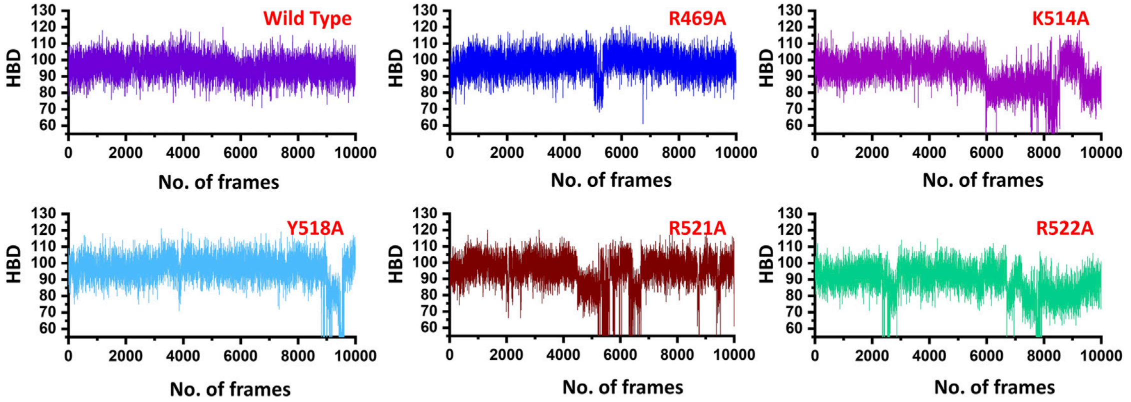


FIGURE 4 | Show the total number of hydrogen bonds in the EBNA1-DNA bound wild type and mutant complexes. All the complexes are coloured differently and tagged. The *x*-axis shows the total number of frames, while the *y*-axis shows the total number of hydrogen bonds. Structural compactness of the wild type and mutant complexes.

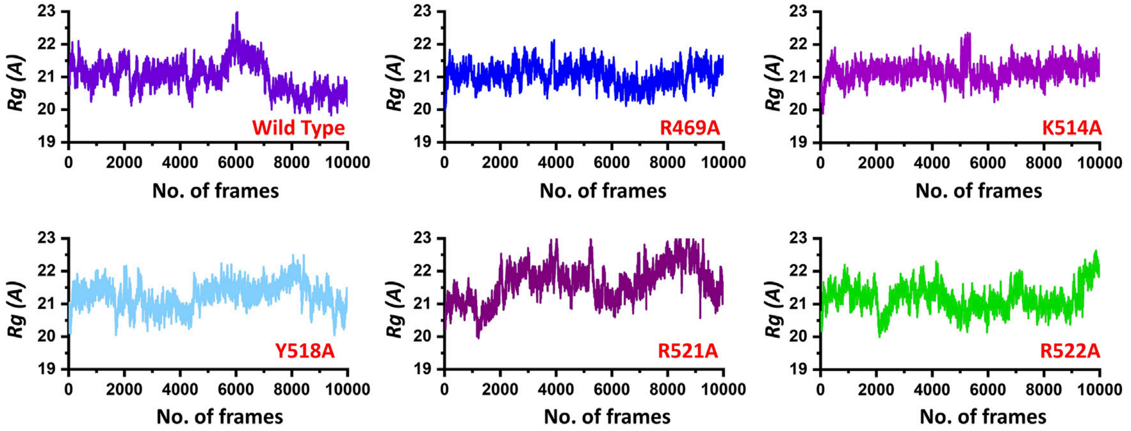


FIGURE 5 | Show the Rg (radius of gyration) of the EBNA1-DNA bound wild type and mutant complexes. All the complexes are coloured differently and tagged. The *x*-axis shows the total number of frames, while the *y*-axis shows *Rg* (radius of gyration).

structural compactness also remained haphazard. Initially, it remained higher but then decreased between 40 and 100 ns while then increased and decreased continuously until the 200 ns. In the case of R521A and R522A, the structural compactness is disrupted significantly. This shows that the loss of structural compactness is due to the binding and unbinding events that occurred during the simulation, and this can be clearly concluded from **Figure 4** as the total number of hydrogen bonds are vary in numbers. Besides the packing of EBNA1 by the DNA terminal and the opening and closing of the loop region 530–560 also demonstrates the compactness variations. The calculated Rg (radius of gyration) results for all the complexes are represented in **Figure 5**.

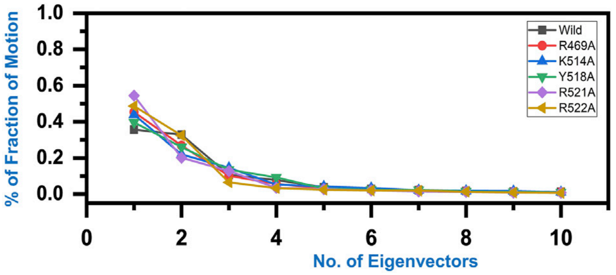


FIGURE 6 | Fraction of the first ten eigenvectors indices generated from the MD trajectory. The percent contribution of each eigenvector is plotted against the corresponding eigenvector.

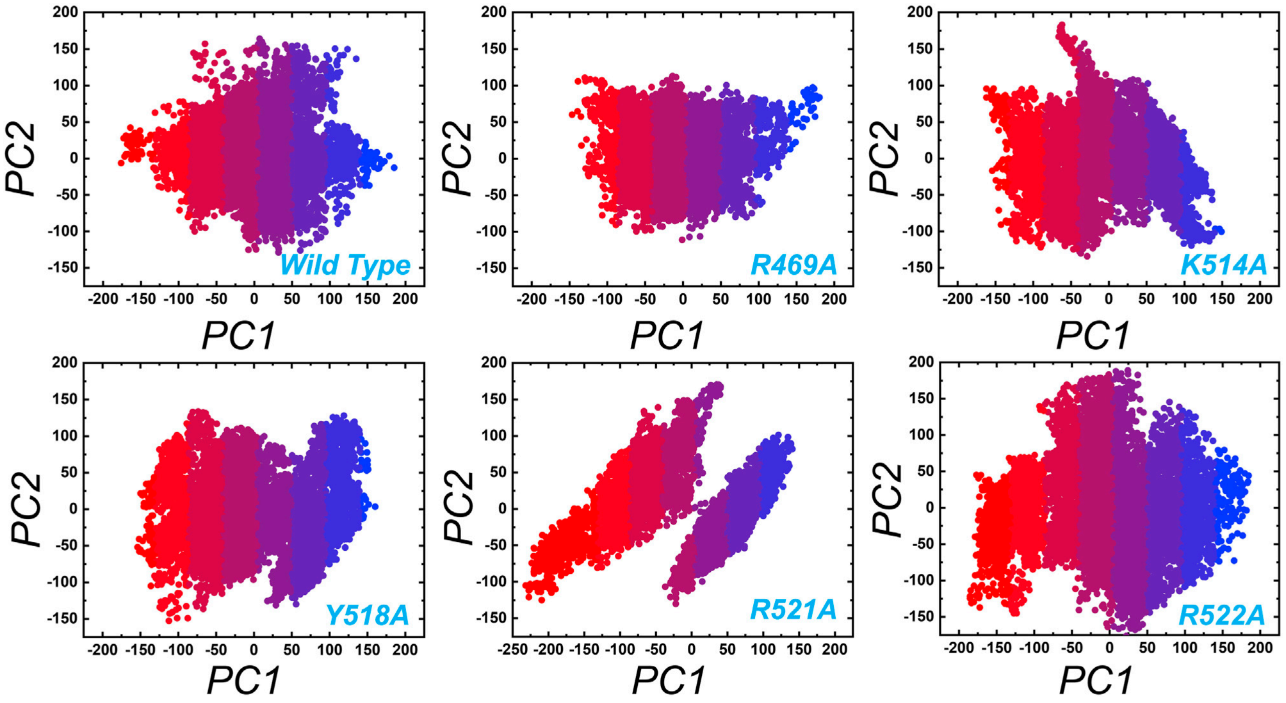


FIGURE 7 | Principal component analysis (PCA) of wild type and the mutant complexes of the EBNA1-DNA. Two PCs i.e. PC1 and PC2 were used for the scattered plot. Each panel represent the respective complex as tagged.

Principal Motions of the Wild Type and Decoy Designed Peptides

Variations in the proteins' trajectories motions were exhibited by each system was captured through PCA. PCA would help to comprehend conformational changes induced variations in the proteins' motion of the wild type and mutant complexes. The internal motion was shown by the first three eigenvectors, while localized fluctuations in the remaining eigenvectors in each complex were observed (**Figure 6**). In the case of wild type peptide complex, the first three eigenvectors contributed 35% variance to the total observed motion, while in R49A, 45%, K514 43%, Y518A 39%, R521A 54% while R522A accounted for 48% variance in motion. This shows the increased motion in the mutated systems and may explain the structural rearrangement due to the mutations in the binding site.

In order to obtain plausibly attributed movements, the first two eigenvectors were projected against one another. The continuous representation of the red to blue colour indicates the transition from one conformation to another over the simulation period. The dots, starting with red and ending in blue, represent each frame. In each complex periodic jumps and continuous overlapping can be observed (**Figure 7**). consequently, all these annotations infer that mutations expressively affected the structure and variations in the internal dynamics of the complexes.

Furthermore, a free energy landscape (FEL) was constructed to relate the structural features and thermodynamics properties. To obtain the energy minima based on probability of given data points of MD trajectories and to map the minimum energy conformation of the all the complexes during the explored time scale, and finally to connect the structural changeovers between these minima. **Figure 8** represent the FEL of all the complexes i.e. wild type, R469A, K514A, Y518A, R521A and R522A. The wild type, K514A, Y518A and R522A shows only one energy minima while R469A and R521A exhibit two lowest energy minima separated by a small subspace, thus explaining global conformational variations adjusted by the mutant complexes in response to mutations. The major variations in these conformations were the loop deviation and nearby beta-sheets conversion. All the variations are highlighted in the **Figure 8**.

Binding Free Energy Calculations

To further connect the protein conformation changes with the binding affinity, we calculated the total binding energy using MM-GBSA approach. This method is considered as the best tool for calculating the real time-binding energy of the biological macromolecules complex. Herein to estimate the impact of the alanine substitutions at a specific position, we estimated the binding free energy. As given in **Table 2**, significant differences in the binding energy can be observed. The electrostatic contribution is significant in each complex. In the wild complex, the total binding free energy was estimated to be −145.18 ± 0.269 kcal/mol. However, in the R469A, this total binding energy was calculated to be −110.75 ± 0.230 kcal/mol. This reduction in the binding energy is due to the disruption of the hydrogen bonding network caused by the alanine

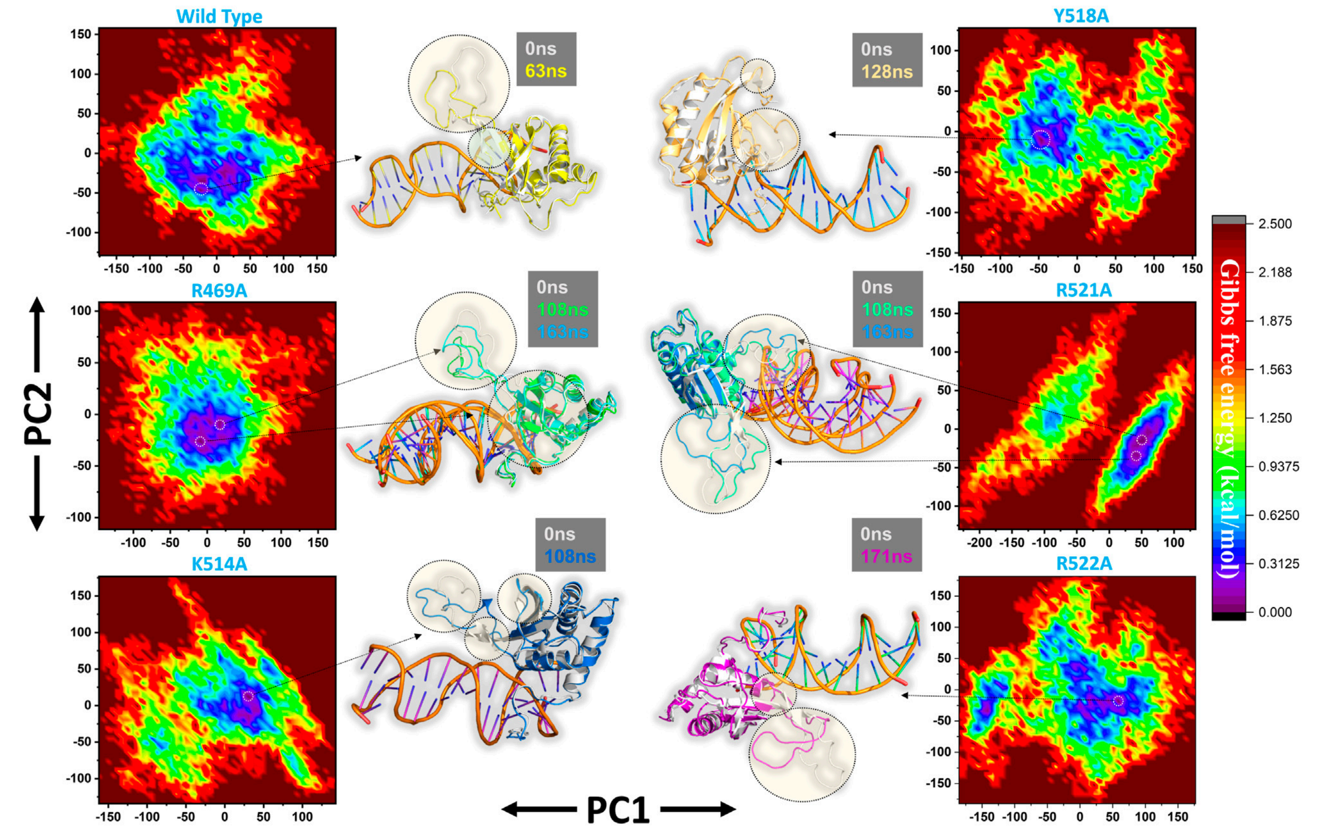


FIGURE 8 | Free Energy Landscape (FEL) of all the complexes i.e. wild type, R469A, K514A, Y518A, R521A and R522A. The first PC1 and second PC2 from the PCA of the backbone carbon were used.

TABLE 2 | Display the total binding energy of the wild type and mutant complexes. Van Der Waal forces, electrostatic energy, generalized born, surface area and the total binding energy values for each complex is given. All the energies are calculated in kcal/mol.

Complex name	vdW	Electrostatic	EGB	ESURF	Total ΔG
Wild type	−132.12 ± 0.142	−4,900.21 ± 2.54	4,904.28 ± 2.43	−17.12 ± 0.017	**−145.18** ± 0.269
R469A	−124.36 ± 0.165	−4,210.70 ± 2.300	4,240.52 ± 2.204	−16.21 ± 0.015	**−110.75** ± 0.230
K514A	−131.48 ± 0.127	−4,331.61 ± 2.0239	4,375.12 ± 1.967	−17.00 ± 0.010	**−104.97** ± 0.175
Y518A	−127.19 ± 0.137	−5,030.86 ± 2.253	5,072.45 ± 2.198	−16.56 ± 0.013	**−102.17** ± 0.190
R521A	−132.93 ± 0.1584	−4,343.05 ± 2.353	4,395.21 ± 2.299	−17.16 ± 0.016	**−97.93** ± 0.226
R522A	−132.66 ± 0.187	−4,420.46 ± 2.280	4,469.95 ± 2.241	−16.87 ± 0.019	**−100.04** ± 0.215

Bold values are the total binding energy.

substitution. A particular interaction formed by the minor groove of DNA with the loop residue R469 causes a significant decline in the total binding energy. The total binding energy results of K514A and Y518A mutant complexes are comparable. For the K514A complex, the total binding energy was reported to be −104.97 ± 0.175 kcal/mol while the total binding energy for Y518A was −102.17 ± 0.190 kcal/mol. This is due to the loss of three hydrogen bonds formed by helix residues with the major groove are diminished upon the substitution. In case of K514A only one hydrogen bond while in case of Y518A, two important hydrogen bonds are lost. Significant drop out was observed in the total binding energy of the R521A mutant complex. The TBE was for R517A was reported to be –97.93 ± 0.226 kcal/mol. On the other hand, the estimated binding energy for R522A was −100.04 ± 0.215 kcal/mol. Overall, these results show that the mutations induced significant energy drop out but the R521A and R522A reduced the binding energy by many folds. Hence these residues play a vital role in recognition of dsDNA and contribute to the infection initiation and progression.

CONCLUSION

In conclusion, herein, we systematically investigated the mechanism of dsDNA recognition by the EBNA1 protein. Our analysis revealed that R469A, K514A, Y518A, R521A and R522A are the key hotspots for drug discovery against the various tumors caused by EBV. In particular, the two arginine substitution R521A and R522A, significantly affected the total binding energy. Thus we hypothesize that these residues impart a critical role in the dsDNA recognition and pathogenesis. This study would help to design structure-based drugs against EBV infections.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.rcsb.org/, the accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

HH, MW, and SA conceptualized the methodology. HH, SA, MW, and ZH did the analysis. HH, SA, and KC revised the methodology, worked on graphics and supervised the study. HH, MW, and KC wrote the manuscript. ZH did the revised work and improved the manuscript.

FUNDING

The work is supported by the National Key Research and Development Program of China under the grant number 2018YFA0902000.

REFERENCES

Ali, A., Khan, A., Kaushik, A. C., Wang, Y., Ali, S. S., Junaid, M., et al. (2019). Immunoinformatic and Systems Biology Approaches to Predict and Validate Peptide Vaccines against Epstein–Barr Virus (EBV). *Scientific Rep.* 9 (1), 1–12. doi:10.1038/s41598-018-37070-z

Balsera, M. A., Wriggers, W., Oono, Y., and Schulten, K. (1996). Principal Component Analysis and Long Time Protein Dynamics. *J. Phys. Chem.* 100 (7), 2567–2572. doi:10.1021/jp9536920

Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Bochkareva, E., Frappier, L., and Edwards, A. M. (1996). Crystal Structure of the DNA-Binding Domain of the Epstein-Barr Virus Origin-Binding Protein, EBNA1, Bound to DNA. *Cell* 84 (5), 791–800. doi:10.1016/s0092-8674(00)81056-9

Bochkarev, A., Bochkareva, E., Frappier, L., and Edwards, A. M. (1998). The 2.2 Å Structure of a Permanganate-Sensitive DNA Site Bound by the Epstein-Barr Virus Origin Binding Protein, EBNA1 1 1Edited by T. Richmond. *J. Mol. Biol.* 284 (5), 1273–1278. doi:10.1006/jmbi.1998.2247

Cruickshank, J., Shire, K., Davidson, A. R., Edwards, A. M., and Frappier, L. (2000). Two Domains of the Epstein-Barr Virus Origin DNA-Binding Protein, EBNA1, Orchestrate Sequence-specific DNA Binding. *J. Biol. Chem.* 275 (29), 22273–22277. doi:10.1074/jbc.m001414200

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* 125 (7), 1731–1737. doi:10.1021/ja026939x

Ernst, M., Sittel, F., and Stock, G. (2015). Contact-and Distance-Based Principal Component Analysis of Protein Dynamics. *J. Chem. Phys.* 143 (24), 244114. doi:10.1063/1.4938249

Garai-Ibabe, G., Grinyte, R., Canaan, A., and Pavlov, V. (2012). Homogeneous Assay for Detection of Active Epstein-Barr Nuclear Antigen 1 by Thrombin Activity Modulation. *Anal. Chem.* 84 (14), 5834–5837. doi:10.1021/ac301250f

Garber, A. C., Hu, J., and Renne, R. (2002). Latency-associated Nuclear Antigen (LANA) Cooperatively Binds to Two Sites within the Terminal Repeat, and Both Sites Contribute to the Ability of LANA to Suppress Transcription and to Facilitate DNA Replication. *J. Biol. Chem.* 277 (30), 27401–27411. doi:10.1074/jbc.m203489200

Gianti, E., Messick, T. E., Lieberman, P. M., and Zauhar, R. J. (2016). Computational Analysis of EBNA1 "druggability" Suggests Novel Insights for Epstein-Barr Virus Inhibitor Design. *J. Comput. Aided Mol. Des.* 30 (4), 285–303. doi:10.1007/s10822-016-9899-y

Haque, T. J., and Crawford, D. H. (1996). Transmission of Epstein - Barr Virus during Transplantation. *Rev. Med. Virol.* 6 (2), 77–84. doi:10.1002/(sici)1099-1654(199606)6:2<77::aid-rmv166>3.0.co;2-c

Hiraku, Y., Kawanishi, S., and Ohshima, H. (2014). *Cancer and Inflammation Mechanisms: Chemical, Biological, and Clinical Aspects*. John Wiley & Sons.

Hussain, I., Pervaiz, N., Khan, A., Saleem, S., Shireen, H., Wei, D.-Q., et al. (2020). Evolutionary and Structural Analysis of SARS-CoV-2 Specific Evasion of Host Immunity. *Genes Immun.*, 1–11.

Junaid, M., Shah, M., Khan, A., Li, C.-D., Khan, M. T., Kaushik, A. C., et al. (2019). Structural-dynamic Insights into the *H. pylori* Cytotoxin-Associated Gene A (CagA) and its Abrogation to Interact with the Tumor Suppressor Protein ASPP2 Using Decoy Peptides. *J. Biomol. Struct. Dyn.* 37 (15), 4035–4050. doi:10.1080/07391102.2018.1537895

Khan, A., Ali, S. S., Khan, M. T., Saleem, S., Ali, A., Suleman, M., et al. (2020). 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.*, 1–12. doi:10.1080/07391102.2020.1779128

Khan, A., Junaid, M., Kaushik, A. C., Ali, A., Ali, S. S., Mehmood, A., et al. (2018). Computational Identification, Characterization and Validation of Potential Antigenic Peptide Vaccines from hrHPVs E6 Proteins Using Immunoinformatics and Computational Systems Biology Approaches. *PloS one* 13 (5). doi:10.1371/journal.pone.0196484

Khan, A., Junaid, M., Li, C.-D., Saleem, S., Humayun, F., Shamas, S., et al. (2020). 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. Biosciences* 6, 159. doi:10.3389/fmolb.2019.00159

Khan, A., Khan, M., Saleem, S., Babar, Z., Ali, A., Khan, A. A., et al. (2020). *Decoding the Structure of RNA-dependent RNA-Polymerase (RdRp), Understanding the Ancestral Relationship and Dispersion Pattern of 2019 Wuhan Coronavirus*.

Khan, A., Khan, M., Saleem, S., Babar, Z., Ali, A., Khan, A. A., et al. (2020). Phylogenetic Analysis and Structural Perspectives of RNA-dependent RNA-Polymerase Inhibition from SARs-CoV-2 with Natural Products. *Interdiscip. Sci. Comput. Life Sci.* 12 (3), 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. (2020). 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.*

Khan, A., Rehman, Z., Hashmi, H. F., Khan, A. A., Junaid, M., Sayaf, A. M., et al. (2020). An Integrated Systems Biology and Network-Based Approaches to Identify Novel Biomarkers in Breast Cancer Cell Lines Using Gene Expression Data. *Interdiscip. Sci. Comput. Life Sci.*, 1–14.

Khan, M. T., Khan, A., Rehman, A. U., Wang, Y., Akhtar, K., Malik, S. I., et al. (2019). Structural and Free Energy Landscape of Novel Mutations in Ribosomal Protein S1 (rpsA) Associated with Pyrazinamide Resistance. *Scientific Rep.* 9 (1), 1–12. doi:10.1038/s41598-019-44013-9

Leight, E. R., and Sugden, B. (2000). EBNA-1: a Protein Pivotal to Latent Infection by Epstein-Barr Virus. *Rev. Med. Virol.* 10 (2), 83–100. doi:10.1002/(sici)1099-1654(200003/04)10:2<83::aid-rmv262>3.0.co;2-t

Lindner, S. E., and Sugden, B. (2007). The Plasmid Replicon of Epstein-Barr Virus: Mechanistic Insights into Efficient, Licensed, Extrachromosomal Replication in Human Cells. *Plasmid* 58 (1), 1–12. doi:10.1016/j.plasmid.2007.01.003

Matsuura, H., Kirschner, A. N., Longnecker, R., and Jardetzky, T. S. (2010). Crystal Structure of the Epstein-Barr Virus (EBV) Glycoprotein H/glycoprotein L (gH/gL) Complex. *Proc. Natl. Acad. Sci.* 107 (52), 22641–22646. doi:10.1073/pnas.1011806108

Morgunova, E., Yin, Y., Jolma, A., Dave, K., Schmierer, B., Popov, A., et al. (2015). Structural Insights into the DNA-Binding Specificity of E2F Family Transcription Factors. *Nat. Commun.* 6 (1), 1–8. doi:10.1038/ncomms10050

Okano, M. (1998). Epstein-Barr Virus Infection and its Role in the Expanding Spectrum of Human Diseases. *Acta Paediatr.* 87 (1), 11–18.

Pearson, K. (1901). LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philosophical Mag. J. Sci.* 2 (11), 559–572. doi:10.1080/14786440109462720

Pope, J. H., Horne, M. K., and Scott, W. (1968). Transformation of Foetal Human Leukocytesin Vitro by Filtrates of a Human Leukaemic Cell Line Containing Herpes-like Virus. *Int. J. Cancer* 3 (6), 857–866. doi:10.1002/ijc.2910030619

Roe, D. R., and Cheatham, T. E., III (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor. Comput.* 9 (7), 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 research, gkw1000.

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23 (3), 327–341. doi:10.1016/0021-9991(77)90098-5

Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3 (2), 198–210.

Sagendorf, J. M., Markarian, N., Berman, H. M., and Rohs, R. (2020). DNAproDB: an Expanded Database and Web-Based Tool for Structural Analysis of DNA-Protein Complexes. *Nucleic Acids Res.* 48 (D1), D277–D287. doi:10.1093/nar/gkz889

Taylor, G. S., Long, H. M., Brooks, J. M., Rickinson, A. B., and Hislop, A. D. (2015). The Immunology of Epstein-Barr Virus-Induced Disease. *Annu. Rev. Immunol.* 33, 787–821. doi:10.1146/annurev-immunol-032414-112326

Tikhmyanova, N., Schultz, D. C., Lee, T., Salvino, J. M., and Lieberman, P. M. (2014). Identification of a New Class of Small Molecules that Efficiently Reactivate Latent Epstein-Barr Virus. *ACS Chem. Biol.* 9 (3), 785–795. doi:10.1021/cb4006326

Tuszynska, I., Magnus, M., Jonak, K., Dawson, W., and Bujnicki, J. M. (2015). NPDock: a Web Server for Protein-Nucleic Acid Docking. *Nucleic Acids Res.* 43 (W1), W425–W430. doi:10.1093/nar/gkv493

Vilar, S., Cozza, G., and Moro, S. (2008). Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Ctmc* 8 (18), 1555–1572. doi:10.2174/156802608786786624

Wang, J., Lindner, S. E., Leight, E. R., and Sugden, B. (2006). Essential Elements of a Licensed, Mammalian Plasmid Origin of DNA Synthesis. *Mol. Cel Biol* 26 (3), 1124–1134. doi:10.1128/mcb.26.3.1124-1134.2006

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal Component Analysis. *Chemometrics Intell. Lab. Syst.* 2 (1-3), 37–52. doi:10.1016/0169-7439(87)80084-9

Yasuda, A., Noguchi, K., Minoshima, M., Kashiwazaki, G., Kanda, T., Katayama, K., et al. (2011). DNA Ligand Designed to Antagonize EBNA1 Represses Epstein-Barr Virus-Induced Immortalization. *Cancer Sci.* 102 (12), 2221–2230. doi:10.1111/j.1349-7006.2011.02098.x

Young, L. S., and Rickinson, A. B. (2004). Epstein-Barr Virus: 40 Years on. *Nat. Rev. Cancer* 4 (10), 757–768. doi:10.1038/nrc1452

Zwanzig, R. (1973). Nonlinear Generalized Langevin Equations. *J. Stat. Phys.* 9 (3), 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 | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fmolb.2022.879875

Novel Disease-Associated Missense Single-Nucleotide Polymorphisms Variants Predication by Algorithms Tools and Molecular Dynamics Simulation of Human TCIRG1 Gene Causing Congenital Neutropenia and Osteopetrosis

Khyber Shinwari[1]*†, *Hafiz Muzzammel Rehman*[2,3], *Guojun Liu*[4]†, *Mikhail A. Bolkov*[1,5]†, *Irina A. Tuzankina*[1,5]† *and Valery. A. Chereshnev*[1,5]†

[1]Institute of Chemical Engineering, Department of Immunochemistry, Ural Federal University, Yekaterinburg, Russia, [2]School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan, [3]Alnoorians Group of Institutes, Shad Bagh, Lahore, Pakistan, [4]School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, China, [5]Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russia

T Cell Immune Regulator 1, ATPase H + Transporting V0 Subunit A3 (TCIRG1 gene provides instructions for making one part, the a3 subunit, of a large protein complex known as a vacuolar H + -ATPase (V-ATPase). V-ATPases are a group of similar complexes that act as pumps to move positively charged hydrogen atoms (protons) across membranes. Single amino acid changes in highly conserved areas of the TCIRG1 protein have been linked to autosomal recessive osteopetrosis and severe congenital neutropenia. We used multiple computational approaches to classify disease-prone single nucleotide polymorphisms (SNPs) in TCIRG1. We used molecular dynamics analysis to identify the deleterious nsSNPs, build mutant protein structures, and assess the impact of mutation. Our results show that fifteen nsSNPs (rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489 and rs116675104) variants are likely to be highly deleterious mutations as by incorporating them into wild protein they destabilize the wild protein structure and function. They are also located in the V-ATPase I domain, which may destabilize the structure and impair TCIRG1 protein activation, as well as reduce its ATPase effectiveness. These mutants have not yet been identified in patients suffering from CN and osteopetrosis while (G405R, R444L, and D517N) reported in our study are already associated with osteopetrosis. Mutation V52L reported in our study was identified in a patient suspected for CN. Finally, these mutants can help to further understand the broad pool of illness susceptibilities associated with TCIRG1 catalytic kinase domain activation and aid in the development of an effective treatment for associated diseases.

Keywords: TCIRG1 gene mutation, congenital neutropenia, osteopetrosis, non-synonymous single nucleotide polymorphisms, molecular dynamics simulation (MD)

Edited by:
Pengfei Tian,
Novozymes, Denmark

Reviewed by:
Weiwei Xue,
Chongqing University, China
Yi He,
University of New Mexico,
United States

Correspondence:
Khyber Shinwari
khybershinwari05@gmail.com

†ORCID:
Khyber Shinwari
https://orcid.org/0000-0002-9800-2797
Liu Guojun
https://orcid.org/0000-0002-1426-8185
Mikhail A. Bolkov
https://orcid.org/0000-0003-2763-9907
IrinaA.Tuzankina
https://orcid.org/0000-0001-7496-0950
Valery A. Chereshnev
https://orcid.org/0000-0003-4329-147X

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

Received: *20 February 2022*
Accepted: *04 April 2022*
Published: *28 April 2022*

Citation:
Shinwari K, Rehman HM, Liu G, Bolkov MA, Tuzankina IA and Chereshnev VA (2022) Novel Disease-Associated Missense Single-Nucleotide Polymorphisms Variants Predication by Algorithms Tools and Molecular Dynamics Simulation of Human TCIRG1 Gene Causing Congenital Neutropenia and Osteopetrosis. Front. Mol. Biosci. 9:879875. doi: 10.3389/fmolb.2022.879875

1 INTRODUCTION

A precise balance between bone creation by osteoblasts and resorption by osteoclasts is required for bone development and homeostasis. Osteopetrosis is a hereditary disease defined by a clinically and genetically heterogeneous group of bone resorption diseases. The three primary types based on hereditary patterns are the age of onset, severity, and type (Tolar et al., 2004). All of these variants have increased bone density, which can cause fractures, osteomyelitis, deformity, dental anomalies, bone marrow failure, and cranial nerve compression, among other phenotypical features (Stark and Savarirayan, 2009). Osteopetrosis is a rare condition, occurring in about one in every 250,000 births, compared to one in every 20,000 births for autosomal dominant osteopetrosis (Loría-Cortés et al., 1977). These conditions are more common in some geographical places, such as Costa Rica, the Middle East, Russia's Chuvash Republic, and Northern Sweden's Västerbotten Province. This spread is aided by the founder effect, geographical isolation, and severe maternal consanguinity (Sobacchi et al., 2013a). Numerous forms of osteopetrosis cases in humans have been linked to changes in at least ten genes (Stark and Savarirayan, 2009). Autosomal recessive osteopetrosis renders bones more sensitive to hematological damage and neurological deficit as a result of a smaller bone marrow cavity and nerve compression (blindness or deafness). In a study published in 2000, changes in T Cell Immune Regulator 1, ATPase H + Transporting V0 Subunit A3 were identified to be a primary source of human autosomal recessive osteopetrosis (Sobacchi et al., 2013a). As a result of the molecular analysis, six new genes (TNFSF11, TNFRSF11A, CLCN7, OSTM1, SNX10, and PLEKHM1) have been discovered to be associated with human ARO. More than half of all autosomal recessive osteopetrosis patients had TCIRG1 mutations (Sobacchi et al., 2001). According to a study, mice with a targeted disruption of Atp6i developed severe osteopetrosis (Li et al., 1999). Despite tremendous progress in our understanding of disease mechanisms in osteoporotic diseases, the genetic basis for 30% of cases is unclear (Sobacchi et al., 2013b). According to the study, TCIRG1 mutations include missense, nonsense, small deletions/insertions, splice-site mutations, significant genomic deletions, and intronic mutations (Frattini et al., 2000; Kornak et al., 2000; Sobacchi et al., 2013b; Sobacchi et al., 2014; Palagano et al., 2015). There is still a link between autosomal recessive osteopetrosis 1 and premature infertility deaths. This issue can be detected as early as the age of 10 days. The most prevalent signs of the illness are pathologic fractures, bone marrow failure, and cranial nerve compression, which are caused by impaired bone turnover, metabolism, and failure to widen cranial nerve foramina (Chávez-Güitrón et al., 2018). High bone density can occur from a bone resorption fault caused by osteoclast dysfunction, which can lead to severe abnormalities. Some of the defects that appear early in fetal development include microcephaly, progressive deafness, blindness, hepatosplenomegaly, and severe anemia. Deafness and blindness are common side effects of secondary cranial nerve hypertension (Susani et al., 2004). Sever Congenital Neutropenia is a hematological condition characterized by low blood neutrophil counts (ANC) of less than 0.5 109/L and recurrent bacterial infections that usually start in childhood. In 1956, Kostmann was the first to describe an autosomal recessive form of sever congenital neutropenia (KOSTMANN, 1956). A recessive type of sever congenital neutropenia is considered to be caused by mutations in HAX1, a gene related to the Bcl-2 family (Carlsson and Fasth, 2001; Klein et al., 2007). Mutations in the ELA2 gene, which codes for the protein neutrophil elastase, an enzyme present in the major granules of neutrophils, are the most common cause of sever congenital neutropenia (Horwitz et al., 1999; Dale et al., 2000). Other genes which can induce neutropenia, include such as those involved in glucose homeostasis (SLC37A4, G6PC3), lysosomal function (LYST, RAB27A, ROBLD3/p14, AP3B1, VPS13B), ribosomal proteins (SBDS, RMRP), mitochondrial proteins (HAX1, AK2, TAZ), immunological functions (STK4, GFI1, CXCR4), and Xlinked (WAS) (Boztug and Klein, 2009). In contrast, many families with autosomal dominant sever congenital neutropenia have no identifiable mutation, showing that there are more sever congenital neutropenia genes. After high-density SNP chips were used to detect IBD regions across affected in a large SCN family, exome sequencing was utilized to find coding single nucleotide variants (SNVs) in the IBD regions (Makaryan et al., 2014). SNPs (single nucleotide polymorphisms) are genetic markers found in the human genome at each 200–300 base pair (Lee et al., 2005). There are roughly 0.5 million SNPs in the human genome's coding region (Rajasekaran et al., 2008). Substituting amino acids is conserved areas can change the structure, stability, and function of proteins. Nonsynonymous SNPs (nsSNPs) are known to alter protein function and have a higher chance of causing disease in humans (George Priya Doss et al., 2008; Chitrala and Yeguvapalli, 2014; Shinwari et al., 2021). Evidently, several studies have shown that nsSNPs are responsible for 50% of the variations related to heredity genetic disorders (Ramensky et al., 2002; Doniger et al., 2008; Radivojac et al., 2010). Alignment methods based on matrix and data tree structure computation are being used by the instruments (Kamatani et al., 2004; Rajasekaran et al., 2008). We described the structural and functional impacts of high-risk nsSNPs on the TCIRG1 protein using a series of prediction algorithms.

2 METHODS

2.1 SNP Retrieval

The nsSNP information for the human TCIRG1 gene was obtained utilizing a variety of web-based data sources, including OMIM (Online Mendelian Inheritance in Man) (Hamosh et al., 2005), NCBI dbSNP (Sherry et al., 2001), and the UniProt database (UniProtKB ID O15072) (UniProt Consortium, 2010).

2.2 Gene Mania

Gene MANIA (https://genemania.org/) (accessed 10 February 2021 using a search strategy for TCIRG1 in the search box) (Warde-Farley et al., 2010) was used to confirm the TCIRG1 gene's linkage and analyze its connection through other genes in order to anticipate the impact of nsSNPs on specific linked genes. GeneMANIA predicts gene-gene connections using pathways, co-expression, co-localization, genetics, protein interaction, and protein domain similarity.

2.3 SIFT and PolyPhen2 Predication

The deleterious/damaging or tolerated nature of isolated nsSNPs will be established first using the SIFT and PolyPhen2 tools. SIFT analyzes protein homology sequences and aligns natural nsSNPs with orthologous and paralogous protein sequences to predict detrimental nsSNPs. If the SIFT score of nonsynonymous SNPs is less than 0.05, they have a deleterious impact on protein function (Ng and Henikoff, 2003). PolyPhen2 assesses a protein's structural and functional effects by analyzing its sequence and amino acid alterations. When an amino acid is substituted or a mutation in a protein domain is discovered, it divides SNPs into three groups: possibly damaging (probabilistic score >0.15), probably damaging (probabilistic score >0.85), and benign (probabilistic score >0.85). PolyPhen2 can determine the PSIC (position-specific independent count) value of protein variations. If mutants have a direct functional impact on protein function, the diversity in PSIC scores among variations implies that (Adzhubei et al., 2010).

2.4 Sequence-Based Prediction and Disease Phenotype Prediction

In-silico tools, PON-P, Mutation Assessor, P-Mut, SNAP2, SNGP-GO, PON-P2, PANTHER, PHD-SNP, SNAP2, PROVEAN, and VarCards algorithms predicted functional implications of the missense mutation as well as confirmatory analysis of the sift and PolyPhen tools. In TCIRG1 protein sequences, to forecast the negative effects of nsSNPS, the PROVEAN algorithm was used. In the case of homologous sequences, a technique like this employs delta alignment scores based on the variant version and a protein sequence comparison. A score of equivalent to or less than 2.5 suggests deleterious nsSNP alignment (Choi et al., 2012). SNAP2 is a neural network-focused classifier. It was used to anticipate how single amino acid alterations in the TCIRG1 protein might affect the protein's function. This server takes a FASTA sequence and produces a prediction score (range from 100 strong neutral predictions to +100 strong effect prediction) that indicates how likely a mutation is to influence native protein function (Bromberg et al., 2008). PMUT uses neural networks to accurately predict the presence of single amino acid point mutations that cause disease (with an 80 percent success rate in humans). When a FASTA sequence was input into the PMut server, the difference between neutral variants and illness-linked protein sequence was discovered. A score of more than 0.5 indicates that nsSNPs are potentially harmful (Ferrer-Costa et al., 2005). SNP-GO, SNP-PhD (Calabrese et al., 2009) (http://snps.biofold.org/phdsnp/phd-snp.html) are a machine-learning-based approach that uses the conservation scores of multiple sequence alignments to make decisions. The ClinVar dataset was used to create and test the PhD-SNP tool, which typically contains 36,000 harmful and benign SNVs, provides an accuracy index score, and assesses if an SNP effect is deleterious or neutral. PANTHER-PSEP (Tang and Thomas, 2016) (PANTHER -position-specific evolutionary preservation, http://pantherdb.org/apparatuses/csnpScoreForm.jsp) employs a metric that is comparable to, but not identical to, "evolutionary preservation," in which homologous proteins are employed to retrieve potential ancestral protein sequences at phylogenetic tree nodes. Each amino acid's roots can be followed to determine how long it has been held in its ancestors in its current state. The PSEP score was categorized into three parts: "probably damaging" (preservation time >450 my), "possibly damaging" (preservation time 200 my), and "probably benign" (preservation time 200 my). VarCARD was used to obtain findings from the MCAP and FATHMM tools. -MKL-coding-pred, LRT, METALR, FATHMM-pred, META SVM, Mutation Assessor, CAAD, DANN, Mutation Taster, META SVM, Mutation Assessor, CAAD, DANN, Mutation Taster. Varcards is a consolidated genetic and medical database that covers human genome coding variants. A number of genomic techniques and databases have been developed to aid in the understanding of genetic variants, notably in nonsynonymous. Varcards, on the other hand, make it easier for scientists, researchers, general practitioners, and geneticists to collect data on a single variant or from a number of different web platforms or databases (Li et al., 2018).

2.5 MutPred Predicts Disease-Related Amino Acid Substitutions and Phenotypes

The MutPred internet server (http://mutpred.mutdb.org/) can be used as a search engine to forecast the molecular mechanism of disease caused by amino acid substitutions in mutant proteins. It makes use of a variety of structural, functional, and evolutionary features of proteins. PSI-BLASAT, SIFT, and Pfam profiles, as well as TMHMM, MARCOIL, and DisProt algorithms, were used with three servers. These are some projections for structural damage. The more the scores of all three servers are aggregated, the higher the forecast accuracy (Pejaver et al., 2020).

2.6 Structure-based Prediction

I-Mutant 3.0 https://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi). The ΔΔG Mut dataset from ProTherm was used to pre-train the algorithm. The ΔΔG value (kcal/mol) can be used to determine a single-site mutation that is dependent on a protein structure or sequence. A ΔΔG value less than zero indicates that the variant alters the structure or sequence of a protein. (Capriotti et al., 2005).

2.7 Identification of Mutant nsSNPs Position in Different Domains

The InterPro (http://www.ebi.ac.uk/interpro/) tool was used for identification of different conserved domains in the TCIRG1 protein and also mapping of nsSNPs positions in different domains (Hunter et al., 2009). Protein sequence in FASTA format or protein ID was inserted as a query to predict domains and motifs.

2.8 Conserved Residues and Sequence Motifs Identification

The human TCIRG1 UniProt protein sequence was BLASTed against the UniprotKB/Swiss-Prot database in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and significant alignment was discovered up to 100 sequences. Clustal Omega was used to perform further computational analysis on sequences having more than 50% identity and an E-value of less than 1.00E-20 (Sievers et al., 2011). The amino acid identities were colored using the Clustal color scheme, and Jalview supplied the conservation index at each alignment site (Waterhouse et al., 2009).

2.9 ConSurf's Conservation Predictions for Amino Acids (ConSurf.tau.ac.il)

The evolutionary conservation of amino acids within a protein sequence is calculated using empirical Bayesian inference. Color palettes and conservation scores are included. A score of 1 was given to variable amino acids, while a score of nine was given to the most conserved amino acid. The FASTA sequence of the TCIRG1 protein was submitted for ConSurf analysis (Berezin et al., 2004).

2.10 Project HOPE Analysis

Project HOPE is a web server that investigates the structural consequences of the desired mutation. The Hope project provides the changed protein in an observable 3D structure by cooperating with UniProt and DAS prediction algorithms. The protein sequence is used as an input source in Project HOPE, and then a structural comparison with the wild type is performed. Project HOPE is a web server that investigates the structural consequences of the desired mutation. The Hope project provides the changed protein in an observable 3D structure by cooperating with UniProt and DAS prediction algorithms. The protein sequence is used as an input source in Project HOPE, and then a structural comparison with the wild type is performed (Venselaar et al., 2010).

2.11 NetSurfP's Secondary Structure Prediction

Information about amino acid surface and solvent accessibility is needed to determine the interaction interfaces or active sites in a fully folded protein. Binding affinity is affected, and if the protein is an enzyme, catalytic activity is disrupted, when amino acid alterations at such sites are detected (Klausen et al., 2019). NetSurfP-2.0 successfully assesses surface and solvent accessibility, structural disorder, backbone dihedral angles, and secondary structure for amino acid residues. The input is FASTA-formatted protein sequences, and the output is deep neural networks trained on solved protein structures (Klausen et al., 2019). NetSurfP-2.0 is available at http://www.cbs.dtu.dk/services/NetSurfP/.

2.12 PTM Sites Prediction

Protein post-translational modifications (PTM) are utilized to predict the protein's function (Deng et al., 2017). GPSMSP v3.0 (http://msp.biocuckoo.org/online.php) was used to predict methylation sites in the TCIRG1 protein. We used NetPhos 3.177 (https://www.cbs.dtu.dk/services/NetPhos/) (Blom et al., 1999) and GPS 5.078 (https://gps.biocuckoo.cn/) (Xue et al., 2005) to predict possible sites for phosphorylation. The NetPhos 3.1 service predicts Serine, Threonine, and Tyrosine phosphorylation sites in proteins using ensembles of neural networks. Residues in the protein with a score greater than 0.5 indicate phosphorylation. A higher GPS 5.0 score, on the other hand, indicates a higher chance of getting phosphorylated. To estimate probable methylation, ubiquitylation sits, we utilized GPS-MSP 1.0 (Xue et al., 2005) (https://msp.biocuckoo.org/), UbPred (Radivojac et al., 2010) (https://www.ubpred.org), and BDMPUB (https://www.bdmpub.biocuckoo.org). Glycosylation is another important method used by NetOglyc4.0 to predict glycosylation sites (Steentoft et al., 2013). (See http://www.cbs.dtu.dk/services/NetOGlyc/for more information.) Glycosylation sites with a score greater than 0.5 are more likely to be glycosylated.

2.13 The FTSite Server (http://FTSite.bu.edu/) Predicts Ligand-Binding Sites

The server FTSite predicted the ligand-binding site in the 3D protein structure. The binding site has been identified in over 94 percent of apoproteins, and the site's prediction is based on energy. PDB data is used as input for ligand-binding hotspot prediction.

2.14 Candidate Variant Filtering

Whole Exome Sequence data of a patient suspected with congenital neutropenia was analyzed for candidate variant filtering and was performed by using BWA, GATK4, and VCF-tools software (Pedersen et al., 2021).

2.15 Predicting the Structure of 3D Proteins

Protein modeling is important in the drug development process. Structure prediction from a given sequence with accuracy similar to experimentally resolved structures is the goal of homology modeling (Cavasotto and Phatak, 2009). The inclusion of inserts and loop sequences, which cannot be reliably anticipated in the absence of a three-dimensional (3D) crystal structure, is a limitation of this technique (Ohlson et al., 2004). In the pharmaceutical sector, computational approaches are frequently used to predict 3D protein models (57). To overcome this problem, these methods aid in the prediction of a protein's tertiary structure based on its amino acid sequence (Katsila et al., 2016). These methods can be classified as either *de novo* or homology modeling, depending on the information available. The most reliable method is template-based modeling, also known as homology modeling or comparative

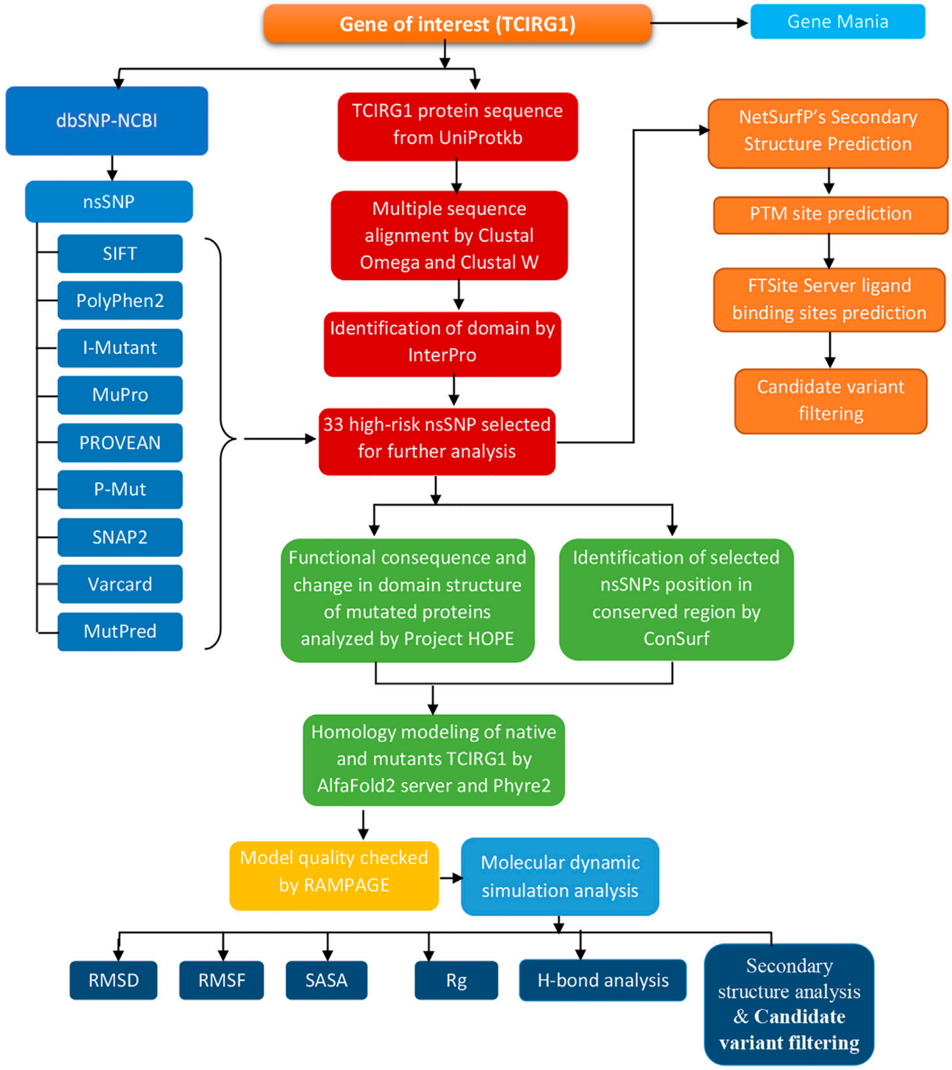

Gene of interest (TCIRG1)
Gene Mania
dbSNP-NCBI
nsSNP
SIFT
PolyPhen2
I-Mutant
MuPro
PROVEAN
P-Mut
SNAP2
Varcard
MutPred
TCIRG1 protein sequence from UniProtkb
Multiple sequence alignment by Clustal Omega and Clustal W
Identification of domain by InterPro
33 high-risk nsSNP selected for further analysis
NetSurfP's Secondary Structure Prediction
PTM site prediction
FTSite Server ligand binding sites prediction
Candidate variant filtering
Functional consequence and change in domain structure of mutated proteins analyzed by Project HOPE
Identification of selected nsSNPs position in conserved region by ConSurf
Homology modeling of native and mutants TCIRG1 by AlfaFold2 server and Phyre2
Model quality checked by RAMPAGE
Molecular dynamic simulation analysis
RMSD
RMSF
SASA
Rg
H-bond analysis
Secondary structure analysis & Candidate variant filtering

FIGURE 1 | Flowchart for methodology.

modeling (Cavasotto and Phatak, 2009). Because there were no resolved crystal structures of TCIRG1 available at the time of this research, SWISS-MODEL and HHPred were used to create a homology model for the mutant protein (Schwede et al., 2003; Hildebrand et al., 2009). The 3D structure for the TCIRG1 was also predicated through Phyre2 which is a 3D homology modeling application that predicts 3D models for proteins (http://www.sbg.bio.ic.ac.uk/phyre 2/html/page.cgi?xml:id=index). As 3D models, the wild type and 22 mutants linked to the most harmful nsSNPs were generated (Kelley et al., 2015). Confirmatory molding was conduct of Wild and Mutant TCIRG1 protein through Alpha fold2 which is a highly accurate protein structure predication (Jumper et al., 2021). To compare wild-type TCIRG1 and selected mutations, researchers employed TMalign (https://zhang lab. ccmb.med.umich.edu/TM-align/). Template Modelling score (TMscore), root mean square deviation (RMSD), and structural superposition are all predicted. The TM scores range from 0 to 1, with a higher value indicating more structural similarity. The higher the RMSD values, the greater the difference between mutant and wild-type structures (Carugo and Pongor, 2001). Three mutants with greater RMSD values were submitted to the ITASSER (https://zhang lab.ccmb.med.umich.edu/I-TASSER R/) for further protein 3D structure comparisons (Zhang, 2008; Roy et al., 2010; Yang et al., 2015). Chimera v1. 11 to investigate molecular characteristics and interactive visualization of the resulting protein structure (Pettersen et al., 2004). PROCHECK was used to validate the 3D models (Laskowski et al., 1993).

2.16 Molecular Dynamic Simulation

For 100 nanoseconds, Desmond, a software from Schrödinger LLC, was used to model molecular dynamics (Bowers et al., 2006; Ferreira et al., 2015). By integrating Newton's classical equation of

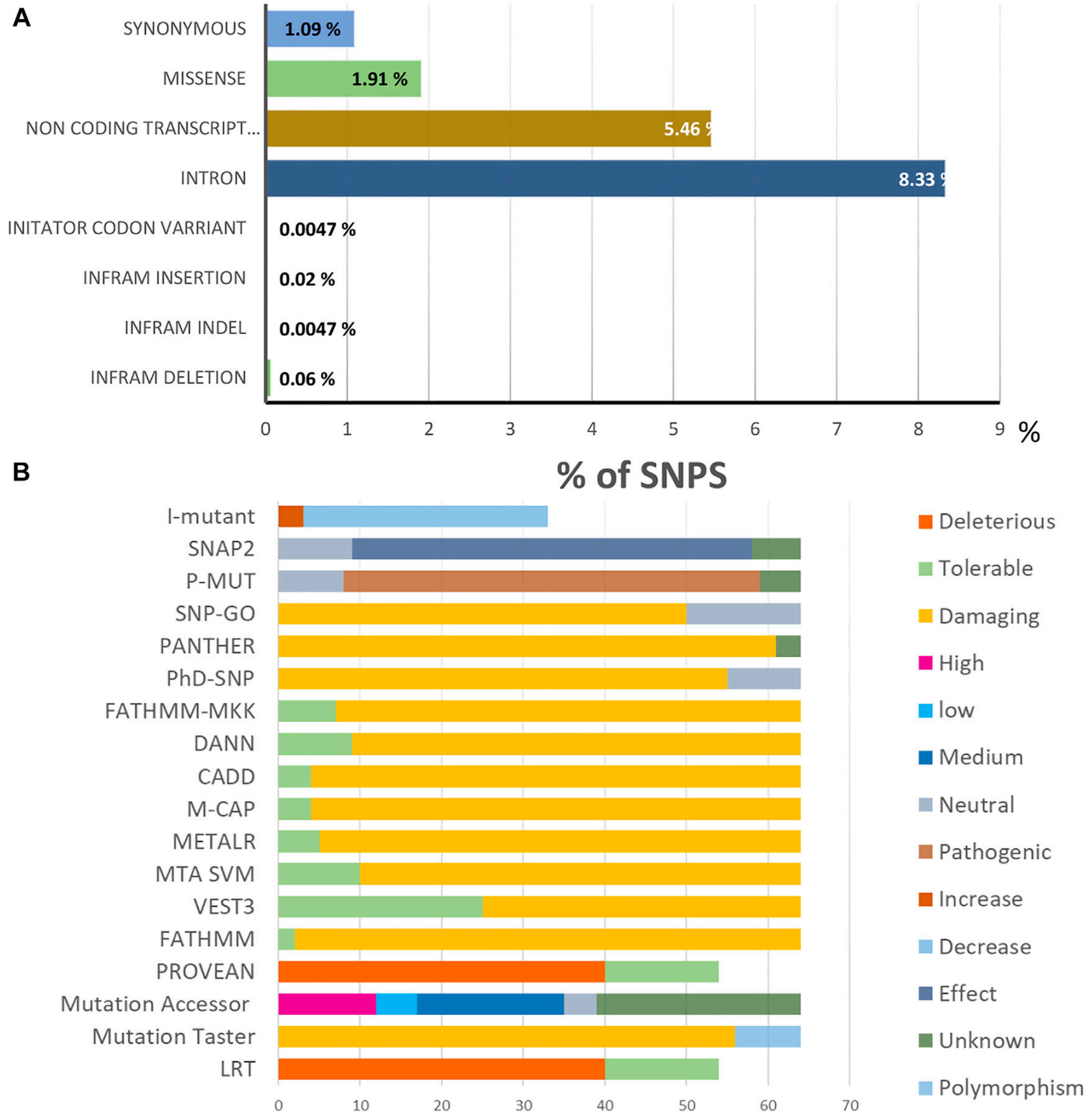


FIGURE 2 | (A). Distribution of SNPs present in the TCIRG1 gene. **(B)**. Prediction results of the 64 SIFT and PolyPhen2 deleterious nsSNPs in the TCIRG1 gene analyzed by the eighteen computational tools.

motion, MD simulations typically compute atom movements over time. Simulations were used to predict the stability of the protein in the physiological environment. (Hildebrand et al., 2019; Rasheed et al., 2021).

2.17 Statistical Analysis

SPSS v23 and MS Excel were used to conduct a correlation study on the predictions made by computational in silico technologies. The significance differences predicted by the various computational techniques were assessed using the Student's t-test. Significant was defined as a *p*-value of less than 0.01.

3 RESULTS

The entire approach, tools, and databases used to discover the harmful SNPs in human TCIRG1 and their structural/functional repercussions owing to mutation are summarized in **Figure 1**.

3.1 SNP Annotation

The NCBI database (http://www.ncbi.nlm.nih.gov/) revealed SNPs in the TCIRG1 gene. It contains 5627 SNPs that were present in Homo sapiens, with 811(1.909%) in coding nonsynonymous regions (missense) and 463 (1.089%) in synonymous sections, as illustrated in **Figures 2A,B**.

TABLE 1 | Gene-mania shows the TCIRG1 gene co-expression and shard domain.

Gene symbol	Description	Co-Expression	Shared Domain
MAN2C1	Mannosidase alpha class 2C member 1	Yes	No
INPPL1	Inositol polyphosphate phosphatase like 1	Yes	No
TRADD	TNFRSF1A associated via death domain	Yes	No
ARPC1B	Actin related protein 2/3 complex subunit 1B	Yes	No
TIMP1	TIMP metallopeptidase inhibitor 1	Yes	No
LSP1	Lymphocyte-specific protein 1	Yes	No
TYMP	Thymidine phosphorylase	Yes	No
HLA-A	Major histocompatibility complex, class I, A	Yes	No
MVP	Major vault protein	Yes	No
ARSA	Arylsulfatase A	Yes	No
PCSK7	Proprotein convertase subtilisin/kexin type 7	Yes	NO
MAP3K11	Mitogen-activated protein kinase kinase kinase 11	Yes	No
ATP6V0A4	ATPase H+ transporting V0 subunit a4	No	Yes
ATP6V0A2	ATPase H+ transporting V0 subunit a2	No	Yes
ATP6V0A1	ATPase H+ transporting V0 subunit a1	No	Yes

3.2 Gene Mania

The TCIRG1 gene codes for a protein that can be found in the extracellular matrix of proteins, as well as other compounds. The TCIRG1 protein plays a crucial role in the formation of the lymphatic system. It helps immature lymphangioblasts grow (differentiate) and migrate (migrate), finally forming the lining (epithelium) of lymphatic channels. Our findings revealed that TCIRG1 is co-expressed with 12 genes (MAN2C1, INPPL1, TRADD, ARPC1B, TIMP1, LSP1, TYMP, HLA-A, MVP, ARSA, PCSK7, and MAP3K11) and shared a domain with only three genes (ATP6VOA4, ATP6VO2A, and ATP6VOA1), Physical interaction with seven genes (KCNK1, TRADD, ERLEC1, SLC30AS, ATP6AP2, ATP6VOA2, ATP6VOA1), and co-localization with two genes (ARSA, TYMP) **Table 1**and **Figures 3A,B**.

3.3 SIFT and POLYPHEN

A total of 5627 nsSNPs were investigated to see if they influenced protein structure or function in any way. The first step is to figure out which of the nsSNPs is causing the amino acid substitution. SIFT calculates the effect of an nsSNP on protein structure and assesses if the induced amino acid is acceptable at that site. SIFT and PolyPhen predicted 64 nsSNPs that produced amino acid substitutions out of a total of 811 nsSNPs (**Table 2** and **Supplementary Table S1**).

3.4 The Most Deleterious SNPs Identified in TCIRG1

3.4.1 Functional SNPs in Coding Areas Were Identified

The various computational prediction tools that were used in this study, are illustrated in **Figure 2B** to identify significant nsSNPs in TCIRG1. The nsSNPs in **table 3** are variations that are predicted to be deleterious by all algorithms. FATHMM-MKL. While they are regarded as high-risk pathogenic nsSNPs, SNP-GO, PHD-SNP, PANTHER, SNAP2, P-MUT PROVEAN, FATHMM, LRT, M-CAP, CAAD, META SVM, METALR, Mutation Assessor, and Mutation Taster are considered high-risk pathogenic nsSNPs. There are a variable number of deleterious SNPs in each technique. SIFT classed 118 and PolyPhen 64 nsSNPs as harmful or deleterious, although PolyPhen did not show any of the 58 nsSNPs that were deleterious. Sift classified deleterious with a threshold of >0.5, and both SIFT and Polyphen confirmed 34 as deleterious. In a total of 118 unique predicted nsSNPs in the TCIRG1 gene, VEST three indicated the fewest six nsSNPs (10%) as destructive or detrimental, and 51 as tolerated. PolyPhn, FATHMM, M-CAP, and PANTHER had the largest percentage of harmful predictions. Using the SNAP2 technique, 41 were found to be harmful (71%) and 16 were found to have no effect (SNAP2 score of 100). The deleterious and damaging effects of 54 (92%) nsSNPs on TCIRG1 protein were predicted using the PANTHER program, with 48 nsSNPs being probably damaging, six nsSNPs being possibly damaging, and three nsSNPs being probably benign (time >450my possibly damaging" (450my > time >200my, "probably benign" (time 200my). PROVEAN is a program that predicts the impact of SNPs on a protein's biological function. 22 (38 percent) nsSNPs in the TCIRG1 gene were projected to be severely detrimental, while 35 nsSNPs were neutral, according to PROVEAN's criterion (>-2.667). With a threshold of (>0.65 (5.545 to 5.975 (higher score > more damaging), the Mutation Assessor classified 24 nsSNPs as deleterious, with 12 high, 17 medium, five low, and 19 as no findings. FATHNMM and FATHMM-MKK (<0.5), CADD (>15) DANN (>0.5), Mutation Taster (<0.5), and with respective scores show all above than (75–90%) nsSNPs as deleterious/damaging. while P-Mut predicated 45 (75.21%) deleterious, 07 neutral, and 5 with no result with a cut off (<0.5). LRT predicted 42 (77%) deleterious nsSNPs with a score (>0.001) and 13 as Neutral. PhD-SNP, SNP-GO, and M-CAP identified 47 (82%), 35 (61%), and 54 (94.73%) as deleterious, respectively. MetalR and MTA-SVM identified 10 (17%) and 37 (64%) nsSNPs as deleterious. Based on the substitution position-specific scores using PANTHER, PROVEAN score, SIFT score, SNPs&GO, FATHMM, LRT, M-CAP, VEST3, CAAD, METALR, Mutation Assessor, Mutation Taster, FATHMM-MKL, PHD-SNP score and PolyPhen server, PSIC score (>0.5). A group of 15 nsSNPs P572L, M546V, I721N, F610S, A732T, F51S, A717D, E722K,

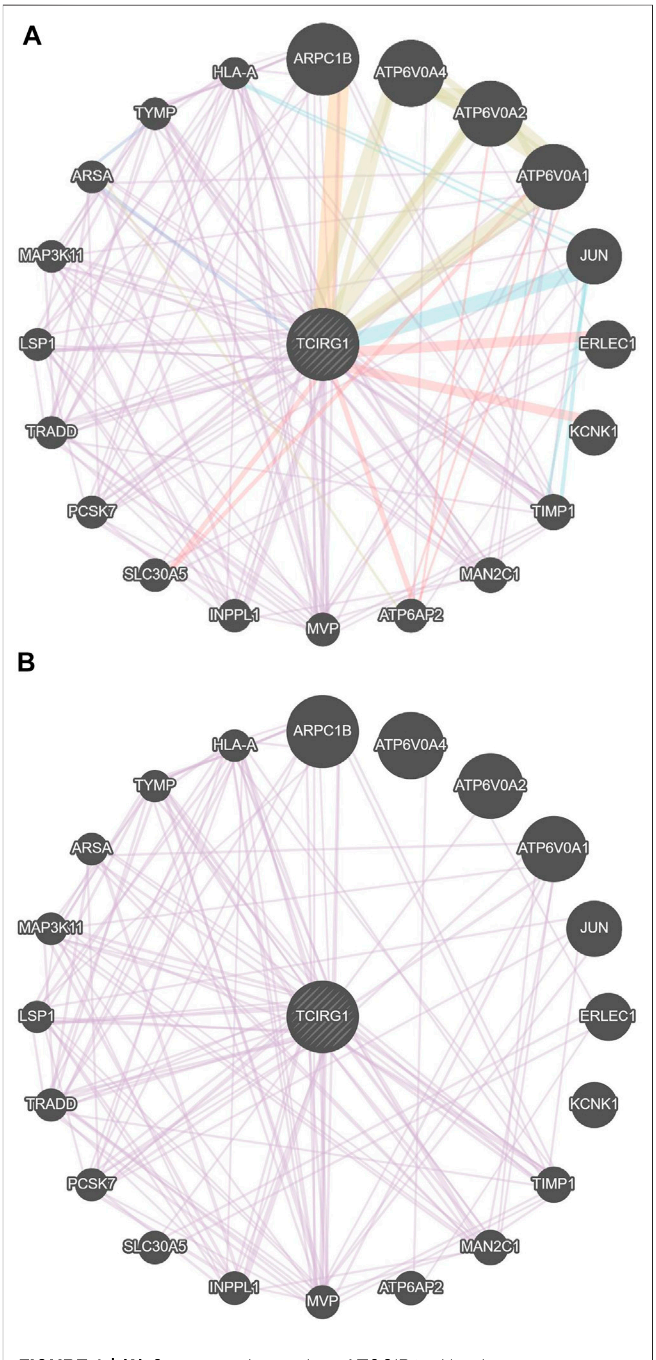


FIGURE 3 | (A) Gene–gene interaction of TCGIR1 with other genes proposed by GeneMANIA. **(B)** Co-expression in GenMANIA.

R57H, R109W, R191H, S532C, G192S, F529L, H804Q were all considered highly deleterious by all state-of-the-art methods. While only LRT disagrees with the result of A717D by other tools. All of the prediction algorithms' findings were found to be statistically significant and strongly correlated. The *p*-value for the Student *t*-test between the tools was 0.001. Results of prediction tools and their significance are shown in (**Supplementary Table S2**).

3.5 MutPred2 Predicts Pathogenic Amino Acid Substitutions

MutPred2 assesses a variety of molecular characteristics of amino acid residues in humans to identify whether a substitution is disease-related or not. It assigns a score based on the chance that a change in amino acid will affect the protein's function. A MutPred2 score of 0.8 or higher is considered highly confidential, while the pathogenicity prediction cutoff is 0.5. The prediction score for all of the substitutions was less than 0.5. The MutPred2 results are available in (**Supplementary Table S6**).

3.6 I-Mutant 3.0 Predicts the Stability of the Mutated Protein due to SNPs

The effects of TCGIR1 high-risk nsSNPs on protein stability and function were predicted using the web program I-Mutant 3.0 (**Supplementary Table S3**) The results showed that (G405R, S474W, and A778V) have increased stability while (P572L, M546V, I730N, F610S, A732T, F51S, A717D, E722K, R57H, R109W, R191W, S532C, G192S, F529L, H804Q, G458S, R444L, R56P, G379S, R757C, N730S, V375M, T314M, D517N, R92W, T368M, A417T, R363C, R56W, and R50C) showed decreased stability.

3.7 Identification of Domains in TCIRG1

InterPro tool was used to locate domain regions in TCIRG1 and to identify the location of nsSNPs in different domains. This tool provides a functional analysis of proteins by classifying them into families. It also predicts the presence of domains and active sites. It has been reported a three domain: such as the V-TYPE PROTON ATPASE 116 KDA SUBUNIT A ISOFORM 3 (1–828), cytoplasmic and non-cytoplasmic are found in TCIRG1. The 33 nsSNPs and fifteen highly deleterious that we have selected are located in V-TYPE PROTON ATPASE and cytoplasmic domains.

3.8 SNPs in TCIRG1 Protein Are Linked to Highly Conserved Buried (Structural) and Exposed (Functional) Amino Acid Residues

TCIRG1 (ATPase H + Transporting V0 Subunit A3, T Cell Immune Regulator 1) is a protein-coding gene that codes for ATPase H + Transporting V0 Subunit A3. Autosomal Recessive 1 and Autosomal Recessive Malignant Osteopetrosis TCIRG1 is associated with disorders like osteopetrosis. The lysosome cycle and the synaptic vesicle cycle are two related pathways. This gene, which is located on chromosome 11, is 830 amino acids long and has a molecular mass of 92968 Da. TCIRG1 sequence-based structural-functional investigation was analyzed using Clustal Omega-based multiple sequence alignment analysis. The Uniprot Knowledgebase was used to retrieve the TCIRG1 protein sequence (Uniprot ID: Q13488). After being BLASTed against UniprotKB/SwissProt entries, the TCIRG1 protein sequence was aligned using Clustal Omega with default settings. This gene, which is located on chromosome 11, is

TABLE 2 | Sift and PolyPhen results of high deleterious nsSNPs in TCIRG1 gene.

ID of nsSNPs	Aa position	SIFT	Score	PolyPhen	Score
rs36027301	R56W	Deleterious	0	Probably damaging	0.999
rs368945298	M546V	Deleterious	0	Probably damaging	0.999
rs115854062	P572L	Deleterious	0	Probably damaging	1
rs150260808	I721N	Deleterious	0	Probably damaging	1
rs137853150	G405R	Deleterious	0	Probably damaging	1
rs137853151	R444L	Deleterious	0	Probably damaging	1
rs147580611	F610S	Deleterious	0	Probably damaging	1.00
rs148921764	E722K	Deleterious	0	Probably damaging	1.00
rs140963213	A417T	Deleterious	0.002	Probably damaging	1
rs144775787	A778V	Deleterious	0.46	Probably damaging	0.883
rs145080707	R213W	Deleterious Low	0.012	Probably damaging	1
rs150648332	R57H	Deleterious	0.001	Probably damaging	1.00
rs150260808	I721N	Deleterious	0	Probably damaging	1
rs201329219	R109W	Deleterious	0.014	Probably damaging	1.00
rs367703865	R191H	Deleterious	0.32	Probably damaging	0.999
rs371214361	S532C	Deleterious	0.001	Probably damaging	1.00
rs199914625	S474W	Deleterious	0	Probably damaging	1
rs200851583	G458S	Deleterious	0	Probably damaging	1
rs371658110	G192S	Deleterious	0.003	Probably damaging	1.00
rs370319355	R50C	Deleterious	0	Probably damaging	1
rs376351835	F529L	Deleterious	0.013	Probably damaging	1.00
rs371004297	G379S	Deleterious	0.011	Probably damaging	1.00
rs200209146	N730S	Deleterious	0.022	Probably damaging	1.00
rs200415611	V375M	Deleterious	0.001	Probably damaging	1.00
rs367818260	**T314M**	**Deleterious**	**0.001**	**Probably Damaging**	**1.00**
rs375809635	R363C	Deleterious	0	Probably damaging	1.00
rs138305091	A732T	Deleterious	0.001	Probably damaging	1.00
rs138308753	F51S	Deleterious	0	Probably damaging	0.996
rs141095902	A717D	Deleterious	0.002	Probably damaging	0.963
rs369264588	D517N	Deleterious	0	Probably damaging	1.00
rs371907380	R92W	Deleterious	0	Probably damaging	1.00
rs373988992	T368M	Deleterious	0	Probably damaging	1.00
rs142606750	R757C	Deleterious	0.003	Probably damaging	1.00
rs367818260	T314M	Deleterious	0.001	Probably damaging	1.00
rs118141250	V52L	Deleterious	0.11	Probably damaging	0.924

Threshold: Sift: < 0.05 Polyphen2: >0.8 (PSIC >0.5) or Benign (PSIC <0.5).

TABLE 3 | TMscore and RMSD values of 56 deleterious nsSNPs in TCIRG1.

SNP-ID	Residual Change	TM-score	RMSD Values	SNP-ID	Residual Change	TM-score	RMSD Values
rs199902030	P572L	0.99626	0.78	rs121908252	R56W	0.99621	0.78
rs200149541	M546V	0.99626	0.78	rs121908254	G379C	0.99435	0.58
rs372499913	I721N	0.99760	0.53	rs147974432	R757C	0.99790	0.48
rs267605221	F610S	0.99312	0.81	rs192224843	N730S	0.99275	0.84
rs374941368	A732T	0.99621	0.78	rs115982879	V375M	0.99743	0.54
rs375717418	F51S	0.99626	0.78	rs139059968	T314M	0.99626	0.78
rs80008675	A717D	0.99661	0.73	rs141125426	D517N	0.99785	0.49
rs149792489	E722K	0.99830	0.46	rs147208835	R92W	0.96213	0.89
rs116675104	R57H	0.99790	0.48	rs147681552	T368M	0.99626	0.78
rs121908250	R109W	0.99626	0.78	rs148498685	A417T	0.99790	0.48
rs121908251	R191H	0.99785	0.49	rs149531418	R363C	0.99626	0.78
rs121908251	S532C	0.99092	0.81	rs149531418	A778V	0.99661	0.76
rs149792489	G192C	0.99626	0.78	rs147208835	R50C	0.99621	0.78
rs116675104	F529L	0.99435	0.58	rs121908250	H804Q	0.99790	0.48
rs121908251	G405R	0.99674	0.62	rs149792489	S474W	0.99760	0.53
rs116675104	G458S	0.99674	0.48	rs121908250	R444L	0.99270	0.84
rs121908251	R56P	0.99657	0.48				

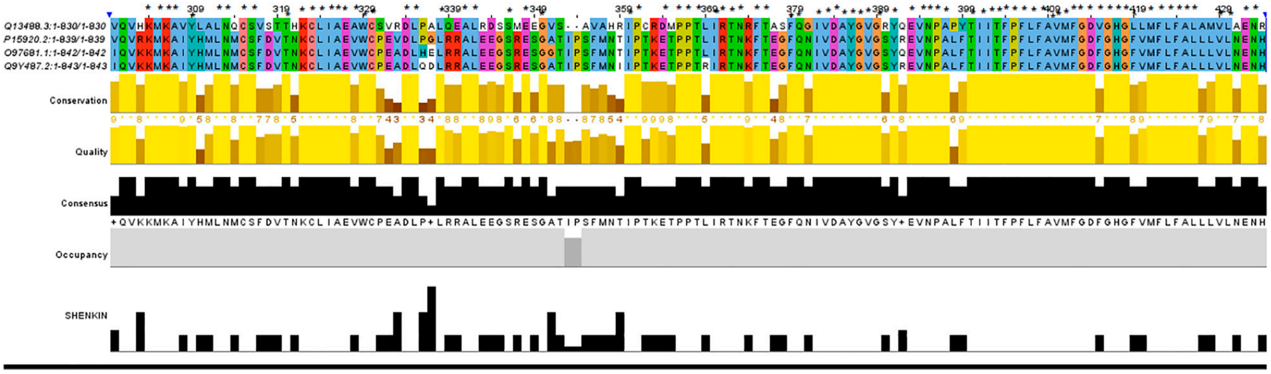


FIGURE 4 | In ABWGB and Q3MI99, amino acid alignment of human TCIRG1 (UniProt ID: Q6UXH8) and homologs in phylogenetically adjacent species. Residues with an asterisk (*) mark indicate evolutionarily conserved amino acids, while solid horizontal bars indicate conserved sequence patterns. The conservation index at each alignment point was provided by Jalview, and the amino acid identities were colored according to the Clustal color scheme.

830 amino acids long and has a molecular mass of 92968 Da. TCIRG1 sequence-based structural-functional investigation was analyzed using Clustal Omega-based multiple sequence alignment analysis. The Uniprot Knowledgebase was used to retrieve the TCIRG1 protein sequence (Uniprot ID: Q13488). After being BLASTed against UniprotKB/SwissProt entries, the TCIRG1 protein sequence was aligned using Clustal Omega with default settings. The highly conserved amino acid residues in human TCIRG1 protein were K304, M305, K306, A307, Y309, L312, N313, C315, S316, T320, K322, K322, C323, L324, I325, A326, E327, W329, C330, D334, L335, L338, A341, L342, S346, E348, S350, I360, P361, P366, P367, T368, I369, R371, T372, N373, F375, F379, Q380, I382, V383, D384, A385, Y386, G387, V388, G389, Y391, E393, V394, N395, P396, A397, T400, I401, I402, I403, F404, P405, F406, L407, F408, A409, V410, M411, F412, G413, D414, G416, H417, G418, L419, M421, F422, L423, F424, A425, L426, V429, L430, and E432. There are eighty-one different conserved residues Results can be seen in **Figure 4**.

3.9 Conservation Analysis

We used the ConSurf web server to look at the conservation of TCIRG1 residues. According to the results of the ConSurf investigation, 22 deleterious missense SNPs are found in highly conserved areas (7-8–9). The other 16 (S7K, V52L, G379S, M403I, G405R, G458S, D517N, F529L, S532C, M546V, A640S, D683H, I732N, N730S, A732T, and H804Q) were predicted as functional and exposed residues, while the other 10 (A20V, R56P, R57H, R191H, G192C, E321K, R366H, T368M, R444L, and E722K) were predicted as functional and exposed residues and the other 16 (S7K, V52L, G379S, M403I, G405R, G458S, D517N, F529L, S532C, M546V, A640S, D683H, I732N, N730S, A732T, and H804Q) were predicted as buried and structural residues. The 18 (S3F, R28W, S45A, R50C, R92W, R109W, R166T, T314M, D328M, S340L, R363C, R382H, R467H, S474W, P572L, Y626S, R628W, and R757C) were predicted as exposed and the other 9 (F51S, V348M, V375M, A417T, T570M, F610S, A717D, A778V, and M783I) were buried residues. The results are shown in **Figure 5**.

3.10 Project Hope

All of the predicating techniques projected negative consequences for 15 high-risk pathogenic TCIRG1 nsSNPs, hence HOPE was utilized to forecast their effects. The hop was based on the size, spatial, charge, hydrophobicity, structure, and function of amino acids. Seven mutant amino acids were smaller than their wild-type counterparts, while eight were larger. The charge was switched from positive to neutral at three different locations. Six alterations exhibited an increase in hydrophobicity, while the other did not. This finding implies that amino acid changes at these locations modify protein structure and interactions with other molecules, influencing protein function. The outcomes can be seen in the graph below (**Supplementary Table S8**).

3.11 TCIRG1 Secondary Structure and Surface & Solvent Accessibility of Residues Analysis by NetSurfP-2.0

The surface accessibility (exposed or buried) of amino acids in a given protein was predicted using NetSurfP-2.0, which determines the relative and absolute accessible surface area of each residue. It can also predict protein secondary structure. Relative Surface Accessibility: With a threshold of 25%, red upward elevation implies residue exposure, whereas sky blue denotes buried residue. A helix is represented by an orange spiral, a strand is represented by an indigo arrow, and a coil is represented by a pink straight line. The disorder is represented as a bloated black line, with the thickness of the line equaling the probability of disordered residue. **Figure 6**: NetSurfP-2.0 results.

3.12 PTMs (Posttranslational Modifications) Predictions

This was done with GPSMSP 3.0, which predicted that no sites in TCIRG1 were methylated. TCIRG1 phosphorylation sites predicted by GPS 3.0 and NetPhos 3.1 are included in **Supplementary Table S1**. NetPhos 3.1 projected phosphorylation potential for 62 residues (Ser23, Thr: 22,

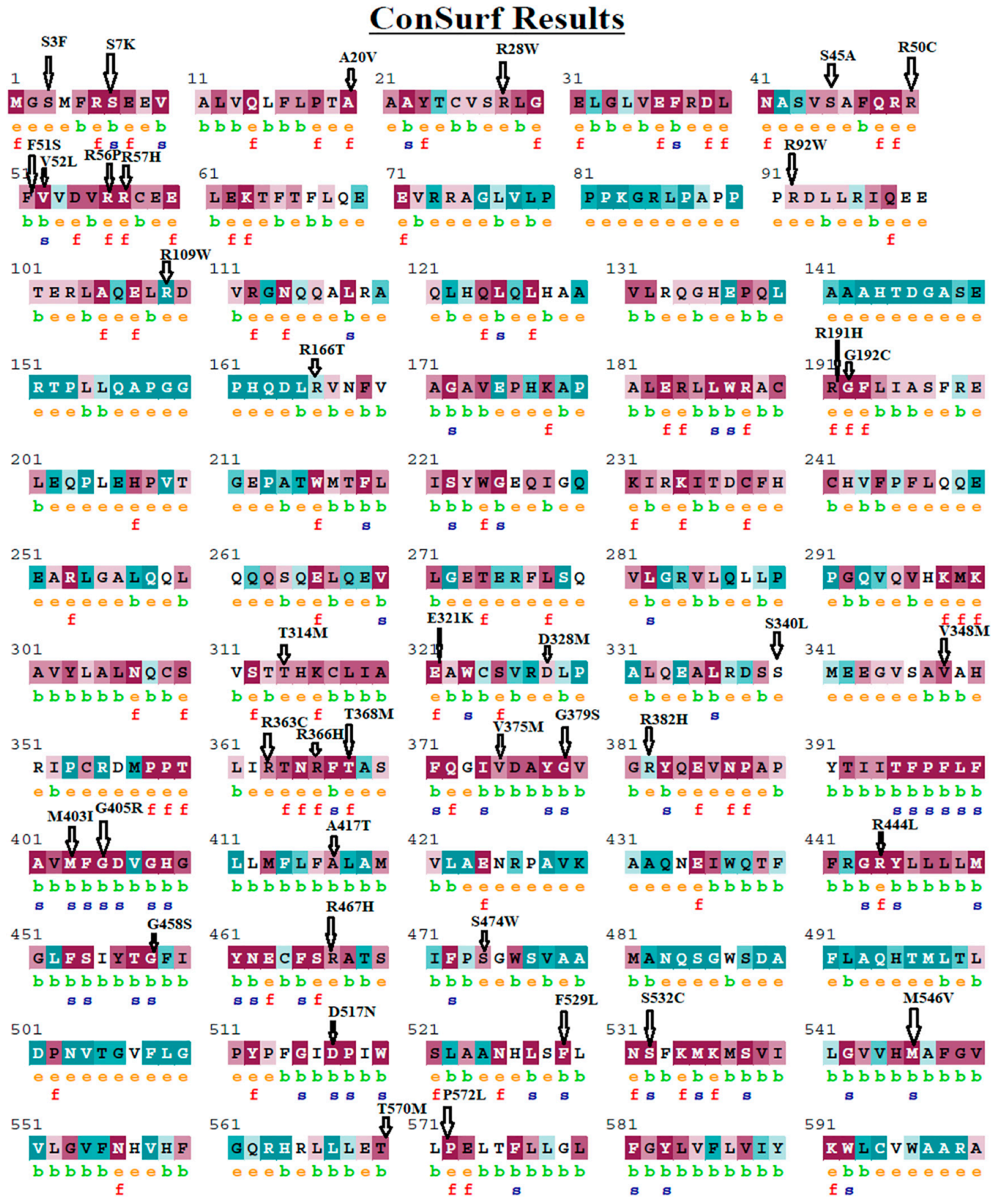


FIGURE 5 | The evolutionary conservation of amino acids in the TCIRG1 gene was assessed using the ConSurf service. A value of 1 indicates a high variability region. The value grows as the region becomes more conserved until it reaches 9.

Tyr: 17). GPS 3.0, on the other hand, suggested that 18 residues (Ser: 12, Thr: 06, Tyr: 00) may be phosphorylated. For ubiquitylation prediction, BDMPUB and UbPred were utilized. UbPred projected that none of the lysine residues would be ubiquitinated, but BDMPUB predicted that none of the lysine residues would be ubiquitinated. None of the BDMPUB predictions were found in a highly conserved or detrimental nsSNP region. **Table 2**, **Supplementary Table S2** shows the results achieved. Potential glycosylation sites were predicted using NetOGlyc4.0. Positions 43, 145, 152, 346, and 474 in wild-type TCIRG1 protein were predicted to be glycosylated with scores of 0.513,032, 0.554,065, 0.884,332, 0.830,233, 0.585,103, and 0.511,937. Interestingly, mutant S532C lost its glycosylation site at position 532, but mutant

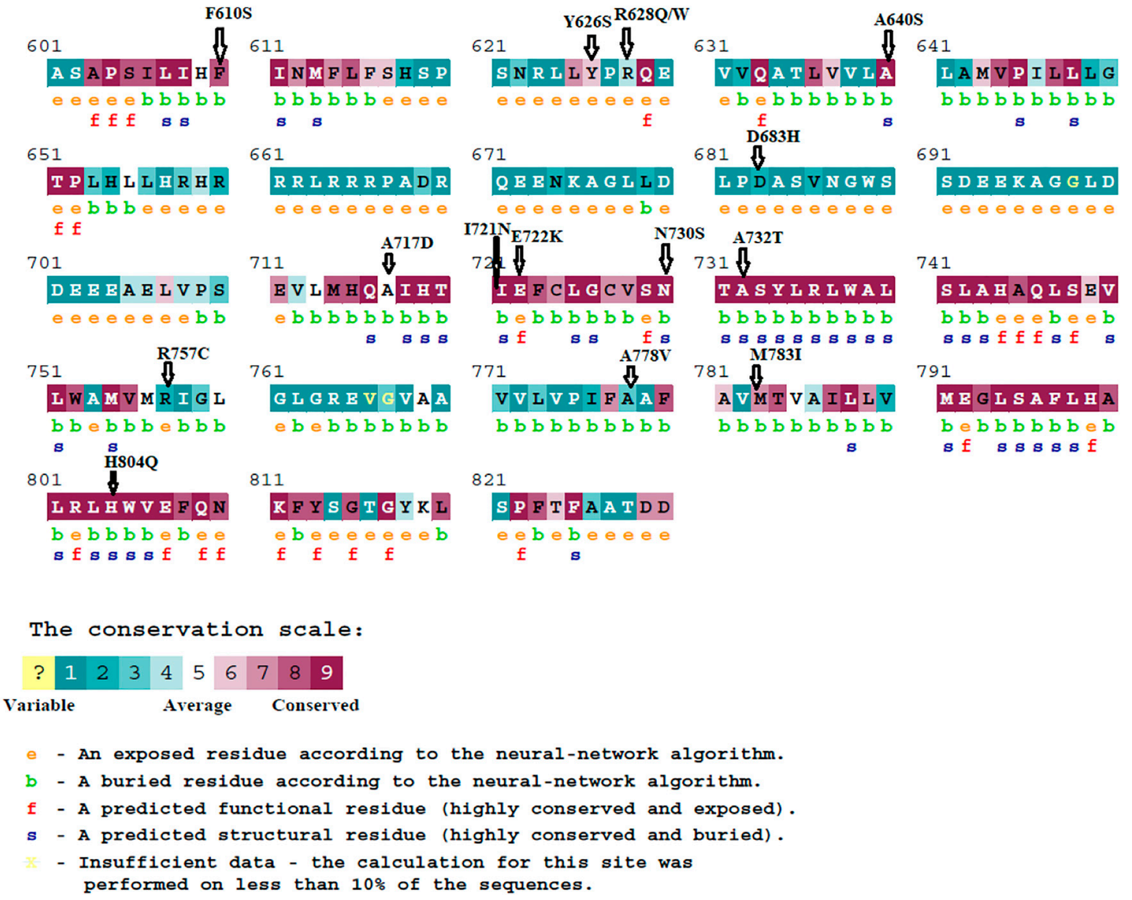

F610S
Y626S
R628Q/W
A640S
D683H
I721N
E722K
A717D
N730S
A732T
R757C
A778V
M783I
H804Q
601
611
621
631
641
651
661
671
681
691
701
711
721
731
741
751
761
771
781
791
801
811
821
The conservation scale:
?
1
2
3
4
5
6
7
8
9
Variable
Average
Conserved
e - An exposed residue according to the neural-network algorithm.
b - A buried residue according to the neural-network algorithm.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).
X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

FIGURE 5 | (Continued).

N730S gained it at position 730. **Supplementary Table S5** contains all of the scores for the wild type and mutants.

3.13 FTSite Predicts Ligand-Binding Sites

The ligand binding sites were predicted using FTSite algorithms, which were then visualized and analyzed using Pymol. Using this technique, three ligand-binding sites in human TCRIG1 protein were found (**Supplementary Figure S11**). Site 1 has 14 residues, while sites two and three each had 9, 13, and so on. In the fifty-six replaced positions, none of the substitutions in the SIFT server's expected ligand-binding sites are detected (**Supplementary Table S7**). In that sequence, the expected binding sites are colored pink, green, and purple. Residues within 5 nm of the binding site are represented using a ball and stick representations of side-chain atoms. The atoms are colored according to their elements, with carbon matching the binding site's color. RaptorX Binding ligand-binding site prediction servers were used to predict ligand-binding sites in the TCIRG1 protein. A pocket multiplicity value of greater than 40, according to the RaptorX Binding server, indicates a precise prediction. The TCIRG1 protein has the maximum pocket multiplicity of 20, with an expected CVM (2+) cation ligand connected to residues L801 H804 W805 D822 D830.

3.14 3D Modeling of TCIRG1 and Its Mutants

The protein 3D model was predicted by HHpred, Phyre2 and AlphaFold2 while the wild-type structure was predicted by AlphaFold2 available in uniport with Q13488 ID. The mutant structures predicted by HHpred and proceed with MD Simulation and similarly, the structure of mutant was also predicted by Alphafold2 and also proceed for 100ns MD simulation for further analysis and validation (**Figure 7A-H**. These structures proceeded with MD simulation for further analysis and validation. Phyre2 was also used to generate 3D structures of the wild-type TCIRG1 protein as well as 56 mutations. nsSNP replacements in the TCIRG1 protein sequence were made separately and then submitted to Phyre2, which predicted the mutant proteins' 3D structures. C6VQ7A was chosen as a template for 3D model prediction by Phyre2 because it was the template with the highest similarity, according to the Phyre2 server. For each mutant model, TM scores and RMSD values were determined. The TM-score measures topological similarity, whereas the RMSD values measure the average distance between the carbon backbones of natural and mutant models. Higher RMSD values indicate that the mutant structure differs from that of the wild type. The mutant R92W (rs371907380) has the highest RMSD value of 0.89B, followed by

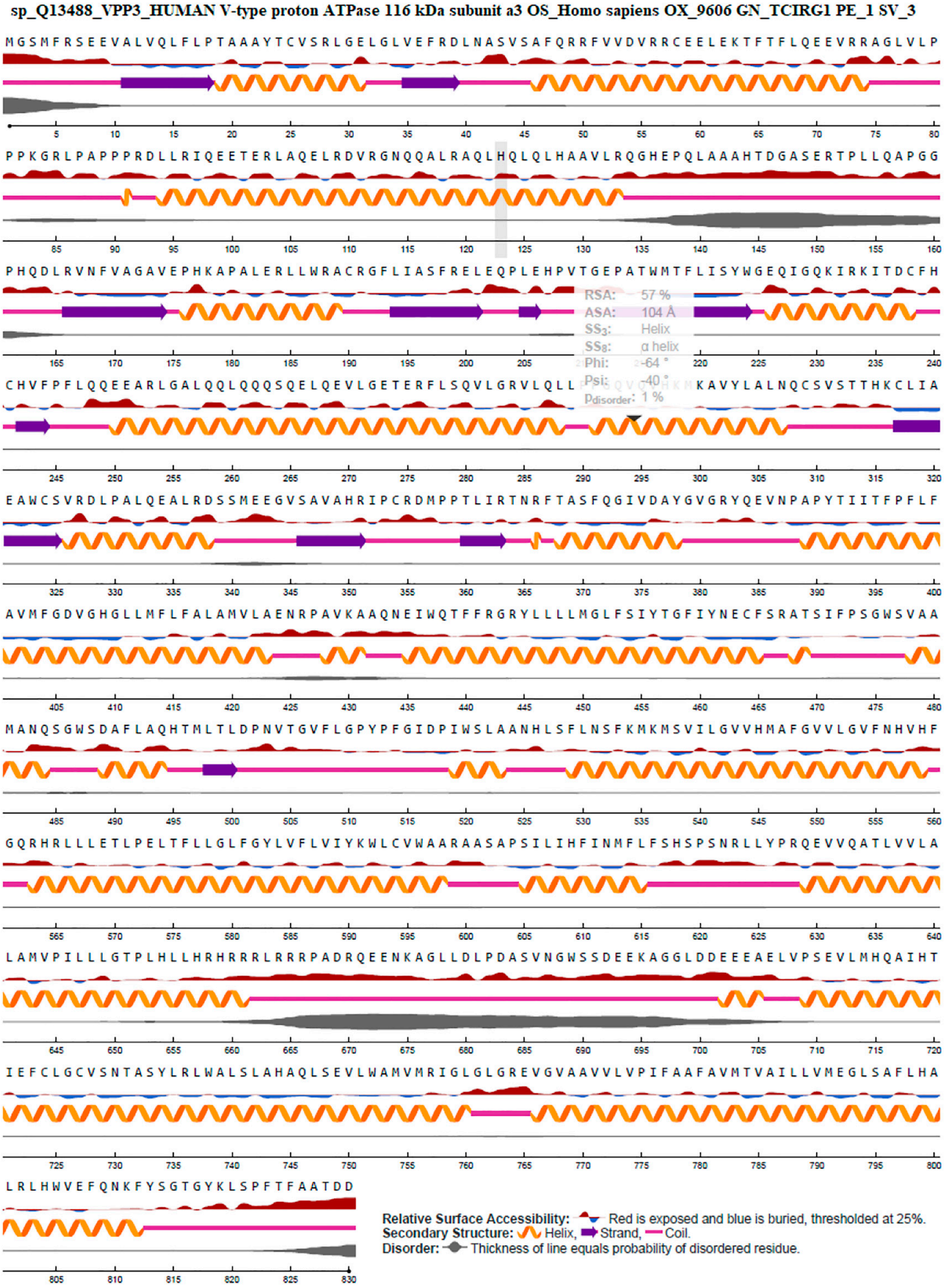


FIGURE 6 | Secondary structure prediction by Net-SurfP-2.0.

R444L (rs137853151), N730S (rs200209146), and S532C (rs371214361) with 0.84B, 0.84B, and 0.81B, respectively. F610S, M546V, and P572L have RMSD values of 0.B, 0.78B, and 0.78B, respectively, indicating no structural differences from wild type. Other nsSNPs showed slight variation which included I721N (0.53B RMSD), A732T (0.78B RMSD), R51C (0.78B RMSD), A717D (0.73B RMSD), E722K (0.46B RMSD), R57H (0.48B RMSD), R109W (0.78B RMSD), R191H (0.49B RMSD),

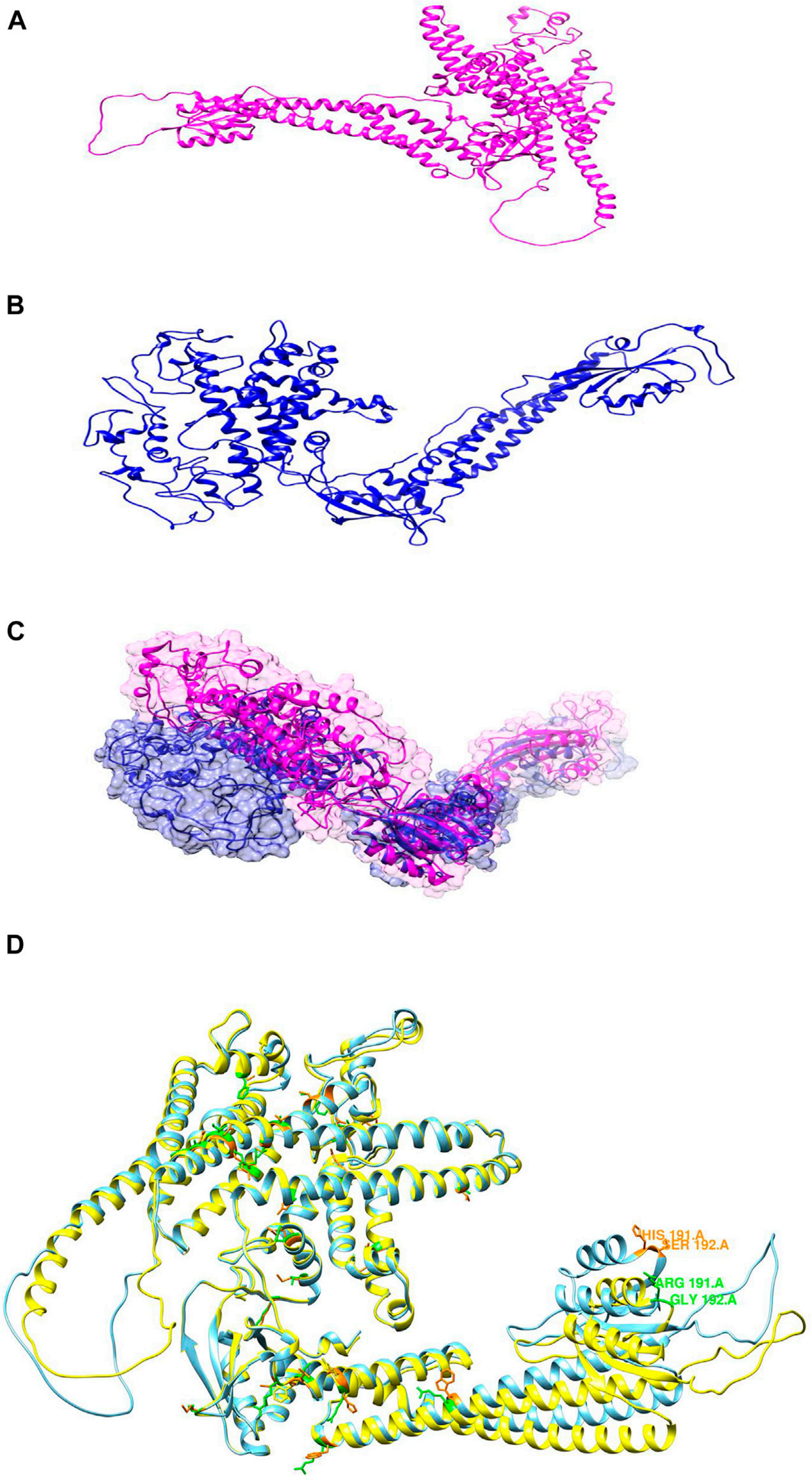


FIGURE 7 | (A) 3D structure of wild type protein predicted by AlphaFold2. **(B)** 3D predicted structure of Mutant protein. **(C)** superimposition of 3D structure of Mutant (blue) and Wild Type Magenta. **(D)** Superimposition of initial 3D structure of Mutant (cyan) and Wild Type (yellow). **(E)** superimposition of 3D structure of Mutant (cyan) and Wild Type (yellow). **(F)** Superimposition of 3D structure of Mutant (cyan) and Wild Type (yellow) at 50 ns. **(G)** 3D structure of Wild type at 100 ns. **(H)** 3D structure of Mutant at 100 ns.

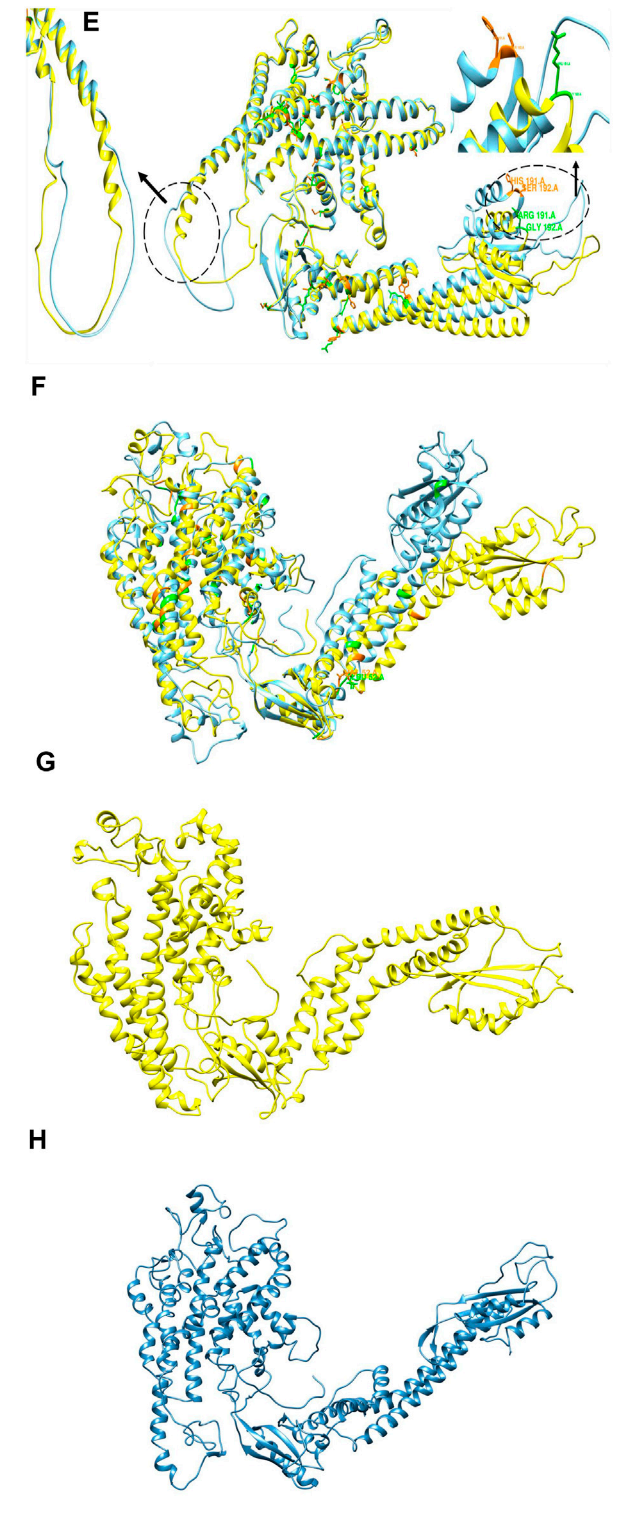


FIGURE 7 | (Continued).

G192C (0.78B RMSD), F529L (0.58B RMSD), H804Q (0.48B RMSD), G405R (0.48B RMSD) S474W (0.53B RMSD), G458S (0.48B RMSD), R56P (0.48B RMSD), R56W (0.78B RMSD), G379C (0.58B RMSD), R757C (0.48B RMSD), V375M (0.54B RMSD), T314M (0.78B RMSD), D517N (0.49B RMSD), T368M (0.78B RMSD), A417T (0.40B RMSD), R363C (0.78B RMSD), A778V (0.76B RMSD) and R50C (0.78B RMSD). **Table 3** shows the TMscores and RMSD values. The four nsSNPs with the greatest RMSD values (R92W, R444L, N730S, and S532C) were chosen and submitted to ITASSER for remodeling. The protein structure produced by the ITASSER is the most dependable since it is the most powerful modeling tool by using Chimera 1.11. Phyre2 Wild type mutant and three mutations superimposed on the wild-type TCIRG1 protein are shown in **Supplementary Figure S9** while validation results for the wild and mutant versions of the 3D models were good, and the Ramachandran plots may be found in the (**Supplementary Figure S10**).

3.15 Clinical Identification of Deleterious V52L nsSNP in a Patients Having Symptoms Related to PID

One of our patient who was a Russian kid 7 years old was suspected for Congenital *Neutropenia*, having symptoms related to chronic infections (right-side catarrhal otitis, acute rhinitis, and chronic tonsillopharyngitis). Whole genome sequencing (WGS) was conducted and the result showed no mutations for the suspected disorder. Analysis of the whole genome sequencing data of the patient was carried out using the BWA, GATK4, VCFtools software. An analysis of the so-called "candidate variant filtering" was performed using the ANNOVAR software and the Combined Annotation Dependent Depletion (CADD) database, and its results are schematically presented in Figure 28. The first filtration step was to remove all synonymous SNV, non-frames InDels and embodiments are marked as "NA" or "unknown". A total of 270 were identified variants or INDEL SNV. Then, the identified variants were filtered by overlaying on the known 351 PID genes and known congenital neutropenia genes. Selected 111 variants were retained to search for more possible ones. After eliminating the common variants, whose Minor allele frequency (MAF)>0.01 for The Exome Aggregation Consortium (ExAC), 1000g and The Genome Aggregation Database (gnomAD), a total of six rare variants remained. To select pathogenic mutations, CADD, the Functional Analysis through Hidden Markov Models (FATHMM), and Protein Variation Effect Analyzer (PROVEAN) models were used, and finally, four mutations that are likely to lead to the development of the disease in this patient were predicted. In particular, a mutation (g. 68041789G > C) was identified in the TCIRG1 gene. The mutation was V52L which in our Insilco analysis this mutation was predicted through many algorithm tools and this mutation was found to disturb the function and structure of TCIRG1 protein.

3.16 Simulation

The wild type and mutant proteins were preprocessed using Protein Preparation Wizard of Maestro, which included complex optimization and minimization. All the systems were prepared using the System Builder tool. TIP3P, a solvent model with an orthorhombic box, was chosen. (Transferable

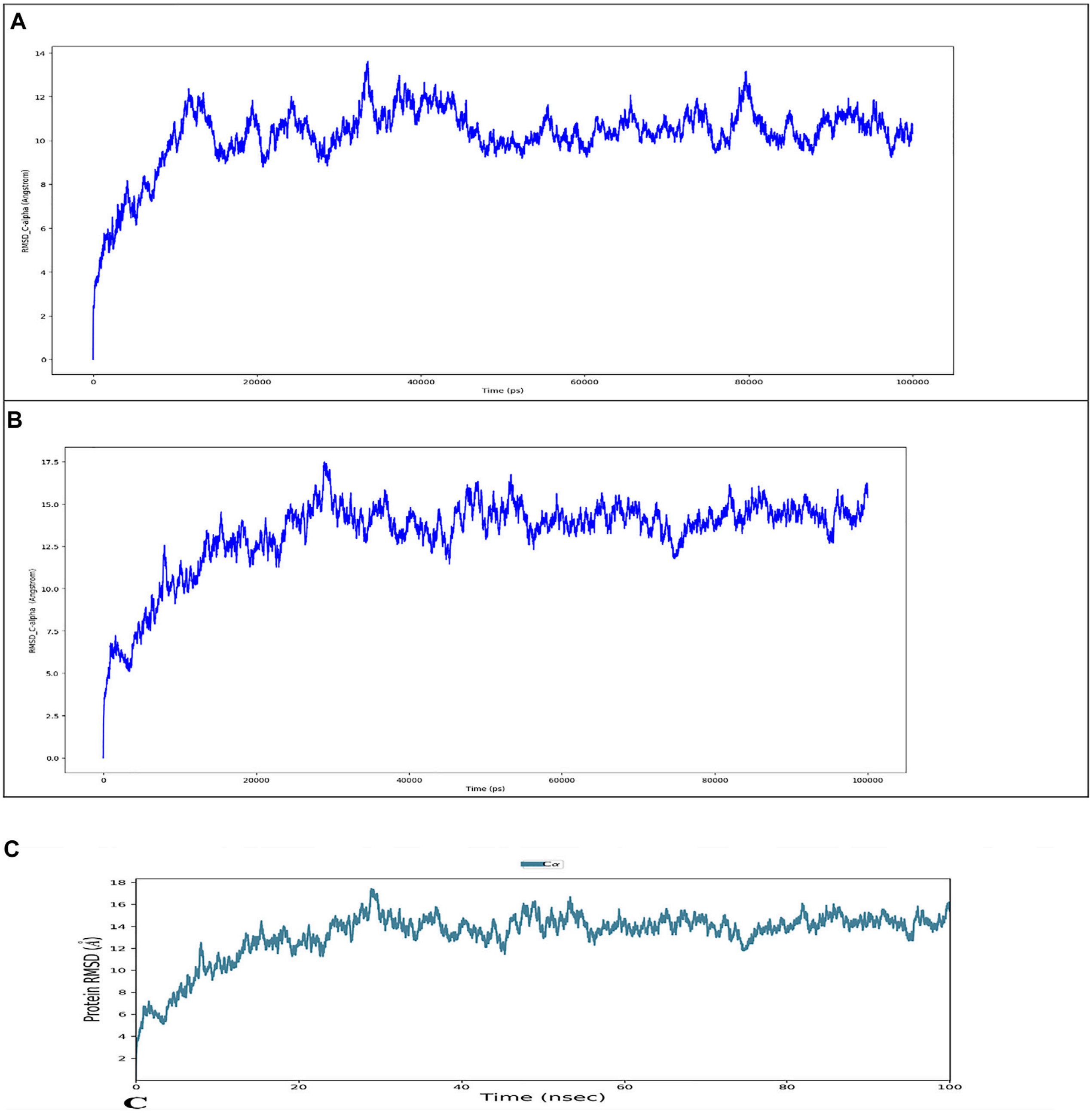


FIGURE 8 | Root mean square deviation (RMSD) of the C-alpha atoms of Wild Type **(A)** and Mutant by HHpred **(B)** and Mutant by Alphafold2 **(C)** with time. The left *Y*-axis shows the variation of proteins RMSD through time.

Intermolecular Interaction Potential three Points). In the simulation, the OPLS 2005 force field was used (Rasheed et al., 2021). To make the models neutral, counter ions were introduced. To mimic physiological conditions, 0.15 M sodium chloride (NaCl) was added. The NPT ensemble with 300 K temperature and 1 atm pressure was chosen for the entire simulation. The models were relaxed before the simulation. The trajectories were saved for examination after every 100 ps, and the simulation's stability was verified by comparing the protein and ligand's root mean square deviation (RMSD) over time.

Figure 8 depicts the evolution of RMSD values for the C-alpha atoms of protein over time. The plot shows that the protein reaches stability at 20,000 ps. After that, for the length of the simulation, fluctuations in RMSD values for wild type remain within 2.0 Angstrom, which is acceptable (Pedersen et al., 2021). The mutant protein RMSD values fluctuate within 3.5 Angstrom after they have been equilibrated. These findings indicate that the mutant protein has higher RMSD throughout the simulation period. On the RMSF graphic (**Figure 9A, B**), peaks represent portions of the proteins that fluctuate the most during the simulation. Protein tails (both N- and C-terminal) typically change more than any other part of the protein. Alpha helices and beta strands, for example, are usually stiffer than the unstructured section of the protein and fluctuate less than loop portions. According to MD trajectories, the residues with greater peaks belong to loop areas or N and C-terminal zones. Alpha-helices and beta-strands are monitored as secondary structure elements during the simulation (SSE). The graph above depicts the distribution of SSE by residue index across the protein structures. The mutant and wild total energy, Vander Waal's energy, and Secondary structure element (SSE are shown in **Figure 9C-J** as mutant show different total energy and Vander Waal's energy from the wild.

3.17 Intramolecular H-Bonds can Be Detected Throughout the Simulation

As seen in **Figure 10**, most of the significant intramolecular interactions discovered by MD are hydrogen bonds. A timeline depicts the interactions and contacts. The distribution of atoms in a protein around its axis is known as the radius of gyration (Rg). Rg is the length that reflects the distance between the rotating point and the place where the energy transfer has the greatest effect. This conceptual idea also aids in the identification of diverse polymer kinds, such as proteins. The two most important markers for forecasting the structural activity of a macromolecule are the calculation of Rg and distance calculations. The pace of folding of a protein is directly related to its compactness, which may be tracked using an advanced computer approach for determining the radius of gyration **Figure 11**.

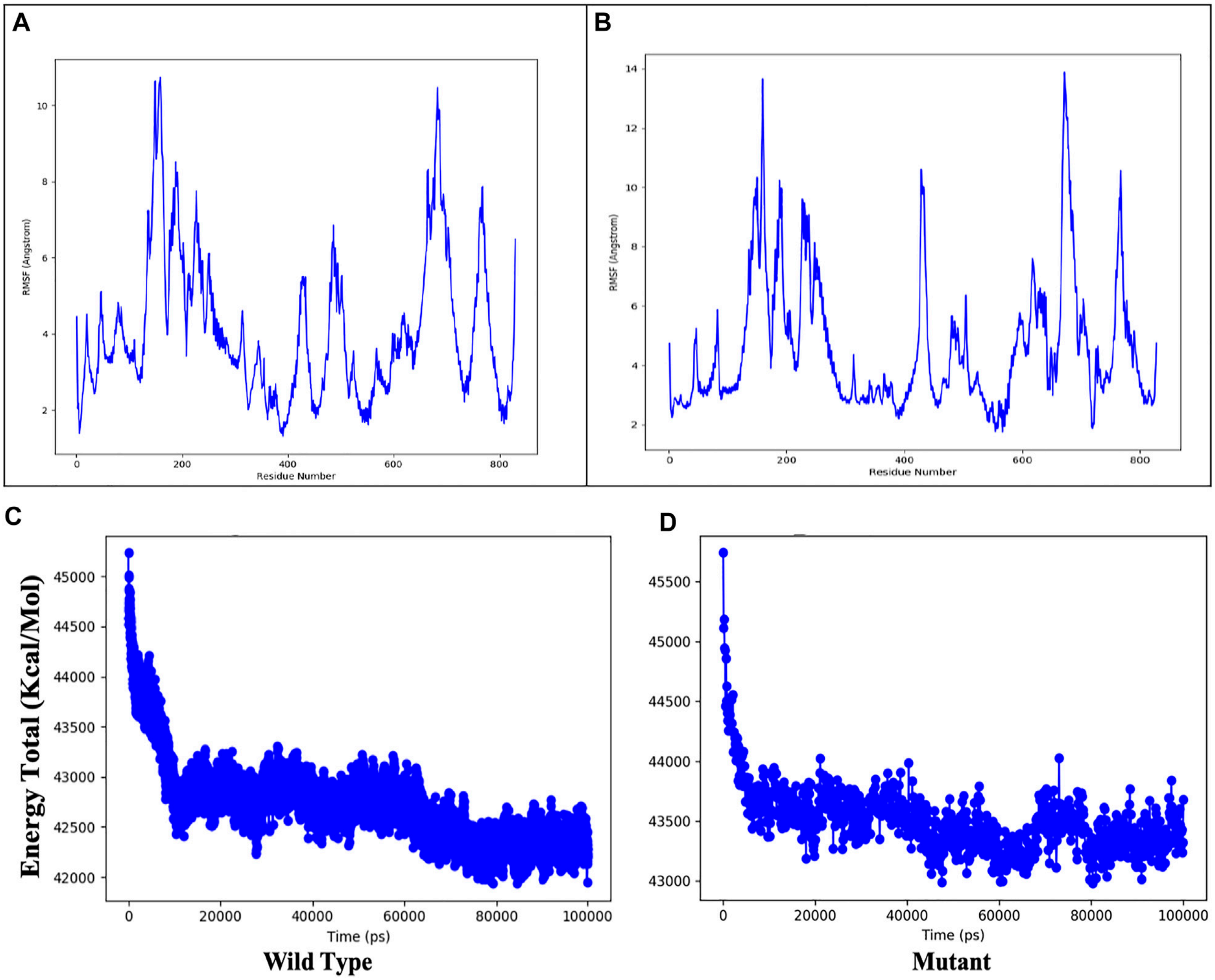


FIGURE 9 | Residue wise Root Mean Square Fluctuation (RMSF) of **(A)** Wild Type protein, **(B)** Mutant protein, total energy of the wild type**(C)** compared to mutant protein (D, Vander Waal's energy of the wild type**(E)** compared to mutant protein **(F)**, Secondary structure element (SSE) percentage of the wild type **(G)** and mutant protein **(H)**, Distribution of SSE by residue index across the protein structures, alpha helices (orange), beta strands (cyan) and loops (white) along the simulated time of 100 ns.

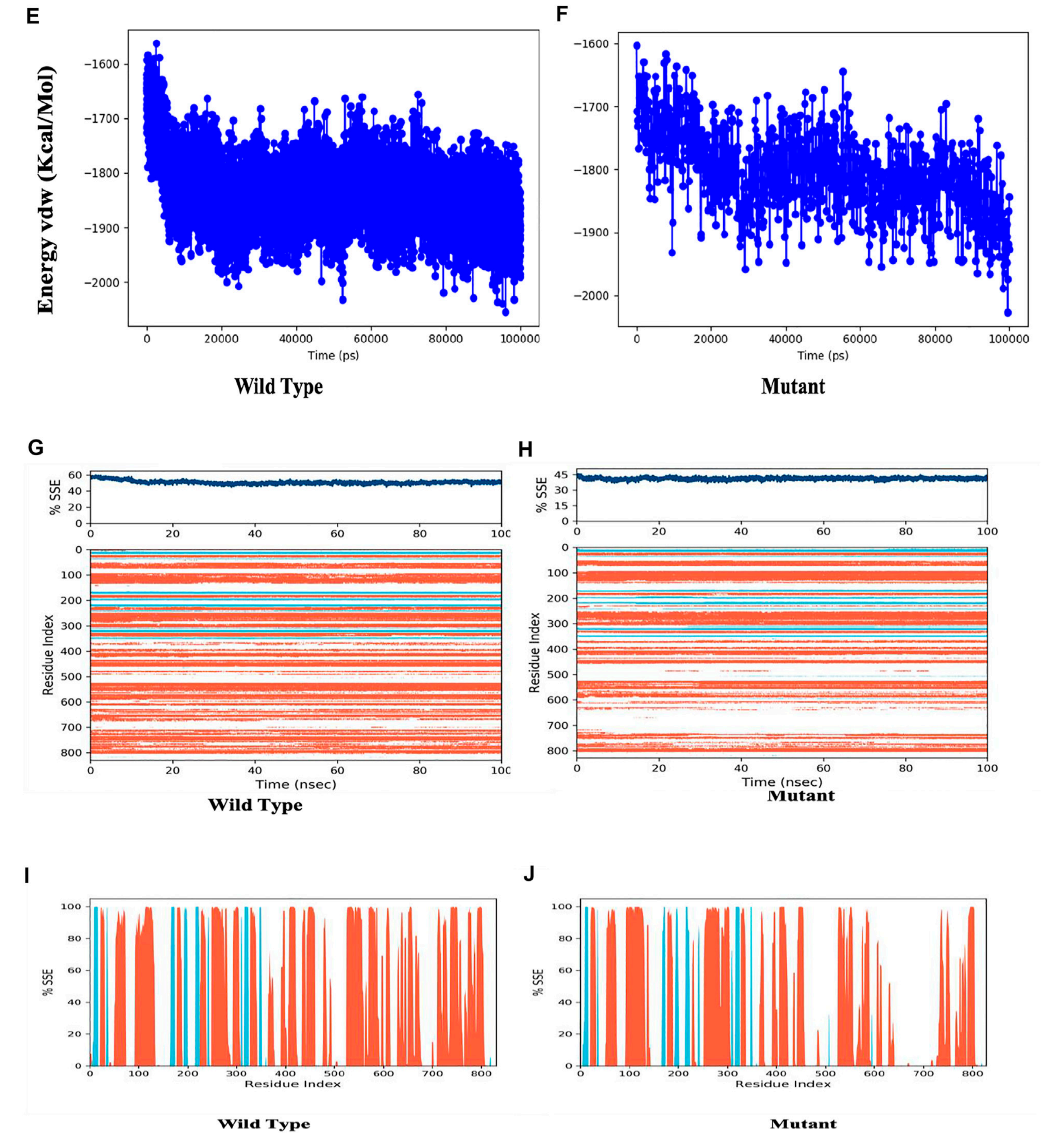


FIGURE 9 | (Continued).

4 DISCUSSION

A number of studies have found a relationship between SNPs in the TCIRG1 gene and osteopetrosis and congenital neutropenia. (Sobacchi et al., 2001; Susani et al., 2004; Makaryan et al., 2014; Scimeca et al., 2003; Rosenthal et al., 2016). TCIRG1 still has far too many SNPs that could play an impact on the disorders caused by this gene. We looked at TCIRG1's nsSNPs to discover which ones were the most detrimental and could be linked to Osteopetrosis, congenital neutropenia, and other immune-related diseases in this study. In this work, the dbSNP database revealed 811 nsSNPs in the TCIRG1 gene. Sixty-four

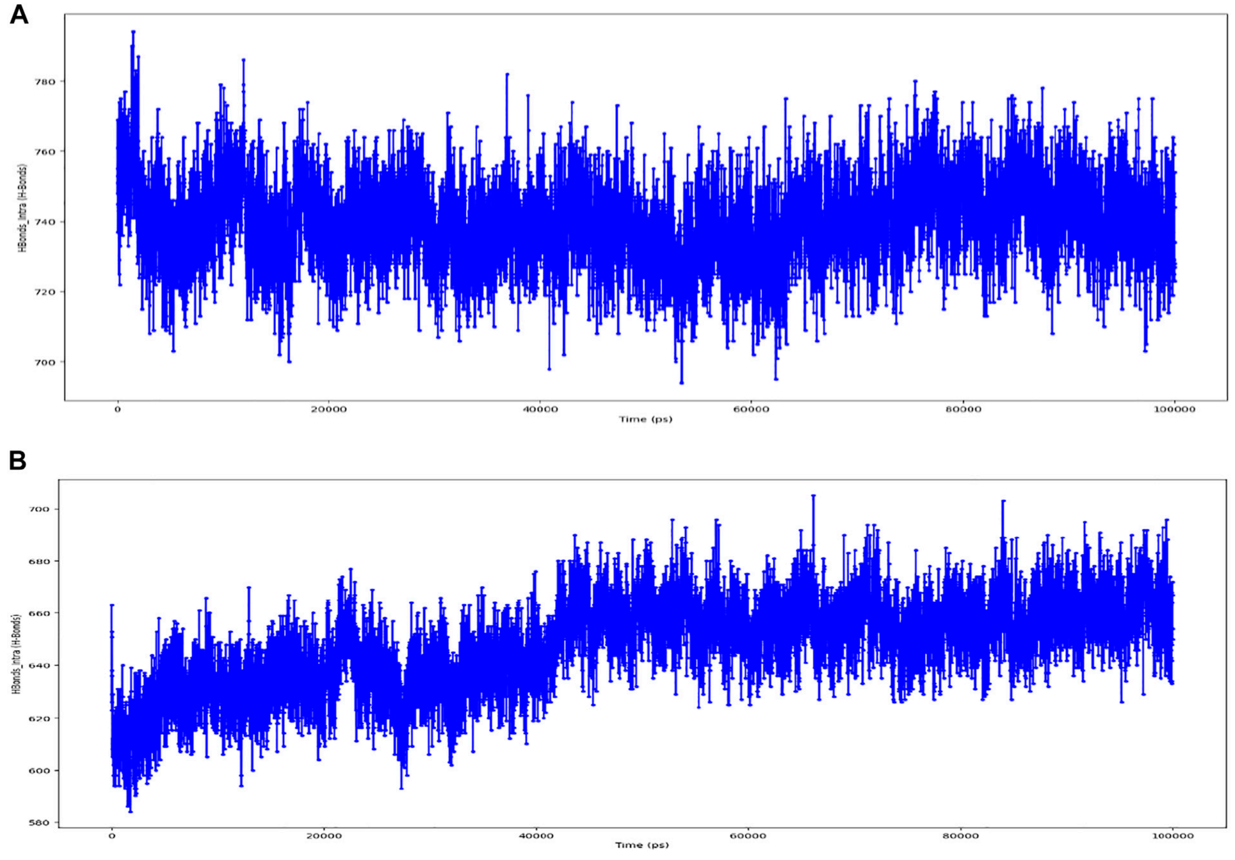


FIGURE 10 | (A) timeline representation of the interactions and contacts (H-bonds) Wild Type. **(B)** timeline representation of the interactions and contacts (H-bonds) of Mutant.

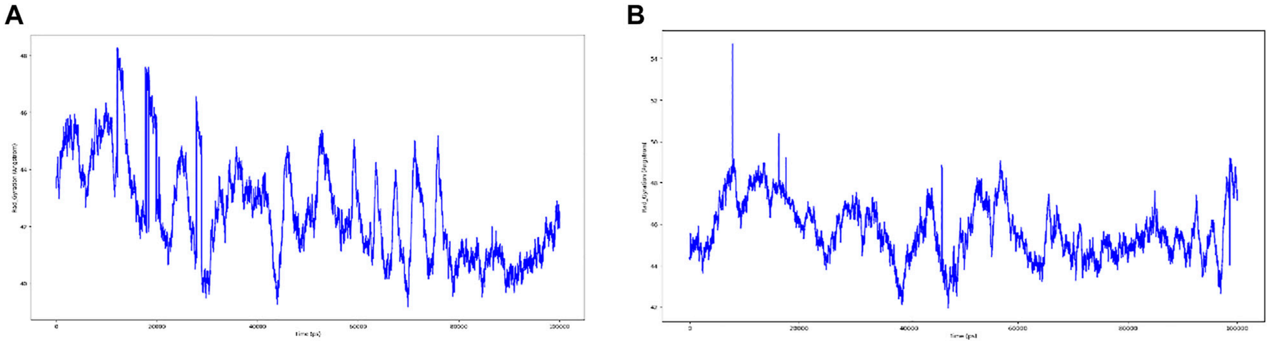


FIGURE 11 | Radius of gyration, **(A)** Wild Type Protein, **(B)** Mutant Protein.

nsSNPs in the TCIRG1 gene were validated as high-risk detrimental by SIFT and PolyPhen. The top fifteen high-risk nsSNPs in (**Table 4**) have been verified as extremely harmful by all state-of-the-art prediction techniques employed in the study. These fifteen nsSNPs (P572L, M546V, I721N, F610S, A732T, F51S, A717D, E722K, R57H, R109W, R191H, S532C, G192S, F529L, and H804Q) have not yet been connected to TCIRG1 gene-related osteopetrosis and congenital neutropenia, however, they could be utilized as a markers nsSNPs variants whenever diagnosing disorders related with TCIRG1 gene. These nsSNPs have been connected to their participation in the pathophysiology of TCIRG1-related

TABLE 4 | confirmation of SIFT and Poly Phen2 predicated highly pathogenic nsSNPs through different predication tools.

Aas	LRT	Mutation taster	Mutation accessor	PROVEAN	FATHMM	VEST3	MTA SVM	METALR	M-CAP	CADD	DANN	FATHMM-MKK	PhD-SNP	PANTHER	SNP-GO	P-MUT	SNAP2
P572L	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
M546V	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
I721N	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F610S	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
A732T	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F51S	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
A717D	N	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
E722K	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R57H	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R109W	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R191H	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
S532C	D	D	H	D	D	D	D	D	D	D	T	D	D	D	D	D	D
G192S	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F529L	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
H804Q	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G405R	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
S474W	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G458S	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R444L	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R56P	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G379S	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R757C	D	D	M	D	D	D	D	D	D	D	T	D	D	D	D	D	D
N730S	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D	D
V375M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
T314M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
D517N	D	D	H	D	D	T	D	D	D	D	D	D	D	D	D	D	D
R92W	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D	D
T368M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
A417T	D	D	H	D	D	D	D	D	D	D	D	T	D	D	D	D	D
R363C	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R56W	D	D	H	D	D	T	D	T	-	D	D	D	D	D	D	D	D
A778V	D	D	M	D	D	D	D	D	D	D	D	D	D	D	N	N	N
R50C	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D	-
V52L	D	D	M	T	D	T	D	D	-	D	D	D	D	D	D	D	D

illnesses such as osteopetrosis and congenital neutropenia. Mutations (G405R, R444L, and D517N) reported in our study are already associated with osteopetrosis. These three mutations are predicated deleterious by all of the algorithm tools except one tool while ConSurf show that they are highly conserved. Our study confirms that these three mutations are shown to destabilize the TCIRG1 protein structure and function. Mutation V52L was identified by analyzing whole genome sequencing data of the patient suspected for congenital neutropenia and was carried out using the BWA, GATK4, VCFtools software. This mutation is shown by our study deleterious, highly conserved and destabilizing the protein structure and function. Fifteen nsSNPs (S474W, G458S, R56P, G379S, R757C, N730S, V375M, T314M, R92W, T368M, A417T, R363C, R56W, A778V and R50C) reported in our study are also deleterious they are not shown damaging by one or two predication tools but they also showed to destabilize protein stability and might be important nSNPs for TCIRG1 gene. ConSurf uses a combination of evolutionary conservation data and solvent accessibility predictions to determine whether an amino acid is conserved, exposed, functional, or structural. Highly conserved residues are predicted to be structurally or functionally relevant based on their positions on the protein surface and core (Ashkenazy et al., 2016). Amino acids involved in protein–protein interactions, for example, are expected to be more conserved. As a result, the nsSNPs that have been found in conserved areas are the most damaging nsSNPs (Miller and Kumar, 2001). Only 26 SNPs out of a total of 56 nsSNPs are found at evolutionary conserved, exposed, and functionally relevant residues (A20V, R56P, R57H, R191H, G192C, E321K, R366H, T368M, R444L, and E722K). There were 16 nsSNPs found at conserved, buried, and structurally significant residues (S7K, V52L, G379S, M403I, G405R, G458S, D517N, F529L, S532C, M546V, A640S, D683H, I732N, N730S, A732T, and H804Q). The remaining nsSNPs were discovered in either exposed or buried residues that were not predicted to have any structural or functional significance in the TCIRG1 protein. The I-Mutant 3.0 web server was used to estimate protein stability, and variations T570M, P572L, M546V, I721N, F610S, A732T, F51S, A717T, R57H, R109W, R191H, G192S, F529L, G458W, R444L, R56P, G379S, N730S, V375M, R92W, and T368 All of these nsSNPs can be important in the diagnosis of the TCIRG1 gene because they reduce the protein's stability. In silico tools have been used to conduct various investigations on genes and proteins such as the CCBE1, ADA, and GJA3 genes (Shinwari et al., 2021; Essadssi et al., 2019; Zhang et al., 2020). Such research may lead to the discovery of novel therapeutic targets. All of the simulated structures were validated using RAMPAGE data. Protein designs with core RAMPAGE values greater than 80% are regarded to be superior (Essadssi et al., 2019). RAMPAGE values for the structure shown in **Figure 5A** (TCGIR1 wild type) were 90.5% preferred residues, 8.8% allowed, 0.6% usually allowed, and 0.2% forbidden. Similarly, for mutants P572L (90.7% favored residues, 8.6% allowed, 0.5% generally allowed, and disallowed 0.2%), R92W (90.5% favored residues, 8.8% allowed, 0.6% generally allowed, and disallowed 0.2%) R444L (90.6% favored residues, 8.8% allowed, 0.3% generally allowed, and disallowed 0.2%), and N730I (90.4% favored residues, 8.8% allowed, 0.3% generally allowed, and disallowed 0.3% and S532C (90.2% favored residues, 8.8% allowed, 0.5% generally allowed, and disallowed 0.6%, and A732T (90.6% favored residues, 8.4% allowed, 0.9% generally allowed, and disallowed 0.2% all the structures were somehow validated. Protein shapes and functions are influenced by PTMs, which have been found to be important in cell signaling, protein–protein interactions, and other essential events in biological systems (Dai and Gu, 2010; Shiloh and Ziv, 2013). We wanted to determine if the selected nsSNPs changed the PTMs of the TCIRG1 protein in this investigation. PTM sites in the protein under research were predicted using a variety of bioinformatics methods. Because lysine residues in certain proteins are methylated, this changes their interaction with DNA and regulates gene expression, methylation is a key PTM. Another essential method for protein regulation is the molecular switch, which adapts the protein to execute functions such as protein structure conformational changes, protein activation and deactivation, and signal transduction pathways (Deutscher and Saier, 2005; Puttick et al., 2008; Cieśla et al., 2011; Sawicka and Seiser, 2014). Among these predictions, the ConSurf Conservation profile shows that rs137 6162684 is highly conserved, exposed, and functionally relevant, indicating its relevance. Phosphorylation capability is demonstrated at position rs137 6162684, which also happens to be structurally essential and highly conserved (ConSurf Prediction), making it incredibly crucial. Ubiquitylation is a protein degradation mechanism that also helps to repair DNA damage (Gallo et al., 2017). Protein function and stability are both dependent on it. In protein–protein interactions, it has a structural role. As revealed by these PTM predictions, phosphorylation is the only PTM that may have a significant impact on TCIRG1 protein structure and function, with residuals rs121908251 and other reported locations in our study having the most significant phosphorylation sites. All of the phosphorylation and ubiquitylation sites identified in our investigation could play a significant role in protein stability and other TCIRG1 gene-related functions. According to GeneMANIA's predictions, TCIRG1 is the most interacting gene in our study and co-expressed with a variety of genes. Any of the most detrimental nsSNPs in the TCIRG1 gene will eventually influence and impair the normal functioning of other linked genes, based on their interaction patterns and coexpression profiles. This highlights the significance of these interconnected and co-expressed genes in congenital neutropenia and other primary immunodeficiency disorders. Our research has all of the essential data and analyses for finding the most damaging nsSNPs because it was thorough. Every study, including ours, is limited in some way. Our research is centered on computer tools and web servers that use mathematical and statistical methodologies. As a result, further research is needed to corroborate these findings. Our findings shed light on TCIRG1 nsSNPs, their conservation, impact on protein stability and functions, protein 3D structure, PTM potential sites, ligand binding sites, and gene-gene interactions with other genes, all of which could be useful in future TCIRG1 research to better understand its role in diseases such as osteopetrosis and congenital neutropenia. The effect of substitutions on protein function was investigated using FTSite. Three ligand-binding sites were predicted by the FTSite server, each having 14.9 and 13 residues. We discovered that several alterations are involved in the ligand-binding region and form the catalytic coordination sphere, which could affect the binding affinity of the TCIRG1 protein. As predicted by SIFT software and other prediction approaches, these changes had an impact on the TCIRG1 structure and decreased its stability.

5 CONCLUSION

Out of 64 SIFT and PolyPhen deleterious predicted nsSNPs variants, this study identified 33 novel sites which are deleterious, while 15 of which were highly deleterious variants predicted damaging/deleterious by all of the algorithms tools used in the study, and these variant mutations may lead to disruption of the original conformation of the native protein. When compared to the original protein structure, our molecular dynamics technique revealed a shift in deviation in critical locations of the mutant structures. These discrepancies can compromise the confirmation of the secondary structure and, as a result, the protein's stability. We also noticed that the ATP binding capability of the mutant proteins was less than that of the native protein. Although the G405R, R444L, and D517N mutant has been previously associated with osteopetrosis according to the literature, no one has predicted the other 12 mutants to be linked with any diseases. As a result, it is conceivable that the unreported nsSNP can cause disease by affecting protein activation or efficiency. The findings of this study will aid future genome association studies in distinguishing harmful SNPs linked with various individual individuals with osteopetrosis and congenital neutropenia. As a result, comprehensive clinical-trial-based investigations on a broad population are required to characterize this data on SNPs, as are experimental mutational research to validate the findings.

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

KS (Conception or design of the work, Data collection, Data analysis and interpretation, critical revision of the article, and Drafting the article), HMR (Data analysis and interpretation), GL (Data analysis and interpretation), MAB (Drafting the article), IAT (Supervision and critical revision of the article) VAC (Supervision and critical revision of the article).

ACKNOWLEDGMENTS

The work was carried out within the framework of research at the Institute of Immunology and Physiology, Ural Branch of the Russian Academy of Sciences, project number AAAA-A21-121012090091-6.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.879875/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

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al. (2016). ConSurf 2016: an Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules. *Nucleic Acids Res.* 44 (W1), W344–W350. doi:10.1093/nar/gkw408

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 (Oxford, England)* 20 (8), 1322–1324. doi:10.1093/bioinformatics/bth070

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

Bowers, K. J. a. C., David, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). "Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters," in SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing (USA: IEEE), 43. doi:10.1109/sc.2006.54

Boztug, K., and Klein, C. (2009). Novel Genetic Etiologies of Severe Congenital Neutropenia. *Curr. Opin. Immunol.* 21 (5), 472–480. doi:10.1016/j.coi.2009.09.003

Bromberg, Y., Yachdav, G., and Rost, B. (2008). SNAP Predicts Effect of Mutations on Protein Function. *Bioinformatics (Oxford, England)* 24 (20), 2397–2398. doi:10.1093/bioinformatics/btn435

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, W306–W310. doi:10.1093/nar/gki375

Carlsson, G., and Fasth, A. (2001). Infantile Genetic Agranulocytosis, Morbus Kostmann: Presentation of Six Cases from the Original "Kostmann Family" and a Review. *Acta Paediatr.* 90 (7), 757–764. doi:10.1080/080352501750315663

Carugo, O., and Pongor, S. (2001). A Normalized Root-Mean-Square Distance for Comparing Protein Three-Dimensional Structures. *Protein Sci. : a Publ. Protein Soc.* 10 (7), 1470–1473. doi:10.1110/ps.690101

Cavasotto, C. N., and Phatak, S. S. (2009). Homology Modeling in Drug Discovery: Current Trends and Applications. *Drug Discov. Today* 14, 676–683. doi:10.1016/j.drudis.2009.04.006

Chávez-Güitrón, L. E., Cerón-Torres, T., Sobacchi, C., Ochoa-Ruiz, E., and Villegas-Huesca, S. (2018). Autosomal recessive osteopetrosis type I: description of pathogenic variant of TCIRG1 gene. Osteopetrosis infantil maligna: descripción de una nueva mutación patogénica de

TCIRG1. *Boletin Medico Del. Hosp. Infantil de Mexico* 75 (4), 255–259. doi:10.24875/BMHIM.M18000028

Chitrala, K. N., and Yeguvapalli, S. (2014). Computational Screening and Molecular Dynamic Simulation of Breast Cancer Associated Deleterious Non-synonymous Single Nucleotide Polymorphisms in TP53 Gene. *PloS one* 9 (8), e104242. doi:10.1371/journal.pone.0104242

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PloS one* 7 (10), e46688. doi:10.1371/journal.pone.0046688

Cieśla, J., Frączyk, T., and Rode, W. (2011). Phosphorylation of Basic Amino Acid Residues in Proteins: Important but Easily Missed. *Acta Biochim. Pol.* 58 (2), 137–148.

Dai, C., and Gu, W. (2010). p53 post-translational Modification: Deregulated in Tumorigenesis. *Trends Molecular Medicine* 16 (11), 528–536. doi:10.1016/j.molmed.2010.09.002

Dale, D. C., Person, R. E., Bolyard, A. A., Aprikyan, A. G., Bos, C., Bonilla, M. A., et al. (2000). Mutations in the Gene Encoding Neutrophil Elastase in Congenital and Cyclic Neutropenia. *Blood* 96 (7), 2317–2322. doi:10.1182/blood.v96.7.2317

Deng, W., Wang, Y., Ma, L., Zhang, Y., Ullah, S., and Xue, Y. (2017). Computational Prediction of Methylation Types of Covalently Modified Lysine and Arginine Residues in Proteins. *Brief. Bioinformatics* 18 (4), 647–658. doi:10.1093/bib/bbw041

Deutscher, J., and Saier, M. H., Jr (2005). Ser/Thr/Tyr Protein Phosphorylation in Bacteria - for Long Time Neglected, Now Well Established. *J. Mol. Microbiol. Biotechnol.* 9 (3-4), 125–131. doi:10.1159/000089641

Doniger, S. W., Kim, H. S., Swain, D., Corcuera, D., Williams, M., Yang, S. P., et al. (2008). A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLoS Genet.* 4 (8), e1000183. doi:10.1371/journal.pgen.1000183

Essadssi, S., Krami, A. M., Elkhattabi, L., Elkarhat, Z., Amalou, G., Abdelghaffar, H., et al. (2019). Computational Analysis of nsSNPs of *ADA* Gene in Severe Combined Immunodeficiency Using Molecular Modeling and Dynamics Simulation. *J. Immunol. Res.* 2019, 5902391. doi:10.1155/2019/5902391

Ferreira, L. G., Dos Santos, R. N., Oliva, G., and Andricopulo, A. D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* 20 (7), 13384–13421. doi:10.3390/molecules200713384

Ferrer-Costa, C., Gelpí, J. L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. (2005). PMUT: a Web-Based Tool for the Annotation of Pathological Mutations on Proteins. *Bioinformatics (Oxford, England)* 21 (14), 3176–3178. doi:10.1093/bioinformatics/bti486

Frattini, A., Orchard, P. J., Sobacchi, C., Giliani, S., Abinun, M., Mattsson, J. P., et al. (2000). Defects in TCIRG1 Subunit of the Vacuolar Proton Pump Are Responsible for a Subset of Human Autosomal Recessive Osteopetrosis. *Nat. Genet.* 25 (3), 343–346. doi:10.1038/77131

Gallo, L. H., Ko, J., and Donoghue, D. J. (2017). The Importance of Regulatory Ubiquitination in Cancer and Metastasis. *Cell Cycle (Georgetown, Tex.)* 16 (7), 634–648. doi:10.1080/15384101.2017.1288326

George Priya Doss, C., Rajasekaran, R., Sudandiradoss, C., Ramanathan, K., Purohit, R., and Sethumadhavan, R. (2008). A Novel Computational and Structural Analysis of nsSNPs in CFTR Gene. *Genomic Med.* 2 (1-2), 23–32. doi:10.1007/s11568-008-9019-8

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders. *Nucleic Acids Res.* 33 (Database issue), D514–D517. doi:10.1093/nar/gki033

Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009). Fast and Accurate Automatic Structure Prediction with HHpred, *Proteins.* 77 (Suppl. 9), 128–132. doi:10.1002/prot.22499

Hildebrand, P. W., Rose, A. S., and Tiemann, J. K. S. (2019). Bringing Molecular Dynamics Simulation Data into View. *Trends Biochem. Sci.* 44 (11), 902–913. doi:10.1016/j.tibs.2019.06.004

Horwitz, M., Benson, K. F., Person, R. E., Aprikyan, A. G., and Dale, D. C. (1999). Mutations in ELA2, Encoding Neutrophil Elastase, Define a 21-day Biological Clock in Cyclic Haematopoiesis. *Nat. Genet.* 23 (4), 433–436. doi:10.1038/70544

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: The Integrative Protein Signature Database. *Nucleic Acids Res.* 37, 211–215. doi:10.1093/nar/gkn785

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Kamatani, N., Sekine, A., Kitamoto, T., Iida, A., Saito, S., Kogame, A., et al. (2004). Large-scale Single-Nucleotide Polymorphism (SNP) and Haplotype Analyses, Using Dense SNP Maps, of 199 Drug-Related Genes in 752 Subjects: the Analysis of the Association between Uncommon SNPs within Haplotype Blocks and the Haplotypes Constructed with Haplotype-Tagging SNPs. *Am. J. Hum. Genet.* 75 (2), 190–203. doi:10.1086/422853

Katsila, T., Spyroulias, G. A., Patrinos, G. P., and Matsoukas, M. T. (2016). Computational Approaches in Target Identification and Drug Discovery. *Comput. Struct. Biotechnol. J.* 14, 177–184. doi:10.1016/j.csbj.2016.04.004

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 Web portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* 10 (6), 845–858. doi:10.1038/nprot.2015.053

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., et al. (2019). NetSurfP-2.0: Improved Prediction of Protein Structural Features by Integrated Deep Learning. *Proteins* 87 (6), 520–527. doi:10.1002/prot.25674

Klein, C., Grudzien, M., Appaswamy, G., Germeshausen, M., Sandrock, I., Schäffer, A. A., et al. (2007). HAX1 Deficiency Causes Autosomal Recessive Severe Congenital Neutropenia (Kostmann Disease). *Nat. Genet.* 39 (1), 86–92. doi:10.1038/ng1940

Kornak, U., Schulz, A., Friedrich, W., Uhlhaas, S., Kremens, B., Voit, T., et al. (2000). Mutations in the A3 Subunit of the Vacuolar H(+)-ATPase Cause Infantile Malignant Osteopetrosis. *Hum. Mol. Genet.* 9 (13), 2059–2063. doi:10.1093/hmg/9.13.2059

KOSTMANN, R. (1956). Infantile Genetic Agranulocytosis; Agranulocytosis Infantilis Hereditaria. *Acta Paediatr.* 45 (Suppl. 105), 1–78. doi:10.1111/j.1651-2227.1956.tb06875.x

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. App. Cryst.* 26, 283–291. doi:10.1107/s0021889892009944

Lee, J. E., Choi, J. H., Lee, J. H., and Lee, M. G. (2005). Gene SNPs and Mutations in Clinical Genetic Testing: Haplotype-Based Testing and Analysis. *Mutat. Res.* 573 (1-2), 195–204. doi:10.1016/j.mrfmmm.2004.08.018

Li, J., Shi, L., Zhang, K., Zhang, Y., Hu, S., Zhao, T., et al. (2018). VarCards: an Integrated Genetic and Clinical Database for Coding Variants in the Human Genome. *Nucleic Acids Res.* 46 (D1), D1039–D1048. doi:10.1093/nar/gkx1039

Li, Y. P., Chen, W., Liang, Y., Li, E., and Stashenko, P. (1999). Atp6i-deficient Mice Exhibit Severe Osteopetrosis Due to Loss of Osteoclast-Mediated Extracellular Acidification. *Nat. Genet.* 23 (4), 447–451. doi:10.1038/70563

Loría-Cortés, R., Quesada-Calvo, E., and Cordero-Chaverri, C. (1977). Osteopetrosis in Children: a Report of 26 Cases. *J. Pediatr.* 91 (1), 43–47. doi:10.1016/s0022-3476(77)80441-1

Maiorov, V. N., and Crippen, G. M. (1994). Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins. *J. Mol. Biol.* 235 (2), 625–634. doi:10.1006/jmbi.1994.1017

Makaryan, V., Rosenthal, E. A., Bolyard, A. A., Kelley, M. L., Below, J. E., Bamshad, M. J., et al. (2014). TCIRG1-associated Congenital Neutropenia. *Hum. Mutat.* 35 (7), 824–827. doi:10.1002/humu.22563

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

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., and Thornton, J. M. (1992). Stereochemical Quality of Protein Structure Coordinates. *Proteins* 12 (4), 345–364. doi:10.1002/prot.340120407

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Res.* 31 (13), 3812–3814. doi:10.1093/nar/gkg509

Ohlson, T., Wallner, B., and Elofsson, A. (2004). Profile-profile Methods Provide Improved Fold-Recognition: a Study of Different Profile-Profile Alignment Methods. *Proteins* 57, 188–197. doi:10.1002/prot.20184

Palagano, E., Blair, H. C., Pangrazio, A., Tourkova, I., Strina, D., Angius, A., et al. (2015). Buried in the Middle but Guilty: Intronic Mutations in the TCIRG1 Gene Cause Human Autosomal Recessive Osteopetrosis. *J. bone mineral Res. official J. Am. Soc. Bone Mineral Res.* 30 (10), 1814–1821. doi:10.1002/jbmr.2517

Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., et al. (2021). Effective Variant Filtering and Expected Candidate Variant Yield in Studies of Rare Human Disease. *NPJ Genomic Med.* 6 (1), 60. doi:10.1038/s41525-021-00227-3

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

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera-Aa Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25 (13), 1605–1612. doi:10.1002/jcc.20084

Puttick, J., Baker, E. N., and Delbaere, L. T. (2008). Histidine Phosphorylation in Biological Systems. *Biochim. Biophys. Acta* 1784 (1), 100–105. doi:10.1016/j.bbapap.2007.07.008

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., et al. (2010). Identification, Analysis, and Prediction of Protein Ubiquitination Sites. *Proteins* 78 (2), 365–380. doi:10.1002/prot.22555

Rajasekaran, R., Doss, G. P., Sudandiradoss, C., Ramanathan, K., Rituraj, P., and Sethumadhavan, R. (2008). Computational and Structural Investigation of Deleterious Functional SNPs in Breast Cancer BRCA2 Gene. *Sheng Wu Gong Cheng Xue Bao = Chin. J. Biotechnol.* 24 (5), 851–856. doi:10.1016/s1872-2075(08)60042-4

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human Non-synonymous SNPs: Server and Survey. *Nucleic Acids Res.* 30 (17), 3894–3900. doi:10.1093/nar/gkf493

Rasheed, M. A., Iqbal, M. N., Saddick, S., Ali, I., Khan, F. S., Kanwal, S., et al. (2021). Identification of Lead Compounds against Scm (Fms10) in Enterococcus Faecium Using Computer Aided Drug Designing. *Life (Basel)* 11 (2). doi:10.3390/life11020077

Rosenthal, E. A., Makaryan, V., Burt, A. A., Crosslin, D. R., Kim, D. S., Smith, J. D., et al. (2016). Association between Absolute Neutrophil Count and Variation at TCIRG1: The NHLBI Exome Sequencing Project. *Genet. Epidemiol.* 40 (6), 470–474. doi:10.1002/gepi.21976

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* 5 (4), 725–738. doi:10.1038/nprot.2010.5

Sawicka, A., and Seiser, C. (2014). Sensing Core Histone Phosphorylation - a Matter of Perfect Timing. *Biochim. Biophys. Acta* 1839 (8), 711–718. doi:10.1016/j.bbagrm.2014.04.013

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: an Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* 31, 3381–3385. doi:10.1093/nar/gkg520

Scimeca, J. C., Quincey, D., Parrinello, H., Romatet, D., Grosgeorge, J., Gaudray, P., et al. (2003). Novel Mutations in the TCIRG1 Gene Encoding the A3 Subunit of the Vacuolar Proton Pump in Patients Affected by Infantile Malignant Osteopetrosis. *Hum. Mutat.* 21 (2), 151–157. doi:10.1002/humu.10165

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 (1), 308–311. doi:10.1093/nar/29.1.308

Shiloh, Y., and Ziv, Y. (2013). The ATM Protein Kinase: Regulating the Cellular Response to Genotoxic Stress, and More. *Nat. Rev. Mol. cell Biol.* 14 (4), 197–210. doi:10.1038/nrm3546

Shinwari, K., Guojun, L., Deryabina, S. S., Bolkov, M. A., Tuzankina, I. A., and Chereshnev, V. A. (2021). Predicting the Most Deleterious Missense Nonsynonymous Single-Nucleotide Polymorphisms of Hennekam Syndrome-Causing CCBE1 Gene. *Scientific World J.* 2021, 6642626. doi:10.1155/2021/6642626

Shivakumar, D., Williams, J., Wu, Y., Damm, W., Shelley, J., and Sherman, W. (2010). Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theor. Comput.* 6 (5), 1509–1519. doi:10.1021/ct900587b

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi:10.1038/msb.2011.75

Sobacchi, C., Schulz, A., CoxonVilla, F. P. A., and Helfrich, M. H. (2013). Osteopetrosis: Genetics, Treatment and New Insights into Osteoclast Function. *Nat. Rev. Endocrinol.* 9 (9), 522–536. doi:10.1038/nrendo.2013.137

Sobacchi, C., Pangrazio, A., Lopez, A. G., Gomez, D. P., Caldana, M. E., Susani, L., et al. (2014). As Little as Needed: the Extraordinary Case of a Mild Recessive Osteopetrosis Owing to a Novel Splicing Hypomorphic Mutation in the TCIRG1 Gene. *J. bone mineral Res. official J. Am. Soc. Bone Mineral Res.* 29 (7), 1646–1650. doi:10.1002/jbmr.2203

Sobacchi, C., Schulz, A., Coxon, F. P., Villa, A., and Helfrich, M. H. (2013). Osteopetrosis: Genetics, Treatment and New Insights into Osteoclast Function. *Nat. Rev. Endocrinol.* 9 (9), 522–536. doi:10.1038/nrendo.2013.137

Sobacchi, C., Frattini, A., Orchard, P., Porras, O., Tezcan, I., Andolina, M., et al. (2001). The Mutational Spectrum of Human Malignant Autosomal Recessive Osteopetrosis. *Hum. Mol. Genet.* 10 (17), 1767–1773. doi:10.1093/hmg/10.17.1767

Stark, Z., and Savarirayan, R. (2009). Osteopetrosis. *Orphanet J. rare Dis.* 4, 5. doi:10.1186/1750-1172-4-5

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T., et al. (2013). Precision Mapping of the Human O-GalNAc Glycoproteome through SimpleCell Technology. *EMBO J.* 32 (10), 1478–1488. doi:10.1038/emboj.2013.79

Susani, L., Pangrazio, A., Sobacchi, C., Taranta, A., Mortier, G., Savarirayan, R., et al. (2004). TCIRG1-dependent Recessive Osteopetrosis: Mutation Analysis, Functional Identification of the Splicing Defects, and *In Vitro* rescue by U1 snRNA. *Hum. Mutat.* 24 (3), 225–235. doi:10.1002/humu.20076

Tang, H., and Thomas, P. D. (2016). PANTHER-PSEP: Predicting Disease-Causing Genetic Variants Using Position-specific Evolutionary Preservation. *Bioinformatics (Oxford, England)* 32 (14), 2230–2232. doi:10.1093/bioinformatics/btw222

Tolar, J., Teitelbaum, S. L., and Orchard, P. J. (2004). Osteopetrosis. *New Engl. J. Med.* 351 (27), 2839–2849. doi:10.1056/NEJMra040952

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148. doi:10.1093/nar/gkp846

Venselaar, H., Te Beek, T. A., Kuipers, R. K., Hekkelman, M. L., and Vriend, G. (2010). Protein Structure Analysis of Mutations Causing Inheritable Diseases. An E-Science Approach with Life Scientist Friendly Interfaces. *BMC bioinformatics* 11, 548. doi:10.1186/1471-2105-11-548

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function. *Nucleic Acids Res.* 38, W214–W220. doi:10.1093/nar/gkq537

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics (Oxford, England)* 25 (9), 1189–1191. doi:10.1093/bioinformatics/btp033

Xue, Y., Zhou, F., Zhu, M., Ahmed, K., Chen, G., and Yao, X. (2005). GPS: a Comprehensive Www Server for Phosphorylation Sites Prediction. *Nucleic Acids Res.* 33, W184–W187. doi:10.1093/nar/gki393

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

Zhang, M., Huang, C., Wang, Z., Lv, H., and Li, X. (2020). In Silico analysis of Non-synonymous Single Nucleotide Polymorphisms (nsSNPs) in the Human GJA3 Gene Associated with Congenital Cataract. *BMC Mol. cell Biol.* 21 (1), 12. doi:10.1186/s12860-020-00252-7

Zhang, Y. (2008). I-TASSER Server for Protein 3D Structure Prediction. *BMC bioinformatics* 9, 40. doi:10.1186/1471-2105-9-40

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 | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fmolb.2022.878017

Metal Ion Binding in Wild-Type and Mutated Frataxin: A Stability Study

S. Morante[1,2]*, S. Botticelli[1,2], R. Chiaraluce[3], V. Consalvi[3], G. La Penna[2,4], L. Novak[3], A. Pasquo[5], M. Petrosino[6], O. Proux[7], G. Rossi[1,2,8], G. Salina[2] and F. Stellato[1,2]

[1]Dipartimento di Fisica, Universitá di Roma Tor Vergata, Rome, Italy, [2]INFN, Sezione di Roma Tor Vergata, Rome, Italy, [3]Dipartimento di Scienze Biochimiche "A. Rossi Fanelli", Sapienza Universitá di Roma, Rome, Italy, [4]CNR—Istituto di Chimica dei Composti Organometallici, Firenze, Italy, [5]ENEA CR Frascati, Diagnostics and Metrology Laboratory FSN-TECFIS-DIM, Frascati, Italy, [6]Chair of Pharmacology, Section of Medicine, University of Fribourg, Fribourg, Switzerland, [7]Observatoire des Sciences de L'Univers de Grenoble, UAR 832 CNRS, Université Grenoble Alpes, Grenoble, France, [8]Museo Storico della Fisica e Centro Studi e Ricerche E. Fermi, Roma, Italy

OPEN ACCESS

Edited by:
Bikash Ranjan Sahoo,
Howard Hughes Medical Institute (HHMI), United States

Reviewed by:
Jagadeeswara Rao Bommi,
Case Western Reserve University, United States
Ernesto Andrès Roman,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

****Correspondence:***
S. Morante
silvia.morante@roma2.infn.it

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *17 February 2022*
Accepted: *12 April 2022*
Published: *31 May 2022*

Citation:
Morante S, Botticelli S, Chiaraluce R, Consalvi V, La Penna G, Novak L, Pasquo A, Petrosino M, Proux O, Rossi G, Salina G and Stellato F (2022) Metal Ion Binding in Wild-Type and Mutated Frataxin: A Stability Study. Front. Mol. Biosci. 9:878017. doi: 10.3389/fmolb.2022.878017

This work studies the stability of wild-type frataxin and some of its variants found in cancer tissues upon Co^{2+} binding. Although the physiologically involved metal ion in the frataxin enzymatic activity is Fe^{2+}, as it is customarily done, Co^{2+} is most often used in experiments because Fe^{2+} is extremely unstable owing to the fast oxidation reaction $Fe^{2+} \rightarrow Fe^{3+}$. Protein stability is monitored following the conformational changes induced by Co^{2+} binding as measured by circular dichroism, fluorescence spectroscopy, and melting temperature measurements. The stability ranking among the wild-type frataxin and its variants obtained in this way is confirmed by a detailed comparative analysis of the XAS spectra of the metal-protein complex at the Co K-edge. In particular, a fit to the EXAFS region of the spectrum allows positively identifying the frataxin acidic ridge as the most likely location of the metal-binding sites. Furthermore, we can explain the surprising feature emerging from a detailed analysis of the XANES region of the spectrum, showing that the longer 81-210 frataxin fragment has a smaller propensity for Co^{2+} binding than the shorter 90-210 one. This fact is explained by the peculiar role of the N-terminal disordered tail in modulating the protein ability to interact with the metal.

Keywords: frataxin, metal ions (Co^{2+}), frataxin mutants, XAS (XAFS, XANES), termal stability

1 INTRODUCTION

The study of the role played by metal ions in biology dates back several decades. Among the metals normally present in living systems, transition metals, such as Fe, Co, Cu, and Zn, are the most noteworthy. They usually exist as electropositive chemical elements that can replace the proton of an acid and form complexes with hydroxyl anions. Metals are highly reactive and are involved in many important biological processes (Shi and Chance, 2008; Bertini et al., 2007; Crichton, 2019). An unbalanced concentration of metals can result in severe threats to living systems.

Many mechanisms make metal ions easily available when needed but harmless when stored. In many instances, whether metals are useful or harmful crucially depends on their local concentration. Thus, storing, metabolism, and trafficking of metals must be accurately regulated. The failure of the tuning mechanism of metal delivery is possibly also involved in the development of severe neurodegenerative pathologies such as Alzheimer's disease, Transmissible Spongiform Encephalopathies, and Parkinson's disease (De Santis et al., 2015).

One of the proteins required for the fine and correct tuning of cellular iron homeostasis is frataxin (FXN). FXN is highly conserved in eukaryotes and prokaryotes and plays a role in different cellular pathways. FXN specific function(s) is (are) still a matter of debate, while there is a general agreement that binding of iron to FXN is required for the cellular control of iron homeostasis. Nevertheless, the FXN role in helping cells regulate iron chemistry and availability has not been fully elucidated. Among many others, functions proposed for FXN are that it may act as an iron chaperone during heme and iron-sulfur (Fe-S) cluster assembly, serve as an iron-storage reservoir during iron overloading, and/or serve as a factor in lowering the concentration of reactive oxygen species (ROS), thus controlling the cellular oxidative stress (Bencze et al., 2006).

It is unlikely that only a single protein can control, in different cells, so many different pathways. In contrast, it is possible that different functions are not specific for FXN. Nevertheless, a breakdown in FXN production may lead to a complete failure in controlling the cellular iron availability. In particular, the deletion of the FXN gene has been seen to be followed by overloading of mitochondrial iron deposits and a general default of the Fe-S cluster assembly (Babcock et al., 1997; Foury and Cazzalini, 1997; Foury, 1999; Rötig et al., 1991). On these premises, FXN has been proposed as a candidate for the control of the mitochondrial iron efflux (Radisky et al., 1999).

Reduced expression levels of FXN in humans are linked to the development, progression, and severity of Friedreich's ataxia (FA) (Pastore and Puccio, 2013), a neurodegenerative disease that is supposed to be linked to disproportion in iron concentration (Gentry et al., 2013).

The 210 amino acids long human FXN precursor is imported into the mitochondrion where it undergoes a two-step proteolytic maturation: first into a 19 kDa (42–210) intermediate and then into the final 14 kDa (81–210) form[1], in both healthy individuals and FA patients (Pastore and Puccio, 2013).

FXN is known to possess non-structured regions (of which there is no 3D information from X-ray crystallography) that are supposed to play quite an important (direct/indirect) role in iron binding.

To date, FXN is not considered a cancer driver gene (Davoli et al., 2013; Martínez-Jiménez et al., 2020). However, it has been reported by Pastore and Puccio (2013) that, in cancer tissues, several FXN variants, whose stability and functional activity are reduced with respect to the wild-type (wt), are present. The importance of mutations in the onset and progression of neoplastic diseases is not yet clarified. However, mitochondrial functions and the complex process of tumorigenesis are connected at multiple levels. Impaired FXN functions may lead to defective mitochondria, which leads to a reduced assembly of Fe-S clusters, thus enhancing carcinogenesis (Rychtarcikova et al., 2017).

The Fe-FXN coordination mode is not fully elucidated. However, of the three putative ion binding sites, one is found to bring into play the His_{86} residue located in the disordered N-terminal region, while the other two involve aspartate and glutamate residues in a region of the protein called the "acidic ridge" (Gentry et al., 2013).

In this complex biological framework, the main goal of our research is to help understand the key structural aspects of the interaction of FXN and the three somatic missense variants of D104G, Y123S, and S161I (described in **Section 2.2**) with Fe^{2+} mimicked by its akin Co^{2+} (Maiti et al., 2017; Söderberg et al., 2011; Söderberg et al., 2013).

Following the same strategy already successfully used in by Gentry et al. (2013), Noguera et al. (2015), and Pastore et al. (2007), in the present study, we have replaced Fe^{2+} with the chemically very similar Co^{2+} ion. In fact, although the FXN natural ligand ion is Fe^{2+}, the latter is known to be very unstable, spontaneously switching to the oxidized Fe^{3+} species in aerobic conditions. Instead, Co^{2+} is not sensitive to aerobic conditions and has been proved to be a reliable surrogate to probe Fe^{2+} sites in proteins (Maiti et al., 2017) and particularly in frataxin (Söderberg et al., 2011; Söderberg et al., 2013). The use of Co^{2+} in place of Fe^{2+} to study the structural properties is also supported by the observation that Co^{2+} has an ionic radius very similar to that of Fe^{2+} (Shannon, 1976).

The experimental technique of election for the study of protein-metal interactions is X-ray absorption spectroscopy (XAS) because, even in the case of very diluted metal-protein complexes, it allows providing accurate structural information about the environment of the metal-binding site within a radius of up to about 5 Å (Strange et al., 1993; Ortega et al., 2012)[2].

In particular, XAS has shown its valuable power in elucidating the mechanisms by which metal ions interacting with intrinsically disordered proteins can push the latter to either the correct folding or a potentially dangerous misfolding, according to specific physicochemical circumstances (Morante et al., 2004; Morante and Rossi, 2014).

This work compares structural XAS information with data obtained from other complementary experimental techniques: circular dichroism, fluorescence spectroscopy and melting temperature measurements. A set of fairly good consistent results for the structure of the protein-metal complex and the stability ranking among wt-FXN and the three variants of D104G, Y123S, and S161I upon metal binding is obtained in this way.

[1] For the readers' convenience, we report here the amino acid sequence of the mature, 81-210, FXN. SGTLGHPGSL DETTYERLAE ETLDSLAEFF EDLADKPYTF EDYDVSFGSG VLTVKLGGDL GTYVINKQTP NKQIWLSSPS SGPKRYDWTG. KNWVYSHDGV SLHELLAAEL TKALKTKLDL SSLAYSGKDA. The α1-helix, from Glu92 to Leu113, is in red. The α2-helix, from Leu182 to Ala193, is in green. The β-sheet core, from Asp124 to Trp168 is in blue.

[2] We may remark here that Co^{2+} has another property that is extremely useful for our XAS measurements: unlike Fe^{2+}, it is a very resilient species under radiation damage.

2 MATERIALS AND METHODS

2.1 Circular Dichroism and Fluorescence Spectroscopy

Circular dichroism (CD) and fluorescence spectroscopy (FS) techniques (Ősz et al., 2002) are used to monitor conformational changes induced by Co^{2+} binding in wt-FXN and in its variants D104G, Y123S, and S161I.

CD measurements were carried out with a JASCO-815 spectropolarimeter (Jasco, Easton, MD, United States), and the results are expressed as the mean residue ellipticity (Θ), assuming a mean residue molecular mass of 110 per amino acid residue.

Far-UV (190–250 nm) CD spectra were monitored at 20°C at an FXN concentration ranging from 100 to 140 μg/ml, using a 0.1 cm path length quartz cuvette. Near-UV (350–420 nm) CD spectra were recorded in a 1.0 cm path length quartz cuvette at a protein concentration ranging between 1.00 and 1.30 mg/ml. In both the far- and near-UV CD spectra measurements, the protein was dissolved in 20 mM Hepes, pH 8.0, 20 mM Na_2SO_4, and 20% glycerol.

Wild-type FXN and variants (with protein concentrations ranging from 100 to 140 μg/ml) were heated from 20 to 95°C and then cooled from 95 to 20°C in a 0.1 cm quartz cuvette with a heating rate of 1.0°/min controlled by a Jasco programmable Peltier element. The dichroic activity at 222 nm and the PMTV were continuously monitored in parallel every 0.5°C (Benjwal et al., 2006). All thermal scans were corrected for the solvent contribution at different temperatures. Melting temperature (Tm) values were calculated by taking the first derivative of the ellipticity at 222 nm with respect to temperature. All denaturation experiments were performed in triplicate.

Intrinsic fluorescence emission spectra were monitored at 50 μg/ml under the identical excitation and emission conditions for all the samples. Right angle light scattering was measured at 20°C with both excitation and emission wavelengths set at 480 nm. Measurements were performed using an LS50B spectrofluorometer (PerkinElmer) with a 1.0 cm path length quartz cuvette. Spectra were recorded in a range from 300 to 450 nm, with a 1 nm sampling interval, and the excitation wavelength was set at 295 nm.

2.2 X-Ray Absorption Spectroscopy

XAS measurements at the Co K-edge were performed at the BM30 beamline of the European Synchrotron Radiation Facility (ESRF; Grenoble, France) Proux et al. (2005).

The beam energy was selected using a Si(220) double-crystal monochromator with a resolution of 0.5 eV. The beam spot on the sample was approximately 240, ×, 150 μm^2 (H × V, FWHM). Spectra were recorded in fluorescence mode using a 13-element solid-state Ge detector. To avoid photo-degradation and undesired spectra evolution during data taking, all the samples, held at 77 K since their preparation, were cooled down and kept at 13 K in a helium cryostat during the whole XAS measurements. For the same reason, in the process of data acquisition, the hitting X-ray beam was systematically moved to a different position on the sample after each scan.

TABLE 1 | In the first column, we report the names by which the analyzed samples are denoted in the study.

Name	81–210	90–210	(Co) eq
wt_90_none		*	None
wt_90_0.8		*	0.8
wt_90_1.6		*	1.6
wt_81_none	*		None
wt_81_0.3	*		0.3
wt_81_0.8	*		0.8
D104G_none	*		None
D104G_08	*		0.8
Y123S_none	*		None
Y123S_08	*		0.8
S161I_none	*		None
S161I_08	*		0.8

Asterisks in the following two columns identify the portion of the amino acid sequence of the FXN protein as either 81-210 (first column) or 90-210 (second column). In the last column, the Co^{2+} concentration in protein equivalent (0.3, 0.8, and 1.6) is reported, while "none" stays for the absence of metal. Of course, this last set of samples, lacking the metal absorber, was not submitted to any XAS measurement.

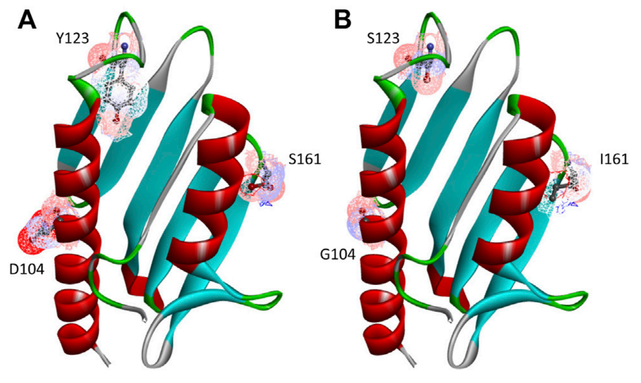


FIGURE 1 | Human FXN folded structure. The three point-mutations, D104G, Y123S, and S161I, which are the object of this investigation, are highlighted. **(A)** Wild-type, **(B)** variants.

As usual, XAS spectra are analyzed by separating the low photons energy XANES (X-ray absorption near edge spectroscopy) region from the high photons energy EXAFS (extended X-ray absorption fine structure) one. The XANES spectra are normalized using the standard software Athena (Ravel and Newville, 2005). EXAFS data are extracted with the help of cubic splines interpolation as implemented in the AUTOBKG algorithm (Newville et al., 1993) of Athena. EXAFS data analysis was performed using the EXCURV98 code (Binsted et al., 1998).

As specified in **Table 1**, besides the wt-FXN in two different lengths, we considered three somatic missense variants (Petrosino et al., 2019): D104G, Y123S, and S161I. These mutations, whose location in the folded protein is highlighted in **Figure 1**, have different effects on the Fe-S cluster formation, as shown by absorbance measurements by Petrosino et al. (2019). The reasons for focusing on these somatic missense variants are summarized as follows:

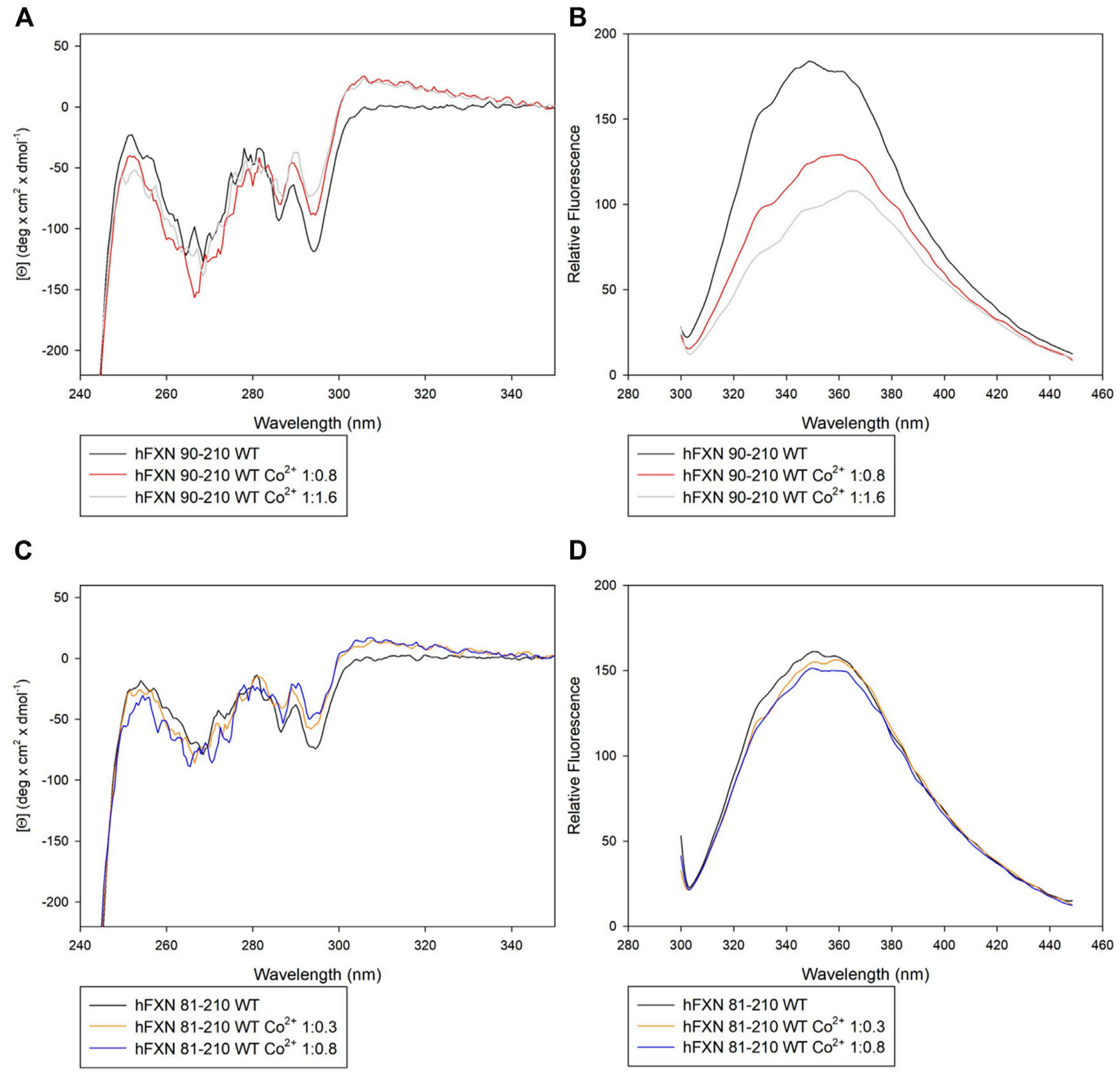


FIGURE 2 | Spectroscopic characterization of FXN isoforms with and without Co^{2+}. **(A)** Near-UV CD spectra of FXN 90-210 in the absence of Co^{2+} (black) and in the presence of 0.8 (red line) and 1.6 (grey line) Co^{2+} equiv. **(B)** Intrinsic fluorescence emission spectra of FXN 90-210 in the absence of Co^{2+} and in the presence of 0.8 and 1.6 Co^{2+} equiv (code color is as in panel **A**). **(C)** Near-UV CD spectra of FXN 81-210 in the absence of Co^{2+} (dark purple) and in the presence of 0.3 (orange) and 0.8 (blue) Co^{2+} equiv. **(D)** Intrinsic fluorescence emission spectra of FXN 81-210 in the absence of Co^{2+} and in the presence of 0.3 and 0.8 Co^{2+} equiv (code color is as in panel **C**). Fluorescence spectra were monitored at 50 μg/ml under identical excitation and emission conditions.

- D104G: this variant is present in liver carcinoma cells. As shown in **Figure 1**, residue D104 is located in the α_1 helix region[3] of the protein. In the case of this variant, Fe-S clusters formation is totally inhibited, although the thermodynamic stability and the secondary and tertiary structure arrangements are similar to those of the wt-FXN (Petrosino et al., 2019).
- Y123S: this variant is present in the cells of the digestive tract carcinoma. Residue Y123 is located in the coil region between the β-sheet region and the α_1 helix (**Figure 1**). In this case, activity measurements show that Fe-S clusters formation is only partially inhibited. This FXN variant shows a Tm of about 14°C lower than that of the wt-FXN, suggesting an enhanced unfolding propensity and a significant decrease in thermal stability (Petrosino et al., 2019).
- S161I: this variant has been found in the cells of the endometrium carcinoma. Residue S161 is located in the coil region between the β-sheet region and the α_2 helix. Again the thermal stability is decreased, as indicated by a Tm lower, by about 11°C, compared to that of the wt-FXN. Interestingly, with this missense mutation, the Fe-S cluster formation is totally inhibited (Petrosino et al., 2019).

For the purpose of XAS measurements, wt-FXN and variants have been expressed as N-terminal His-tagged proteins using a

[3] This nomenclature is defined in the footnote in the introductory section.

TABLE 2 | Melting temperature of wt-FXN and D104G, Y123S, and S161I variants, in the absence of Co^{2+} and in the presence of 0.8 and 1.6 Co^{2+} equiv. Sample names are as shown in **Table 1**.

Sample name	Tm (°C)
wt_81_none	65
wt_81_0.3	65
wt_81_0.8	65
wt_90_none	66
wt_90_0.8	60
wt_90_1.6	60
D104G_90_none	68
D104G_90_0.8	63
Y123S_90_none	53
Y123S_90_0.8	Tm_1 = 43; Tm_2 = 59
S161I_90_none	56
S161I_90_0.8	Tm_1 = 46; Tm_2 = 54

pET28a vector in *E. coli* Rosetta cells transformed with the selected plasmid, grown in LB medium, and purified as described by Petrosino et al. (2019).

The samples listed in **Table 1** were prepared for the XAS measurements according to the following protocol. For the wt protein and its variants, stocks of 45 µL of 0.9 mM protein solutions were obtained by dissolving the appropriate amount of protein in 20 mM Hepes, pH 8.0, containing 20 mM Na_2SO_4 and 20% glycerol. A Co^{2+} 16 mM stock solution was prepared by dissolving $CoSO_4$ in a 20 mM Hepes, pH 8.0, containing 20 mM Na_2SO_4 with 20% glycerol. In order to get the desired Co^{2+} final concentrations (see **Table 1**), to each 45 µL sample holder, the appropriate amount of a Co^{2+} concentrated stock solution was finally added. The solutions in each sample holder were mixed by pipetting and then immediately frozen in liquid N_2, shipped overnight from Rome to Grenoble in a dry shipping Dewar cooled with liquid $N_{2,}$ and immediately transferred to the ESRF beamline upon arrival.

3 RESULTS

3.1 Circular Dichroism, Fluorescence Spectroscopy, and Thermal Shift Analysis

Both near-UV CD and FS suggest that Co^{2+} interacts preferentially with the short FXN isoform 90-210 rather than with the long isoform 81-210.

In particular, panel **A** of **Figure 2** shows that, by adding Co^{2+} to FXN 90-210, important changes occur in the regions of 252–256, 265–270, and 288–295 nm. At the same time, fluorescence intensity reduction and redshift are observed (see panel **B** of **Figure 2**), indicating consistent conformational changes that may be related to the Co^{2+} presence with protein-metal complex formation. Right angle light scattering data indicate that aggregation never occurred in any measured samples.

Furthermore, the longer, FXN 81-210, isoform shows changes in the near-UV CD spectrum in the regions 252–256 and 288–295 nm (see panel **C** of **Figure 2**). However, neither fluorescence intensity modification nor redshift is visible (panel **D**), thus hinting at a lack of protein-metal complex formation.

These indications are confirmed by thermal stability analysis performed on the protein in the presence of Co^{2+} at the concentrations specified in **Table 2** (see **Figure 3**). Thermal stability measurements are considered a powerful and reliable way to study metal-binding interactions in proteins (Layton and Hellinga, 2011).

In order to measure thermal stability, samples with the wt-FXN and samples with the D104G, Y123S, and S161I variants at concentrations ranging from 100 to 140 µg/ml, in the absence of Co^{2+} and the presence of different concentrations of Co^{2+}, were heated from 20°C to 95°C in a quartz cuvette with a heating rate of 1°C/min, controlled by a Jasco programmable Peltier element as described by Petrosino et al. (2019). We stress that thermal transitions appeared reversible in all the conditions we considered, with no hysteresis during the cooling phase (see **Supplementary Material**).

Looking at the behavior of the molar ellipticity changes at 222 nm between 20°C and 95°C, we observe a decrease in the Tm values of the wt sample by 6° upon adding 0.8 equiv or 1.6 Co^{2+} equiv to FXN 90-210 (**Table 2**; **Figure 3A**). Instead, Tm does not change when the metal ion is added to FXN 81-210 (see **Table 2**; **Figure 3B**). The three variants, D104G, Y123S, and S161I, show conformational changes and a decrease in Tm in the presence of Co^{2+} (see **Figures 4**, **5**).

In the case of the Y123S and S161I variants, two melting temperatures are reported in **Table 2**. They correspond to the two local maxima displayed by the dotted curves in the insets of **Figures 5B,C**. The presence of two slopes in the Θ_{222} curve is interpreted as an indication of the existence of a two-step thermal denaturation process.

The most significant structural modification induced by Co^{2+} was observed in the case of the Y123S variant, which involves the tyrosine residue in position 123, in a region usually considered a putative FXN iron-binding site (Nair et al., 2004; Gentry et al., 2013).

In the case of the Y123S variant, the presence of Co^{2+} induces a dramatic modification in both the near-UV CD and the fluorescence spectra (**Figures 4B,E**). Indeed, in the presence of 0.8 Co^{2+} equiv, a significant loss of cooperativity occurs in the thermal unfolding process, as shown by the far-UV CD data of $[\Theta]_{222}$ as a function of the temperature reported in **Figures 5B,C** and **Table 2**.

3.2 X-Ray Absorption Spectroscopy

Figure 1 displays the structured region of wt-FXN and the short non-structured region at the C-terminal of the amino acid sequence. A second disordered region is located between residues 81 and 90[4]. For the latter, we propose a peculiar role in modulating metal binding (see **Section 4**) based on the analysis of XAS data (see below) and the melting temperature results reported in **Table 2**.

[4] The amino acid sequence of the mature, 81-210, wt-FXN is given in the footnote in the introductory section.

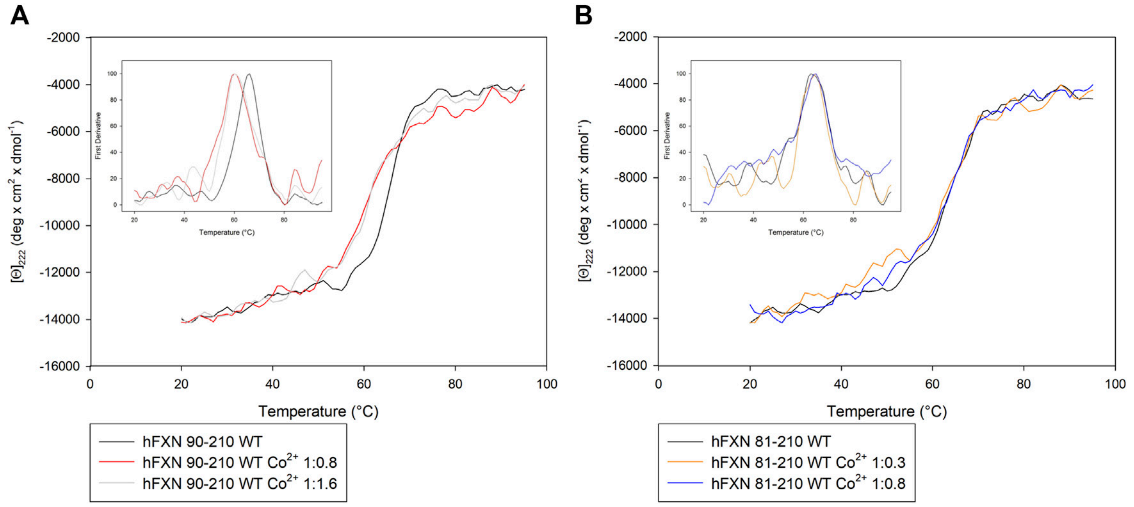


FIGURE 3 | Thermal unfolding transition of FXN isoforms with and without Co^{2+}. **(A)** Thermal unfolding curves of FXN 90-210 in the absence of Co^{2+} (black) and in the presence of 0.8 (red) or 1.6 (grey) Co^{2+} equiv. **(B)** Thermal unfolding curves of FXN 81-210 in the absence of Co^{2+} (dark purple) and in the presence of 0.3 (orange) or 0.8 (blue) Co^{2+} equiv. The curves are obtained by monitoring the molar ellipticity at 222 nm $[\Theta]_{222}$, in the range between 20 and 95°C. Insets show the first derivative of the experimental curve.

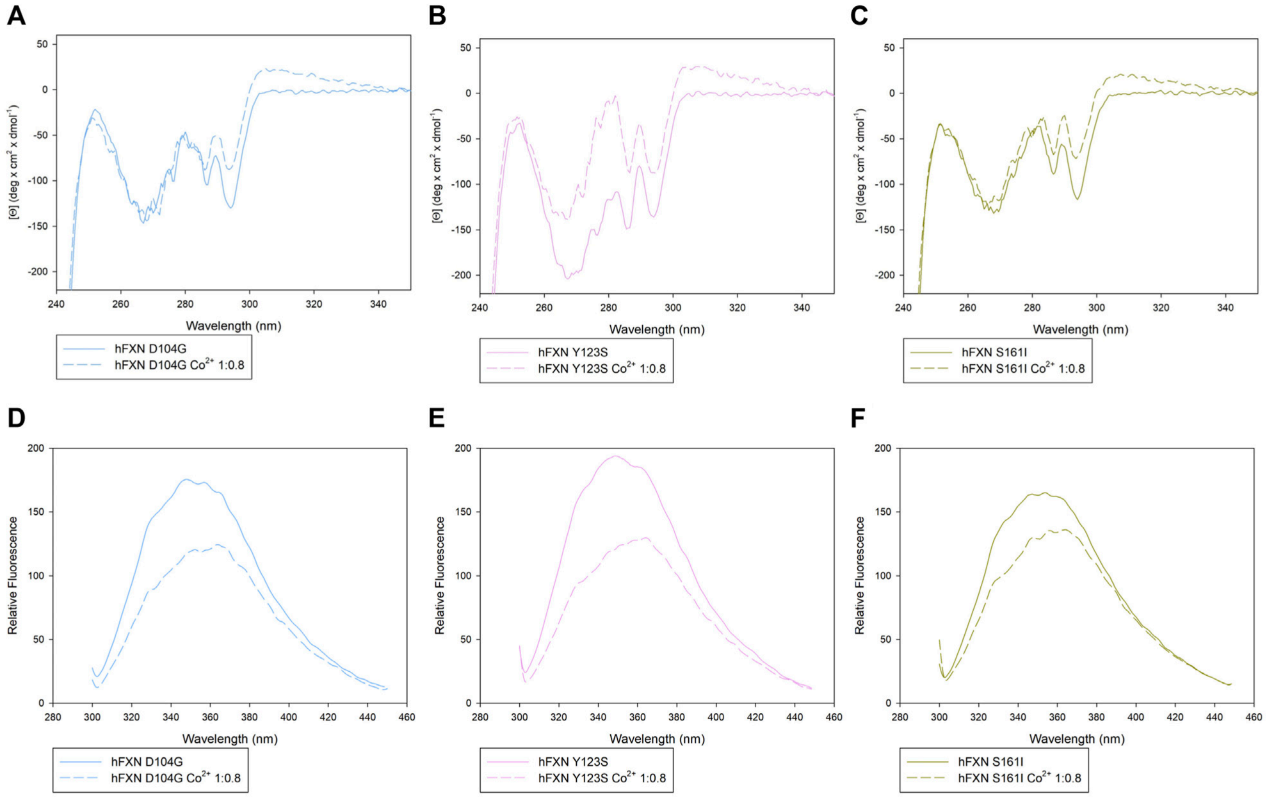


FIGURE 4 | Spectroscopic characterization of the D104G, Y123S, and S161I variants with and without Co^{2+}. Upper panels: near-UV CD spectra of D104G (light blue, **A**), Y123S (light pink, **B**), and S161I (dark yellow, **C**) in the absence of Co^{2+} (continuous line) and in the presence of 0.8 Co^{2+} equiv (dotted line). Lower panels: intrinsic fluorescence emission spectra of D104G (light blue, **D**), Y123S (light pink, **E**), and S161I (dark yellow, **F**) in the absence of Co^{2+} (continuous line) and in the presence of 0.8 Co^{2+} equiv (dotted line). Fluorescence spectra were monitored at 50 µg/ml under identical excitation and emission conditions.

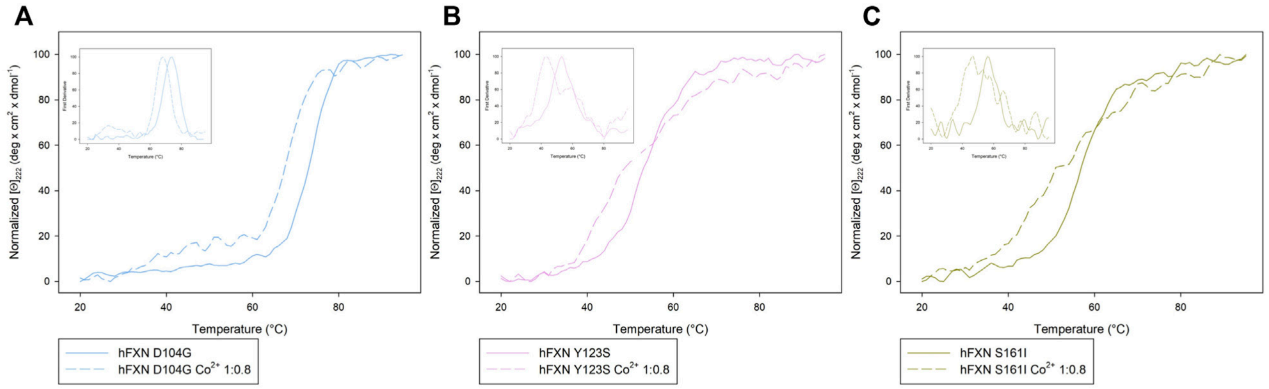


FIGURE 5 | Comparison of the Near-UV CD data of $[\Theta]_{222}$ as a function of the temperature of the D104G (light blue), Y123S (light pink), and S161I (dark yellow) variants **(A–C)** in the absence of Co^{2+} (continuous curves) and in the presence of 0.8 Co^{2+} equiv (dotted lines). The insets show the first derivative of the experimental curve with respect to temperature.

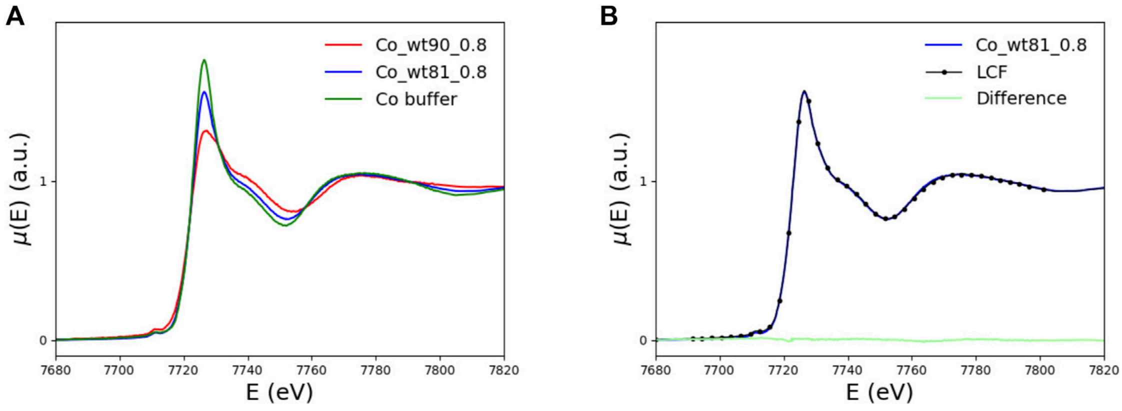


FIGURE 6 | (A) XANES spectra of the FXN samples including (wt_81_0.8, blue curve) and not including (wt_90_0.8, red curve) the N-terminal 81-90 disordered region, compared with the buffer spectrum (green curve). **(B)** XANES spectrum of the long, wt_81_0.8, fragment (blue curve) compared to the spectrum obtained as a linear combination of 43% of wt_90_0.8 and 57% of buffer spectra (black dotted). In light green, we plot the difference spectrum.

The first step toward a correct interpretation of the physicochemical implications of the FXN-metal interaction is to clarify the structure of the accessible metal-binding sites located along the protein amino acid sequence. For this task, XAS is the technique of election as it allows accurately "taking a picture" of the metal ion atomic environment within a region of up to about 5 Å around the metal absorber.

As customarily done, we shall separately discuss the features of the low energy (XANES) and the high energy (EXAFS) region of the XAS spectrum.

The XANES region is dominated by multiple scattering processes, where the photoelectron undergoes more than one scattering event against the atomic scatterers located within a sphere of a radius of about 5 Å from the absorber. Although extracting accurate geometrical information from the XANES region of the spectrum is computationally extremely difficult, fingerprints of specific geometries can still be identified (see **Section 3.2.1**). In the EXAFS region, where the extracted photoelectron owns a quite high kinetic energy, single scattering events prevail, thus making, in principle, the extraction of geometrical information comparatively much easier (see **Section 3.2.2**).

3.2.1 XANES

In this section, we discuss the information one can get from comparing the XANES spectrum of the wt-FXN with the three variants we are focusing our attention on.

In the left panel of **Figure 6**, the XANES spectra of the wt_81_0.8 and wt_90_0.8 samples are compared with the buffer spectrum. The two peptides differ because of the presence (wt_90_0.8) or absence (wt_81_0.8) of the disordered 81-90 tail.

We recall that the His_{86} amino acid, which is supposed to be a binding site for Co^{2+} (as well as Fe^{2+}) (Gentry et al., 2013), is missing in the shorter fragment wt_90_0.8.

By comparing, in the left panel of **Figure 6**, the XANES data for the wt_81_0.8 sample with those of the wt_90_0.8 sample and the buffer, one sees that the spectrum of the longer fragment,

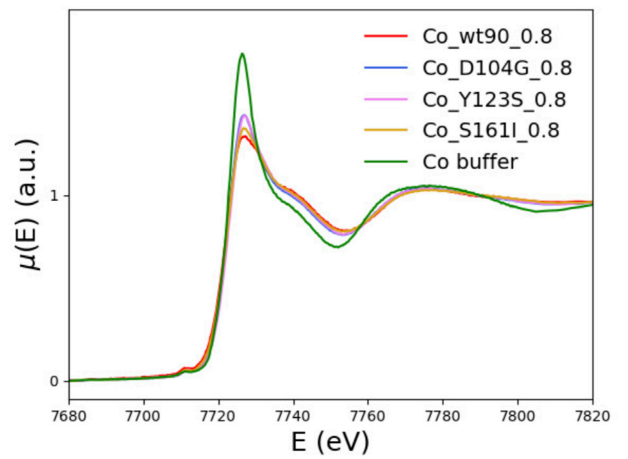


FIGURE 7 | The XANES spectra of the D104G (blue), Y123S (violet), and S161I (gold) variants are plotted together with the wt-FXN (red) and the buffer (green) spectra.

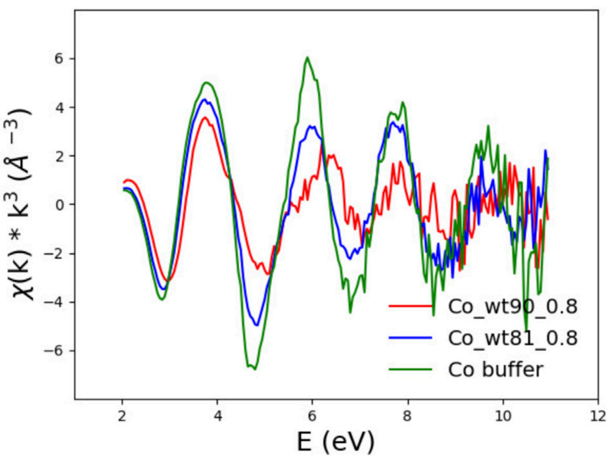


FIGURE 8 | EXAFS spectra of FXN wt_90_0.8 (red) and wt_81_0.8 (blue) in the presence of 0.8 Co^{2+} equiv compared to the buffer spectrum (green).

wt_81_0.8 sample, looks much more similar to the buffer than the spectrum of the wt_90_0.8 sample. This feature suggests that, in the case of wt_81_0.8, a non-negligible amount of Co^{2+} remains free in the solution.

In order to make this observation more quantitative, we perform a minimization of the difference between the wt_81_0.8 XANES spectrum and the spectrum obtained as a linear combination, point by point, of the wt_90_0.8 and buffer spectra. By fitting the mixing coefficient, we find that the wt_81_0.8 XANES spectrum is well reproduced by a linear combination of 43% of the wt_90_0.8 spectrum plus 57% of the buffer.

We conclude that, in the case of the wt_81_0.8 sample, more than half (57%) of the available Co^{2+} does not bind to FXN but remains free in solution. This observation is the reason why we did not explore higher metal concentrations. In fact, going at 1.6 eq (or higher) would only increase the uninteresting contribution of Co^{2+} in the solution.

This finding seems somewhat surprising because it says that, in the presence of an extra putative metal-binding site, a smaller amount of Co^{2+} is bound to the protein.

We will show in **Section 4** that the reason for this, at first sight, strange feature has to do with the peculiar arrangement of the disordered 81-90 tail belonging to the longer wt_81_0.8 fragment. The point is that this tail can "obscure" some of the protein metal-binding sites otherwise available for binding in the case of the shorter wt_90_0.8 peptide (see **Figure 13**). This interesting interpretation of the puzzling XANES data in **Figure 6** is nicely confirmed by the results of melting temperature measurements displayed in **Table 2**, which show that, in the presence of the metal, the longer 81-210 fragment is more stable than the shorter one.

In **Figure 7**, the XANES spectra of the D104G (blue), Y123S (pink), and S161I (gold) variants are compared with those of the wt_90_0.8 sample (red) and the buffer (green). We see that the buffer spectrum appears to be significantly different from the spectra of all the variants, thus concluding that all the variants can bind Co^{2+}.

The tendency of XANES spectra of D104G and Y123S variants toward that of the buffer indicates a slightly larger contribution to spectra of dissociated forms in these variants than in the wild-type sequence.

Qualitatively, we see that the most significant difference with the wt-FXN spectrum occurs in the case of variants, in which the destabilizing effect due to metal binding (measured by the difference of the melting temperature with and without metal, see **Table 2**) is large. While destabilization upon metal binding is somewhat surprising in the case of the D104G mutation, as the latter just occurs well inside the alleged metal-binding acidic ridge, the other two mutations, Y123S and S161I, are located far away from it, and the reasons at the basis of the metal-induced destabilization are more difficult to identify. They would require a much more refined investigation, which is outside the scope of the present work.

3.2.2 EXAFS

The first simple observation is that EXAFS spectra show the same ranking of similarity observed for the XANES spectra. In particular, one notices that although the EXAFS spectra of the wt_90_210 and wt_81_210 samples are both definitely different from that of the buffer (see **Figure 8**), they are also quite different between themselves, with the spectrum of the wt_81_210 sample visibly more similar to that of the buffer than to that of the wt_90_210 sample.

Structural information about the atomic arrangement around the absorbing metal can be extracted from a fit to the EXAFS region of the spectrum, provided that some "reasonable" assumption about the atomic structure around the Co^{2+} absorber is available. Unfortunately, not much is known about the precise location of the possible Fe/Co-FXN binding sites (Gentry et al., 2013). In order to

TABLE 3 | Pairs of residues to which MIB assigns the highest score as Fe^{2+}, Fe^{3+}, and Co^{2+} metal-binding sites in wt-FXN 90-210.

FXN 90-210		
Fe^{2+}	Fe^{3+}	Co^{2+}
Asp_{104}-Glu_{108}	Glu_{108}-Glu_{111}	Glu_{100}-Asp_{104}
Glu_{96}-Glu_{100}	Asp_{104}-Glu_{108}	Glu_{108}-Asp_{112}

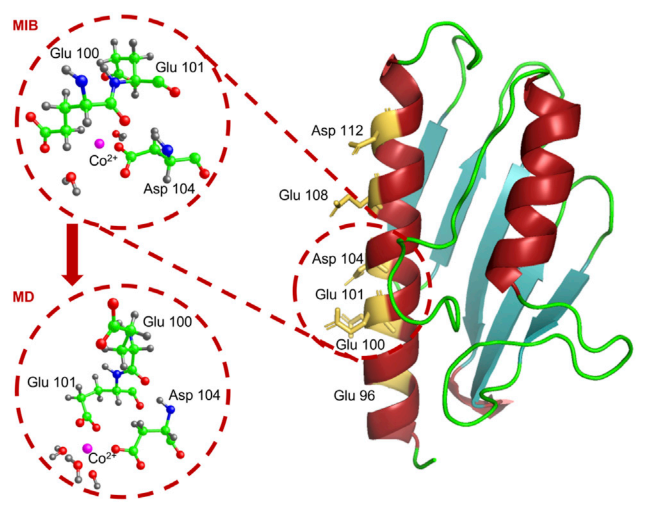


FIGURE 9 | We show the atomic structure of the Glu_{100}-Asp_{104} site around the Co^{2+} ion, as proposed by MIB (upper circle) and at the end of the MD simulation (lower circle). Co^{2+} is in pink, carbon in green, oxygen in red, nitrogen in blue, and hydrogen in grey.

overcome this lack of information, we decided to take advantage of the metal ion-binding (MIB) site prediction and docking server (Lin et al., 2016) that, given the protein 3D structure, provides the most probable binding sites for a specific metal ion.

The amino acid patterns to which MIB has assigned the highest scores for Co^{2+}, Fe^{2+}, and Fe^{3+} binding in FXN 90-210 are reported in **Table 3**. As expected, the so-called acid ridge is signaled as the most probable region for Co^{2+}, Fe^{2+}, and Fe^{3+} binding.

We limited the MIB search to FXN 90-210 because the FXN 81-210 longer fragment differs from the former for the presence of the disordered 81-90 N-terminal tail, which is not supposed to be able to host a metal. Despite this fact, as discussed in **Section 4**, this unstructured protein region will be argued to play an important role in the FXN 81-210 ability to bind the metal.

Building on the structural information provided by the MIB search, we are in a position to construct a physically sensible Co^{2+} binding model configuration from which one can start the theoretical calculation of the EXAFS spectrum. Fitting model parameters against the measured spectrum will allow for accurately determining the atomic structure of the metal-binding site.

TABLE 4 | Co^{2+}-ligand distances and Fermi energy shift at the end of the EXAFS fit. Note that the carboxylic group of Glu_{101} contributes to the octahedral coordination with two oxygen atoms.

Coordinated atoms	$r \pm \Delta r$ (Å)
4 O (Glu_{101} + Asp_{104} + 2 water molecules)	2.02 ± 0.02
1 O (Glu_{101})	2.21 ± 0.01
1 O (water)	1.85 ± 0.02
E_F = (3 ± 1) eV	R-factor = 35%

As a starting configuration, we took the one with the highest score among those listed in the last column of **Table 3**. It corresponds to a situation in which Co^{2+} is coordinated to the residues Glu_{100} and Asp_{104}.

The structure provided by MIB (see the upper circled inset in the left part of **Figure 9**) is relaxed and equilibrated at 300 K with classical MD simulations.

MD simulations were performed using the code NAMD (Phillips et al., 2020) with the Co^{2+} ion forced in its preferred octahedral geometry by employing the *dummy atoms* method (Pang, 2001; Furlan et al., 2010; La Penna et al., 2013; Duarte et al., 2014; La Penna et al., 2015). Indeed, the octahedral Co^{2+} geometry, which is at the basis of our MD simulations, is confirmed by the pre-edge feature visible in the XANES spectra of **Figure 7**, around 7,710 eV (signaling a 1 s → 3 d transition), typical of an octahedral (or pseudo octahedral) site geometry (Bresson et al., 2006).

The *dummy atoms* method amounts to perform MD simulations where the positive density of the divalent cation is distributed in a blocked octahedral geometry to prevent the negatively charged groups from collapsing over the charged point-like particle of the opposite sign (La Penna and Chelli, 2018).

The metal-protein complex was placed in a 64 × 62 × 74 $Å^3$ box and solvated by adding 9259 TIP3P water molecules to have a layer of water of at least 15 Å around the protein. Na^+ and Cl^- ions were inserted in the box to get an overall neutral system with the same ion concentration as in the *in vitro* experiments. At the end of a 4 ns MD simulation, Co^{2+} is found to be stably located at a binding distance from the Glu_{101} and Asp_{104} residues (see lower circled inset in the left panel of **Figure 9**), thus slightly displaced from its initial position.

A sphere of a 5 Å radius, centered on the Co^{2+} ion, is excised from the simulation box and used as the starting configuration for the fit of the EXAFS data of the *wt*_90_08 sample.

The fit to the EXAFS data is performed by employing the program EXCURV98 Binsted et al. (1998). The code implements the so-called constrained refinement strategy, which consists of treating molecules as rigid bodies. In this way, the number of fitting parameters can be significantly reduced (Binsted et al., 1992). In the present case, we considered the amino acids bound to the Co^{2+} ion rigid bodies, namely, Glu_{101} and Asp_{104}. Starting from the configuration depicted in the lower circled inset in **Figure 9**, we

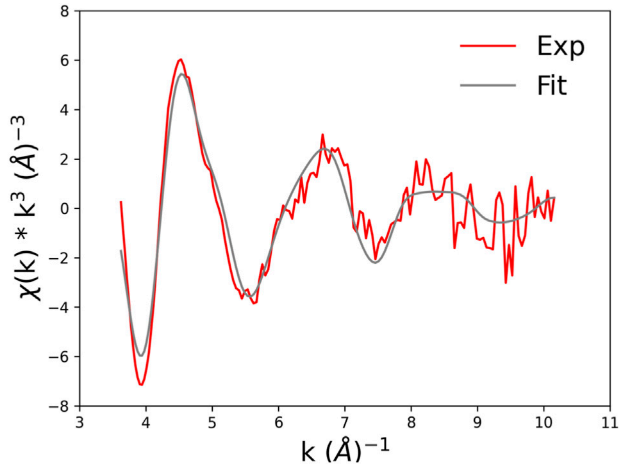


FIGURE 10 | Experimental EXAFS spectrum of wt_90_0.8 (red curve) and best-fit model (grey curve).

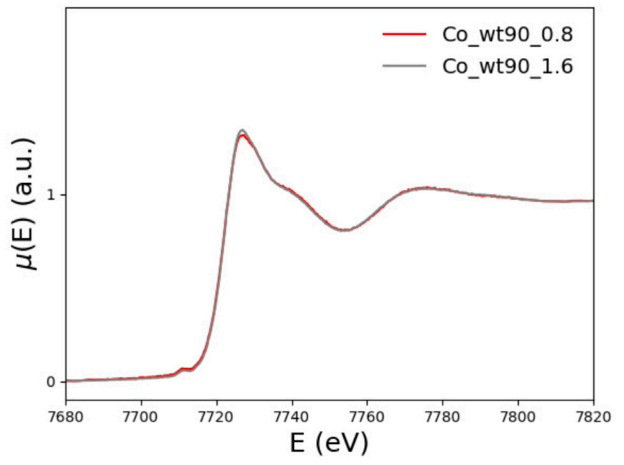


FIGURE 12 | XANES spectra of wt_90_0.8 (red) and wt_90_1.6 (grey) in the presence of 0.8 and 1.6 Co^{2+} equiv, respectively.

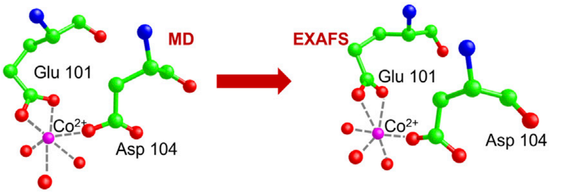


FIGURE 11 | Comparison of the Co. binding site configuration at the end of the MD simulation and after the EXAFS fitting step. The color code is as shown in **Figure 9**. Hydrogen atoms are not shown as they are standardly considered not to contribute to the XAS signal. Broken lines represent the distances of Co^{2+} from its six oxygen ligands.

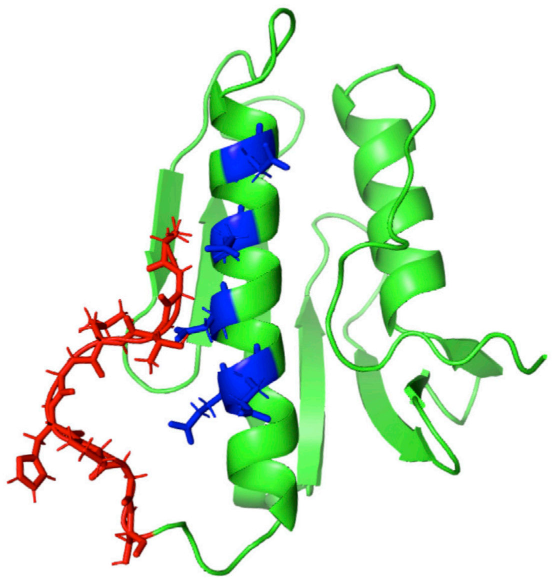


FIGURE 13 | We display the configuration of the 81-90 disordered region in red as it results after an elaborated MD relaxation step that was started from an *all-trans* geometry. The potential metal-binding site locations are highlighted in blue.

refined the Fermi energy shift and the distances of the six oxygen atoms located in the first coordination sphere. These are the two oxygen atoms belonging to the carboxylic group of Glu_{101}, the oxygen belonging to Asp_{104}, and three oxygen atoms belonging to the three nearby water molecules (see **Table 4**).

The best fit is displayed in **Figure 10**, where we show the experimental EXAFS data (in red) superimposed on the fitted curve (in grey). **Table 4** gives the O-Co^{2+} best fit distances. **Figure 11** compares the metal-binding site configuration at the end of the MD simulation and after the EXAFS fitting step. We remark that the MD simulation was able to give a pretty good picture of the atomic environment around the metal, as one finds that both before and after the EXAFS fit the six metal ligands (two oxygen atoms from Glu_{101}, one atom from Asp_{104}, and three atoms from water molecules) all fall in a shell between 1.8 and 2.2 Å.

We end this section with an observation concerning the second (putative) metal-binding site (Glu_{108}-Asp_{112}) identified by the MIB prediction and docking server (see **Table 3**). We note that an analysis completely analog to the one performed in the case of the Glu_{100}-Asp_{104} site shows that the atomic environments of the two sites around the metal are practically identical (data not shown). The main reason is that, in the two sites, the same two ligands, namely, Glu and Asp, *via* the carboxylate group of the acid chain, are involved in metal binding. Thus, we expect that although the two ligands at the end of the MD simulation can be differently positioned in space, the two binding sites will equally contribute to the EXAFS experimental signal.

4 THE ROLE OF THE DISORDERED 81-90 N-TERMINAL REGION

It is very instructive to look at the XANES spectral regions (see **Figure 6**) in Section 3.2.1 in conjunction with the pattern of the measured melting temperature of wt-FXN and D104G, Y123S, and S161I variants (see **Table 2**), as quite an interesting feature emerges.

To start the discussion, we recall that the XANES spectrum of the 81-210 fragment is very well reproduced by the sum, point by point, of 43% of the wt_90_0.8 spectrum plus 57% of the buffer spectrum (see **Section 3.2.1**). This suggests that more than half of Co^{2+} remains in the solution; it is not bound to the protein.

We also note that, on the one hand, as **Figure 12** shows, the XANES spectra of the 90-210 wt-FXN samples with 0.8 and 1.6 Co^{2+} equiv are well superimposable, proving that, even when the Co^{2+} is added to the solution at two times the protein concentration, still there is no Co^{2+} in the solution. This hints at the conclusion that at least two (more or less equivalent) Co^{2+} binding sites are available in the wt-FXN protein. On the other hand, it is necessary to assume that about half of the amount of Co^{2+} in the sample has remained free in solution (i.e., not bound to the protein) to get a good fitting of the XANES and EXAFS spectra of the 81-210 FXN.

A suggestive explanation for this rather puzzling behavior could be that the 81-90 tail present in 81-210 FXN somehow prevents Co^{2+} from binding to one of the two available coordination sites present in the acidic ridge.

4.1 Experimental Evidence

The disordered 81-90 N-terminal tail in one of its relaxed configurations is geometrically long enough to be able to somehow "obscure" part of the acidic ridge in the α_1 helix. In **Figure 13**, we show one of these typical configurations. They are obtained at the end of a long and careful MD simulation started from an initial *all-trans* arrangement.

This elongated configuration can result from the "sealing" of the Co^{2+} ion acting between the acidic ridge and the carboxylic end, which helps stabilize the structure. Such a model would explain why, in the case of the long 81-210 wt-FXN, half of Co^{2+} is found free in solution when the latter is added to the solution at a concentration of twice that of the protein.

The scenario we propose is confirmed by the pattern of the Tm values shown in **Table 2**. In fact, while the presence of Co^{2+} does not affect the Tm of the long, 81-210, wt-FXN fragment, it significantly destabilizes the short one, 90-210[5]. This observation reinforces our hypothesis, according to which, in the presence of the 81-90 disordered tail, metal ion binding sites in the acidic ridge are not fully available. As a result, the destabilizing effect due to metal binding is reduced or not occurring anymore.

5 CONCLUSION

Based on information extracted from some experimental techniques (CD, FS, XAS) and numerical simulations, we confirm the ability of the wt-FXN and its variants to stably bind metal ions (in our investigation Co^{2+}), strengthening in this way the conjecture according to which FXN is involved in Fe^{2+} storage and transport.

In particular, XAS data consistently confirm the ranking of thermal stability established by measuring the melting temperatures when wt-FXN is compared to D104G, Y123S, and S161I variants.

Starting from the crystallographic structure of the putative Co^{2+} binding sites identified by the MIB site prediction and docking server (Lin et al., 2016), we have performed long (up to 4 ns) MD simulations of FXN in the presence of Co^{2+} to have a reliable atomic geometry around the metal that could be used as an initial configuration in the analysis of our newly collected XAS data. A fit to the EXAFS region of the spectrum allows us to positively identify the FXN acidic ridge as the location of the most likely metal-binding sites.

Furthermore, we can explain the surprising feature emerging from a detailed analysis of the XANES region of the spectrum, according to which the longer 81-210 FXN fragment has a smaller propensity for Co^{2+} binding than the shorter 90-210 one, despite the presence in the former of the His_{86} residue, which is supposed to be a specific Co^{2+} binding site (Gentry et al., 2013).

Indeed, the difference between the XAS spectrum of the long 81-210 FXN fragment compared to that of the shorter 90-210 one that we understand as due to a large fraction (almost a half) of Co^{2+} remaining in solution in the case of 81-210 FXN, is explained as due to a peculiar behavior of the "disordered" N-terminal region of 81-90 FXN. Our conjecture, which to our knowledge was never considered in the literature, is that the disordered N-terminal tail in the 81-210 FXN sample gets locked in an extended configuration by binding a Co^{2+} ion, which acts as a bridge between the 81-90 tail and the α_1-helix. In this configuration, after a first metal ion is bound, the other metal-binding site allegedly present in the acid ridge is "obscured" by the extended N-terminal tail and becomes unattainable by a second metal ion. The Co^{2+} "sealing" ability we are invoking here is in agreement with the large body of experimental work that assigns a role to metal ions as structural stabilizers in disordered peptides and protein folding processes (see La Penna and Morante (2021) for a recent review on this issue).

In the future, we intend to develop our investigations along two lines. The first is a simulation study of the relative stability of wt-FXN compared to that of a more extended set of FXN single-point variants (Botticelli et al., 2022) than the ones considered here. The computation is carried out using the MD methods where altruistic metadynamics is employed to generate maximally unbiased sets of protein configurations (Barducci et al., 2006; Laio and Gervasio, 2008). The

[5] In this analysis, when two Tm are reported (see **Table 2**), the smallest is considered as already at this temperature unfolding is starting to take place.

constrained maximal entropy principle is then exploited to appropriately reweigh the collected configurations in thermal average computations (La Penna et al., 2004).

The second line of investigation is of a more experimental nature and aims to extend to the FXN variants considered by Botticelli et al. (2022), the CD, FS, and XAS measurements described in the present study.

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

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

FUNDING

This work was partially supported by the BIOPHYS initiative (INFN, Italy). We also acknowledge partial financial support from the MIUR (Italy) PRIN-201744NR8S project.

ACKNOWLEDGMENTS

Numerical calculations have been made possible owing to a CINECA-INFN agreement, providing us access to resources on MARCONI at CINECA. The ESRF program committee is acknowledged for beamtime allocation (proposal LS-2968).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.878017/full#supplementary-material

REFERENCES

Babcock, M., de Silva, D., Oaks, R., Davis-Kaplan, S., Jiralerspong, S., Montermini, L., et al. (1997). Regulation of Mitochondrial Iron Accumulation by Yfh1p, a Putative Homolog of Frataxin. *Science* 276, 1709–1712. doi:10.1126/science.276.5319.1709

Barducci, A., Chelli, R., Procacci, P., Schettino, V., Gervasio, F. L., and Parrinello, M. (2006). Metadynamics Simulation of Prion Protein: β-Structure Stability and the Early Stages of Misfolding. *J. Am. Chem. Soc.* 128, 2705–2710. doi:10.1021/ja057076l

Bencze, K. Z., Kondapalli, K. C., Cook, J. D., McMahon, S., Millán-Pacheco, C., Pastor, N., et al. (2006). The Structure and Function of Frataxin. *Crit. Rev. Biochem. Mol. Biol.* 41, 269–291. doi:10.1080/10409230600846058

Benjwal, S., Verma, S., Röhm, K.-H., and Gursky, O. (2006). Monitoring Protein Aggregation during thermal Unfolding in Circular Dichroism Experiments. *Protein Sci.* 15, 635–639. doi:10.1110/ps.051917406

Binsted, N., Gurman, S., and Campbell, J. (1998). *Daresbury Laboratory EXCURV98 Program*. Warrington: CLRC Daresbury Laboratory.

Binsted, N., Strange, R. W., and Hasnain, S. S. (1992). Constrained and Restrained Refinement in EXAFS Data Analysis with Curved Wave Theory. *Biochemistry* 31, 12117–12125. doi:10.1021/bi00163a021

Botticelli, S., La Penna, G., Nobili, G., Rossi, G., Stellato, F., and Morante, S. (2022). Modelling Protein Plasticity: The Example of Frataxin and its Variants. *Molecules* 27, 1955. doi:10.3390/molecules27061955

Bresson, C., Esnouf, S., Lamouroux, C., Solari, P. L., and Den Auwer, C. (2006). Xas Investigation of Biorelevant Cobalt Complexes in Aqueous media. *New J. Chem.* 30, 416–424. doi:10.1039/b514454j

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., et al. (2013). Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* 155, 948–962. doi:10.1016/j.cell.2013.10.011

De Santis, E., Minicozzi, V., Proux, O., Rossi, G., Silva, K. I., Lawless, M. J., et al. (2015). Cu(II)-Zn(II) Cross-Modulation in Amyloid-Beta Peptide Binding: An X-ray Absorption Spectroscopy Study. *J. Phys. Chem. B* 119, 15813–15820. doi:10.1021/acs.jpcb.5b10264

Duarte, F., Bauer, P., Barrozo, A., Amrein, B. A., Purg, M., Åqvist, J., et al. (2014). Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *J. Phys. Chem. B* 118, 4351–4362. doi:10.1021/jp501737x

Foury, F., and Cazzalini, O. (1997). Deletion of the Yeast Homologue of the Human Gene Associated with Friedreich's Ataxia Elicits Iron Accumulation in Mitochondria. *FEBS Lett.* 411, 373–377. doi:10.1016/s0014-5793(97)00734-5

Foury, F. (1999). Low Iron Concentration and Aconitase Deficiency in a Yeast Frataxin Homologue Deficient Strain. *FEBS Lett.* 456, 281–284. doi:10.1016/s0014-5793(99)00961-8

Furlan, S., Hureau, C., Faller, P., and La Penna, G. (2010). Modeling the Cu+ Binding in the 1–16 Region of the Amyloid-β Peptide Involved in Alzheimer's Disease. *J. Phys. Chem. B* 114, 15119–15133. doi:10.1021/jp102928h

Gentry, L. E., Thacker, M. A., Doughty, R., Timkovich, R., and Busenlehner, L. S. (2013). His86 from the N-Terminus of Frataxin Coordinates Iron and Is Required for Fe-S Cluster Synthesis. *Biochemistry* 52, 6085–6096. doi:10.1021/bi400443n

I. Bertini, H. B. Gray, E. I. Stiefel, and J. S. Valentine (Editors) (2007). *Biological Inorganic Chemistry: Structure and Reactivity* (Sausalito, California: University Science Books).

La Penna, G., and Chelli, R. (2018). Structural Insights into the Osteopontin-Aptamer Complex by Molecular Dynamics Simulations. *Front. Chem.* 6, 2. doi:10.3389/fchem.2018.00002

La Penna, G., Hureau, C., Andreussi, O., and Faller, P. (2013). Identifying, by First-Principles Simulations, Cu[Amyloid-β] Species Making Fenton-Type Reactions in Alzheimer's Disease. *J. Phys. Chem. B* 117, 16455–16467. doi:10.1021/jp410046w

La Penna, G., Minicozzi, V., Morante, S., Rossi, G. C., and Stellato, F. (2015). A First-Principle Calculation of the XANES Spectrum of Cu2+ in Water. *J. Chem. Phys.* 143, 124508. doi:10.1063/1.4931808

La Penna, G., and Morante, S. (2021). Aggregates Sealed by Ions. *Methods Mol. Biol.* 2340, 309–341. doi:10.1007/978-1-0716-1546-1_14

La Penna, G., Morante, S., Perico, A., and Rossi, G. C. (2004). Designing Generalized Statistical Ensembles for Numerical Simulations of Biopolymers. *J. Chem. Phys.* 121, 10725–10741. doi:10.1063/1.1795694

Laio, A., and Gervasio, F. L. (2008). Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Rep. Prog. Phys.* 71. doi:10.1088/0034-4885/71/12/126601

Layton, C. J., and Hellinga, H. W. (2011). Quantitation of Protein-Protein Interactions by thermal Stability Shift Analysis. *Protein Sci.* 20, 1439–1450. doi:10.1002/pro.674

Lin, Y.-F., Cheng, C.-W., Shih, C.-S., Hwang, J.-K., Yu, C.-S., and Lu, C.-H. (2016). MIB: Metal Ion-Binding Site Prediction and Docking Server. *J. Chem. Inf. Model.* 56, 2287–2291. doi:10.1021/acs.jcim.6b00407

Maiti, B. K., Almeida, R. M., Moura, I., and Moura, J. J. G. (2017). Rubredoxins Derivatives: Simple sulphur-rich Coordination Metal Sites and its Relevance for Biology and Chemistry. *Coord. Chem. Rev.* 352, 379–397. doi:10.1016/j.ccr.2017.10.001

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A Compendium of Mutational Cancer Driver Genes. *Nat. Rev. Cancer* 20, 555–572. doi:10.1038/s41568-020-0290-x

Morante, S., González-Iglesias, R., Potrich, C., Meneghini, C., Meyer-Klaucke, W., Menestrina, G., et al. (2004). Inter- and Intra-octarepeat Cu(II) Site Geometries in the Prion Protein. *J. Biol. Chem.* 279, 11753–11759. doi:10.1074/jbc.m312860200

Morante, S., and Rossi, G. (2014). Metals in Alzheimer's Disease: A Combined Experimental and Numerical Approach. *Advanc. Alz. Res.* 2, 100–147.

Nair, M., Adinolfi, S., Pastore, C., Kelly, G., Temussi, P., and Pastore, A. (2004). Solution Structure of the Bacterial Frataxin Ortholog, CyaY. *Structure* 12, 2037–2048. doi:10.1016/j.str.2004.08.012

Newville, M., Līviņš, P., Yacoby, Y., Rehr, J. J., and Stern, E. A. (1993). Near-edge X-ray-absorption fine Structure of Pb: A Comparison of Theory and experiment. *Phys. Rev. B* 47, 14126–14131. doi:10.1103/physrevb.47.14126

Noguera, M. E., Roman, E. A., Rigal, J. B., Cousido-Siah, A., Mitschler, A., Podjarny, A., et al. (2015). Structural Characterization of Metal Binding to a Cold-Adapted Frataxin. *J. Biol. Inorg. Chem.* 20, 653–664. doi:10.1007/s00775-015-1251-9

Ortega, R., Carmona, A., Llorens, I., and Solari, P. L. (2012). X-ray Absorption Spectroscopy of Biological Samples. A Tutorial. *J. Anal. Spectrom.* 27, 2054–2065. doi:10.1039/c2ja30224a

Ősz, K., Bóka, B., Várnagy, K., Sóvágó, I., Kurtán, T., and Antus, S. (2002). The Application of Circular Dichroism Spectroscopy for the Determination of Metal Ion Speciation and Coordination Modes of Peptide Complexes. *Polyhedron* 21, 2149–2159.

Pang, Y.-P. (2001). Successful Molecular Dynamics Simulation of Two Zinc Complexes Bridged by a Hydroxide in Phosphotriesterase Using the Cationic Dummy Atom Method. *Proteins* 45, 183–189. doi:10.1002/prot.1138

Pastore, A., and Puccio, H. (2013). Frataxin: a Protein in Search for a Function. *J. Neurochem.* 126, 43–52. doi:10.1111/jnc.12220

Pastore, C., Franzese, M., Sica, F., Temussi, P., and Pastore, A. (2007). Understanding the Binding Properties of an Unusual Metal-Binding Protein – a Study of Bacterial Frataxin. *FEBS J.* 274, 4199–4210. doi:10.1111/j.1742-4658.2007.05946.x

Petrosino, M., Pasquo, A., Novak, L., Toto, A., Gianni, S., Mantuano, E., et al. (2019). Characterization of Human Frataxin Missense Variants in Cancer Tissues. *Hum. Mutat.* 40, 1400–1413. doi:10.1002/humu.23789

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on Cpu and Gpu Architectures with NAMD. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475

Proux, O., Biquard, X., Lahera, E., Menthonnex, J. J., Prat, A., Ulrich, O., et al. (2005). FAME A New Beamline for XRay Absorption Investigations of VeryDiluted Systems of Environmental, Material and Biological Interests. *Physica Scripta*, 970. doi:10.1238/physica.topical.115a00970

Radisky, D. C., Babcock, M. C., and Kaplan, J. (1999). The Yeast Frataxin Homologue Mediates Mitochondrial Iron Efflux. *J. Biol. Chem.* 274, 4497–4499. doi:10.1074/jbc.274.8.4497

Ravel, B., and Newville, M. (2005). ATHENA,ARTEMIS,HEPHAESTUS: Data Analysis for X-ray Absorption Spectroscopy usingIFEFFIT. *J. Synchrotron Radiat.* 12, 537–541. doi:10.1107/s0909049505012719

R. Crichton (Editor) (2019). *Biological Inorganic Chemistry: A New Introduction to Molecular Structure and Function*. third edition edn (Academic Press). doi:10.1016/B978-0-12-811741-5.00030-8

Rötig, A., de Lonlay, P., Chretien, D., Foury, F., Koenig, M., Sidi, D., et al. (1991). Aconitase and Mitochondrial Iron-sulphur Protein Deficiency in Friedreich Ataxia. *Nat. Genet.* 17, 215–217. doi:10.1038/ng1097-215

Rychtarcikova, Z., Lettlova, S., Tomkova, V., Korenkova, V., Langerova, L., Simonova, E., et al. (2017). Tumor-initiating Cells of Breast and Prostate Origin Show Alterations in the Expression of Genes Related to Iron Metabolism. *Oncotarget* 8, 6376–6398. doi:10.18632/oncotarget.14093

Shannon, R. D. (1976). Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides. *Acta Cryst. Sect A.* 32, 751–767. doi:10.1107/s0567739476001551

Shi, W., and Chance, M. R. (2008). Metallomics and Metalloproteomics. *Cell. Mol. Life Sci.* 65, 3040–3048. doi:10.1007/s00018-008-8189-9

Söderberg, C. A. G., Rajan, S., Shkumatov, A. V., Gakh, O., Schaefer, S., Ahlgren, E.-C., et al. (2013). The Molecular Basis of Iron-Induced Oligomerization of Frataxin and the Role of the Ferroxidation Reaction in Oligomerization. *J. Biol. Chem.* 288, 8156–8167. doi:10.1074/jbc.M112.442285

Söderberg, C. A. G., Shkumatov, A. V., Rajan, S., Gakh, O., Svergun, D. I., Isaya, G., et al. (2011). Oligomerization Propensity and Flexibility of Yeast Frataxin Studied by X-ray Crystallography and Small-Angle X-ray Scattering. *J. Mol. Biol.* 414, 783–797. doi:10.1016/j.jmb.2011.10.034

Strange, R., Morante, S., Stefanini, S., Chiancone, E., and Desideri, A. (1993). Nucleation of the Iron Core Occurs at the Three-fold Channels of Horse Spleen Apoferritin: an EXAFS Study on the Native and Chemically-Modified Protein. *Biochim. Biophys. Acta (Bba) - Protein Struct. Mol. Enzymol.* 1164, 331–334. doi:10.1016/0167-4838(93)90267-u

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.

METHODS
published: 16 June 2022
doi: 10.3389/fmolb.2022.867572

PON-All: Amino Acid Substitution Tolerance Predictor for All Organisms

Yang Yang[1,2], *Aibin Shao*[1] *and Mauno Vihinen*[3]*

[1]*School of Computer Science and Technology, Soochow University, Suzhou, China,* [2]*Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China,* [3]*Department of Experimental Medical Science, Lund University, Lund, Sweden*

Genetic variations are investigated in human and many other organisms for many purposes (e.g., to aid in clinical diagnosis). Interpretation of the identified variations can be challenging. Although some dedicated prediction methods have been developed and some tools for human variants can also be used for other organisms, the performance and species range have been limited. We developed a novel variant pathogenicity/tolerance predictor for amino acid substitutions in any organism. The method, PON-All, is a machine learning tool trained on human, animal, and plant variants. Two versions are provided, one with Gene Ontology (GO) annotations and another without these details. GO annotations are not available or are partial for many organisms of interest. The methods provide predictions for three classes: pathogenic, benign, and variants of unknown significance. On the blind test, when using GO annotations, accuracy was 0.913 and MCC 0.827. When GO features were not used, accuracy was 0.856 and MCC 0.712. The performance is the best for human and plant variants and somewhat lower for animal variants because the number of known disease-causing variants in animals is rather small. The method was compared to several other tools and was found to have superior performance. PON-All is freely available at http://structure.bmc.lu.se/PON-All and http://8.133.174.28:8999/.

Keywords: variation interpretation, mutation, animal variants, plant variants, amino acid substitution, prediction, pathogenicity, machine learning

OPEN ACCESS

Edited by:
Tiziana Sanavia,
University of Torino, Italy

Reviewed by:
Emidio Capriotti,
University of Bologna, Italy
Rita Casadio,
University of Bologna, Italy

Correspondence:
Mauno Vihinen
mauno.vihinen@med.lu.se

Specialty section:
This article was submitted to Protein Biochemistry for Basic and Applied Sciences, a section of the journal Frontiers in Molecular Biosciences

Received: *01 February 2022*
Accepted: *02 May 2022*
Published: *16 June 2022*

Citation:
Yang Y, Shao A and Vihinen M (2022) PON-All: Amino Acid Substitution Tolerance Predictor for All Organisms. Front. Mol. Biosci. 9:867572. doi: 10.3389/fmolb.2022.867572

INTRODUCTION

Genome and exome sequencing are frequently used techniques in biology and clinical settings. Efficient resequencing has moved the bottleneck from obtaining sequence and variation information to variation interpretation. Many tools have been released for variant pathogenicity, also called variant tolerance and prediction (Adzhubei et al., 2010; Choi et al., 2012; Olatubosun et al., 2012; Capriotti et al., 2013; Kircher et al., 2014; Schwarz et al., 2014; Dong et al., 2015; Niroula et al., 2015; Vaser et al., 2016; Rogers et al., 2018). These methods are also used for clinical diagnosis in many countries and laboratories according to American College for Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards et al., 2015) guidelines. These guidelines state that predictions could support the diagnosis if several methods agree. This recommendation is problematic and should be reconsidered as it reduces the number of cases that can be predicted because the method with the poorest performance dictates the outcome (Vihinen, 2020).

Variant interpretation methods have been divided into three main categories: those based on evolutionary information, those utilizing many types of features, including evolutionary details, and meta-predictors that use predictions from other predictors as the starting point (Niroula and

Vihinen, 2016). Methods in the last two categories are typically based on machine learning (ML). Several different ML algorithms have been applied, there is not a single best one among them. The predictor performance depends on the quality of data, used features and their selection, implementation of the predictor, and other factors.

Most pathogenicity/tolerance predictors classify variants into two classes (pathogenic/benign), while some have three or more categories. Additional categories could be useful if the predictions are reliable because diseases are not simple binary states, as indicated by the pathogenicity model (Vihinen, 2017). These tools do not explain the cause and mechanism of diseases due to harmful variants. Many other types of predictors are available for various effects and mechanisms, including RNA splicing, protein stability, solubility, disorder, aggregation, and localization.

In variation interpretation, most of the work has been devoted to explaining human variants; however, there is increasing interest and need for interpretation of variants and their consequences also in other organisms. This knowledge is essential for understanding diseases in non-human organisms, obtaining insight into genetic disease mechanisms, genetic diagnosis in veterinary and botany, and scientific inquiry and comparison, among others. Although several predictors trained on human data are applicable to (or at least capable of) accepting variants from other organisms, they have not been systematically developed and tested for alterations from other organisms. Evolutionary methods could be easily adapted for this purpose. However, evolutionary data alone are of limited significance as they do not allow the development of the most reliable predictors. Variation interpretation is a very complex problem, and many features are needed to achieve high prediction performance.

Some predictors have been developed and trained on plant (Kono et al., 2018; Kovalev et al., 2018) and animal variant data (Plekhanova et al., 2019). In animal and plant experiments, it would be important to know whether the used strains contain harmful variants since they may act as confounding factors in various studies. In veterinary medicine, there is increased interest in variants (e.g., in pet animals) also outside the most common species of cats and dogs. Variation data and even genetic data are scarce for many of these species. Experimental validation of variation effects is laborious and often outside the available resources. Therefore, in many cases, the only means to assess the harmfulness of identified variants is to perform computational predictions. As there are not many special tools and even those available as generic methods have not been systematically tested, there is no way of knowing the reliability of the predictions. Some databases, especially the Online Mendelian Inheritance in Animals (OMIA), are valuable. However, there are currently data only for nine named species (and others). Further, the number of likely disease-causal variants is only 1,381. The best performing human variant effect predictors have been trained on tens of thousands of variants.

We have developed several methods for variation interpretation, mainly based on ML. These include PON-P (Olatubosun et al., 2012) and PON-P2 (Niroula et al., 2015) for human pathogenicity prediction of amino acid substitutions, PON-Tstab (Yang et al., 2018) for variants affecting protein stability, PON-Sol (Yang et al., 2016) and PON-Sol2 (Yang et al., 2021) for solubility affecting variants, PON-Diso (Ali et al., 2014) for protein disorder affecting alterations, and PON-mt-tRNA (Niroula et al., 2016) for variants in mitochondrial tRNA molecules. These tools are highly accurate and among the best in their application areas. Several aspects have to be considered in method development: data collection, feature selection, method training, and systematic performance benchmarking (Niroula and Vihinen, 2016).

We collected a data set of human, animal, and plant variants and trained an ML predictor using a gradient boosting algorithm and exhaustive feature selection. The method is called PON-All as it can predict the consequences of amino acid substitutions in proteins from any organism. Several predictors were developed and extensively tested by reporting a full set of performance measures. PON-All was systematically trained and tested and found to have very high performance in predictions for all three types of organisms. The method is fast and freely available as a web resource.

MATERIALS AND METHODS

Data Sets

Amino acid substitutions in human, animal, and plant sequences were collected from databases and publications. The human variants were obtained from VariBench (Nair et al., 2013; Sarkar et al., 2020), including 13,885 harmful variants originally used to train PON-P2 (Niroula et al., 2015). Additional 6369 verified clinical cases were obtained from ClinVar (Landrum et al., 2014) and 2,058 variants in membrane proteins (Orioli and Vihinen, 2019) from VariBench. Only amino acid substitutions with harmful clinical effects were collected. Duplicate cases were removed.

Human neutral variations with minor allele frequency (MAF) 1%<MAF<25% were from ExAC and obtained from VariBench (http://structure.bmc.lu.se/VariBench/ExAC_AAS_20171214.xlsx). The data set had originally been used to test the sensitivity of several predictors (Niroula and Vihinen, 2019). Because these variants have high MAF in populations, they are considered benign. Benign variations used for training and testing were randomly selected. The numbers of variations used in different stages are indicated in **Table 1**. An additional set of 370 benign variants obtained from ClinVar was used to assess specificity.

There were two sources for variations in animals. Cases with the notation “likely causal variants” were obtained from OMIA (Nicholas, 2003). Additional mammalian deleterious variants were obtained from Plekhanova et al. (2019). The main species included dogs, mice, and cattle. Plant data were taken from the data set used to develop a random forests pathogenicity predictor for plant protein variations (species included *Arabidopsis*, *Oryza sativa*, and *Pisum sativum*) (Kovalev et al., 2018). Altogether, there were 23,138 pathogenic variations and 27,816 neutral variations in 16,026 proteins in the three types of species.

TABLE 1 | Division of cases to data sets for cross-validation and blind testing. The first number is for proteins and the second for variants.

	10-fold cross-validation			Blind test			Total		
	Pathogenic	Neutral	Total	Pathogenic	Neutral	Total	Pathogenic	Neutral	Total
Humans	2,173/17,504	12,141/23,600	13,383/41,104	170/1,980	669/1,967	740/3,926	2,343/19,484	12,810/25,567	14,123/45,030
Animals	117/162	116/144	232/306	109/155	125/169	233/324	226/317	241/313	465/630
Plants	913/2,601	629/1,562	1,150/4,163	228/736	152/374	288/1,110	1,141/3,337	781/1,936	1,438/5,273
Total	3,203/20,267	12,886/25,306	14,765/45,573	507/2,871	946/2,510	1,261/5,360	3,710/23,138	13,832/27,816	16,026/50,933

As some features used in training were protein-specific, it was necessary to partition the cases so that all variants in the same protein were in the same data set (either training or test set) to ensure the universality of the classification and avoid bias. Further, we balanced the numbers of pathogenic and neutral cases.

The blind test data set contained cases used to test PON-P2 (Niroula et al., 2015). Half of the animal variants were randomly distributed to the blind test set. Because there were substantially more plant variants, we randomly selected 20% of the variants for the blind test set. In addition, the division of 10-fold cross-validation (CV) training sets and blind test sets ensured that the variations in any protein were always in either the training set or blind test set. Another principle of data division was that the numbers of harmful and neutral variations in each data set were balanced (1:1). The data sets used for training and testing are available in VariBench (Nair et al., 2013; Sarkar et al., 2020) at http://structure.bmc.lu.se/VariBench/trainingall.php and http://8.133.174.28:8999/.

Features

To train the predictor, we started with 1,085 features: 617 amino acid features, 436 variation type features, 25 neighborhood features, 2 evolutionary conservation features, 1 protein feature, and 1 GO feature.

A total of 617 complete amino acid propensity scales were from AAindex (Kawashima and Kanehisa, 2000). This feature set has been previously used to train PON-P2 (Niroula et al., 2015), PON-PS (Niroula and Vihinen, 2017), PON-Tstab (Yang et al., 2018), and PON-Sol2 (Yang et al., 2021). For each variant, the difference between the score for the original amino and the variant amino acid was calculated.

There were two matrices to obtain variation-type features. A total of 400 features came from the 20*20 matrix, where the two dimensions represented original and variant residues. Another 36 features denote a 6*6 matrix representing the physical and chemical properties of amino acids. The six amino acid categories were hydrophobic (V, I, L, F, M, W, Y, C), negatively charged (D, E), positively charged (R, K, H), conformational (G, P), polar (N, Q, S), and others (A, T) and have been previously described (Shen and Vihinen, 2004).

In order to represent the sequence context of variation sites, 25 neighborhood features were included. A 20-dimensional vector of neighborhood residues counts the occurrences of each amino acid type within a neighborhood in a window of 23 positions, that is, 11 positions before and after the variation site (Lockwood et al., 2011). In addition, we included the frequencies of five groups of amino acids (nonpolar, polar, charged, positively charged, and negatively charged) in the neighborhood window of 23 positions.

For evolutionary conservation, DIAMOND (Buchfink et al., 2015) was used to compare each protein sequence to SwissProt (Shomer, 1997) to find related sequences and calculate the number of hits. DIAMOND was chosen as it is substantially faster than BLAST (Altschul et al., 1997) but with a similar degree of sensitivity. The identified sequences were aligned and then used to calculate SIFT scores for evolutionary conservation of each variant position using SIFT 4G (Vaser et al., 2016).

The protein feature was defined as the length of the protein sequence. Additional features included whether the variation was in the first amino acid in the peptide chain and position within the sequence.

Features derived from Gene Ontology (GO) terms have previously been used for variant classification (Kaminker et al., 2007; Calabrese et al., 2009; Niroula et al., 2015). For the full set of GO terms, we combined results from AmiGO (Carbon et al., 2009) and QuickGO (Munoz-Torres and Carbon, 2017) using the R Bioconductor tool GO.db (https://bioconductor.org/packages/GO.db/). We collected all the ancestors of all GO terms and filtered the GO entries so that each protein contained each GO term once. Two sets of GO terms were created for each category (pathogenic and neutral). The sum of the logarithm ratio of GO frequencies of the pathogenic set and that of the neutral set was calculated as follows:

$$LR = \sum \log \frac{f(P_i)+1}{f(N_i)+1},$$

where LR is the value for the GO annotations and $f(P_i)$ and $f(N_i)$ are the frequencies of the ith GO term in pathogenic and neutral data sets, respectively. To avoid uncertain ratios, we added 1 to all the frequencies. If a protein had not been annotated with GO terms, then $LR = 0$, and this feature was not considered in the prediction. We separately trained predictors with and without GO annotations.

We tested the usefulness of functional annotation features and found that almost all the variation records had functional annotations. Site-specific annotations were determined from UniProtKB/Swiss-Prot. The variations that occurred at such sites were identified. We collected all site terms and filtered them so that each protein contained each site term once. Two sets of site terms were created for the two categories (pathogenic and neutral). The sum of the logarithm ratio of site frequencies of the pathogenic set and that of the neutral set were defined as follows:

$$FS = \sum \log \frac{f\,(P_i)+1}{f\,(N_i)+1},$$

where *FS* is the value for the site annotations and $f(P_i)$ and $f(N_i)$ are the frequencies of the *i*th site term in pathogenic and neutral data sets, respectively. To avoid uncertain ratios, we added 1 to all the frequencies. If a protein had not been annotated with site terms, then *FS = 0,* and this feature was not considered in the prediction.

Algorithms

We trained predictors with three machine learning algorithms: random forests (RF) (Breiman, 2001; Pavey et al., 2017), XGBoost (Chen et al., 2016; Yu et al., 2020), and Light GBM (LGBM) (Wang et al., 2017; Zhang et al., 2019). The default parameters were used in each case. All the algorithms were implemented in Python in the standard learn package (Pedregosa et al., 2011). Random forests is an ensemble algorithm. It applies several decision trees on a subset of the data set and uses the average accuracy of each decision tree to improve the performance and reduce overfitting. The gradient boosting model evaluates the output features based on the combination output result of weak prediction learner models. It minimizes a loss function to optimize the model. Sequential models are constructed using the decision trees until maximum accuracy is achieved.

XGBoost and LightGBM are implementations of gradient boosting. Initial results for LightGBM and XGBoost were similar and better than for random forests. Because of the similar performance, we chose LightGBM which is faster due to Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (Ke et al., 2017).

Reliability Assessment

The probability method was used to identify variations with high confidence. The probability distribution function of self-sampling probability cannot be determined. Therefore, we used Chebyshev's inequality, which is applicable to arbitrary distributions. For random variables *X* with mean μ and standard deviation σ, Chebyshev's inequality guarantees that at least 1 – (1/k^2) values are distributed within *k* standard deviations of the mean values:

$$P\left(\mu - k\sigma < X < \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}.$$

When 1 – (1/k^2) is 0.95, and if the range of μ±*k*σ does not include 0.5, the prediction is marked as credible and classified as pathogenic or neutral; else, the variation is considered as unclassified (UV, unclassified variant, also called VUS, variant of uncertain significance).

Performance Assessment

We used eight measures to evaluate the classification performance (Vihinen, 2012; Vihinen, 2013). The measures included positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy, Matthews correlation coefficient (MCC) and overall performance measure (OPM) (Niroula et al., 2015). The mathematical definitions of these measures are as follows:

$$PPV = \frac{TP}{TP + FP},$$

$$NPV = \frac{TN}{TN + FN},$$

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

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

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

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}},$$

$$nMCC = \frac{1 + MCC}{2},$$

$$OPM = \frac{(PPV + NPV)(Sensitivity + Specificity)(Accuracy + nMCC)}{8}.$$

The area under the curve (AUC) was calculated from the Receiver Operating Characteristics (ROC) curve, where cases are plotted based on sensitivity *versus* 1 − specificity.

TP and TN are the numbers of correctly predicted pathogenic and neutral cases, and FN and FP are the numbers of wrong predictions for pathogenic and neutral cases, respectively.

Coverage measures the ratio of predicted cases among all the instances. *X* indicates the number of cases classified as harmful or neutral, and *Y* is the total number of test variants:

$$\text{Coverage} = \frac{X}{Y}.$$

The reason to measure coverage in this way is that PON-All classifies cases into three categories while the data for training and testing are binary (benign/pathogenic).

Feature Selection

We used the recursive feature elimination (RFE) method (Guyon et al., 2002) to carry out multiple rounds of training. First, the prediction model was trained with all the features, and each feature was assigned a weight. Then, the features with the minimum absolute weight were removed. Recursion was repeated until the preset number of features was achieved. To identify the optimal set of features, we trained methods in addition to the full set of features also with 100, 50, 20, and 10 features.

RESULTS

There is an increasing interest in variation interpretation in several organisms. Many of the current variant tolerance/pathogenicity predictors are either just for humans or have not been systematically benchmarked and/or trained with data from other organisms than human. Therefore, we developed a

TABLE 2 | Comparison of method performance in 10-fold cross-validation when using all the features for training. The numbers are averages.

Measure	RF	XGBoost	LGBM
TP	1,528	1,633.4	1,651.3
TN	2,307.9	2,364.3	2,343.1
FP	222.7	166.3	187.5
FN	498.7	393.3	375.4
PPV	0.87	0.91	0.90
NPV	0.82	0.86	0.86
Sensitivity	0.75	0.81	0.81
Specificity	0.91	0.93	0.93
Accuracy	0.84	0.88	0.88
MCC	0.68	0.75	0.75
OPM	0.59	0.67	0.67
AUC	0.83	0.87	0.87

PON-All tool to predict the consequences of amino acid substitutions in any organism. The method was trained on human, animal, and plant variants with known outcomes, validated on a blind test set, and compared to several other tools. The training and assessment of the performance are described according to published guidelines (Vihinen, 2012; Vihinen, 2013), the method is freely available, and the used data are distributed.

Variations were collected from several sources; see **Table 1**. There is a severe imbalance in the number of variants from different sources. The number of animal variants is clearly smaller than the others. There were 630 variants in 465 proteins in the animal data, while the corresponding numbers were 45,030 variants in 14,123 proteins in human and 5,273 variants in 1438 plant proteins (**Table 1**). The ratio of human, plant, and animal variants was 90:10:1.

The largest numbers of disease-causing animal variations are known in rodents (mice and rats). However, we excluded these variants because they are typical models for human diseases and the corresponding variants are largely included in the human data set. Further, using these cases in the blind test could biased the analysis. The included animal variants originate from animal conditions.

In the case of animal variants, we selected a substantially larger ratio to the test set to facilitate reliable performance assessment. Totally, we had 50,933 variants in 16,026 proteins, thus covering a wide spectrum of different sequences. A total of 23,138 of the variants were pathogenic and 27,816 were benign.

Predictor Training

When training the methods, we followed the principles for systematic ML method training presented earlier (Niroula and Vihinen, 2016). We started by choosing the ML algorithm. Three ML algorithms were tested based on our earlier experience in variation interpretation. We implemented predictors with LGBM, RF, and XGBoost and performed 10-fold cross-validation (CV) **Table 2**. The methods were trained on all the features. Because the two versions of gradient boosting were somewhat better than when using RF, we chose LGBM as it is faster. The OPM was 0.67, accuracy 0.88, and MCC 0.75. Further, the predictors were quite balanced. Due to the use of GOSS and EFB technologies, LightGBM was the fastest to train and test.

Next, we performed feature selection. We collected 1,085 features, including 617 amino acid features, 436 variation type features, 25 neighborhood features, 2 evolutionary conservation features, 3 protein features, 1 Gene Ontology feature, and 1 functional annotation feature. RFE was used to recursively reduce the number of features. To decide the optimal number of features, we tested the performance in 10-fold CV with different numbers of features: all, 100, 50, 20, or 10. We wanted to proceed with the smallest possible number of features as the event space is large. The ratio of human, plant, and animal variations in the 10-fold CV for this purpose was 100:10:1.

Further, the methods were implemented with or without rejection and with or without GO features. Classification with the reject option was found useful in PON-P2 (Niroula et al., 2015) to distinguish the category for UVs and obtain reliable predictions for benign and pathogenic cases. UV variants cannot be classified as pathogenic or benign. This class also implies the heterogeneity of phenotypes in different individuals bearing the same variant and is a normal feature for certain variants.

The results of the performance assessment are in **Supplementary Tables S1, S2**. The performances are clearly better when using the GO feature and when applying the rejection option. The results overall are very similar within the different tests for different numbers of features indicating that the number of features can be substantially reduced without a major impact on the performance. The implementation without rejection and GO feature had the best performance with a predictor trained on 50 features (OPM, 0.479), but differences were minimal for methods trained with different numbers of features, effectively in the third decimal place (**Supplementary Table S2**). Similarly, the differences in the other measures were very small or non-existent.

GO features have been useful in several predictors, such as SNPs&GO (Capriotti et al., 2013) and PON-P2 (Niroula et al., 2015). However, GO annotations are far from complete, and the coverage in non-human organisms can be very low, or the annotations may be completely missing. Therefore, to facilitate as many predictions as possible, we developed methods both with and without GO features.

When using GO features (but without the rejection option) (**Supplementary Table S1**), the overall performance is substantially better than without GO details (**Supplementary Table S2**). The results for 50 features were the best (OPM, 0.673), but those for 20 features were very close (OPM, 0.671). The results for the other measures were also very close irrespective of the number of features, thus indicating that the number of features could be significantly reduced.

Without GO but with rejection, the best OPM was achieved with all features (OPM 0.671). However, differences are marginal in the third decimal. The results with the GO feature (**Supplementary Table S1**) but without rejection are close to those for methods with rejection but without GO annotations (OPM 0.676 without rejection). The performance is further increased when rejection is applied (OPM shifted from 0.812 to 0.832, MCC from 0.865 to 0.880). The coverage of predictions increased substantially when GO features were not used, typically

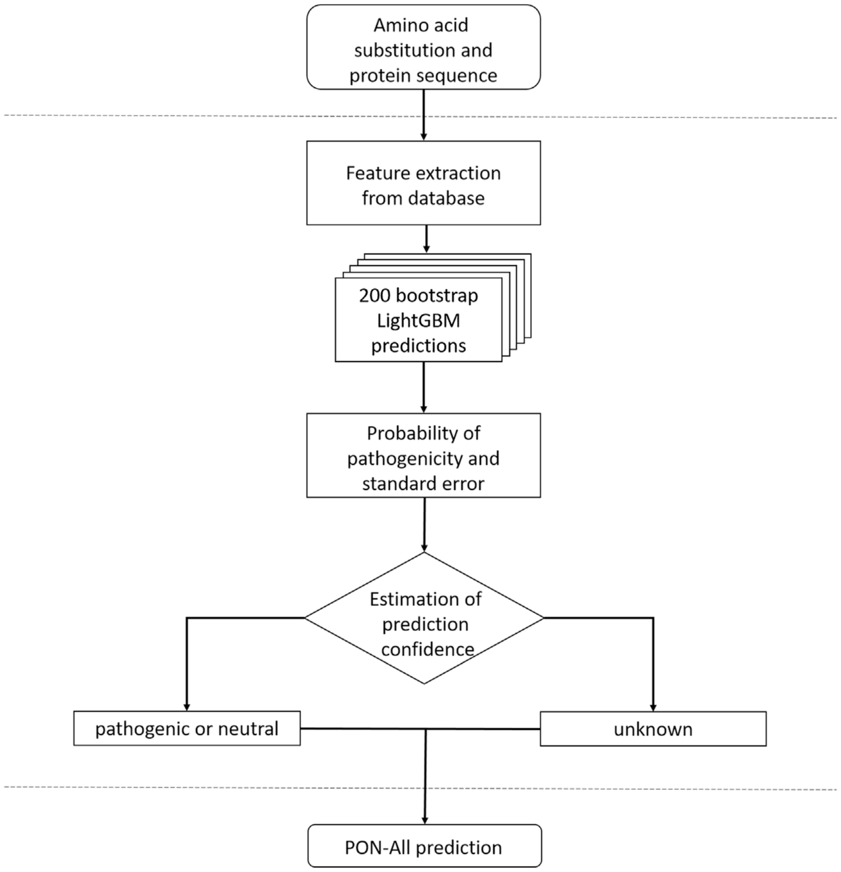

Amino acid substitution and protein sequence
Feature extraction from database
200 bootstrap LightGBM predictions
Probability of pathogenicity and standard error
Estimation of prediction confidence
pathogenic or neutral
unknown
PON-All prediction

FIGURE 1 | Flowchart for PON-All predictor.

by 20%, thus allowing predictions for many more variants. There is thus a balance between the number of cases that can be predicted and the optimal performance.

Based on the results, we chose to train the final predictors with 20 features. It is beneficial to use a smaller set of features to better cover the event space (as it is smaller), thereby increasing representativeness and reducing the risk of overfitting. The flowchart of PON-All is shown in **Figure 1**. We trained two predictors, one with and one without GO terms. The selected features are listed in **Supplementary Tables S3, S4**. Of the 20 features on both lists, 15 were shared by the two methods. The selected features represent different types of features, including amino acid features, variation type and neighborhood features, evolutionary conservation features, and protein feature. The unique features in the method with GO annotations included amino acid propensities, neighborhood features, conservation feature, and GO annotations. In the method without GO annotations, the unique features were for amino acid features and neighborhood feature. The importance of the features is indicated in **Supplementary Tables S3, S4**. The protein feature is the most informative, followed by sums of log odd ratios for GO terms and functional site terms. Sequence conservation features, the number of homologs and SIFT 4G feature, are followed in significance by position within sequence and number of nonpolar amino acids. The other selected features have clearly lower significance in the case of the predictor with the GO feature. The highest scores for features in the case of prediction without GO feature are for protein feature, number of SwissProt homologs, position within a sequence, number of nonpolar amino acids, and SIFT 4G score.

We trained the final predictors with 20 features both when including and excluding GO annotations and named the tool PON-All because it can predict the effects of amino acid substitutions in proteins from any organism, unlike many existing methods. By default, predictions are made using GO features. However, if the annotations are missing, a predictor not requiring these features is used.

Performance Assessment With Blind Test Data Set

The performance of the method was tested with the blind test set, data that were withdrawn in the initial partitioning and not used during method development. **Table 3** shows results for PON-All with and without GO annotations. There are results for the entire test data set and separately for the three groups of organisms. As

TABLE 3 | Performance assessment in the blind test set with and without the GO feature. The results are shown with and without (in brackets) rejection.

Measure	All variants		Humans		Animals		Plants	
	w GO	wo GO	w GO	wo GO	w GO	wo GO	w GO	wo GO
TP	1,945 (2,278)	1,201 (1,928)	1,274 (1,552)	789 (1,327)	72 (102)	64 (112)	603 (624)	341 (489)
TN	1,855 (2,284)	1,344 (2,109)	1,421 (1,780)	1,052 (1,659)	118 (143)	98 (140)	318 (361)	201 (310)
FP	143 (365)	177 (540)	138 (326)	148 (447)	4 (26)	14 (29)	4 (13)	15 (64)
FN	217 (433)	251 (783)	94 (268)	154 (493)	35 (53)	12 (43)	88 (112)	85 (247)
PPV	0.932 (0.862)	0.872 (0.781)	0.902 (0.826)	0.842 (0.748)	0.947 (0.797)	0.821 (0.794)	0.993 (0.980)	0.958 (0.884)
NPV	0.895 (0.841)	0.843 (0.729)	0.938 (0.869)	0.872 (0.771)	0.771 (0.730)	0.891 (0.765)	0.783 (0.763)	0.703 (0.557)
Sensitivity	0.900 (0.840)	0.827 (0.711)	0.931 (0.853)	0.837 (0.729)	0.673 (0.658)	0.842 (0.723)	0.873 (0.848)	0.800 (0.664)
Specificity	0.928 (0.862)	0.884 (796)	0.911 (0.845)	0.877 (0.788)	0.967 (0.846)	0.875 (0.828)	0.988 (0.965)	0.931 (0.829)
Accuracy	0.913 (0.851)	0.856 (0.753)	0.921 (0.849)	0.859 (0.761)	0.830 (0.756)	0.862 (0.778)	0.909 (0.887)	0.844 (0.720)
MCC	0.827 (0.703)	0.712 (0.509)	0.841 (0.697)	0.714 (0.518)	0.678 (0.515)	0.714 (0.555)	0.817 (0.777)	0.695 (0.466)
AUC	0.913 (0.851)	0.856 (0.753)	0.921 (0.85)	0.855 (0.758)	0.818 (0.751)	0.858 (0.775)	0.929 (0.895)	0.842 (0.747)
OPM	0.763 (0.617)	0.628 (0.429)	0.781 (0.611)	0.630 (0.438)	0.588 (0.434)	0.631 (0.470)	0.751 (0.701)	0.608 (0.391)
Coverage	0.776 (1.000)	0.555 (1.000)	0.746 (1.000)	0.546 (1.000)	0.707 (1.000)	0.580 (1.000)	0.913 (1.000)	0.578 (1.000)

the ratios of variations in the groups are widely different, it is important to look at them separately. Otherwise, the largest group, for human variations, would dominate the overall output.

In the results for the entire data set and when using GO annotations, the OPM was 0.763, accuracy 0.913, and MCC 0.827. Overall, the method is well balanced (**Table 3**). Without GO features, the performance dropped somewhat, OPM to 0.628, accuracy to 0.856, and MCC to 0.712. The results for the option without rejection were further reduced. The overall coverage with GO and with rejection was 0.776 and without rejection complete (1.000). The corresponding figures for predictions without GO terms were 0.551 and 1.000. Thus, the increase in coverage comes with reduced overall performance.

ClinVar provides community-assessed variation information. It would have been interesting to train the tool with benign cases from this database, but there were only 370 cases. They were used for an additional test of specificity. After removing variants used for PON-All training, there were 298 variants left. The specificity for this data set was 0.982 with the GO feature and 0.84 without the GO parameter. The coverages were 0.729 and 0.515, respectively. The specificity is very similar to that of PON-P2 on a much larger ExAC data set (Niroula and Vihinen, 2019).

When we compared the results for variants in humans, animals, and plants separately (**Table 3**), predictions for humans were somewhat increased from those for all variants, OPM of 0.781 (vs. 0.763), accuracy of 0.921 (vs. 0.913), and MCC of 0.841 (vs. 0.827). The differences are about the same magnitude also for the other measures. The predictions are about the same degree lower for plants as they are increased for humans in comparison to the total. For example, the best results, those with GO features and rejection in plants, were for OPM 0.751 vs. 0.763 for all variants and, similarly, accuracy 0.909 vs. 0.913 and MCC 0.817 vs. 0.827. The corresponding measures for animals were substantially lower, 0.588, 0.830, and 0.678. The reason for the drop in the scores for animals is that only a small number of animal-specific variants were available. Overall, the results for PON-All were good, and the tool can be used for reliable predictions of unknown cases.

To further test the impact of data sets, we trained separate predictors for human, animal, and plant variants using the PON-All training data. The results of the blind test are shown in **Supplementary Table S5**. The performance scores for humans and plants are close to those for PON-All. Interestingly, the performance of the human-specific predictor is slightly lower than for PON-All. OPM in a blind test with GO is 0.774 while the figure for PON-All is 0.781, the corresponding figures for accuracy are 0.918 and 0.921 and for MCC 0.836 and 0.841. Similarly, all the other scores are also very close to those for PON-All. Thus, the differences are very small in the third decimal. Similar observations were made with plant variants.

The coverage is also slightly lower for the human-specific prediction, whereas the specific predictor has somewhat higher coverage in plants. The coverage of the animal-specific predictor is clearly lower (0.605 vs. 0.707) than the results for PON-All. The training data for animal variants was so small that this is expected. What is somewhat unexpected is that the scores are better for the specific than the generic predictor. MCC of the specific tool with GO feature and rejection is 0.803 *versus* 0.678. Similarly, accuracy is 0.903 *versus* 0.830 and OPM 0.734 *versus* 0.588. One could have expected human variants to increase the performance of animal cases, but that seems not to be the case.

Even the results for animal variant predictors are promising, especially when considering that only 306 variants were used for training. The blind test set for animals contained 324 variants. In conclusion, the performances of the PON-All were close to those for specific predictors, and since the generic predictor has been trained with a large number of cases, the method can predict the effects of variants in all kinds of proteins in all organisms. PON-All was slightly better for human and plant variants. Only in the case of the animal variants, the specific tool was somewhat better. In conclusion, the generic PON-All is overall the best choice. We would argue this to be true also in the case of animal variants, as the large body of cases for humans will allow details for predictions in animals, as well. However, this may be species-dependent.

The largest portion of variations were for humans. Most previous methods that can be used for other organisms have been trained on human data only. Therefore, we tested the performance when animal and plant variants were predicted with a human-specific predictor. The results are in **Supplementary Table S6**. Compared to generic PON-All

TABLE 4 | Blind test performance of PON-All compared to other predictors.

	PON-all wGO	PON-all woGO	PON-P2	SIFT 4G	PolyPhen2	MutationTaster	FATHMM	PROVEAN	MetaSVM	MetaLR	CADD_10[a]	CADD_15[a]	CADD_20[a]
TP	1,274	789	831	1,391	1,530	1,544	1,113	1,364	1,239	1,234	1,630	1,599	1,545
TN	1,421	1,052	1,032	1,197	1,003	1,044	1,472	1,320	1,651	1,639	498	710	1,020
FP	138	148	141	591	790	771	323	490	166	178	1,319	1,107	797
FN	94	154	132	288	149	135	563	313	440	445	49	80	134
PPV	0.902	0.842	0.855	0.702	0.659	0.667	0.775	0.736	0.882	0.874	0.553	0.591	0.660
NPV	0.938	0.872	0.887	0.806	0.871	0.885	0.723	0.808	0.790	0.786	0.910	0.899	0.884
Sens	0.931	0.837	0.863	0.828	0.911	0.920	0.664	0.813	0.738	0.735	0.971	0.952	0.920
Spes	0.911	0.877	0.880	0.669	0.559	0.575	0.820	0.729	0.909	0.902	0.274	0.391	0.561
ACC	0.921	0.859	0.872	0.746	0.730	0.741	0.745	0.770	0.827	0.822	0.609	0.660	0.734
MCC	0.841	0.714	0.742	0.503	0.500	0.523	0.491	0.543	0.659	0.649	0.337	0.410	0.512
OPM	0.781	0.63	0.661	0.423	0.416	0.436	0.414	0.459	0.570	0.559	0.291	0.341	0.426
Coverage	0.746	0.546	0.544	0.883	0.884	0.890	0.884	0.888	0.890	0.890	0.890	0.890	0.890

[a] *For CADD, 10,15, and 20 are three common thresholds.*

(**Table 3**), in the case of animal variants, some scores are better for animal data. In the case of plant variants, the generic predictor provides better results. These results can be explained mainly by the small number of available animal variants. Human cases provide additional strength for the prediction. Plants are so different from humans that a similar effect is not seen. The coverage of animal variants with human-specific predictors is somewhat smaller than for PON-All and substantially reduced for plant variants. Comparison to animal and plant-specific predictors in **Supplementary Table S5** indicates that the measures for the human-based predictions are clearly lower, except for coverage of animal variants. There, the wider distribution of human training cases leads to increased performance.

Comparison to Other Tools

It was not possible to compare the performance to non-human variant predictors because they are not available as predictors or they are based on the same data sets as used herein. The method for mammalian variants (Plekhanova et al., 2019) is an ML tool chosen among several tested algorithms. The data set contained human, mouse, dog, and cattle variants. Fourteen features without selection were used. Two methods have been described for plant variants. One of them is specific for *A. thaliana* (Kono et al., 2018) and was trained on the same data as PON-All. The method is based on the likelihood ratio test implemented with the BAD_Mutations pipeline (Kono et al., 2016). The other plant predictor was trained on *Arabidopsis* cases (Kovalev et al., 2018) using transfer learning based on 18 features but without feature selection.

We compared the performance of PON-All to several widely used generic variant tolerance predictors. The compared tools included CADD (Kircher et al., 2014), FATHMM (Rogers et al., 2018), MetaLR and MetaSVM (Dong et al., 2015), MutationTaster (Schwarz et al., 2014), PolyPhen2 (1), PON-P2 (Niroula et al., 2015), PROVEAN (Choi et al., 2012), and SIFT 4G (Vaser et al., 2016). **Table 4** indicates that, with the GO feature, the scores are better than for PON-P2, which has the closest performance. The other methods have clearly lower performance. These results are in line with many previous benchmarks that have shown PON-P2 to be the best or among the best tools (Niroula and Vihinen, 2019; Orioli and Vihinen, 2019; Sarkar et al., 2020). The coverage of PON-All is almost 20% higher than that for PON-P2, thus providing a significant improvement when also the performance scores are improved.

In the case of CADD, results are provided for three widely used thresholds since the developers did not optimize the threshold. By putting the value to 20, it was possible to increase the performance. However, this came with the cost of increased false-positive hits. As previous benchmarks have indicated (Niroula and Vihinen, 2019; Orioli and Vihinen, 2019), CADD has a substantial false hit rate so that about 1/3 of benign variants are classified as pathogenic. PON-P2, MetaSVM, and MetaLR had the best performances after PON-All (**Table 4**). PON-P2 is the closest to PON-All. The scores for MetaSVM and MetaLR are clearly lower.

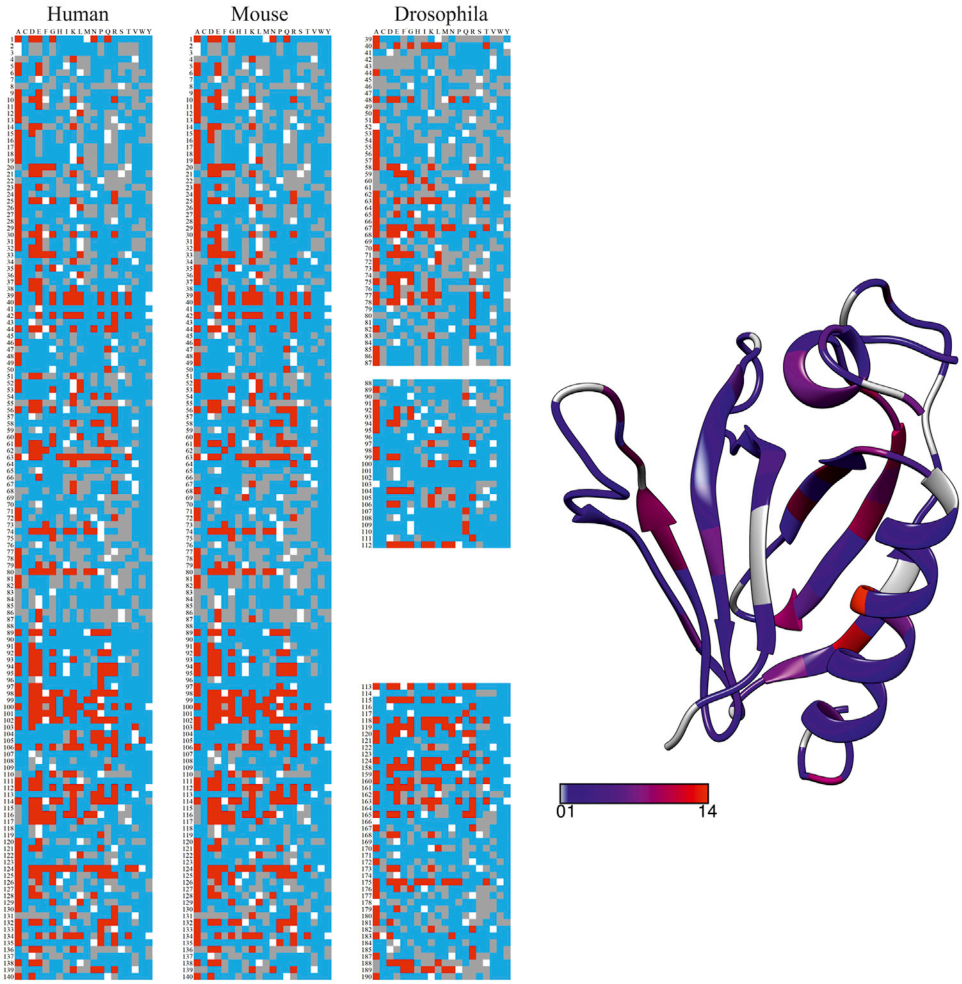


FIGURE 2 | Visualization of PON-All predictions (without GO terms) for human, mouse, and *Drosophila* BTK PH domains. The sequences were adjusted based on multiple sequence alignment so that corresponding amino acid positions in the three proteins are on the same line. Red indicates predicted pathogenic variation, benign variants are blue, UVs are gray, and the original residue is white. The distribution of the number of pathogenic variants is shown to the right in the human PH domain (PDB 6tt2). The color scheme for the numbers of pathogenic variants is the scale in the figure.

Example of Application

To highlight the applicability and performance of the PON-All tool, we predicted all possible amino acid substitutions in three related proteins. Predictions were made for Bruton tyrosine kinase (BTK) pleckstrin homology (PH) domains. The sequences were obtained from UniProtKB for human (Q06187-1), mouse (P35991-1), and *Drosophila melanogaster* (P08630-1) BTK. The sequences were aligned with Clustal Omega (Sievers et al., 2011). Harmful variants in human BTK cause X-linked agammaglobulinemia (XLA), a primary immunodeficiency (Mohamed et al., 2009), in mouse X-linked immunodeficiency (Khan et al., 1995). In *Drosophila*, the related protein, BTK29A, is involved, for example, in survival and male genital development (Hamada et al., 2005). Numerous XLA-causing variants are known in humans and listed in BTKbase (Väliaho et al., 2006). In xid mice, variant E41K in the PH domain is the causative alteration. *Drosophila* fic^P variant is due to intronic alteration and causes alternative splicing and deletion of the PH domain (Baba et al., 1999). Thus, variants in the BTK PH domain are related to important functions in all the three organisms; thereby, it is of interest to investigate the effects of variants in these domains.

All the 19 possible single amino acid substitutions in each position were generated and predicted with PON-All. The results are shown in **Figure 2**, where the predicted pathogenic and benign variants are color-coded. Mouse and *Drosophila* sequences were aligned with human BTK by either deleting amino acids or adding empty lines to keep the sequences in alignment. The human BTK PH domain (PDB id 6tt2) to the right indicates the number of predicted harmful variants by a rainbow coloring scheme. The maximum number of harmful variants in a position was 14, shown in bright red. These residues are in the middle of secondary structural elements. The majority of the variants in tolerant positions, gray for those where no harmful variants were predicted and blue with small numbers of harmful variants, are mainly in the ends of secondary structural elements and in surface loops. Interestingly, positions 7 and 8 in the middle of the first β-strand tolerate all substitutions. The differences in vulnerabilities are also clearly visible in the graphs for the mouse and fruit fly sequences.

The method facilitates the first-time systematic comparison of site vulnerabilities for sequences from various organisms.

PON-All Web Application

PON-All is freely available as a web application at http://structure.bmc.lu.se/PON-All/ and http://8.133.174.28:8999/. The program has a user-friendly web interface that accepts variations in protein sequence, as amino acid substitutions, or in a VCF file (human). Batch submission, including all variants and proteins of interest, is accepted and recommended. PON-All provides a complete report, which is sent to the user by email when ready.

DISCUSSION

We have developed the first generic variant pathogenicity predictor that has been trained and tested on variants also from animals and plants. PON-All shows good performance for the prediction of the three types of organisms. Because the number of animal variants was clearly smaller than that for plants or humans, the drop in the performance is understandable. We could have increased animal variants by including cases from rodent databases. However, we felt that it would have biased the data as lots of these variants are generated to model human conditions. Overfitting is a potential problem in gradient boosting methods. Independent cross-validation and blind test set results are well in line. If the method were overfitted, there would be discrepancies in the performance for the different data sets and partitions. Further, we have used extensive data set and a minimal number of features, which are the classical remedies for overfitting.

PON-All has improved performance in comparison to the other methods. In addition to higher reliability, the tool has also increased coverage, up to 20% in comparison to PON-P2. This is important and facilitates reliable predictions in substantially increased numbers. These methods will never reach 100% coverage because disease-causing variants display a continuum. Some variants can be disease-related in some individuals, but not in all who carry the variant. PON-All is good at recognizing such cases and ranking them as UVs.

The use of GO features and the reject option clearly improved the performance. This is the default mode of prediction; however, apart from human and some well studied model organisms not a feasible option. GO annotations are scarce or missing for less investigated organisms. Even in these cases, predictions are still rather reliable. The coverage of such variants was reduced. Still, the new method makes a significant contribution also in these cases.

Fifteen out of the 20 features per predictor are shared with the tools that have been trained with or without GO features. Thus, in addition to the GO feature, some others differ between the two installations. This indicates the interplay between features and that it is important to perform proper feature selection. Some of the previous tools have been developed without feature selection, just using all the features that were originally collected. When the numbers of variants are small, as in this study, especially for animals, the event space remains very large if feature selection is not applied. A small number of training and test cases cannot cover such a space, and the representativeness is low (Schaafsma and Vihinen, 2018).

Methods like this are used for variant interpretation and recognition of pathogenic or, more generally, harmful variants. PON-All can be used for all organisms. In the case of variants in pathogens, it has to be remembered that harmful variants in such organisms mean harmful variants for that organism, not for human or other target organisms.

DATA AVAILABILITY STATEMENT

The data sets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: VariBench http://structure.bmc.lu.se/VariBench and http://8.133.174.28:8999/.

AUTHOR CONTRIBUTIONS

YY, analysis, resources, software, and funding. AS, software, analysis, and visualization. MV, conceptualization, resources, analysis, visualization, and funding. All authors wrote the manuscript and read and approved the submitted version.

FUNDING

This work was supported by the key project of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJA520010), Collaborative Innovation Center of Novel Software Technology and Industrialization, Vetenskapsrådet (2019-01403), and the Swedish Cancer Society (Grant no. CAN 2017/699 and CAN 20 1350).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.867572/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, 248–249. doi:10.1038/nmeth0410-248

Ali, H., Urolagin, S., Gurarslan, Ö., and Vihinen, M. (2014). Performance of Protein Disorder Prediction Programs on Amino Acid Substitutions. *Hum. Mutat.* 35, 794–804. doi:10.1002/humu.22564

Altschul, S., Madden, T. L., Schaffer, 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

Baba, K., Takeshita, A., Majima, K., Ueda, R., Kondo, S., Juni, N., et al. (1999). The Drosophila Bruton's Tyrosine Kinase (Btk) Homolog Is Required for Adult Survival and Male Genital Formation. *Mol. Cell Biol.* 19, 4405–4413. doi:10.1128/mcb.19.6.4405

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

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, 1237–1244. doi:10.1002/humu.21047

Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P. L., Altman, R. B., Casadio, R., et al. (2013). WS-SNPs&GO: a Web Server for Predicting the Deleterious Effect of Human Protein Variants Using Functional Annotation. *BMC Genomics* 14 (Suppl. 3), S6. doi:10.1186/1471-2164-14-S3-S6

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: Online Access to Ontology and Annotation Data. *Bioinformatics* 25, 288–289. doi:10.1093/bioinformatics/btn615

Chen, T., Guestrin, C., and XGBoost (2016). A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Fransisco, CA, USA: ACM, 785–794.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* 7, e46688. doi:10.1371/journal.pone.0046688

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Hum. Mol. Genet.* 24, 2125–2137. doi:10.1093/hmg/ddu733

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* 46, 389–422. doi:10.1023/a:1012487302797

Hamada, N., Bäckesjö, C.-M., Smith, C. I. E., and Yamamoto, D. (2005). Functional Replacement ofDrosophilaBtk29A with Human Btk in Male Genital Development and Survival. *FEBS Lett.* 579, 4131–4137. doi:10.1016/j.febslet.2005.06.042

Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., et al. (2007). Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms. *Cancer Res.* 67, 465–473. doi:10.1158/0008-5472.can-06-1736

Kawashima, S., and Kanehisa, M. (2000). AAindex: Amino Acid Index Database. *Nucleic Acids Res.* 28, 374. doi:10.1093/nar/28.1.374

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). *A Highly Efficient Gradient Boosting Decision Tree Neural Information Processing Systems*. USA: La Jolla, CA.

Khan, W. N., Alt, F. W., Gerstein, R. M., Malynn, B. A., Larsson, I., Rathbun, G., et al. (1995). Defective B Cell Development and Function in Btk-Deficient Mice. *Immunity* 3, 283–299. doi:10.1016/1074-7613(95)90114-0

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, 310–315. doi:10.1038/ng.2892

Kono, T. J. Y., Fu, F., Mohammadi, M., Hoffman, P. J., Liu, C., Stupar, R. M., et al. (2016). The Role of Deleterious Substitutions in Crop Genomes. *Mol. Biol. Evol.* 33, 2307–2317. doi:10.1093/molbev/msw102

Kono, T. J. Y., Lei, L., Shih, C.-H., Hoffman, P. J., Morrell, P. L., and Fay, J. C. (2018). Comparative Genomics Approaches Accurately Predict Deleterious Variants in Plants. *G3 (Bethesda)* 8, 3321–3329. doi:10.1534/g3.118.200563

Kovalev, M. S., Igolkina, A. A., Samsonova, M. G., and Nuzhdin, S. V. (2018). A Pipeline for Classifying Deleterious Coding Mutations in Agricultural Plants. *Front. Plant Sci.* 9, 1734. doi:10.3389/fpls.2018.01734

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. *Nucl. Acids Res.* 42, D980–D985. doi:10.1093/nar/gkt1113

Lockwood, S., Krishnamoorthy, B., and Ye, P. (2011). Neighborhood Properties Are Important Determinants of Temperature Sensitive Mutations. *PLoS One* 6, e28507. doi:10.1371/journal.pone.0028507

Mohamed, A. J., Yu, L., Bäckesjö, C.-M., Vargas, L., Faryal, R., Aints, A., et al. (2009). Bruton's Tyrosine Kinase (Btk): Function, Regulation, and Transformation with Special Emphasis on the PH Domain. *Immunol. Rev.* 228, 58–73. doi:10.1111/j.1600-065x.2008.00741.x

Munoz-Torres, M., and Carbon, S. (2017). Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools. *Methods Mol. Biol.* 1446, 149–160. doi:10.1007/978-1-4939-3743-1_11

Nair, P. S., and Vihinen, M., (2013). VariBench: A Benchmark Database for Variations. *Hum. Mutat.* 34, 42–49. doi:10.1002/humu.22204

Nicholas, F. W. (2003). Online Mendelian Inheritance in Animals (OMIA): a Comparative Knowledgebase of Genetic Disorders and Other Familial Traits in Non-laboratory Animals. *Nucleic Acids Res.* 31, 275–277. doi:10.1093/nar/gkg074

Niroula, A., Urolagin, S., and Vihinen, M (2015). PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLoS ONE* 10 (2), e0117380. doi:10.1371/journal.pone.0117380

Niroula, A., and Vihinen, M. (2019). How Good Are Pathogenicity Predictors in Detecting Benign Variants? *PLoS Comput. Biol.* 15, e1006481. doi:10.1371/journal.pcbi.1006481

Niroula, A., and Vihinen, M (2016). PON-mt-tRNA: a Multifactorial Probability-Based Method for Classification of Mitochondrial tRNA Variations. *Nucleic Acids Res.* 44, 2020–2027. doi:10.1093/nar/gkw046

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

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

Olatubosun, A., Väliaho, J., Härkönen, J., Thusberg, J., Vihinen, M., and Pon-, P. (2012). Integrated Predictor for Pathogenicity of Missense Variants. *Hum. Mutat.* 33, 1166–1174. doi:10.1002/humu.22102

Orioli, T., and Vihinen, M. (2019). Benchmarking Subcellular Localization and Variant Tolerance Predictors on Membrane Proteins. *BMC Genomics* 20, 547. doi:10.1186/s12864-019-5865-0

Pavey, T. G., Gilson, N. D., Gomersall, S. R., Clark, B., and Trost, S. G. (2017). Field Evaluation of a Random Forest Activity Classifier for Wrist-Worn Accelerometer Data. *J. Sci. Med. Sport* 20, 75–80. doi:10.1016/j.jsams.2016.06.003

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Plekhanova, E., Nuzhdin, S. V., Utkin, L. V., and Samsonova, M. G. (2019). Prediction of Deleterious Mutations in Coding Regions of Mammals with Transfer Learning. *Evol. Appl.* 12, 18–28. doi:10.1111/eva.12607

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi:10.1038/gim.2015.30

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., Campbell, C., et al. (2018). FATHMM-XF: Accurate Prediction of Pathogenic Point Mutations via Extended Features. *Bioinformatics* 34, 511–513. doi:10.1093/bioinformatics/btx536

Sarkar, A., Yang, Y., and Vihinen, M. (2020). Variation Benchmark Datasets: Update, Criteria, Quality and Applications. *Database* 2020, baz117.

Schaafsma, G. C. P., and Vihinen, M. (2018). Representativeness of Variation Benchmark Datasets. *BMC Bioinforma.* 19 (1), 461. doi:10.1186/s12859-018-2478-6

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: Mutation Prediction for the Deep-Sequencing Age. *Nat. Methods* 11, 361–362. doi:10.1038/nmeth.2890

Shen, B., and Vihinen, M. (2004). Conservation and Covariance in PH Domain Sequences: Physicochemical Profile and Information Theoretical Analysis of

XLA-Causing Mutations in the Btk PH Domain. *Protein Eng. Des. Sel.* 17, 267–276. doi:10.1093/protein/gzh030

Shomer, B. (1997). Seqalert-a Daily Sequence Alertness Server for the EMBL and SWISSPROT Databases. *Bioinformatics* 13, 545–547. doi:10.1093/bioinformatics/13.5.545

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi:10.1038/msb.2011.75

Väliaho, J., Smith, C. I. E., and Vihinen, M. (2006). BTKbase: the Mutation Database for X-Linked Agammaglobulinemia. *Hum. Mutat.* 27, 1209–1217.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT Missense Predictions for Genomes. *Nat. Protoc.* 11, 1–9. doi:10.1038/nprot.2015.123

Vihinen, M. (2012). How to Evaluate Performance of Prediction Methods? Measures and Their Interpretation in Variation Effect Analysis. *BMC Genomics* 13 (Suppl. 4), S2. doi:10.1186/1471-2164-13-S4-S2

Vihinen, M. (2020). Problems in Variation Interpretation Guidelines and in Their Implementation in Computational Tools. *Mol. Genet. Genomic Med.* 8, e1206. doi:10.1002/mgg3.1206

Vihinen, M. (2013). Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis. *Hum. Mutat.* 34, 275–282. doi:10.1002/humu.22253

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

Wang, D., Zhang, Y., Zhao, Y., and LightGBM (2017). Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics - ICCBB. October 18 - 20, 2017. New York, NY, United States. Association for Computing Machinery, 7–11.

Yang, Y., Urolagin, S., Niroula, A., Ding, X., Shen, B., and Vihinen, M. (2018). PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. *Int. J. Mol. Sci.* 19, 19. doi:10.3390/ijms19041009

Yang, Y., Zeng, L., and Vihinen, M. (2021). Prediction of Effects of Variants on Protein Solubility. *Int. J. Mol. Sci.* 22, 8027.

Yang, Y., Niroula, A., Shen, B., and Vihinen, M. (2016). PON-sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. *Bioinformatics* 32, 2032–2034. doi:10.1093/bioinformatics/btw066

Yu, B., Qiu, W., Chen, C., Ma, A., Jiang, J., Zhou, H., et al. (2020). SubMito-XGBoost: Predicting Protein Submitochondrial Localization by Fusing Multiple Feature Information and eXtreme Gradient Boosting. *Bioinformatics* 36, 1074–1081. doi:10.1093/bioinformatics/btz734

Zhang, J., Mucs, D., Norinder, U., Svensson, F., and LightGBM (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model.* 59, 4150–4158. doi:10.1021/acs.jcim.9b00633

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 | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fmolb.2022.1063620

Check for updates

OPEN ACCESS

EDITED BY
Daniele Dell'Orco,
University of Verona, Italy

REVIEWED BY
Roberta Chiaraluce,
Sapienza University of Rome, Italy
Carlo Travaglini-Allocatelli,
Sapienza University of Rome, Italy

*CORRESPONDENCE
Angel L. Pey,
angelpey@ugr.es

†These authors have contributed equally to this work and share first authorship

‡These authors share senior authorship

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

RECEIVED 07 October 2022
ACCEPTED 03 November 2022
PUBLISHED 24 November 2022

CITATION
Pacheco-Garcia JL, Cagiada M,
Tienne-Matos K, Salido E,
Lindorff-Larsen K and L. Pey A (2022),
Effect of naturally-occurring mutations on the stability and function of cancer-associated NQO1: Comparison of experiments and computation.
Front. Mol. Biosci. 9:1063620.
doi: 10.3389/fmolb.2022.1063620

Effect of naturally-occurring mutations on the stability and function of cancer-associated NQO1: Comparison of experiments and computation

Juan Luis Pacheco-Garcia[1†], Matteo Cagiada[2†], Kelly Tienne-Matos[1], Eduardo Salido[3], Kresten Lindorff-Larsen[2‡] and Angel L. Pey[4*‡]

[1]Departamento de Química-Física, Universidad de Granada, Granada, Spain, [2]Department of Biology, Linderstrøm-Lang Centre for Protein Science, University of Copenhagen, Copenhagen, Denmark, [3]Center for Rare Diseases (CIBERER), Hospital Universitario de Canarias, Universidad de la Laguna, La Laguna, Tenerife Tenerife, Spain, [4]Departamento de Química Física, Unidad de Excelencia en Química Aplicada a Biomedicina y Medioambiente e Instituto de Biotecnología, Universidad de Granada, Granada, Spain

Recent advances in DNA sequencing technologies are revealing a large individual variability of the human genome. Our capacity to establish genotype-phenotype correlations in such large-scale is, however, limited. This task is particularly challenging due to the multifunctional nature of many proteins. Here we describe an extensive analysis of the stability and function of naturally-occurring variants (found in the COSMIC and gnomAD databases) of the cancer-associated human NAD(P)H:quinone oxidoreductase 1 (NQO1). First, we performed *in silico* saturation mutagenesis studies (>5,000 substitutions) aimed to identify regions in NQO1 important for stability and function. We then experimentally characterized twenty-two naturally-occurring variants in terms of protein levels during bacterial expression, solubility, thermal stability, and coenzyme binding. These studies showed a good overall correlation between experimental analysis and

Abbreviations: CTD, C-terminal domain; COSMIC, Catalogue Of Somatic Mutations In Cancer cell lines database; Dic, dicoumarol; FAD, Flavin adenine dinucleotide; GEMME: Global Epistatic Model for predicting Mutational Effects; gnomAD, Genome Aggregation Database; HDX, hydrogen/deuterium exchange; HEPES, 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid; IMAC, Immobilized metal affinity chromatography; K_d, apparent dissociation constant; k_{obs}, observed rate constant for proteolysis; k_{prot}, second-order rate constant for proteolysis; LoF, Loss-of-function; MMI, monomer:monomer interface; MSA, multiple sequence alignment; NTD, N-terminal domain; NQO1, NAD(P)H:quinone oxidoreductase 1; S, soluble protein levels; SASA, solvent accessible surface area; SDS-PAGE, polyacrylamide gel electrophoresis in the presence of sodium dodecylsulphate; S/T, Soluble/Total ratio of NQO1 protein; T_m, apparent half-denaturation temperature; WT, wild-type; ΔE, Normalized predicted effect of an amino acid substitution from the GEMME model; $\Delta\Delta G$, computationally calculated change in unfolding Gibbs free energy between a variant and the WT protein; $\Delta\Delta G_{FAD}$, difference in binding Gibbs free energy between a variant and the WT protein; $\Delta\Delta G_{melting}$, operational metric for the estimation of the unfolding Gibbs free energy changes between a variant and the WT protein from ΔT_m; $\Delta\Delta G_{PROT}$, difference in free energy of cleavable and non-cleavable states for proteolysis between a variant and the WT protein; ΔT_m, difference in T_m between a variant and the WT protein.

computational predictions; also the magnitude of the effects of the substitutions are similarly distributed in variants from the COSMIC and gnomAD databases. Outliers in these experimental-computational genotype-phenotype correlations remain, and we discuss these on the grounds and limitations of our approaches. Our work represents a further step to characterize the mutational landscape of NQO1 in the human genome and may help to improve high-throughput *in silico* tools for genotype-phenotype correlations in this multifunctional protein associated with disease.

KEYWORDS

protein function, protein stability, genotype-phenotype correlations, computational prediction, sequence conservation

1 Introduction

Advances in technologies for whole-genome or exome sequencing and high-throughput functional assays have increased our knowledge on the consequences of the genetic variability in humans and the relationship to disease (McInnes et al., 2021; Arnedo-Pac et al., 2022; Høie et al., 2022; Katsonis et al., 2022). However, our capacity to predict the pathogenicity of single amino acid variants is still limited, with some approaches providing good overall results but failing to predict correlation for some individual mutations or phenotypes (Katsonis et al., 2022).

Current approaches for correlating genotype and phenotype can broadly be classified into two classes. First, experimental approaches are based on the characterization of one or several functional features (for example enzymatic function and regulation, protein-protein interactions, transport to different intracellular or extracellular locations, protein turnover, ligand binding) (Xu et al., 2017; Abildgaard et al., 2019; Pacheco-Garcia et al., 2021; Høie et al., 2022). In the case of high-throughput experimental approaches typically only one or two aspects of protein function are analysed (for example protein abundance or ability to rescue a growth phenotype) (Cagiada et al., 2021). Second, the use of structure- or sequence-based methods to predict pathogenicity are becoming increasingly robust (Abildgaard et al., 2019; Arnedo-Pac et al., 2022; Høie et al., 2022). Although experiments may be implemented in a high-throughput fashion, it has until now been limited to a relatively small set of proteins and assays (Arnedo-Pac et al., 2022; Høie et al., 2022). Thus, while computational approaches also have limitations, they may be appealing due to their potential application on a proteomic scale (Arnedo-Pac et al., 2022; Høie et al., 2022).

In this work, we apply both types of approaches to increase our understanding of the correlation between genotype and phenotype for missense variants of the human NAD(P)H: quinone oxidoreductase 1 (NQO1) protein. NQO1 is associated with several diseases including cancer, Alzheimer´s and Parkinson’s disease (Beaver et al., 2019; Luo et al., 2019). NQO1 is a multifunctional protein, displaying both enzymatic and non-enzymatic functions. As an enzyme, it catalyzes the FAD-dependent (two-electron) reduction of a large set of quinone substrates, with functions including redox maintenance of vitamins, detoxification of xenobiotics and activation of cancer pro-drugs (Ross and Siegel, 2018; Beaver et al., 2019; Anoz-Carbonell et al., 2020; Salido et al., 2022). Among non-enzymatic functions, NQO1 may interact with proteins and RNA, controlling their stability and function (Beaver et al., 2019; Asher et al., 2005; Ben-Nissan and Sharon, 2014; di Francesco et al., 2016; Oh et al., 2016). Many of these functions are associated with the catalytic competence and FAD binding, such as protein-protein interactions, intracellular stability and association with microtubules (Asher et al., 2005; Martínez-Limón et al., 2016; Martínez-Limón et al., 2020; Siegel et al., 2021). The native form of NQO1 is dimeric, containing two different domains: an N-terminal domain (NTD, residues 1–225), that tightly binds one FAD molecule/domain required for catalysis and contains most of the monomer-monomer interface (MMI), whereas the C-terminal domain (CTD, residues 225–274) complete the active site and the MMI (Li et al., 1995; Faig et al., 2000; Lienhart et al., 2014; Medina-Carmona et al., 2017a; Pacheco-Garcia et al., 2021).

We have recently shown that ligand binding and variant effects on stability propagate to long distances in the native state, affecting different functional features in a counterintuitive fashion (Medina-Carmona et al., 2016; Pey, 2018; Medina-Carmona et al., 2019a; Vankova et al., 2019; Pacheco-García et al., 2020; Pacheco-Garcia et al., 2021; Pacheco-Garcia et al., 2022a). Therefore, NQO1 represents a challenging and biomedically relevant system to compare the performance of computational and experimental methods to explain and to predict genotype-phenotype in a large-scale for a multi-functional protein. Here, we use computational tools to probe 5,187 variants of NQO1 that includes a set of clinically relevant missense variants which we then experimentally characterized. In this set, thirteen variants come from large-scale human sequencing data (gnomAD) and nine from the catalogue of somatic mutations in human cancer lines (COSMIC)

TABLE 1 Set of NQO1 variants experimentally characterized in this work.

Mutation	Database	% ASA[a]	Variant class	Residue class
G3S	gnomAD	6.0 ± 4.7	WT-like	WT-like
G3D	COSMIC	6.0 ± 4.7	WT-like	WT-like
L7P	COSMIC	0.3 ± 0.2	Total-loss	Total-loss
L7R	gnomAD	0.3 ± 0.2	Total-loss	Total-loss
V9I	gnomAD	0.0 ± 0.0	WT-like	Total-loss
T16M	gnomAD	43 ± 14	Stable but inactive	WT-like
Y20N	gnomAD	21 ± 5	WT-like	WT-like
A29T	COSMIC	2.2 ± 0.5	WT-like	Unstable but active
K32N	gnomAD	79 ± 11	WT-like	WT-like
G34V	gnomAD	54 ± 11	Total-loss	Stable but Inactive
E36K	gnomAD	64 ± 3	WT-like	WT-like
S40L	gnomAD	0.0 ± 0.0	WT-like	Total-loss
D41G	gnomAD	14 ± 2	Total-loss	Stable but inactive
D41Y	COSMIC	14 ± 2	Stable but inactive	Stable but inactive
M45L	COSMIC	28 ± 5	WT-like	WT-like
M45I	COSMIC	28 ± 5	WT-like	WT-like
I51V	gnomAD	14 ± 1	WT-like	Stable but inactive
W106R	gnomAD	10 ± 1	Total-loss	Total-loss
W106C	COSMIC	10 ± 1	Total-loss	Total-loss
F107C	gnomAD	7.6 ± 0.5	Unstable but active	Unstable but active
M155I	COSMIC	13 ± 3	WT-like	WT-like
H162N	COSMIC	6.6 ± 0.5	WT-like	WT-like

The table indicates whether the variants are found in the COSMIC/gnomAD databases as well as the solvent exposure (as % ASA) determined using a crystallographic model of WT NQO1 [PDB code 2F1O (Asher et al., 2006)], the software Getarea and the computational classification at variant and residue level using a combination of predictions of thermodynamic stability change upon mutation and evolutionary conservation.

[a] Using GetArea (http://curie.utmb.edu/getarea.html) and the structure with PDB code 2F1O (Asher et al., 2006). Data are the average ±s.d. from eight monomers.

(Table 1). As of ninth of January 2022, none of these variants were found in both databases. Whether variants found in COSMIC or gnomAD databases are associated with disease (e.g. predisposition to cancer development) is unknown. The set of variants we studied experimentally comprises very different amino acid side chain characteristics and display different levels of solvent exposure (Table 1).

2 Materials and methods

2.1 Experimental methods

2.1.1 Protein expression and purification

Mutations were introduced by site-directed mutagenesis in the wild-type (WT) NQO1 cDNA cloned into the pET-15b vector (pET-15b-NQO1) by GenScript (Leiden, Netherlands). Mutated codons were optimized for expression in *E. coli* and mutagenesis was confirmed by sequencing the entire cDNA. The plasmids were transformed in *E. coli* BL21 (DE3) cells (Agilent Technologies, Santa Clara, CA, United States) for protein expression.

To determine the amount of soluble NQO1 at 37°C, 5 ml of LB medium containing 0.1 mg mL^{-1} ampicillin (purchased from Canvax Biotech, Córdoba, Spain) was inoculated with transformed cells and grown for 16 h at 37°C. 0.5 ml of these cultures were diluted into 10 ml of LB containing 0.1 mg mL^{-1} ampicillin (LBA) and grown at 37°C for 3 h. After that, cultures were induced with 0.5 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG, Canvax Biotech) at 37 °C for 4 h. Cells were harvested by centrifugation at 2,900 *g* in a bench centrifuge at 4°C and frozen at −80°C for 16 h. Pellets were resuspended in binding buffer (20 mM Na-phosphate, 300 mM NaCl, 50 mM imidazole, pH 7.4) with 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma-Aldrich, Madrid, Spain) and sonicated in an ice bath. These *total extracts* were centrifugated (24,000 *g*, 30 min, 4°C in a bench centrifuge) to obtain the *soluble extracts*. The amount of NQO1 in total and soluble extracts was determined by Western-blotting providing the S/T (soluble/total) ratio for each variant. Samples were denatured using Laemmli's buffer, resolved in polyacrylamide gel electrophoresis in the presence of sodium dodecylsulphate (SDS-PAGE, 12% acrylamide) gels and transferred to PVDF membranes (GE Healthcare, Chicago, IL, United States) using standard procedures. Immunoblotting was carried out using primary monoclonal antibody against NQO1 (sc-393736, Santa Cruz Biotechnology, Dallas, TX, United States) at 1: 500 dilution and, as secondary antibody, an anti-mouse IgGκ BP-HRP (sc-516102, Santa Cruz Biotechnology) at 1: 2000 dilution was used. Samples were visualized using luminol-based enhanced chemiluminescence (from BioRad Laboratories, Hercules, CA, United States), scanned and analysed using Image Lab (from BioRad Laboratories).

For large-scale purifications, a preculture (100 ml) was prepared from a single clone for each variant and grown for 16 h at 37°C in LBA and diluted into 2.4–4.8 L of LBA. After 3 h at 37°C, NQO1 expression was induced by the addition of 0.5 mM IPTG for 6 h at 25°C. Cells were harvested by centrifugation at 8,000 *g* and frozen overnight at −80 °C. NQO1 proteins were purified using immobilized nickel affinity chromatography columns (IMAC, GE Healthcare) as described (Anoz-Carbonell et al., 2020). Isolated dimeric fractions of NQO1 variants were exchanged to HEPES-KOH buffer 50 mM pH 7.4 using PD-10 columns (GE Healthcare). The UV–visible spectra of purified NQO1 proteins were measured in a Cary spectrophotometer (Agilent Technologies, Waldbronn, Germany) and used to quantify NQO1 concentration and the content of FAD as described in (Anoz-Carbonell et al., 2020). Apo-proteins were obtained as described in (Vankova et al., 2019). Briefly, holo-proteins were incubated with 2 M urea and 2 M KBr in HEPES-KOH 50 mM pH 7.4 in the presence of 2 mM β-mercaptoethanol and 1 mM PMSF and loaded into IMAC columns, washed with 2 M urea and 2 M KBr in HEPES-KOH 50 mM pH 7.4, 2 mM β-mercaptoethanol, then with HEPES-KOH 50 mM pH 7.4, 2 mM β-mercaptoethanol eluted with 20 mM Na-Phosphate 300 mM NaCl 500 mM imidazole pH 7.4 and finally exchanged to HEPES-KOH 50 mM pH 7.4. These apo-proteins contained less than 2% bound FAD based on UV-visible spectra. Samples were stored at −80°C upon flash freezing in liquid N_2. Protein purity and integrity was checked by SDS-PAGE.

2.1.2 *In vitro* characterization of NQO1 variants

Thermal denaturation of NQO1 proteins, as holo-proteins (2 μM in monomer +100 μM FAD) was monitored by following changes in tryptophan emission fluorescence in HEPES-KOH 50 mM at pH 7.4 as described in (Medina-Carmona et al., 2017b). T_m values were reported as mean ± s.d. of four independent measurements.

Fluorescence titrations were carried out at 25°C using 1 cm × 0.3 cm path-length cuvettes in a Cary Eclipse spectrofluorimeter (Agilent Technologies, Waldbronn, Germany). Experiments were performed in 20 mM K-phosphate, pH 7.4, essentially as described in (Pacheco-García et al., 2020). Briefly, apo-NQO1 (0.25 μM in subunit) was mixed with 0–2 μM FAD in K-phosphate 20 mM pH 7.4. Samples were incubated at 25°C in the dark for at least 10 min before measurements. Spectra were acquired in the 340–360 nm range upon excitation at 280 nm (slits 5 nm), and spectra were averaged over 10 scans registered at a scan rate of 200 nm min^{-1}. FAD binding fluorescence intensities at 350 nm were fitted using a single and identical type of binding sites as described in (Pacheco-García et al., 2020). Variant effects on the FAD binding free energy ($\Delta\Delta G_{FAD}$) were calculated as:

$$\Delta\Delta G_{FAD} = R \cdot T \cdot \ln \frac{K_{d(mut)}}{K_{d(WT)}} \tag{1}$$

Where R is the ideal gas constant (1.987 cal mol·K^{-1}), T is the experimental temperature (298.15 K), and $K_{d(mut)}$ and $K_{d(WT)}$ are the FAD binding dissociation constant of the mutant and WT protein variants, respectively. A positive value of $\Delta\Delta G_{FAD}$ indicates that the mutation reduces the affinity for FAD.

For proteolysis studies, NQO1 samples (10 μM in monomer) were prepared in HEPES-KOH 50 mM at pH 7.4 in the presence of 100 μM FAD (NQO1_{holo}) and incubated at 25ºC for 5 min. The proteolysis reaction was initiated upon addition of 0.02–1.2 μM thermolysin (from *Geobacillus stearothermophilus*, Sigma-Aldrich) and a final concentration of 10 mM $CaCl_2$. Samples were incubated at 25°C and aliquots were collected over time and the reaction quenched by addition of EDTA pH 8 (final concentration of 20 mM) and Laemmli´s buffer (2x). Controls (time 0) were prepared likewise but without thermolysin. Samples were resolved by SDS-PAGE under reducing conditions in gels containing 12% acrylamide. Gels were stained with Coomassie blue G-250. Densitometry was carried out using ImageJ. Data were analyzed using an exponential function to provide the apparent rate constant (k_{obs}). From the linear dependence of k_{obs} vs. [thermolysin], we obtained the second-order rate constant k_{prot}. Linearity in these plot indicate that proteolysis occurs under a EX2 mechanism, thus reflecting the thermodynamic stability of the thermolysin cleavage site (Ser72-Val73) between non-cleavable and cleavable states (Park and Marqusee, 2004). These k_{prot} values were used to determine mutational effects on the local stability of thermolysin cleave site ($\Delta\Delta G_{PROT}$) using Eq. 2:

$$\Delta\Delta G_{PROT} = R \cdot T \cdot ln \frac{k_{prot\,(mut)}}{k_{prot\,(WT)}} \tag{2}$$

Where R is the ideal gas constant (1.987 cal mol·K^{-1}), T is the experimental temperature (298.15 K), and $k_{prot(mut)}$ and $k_{prot(WT)}$ are the second-order proteolysis rate constants of the mutant and WT protein variants, respectively. A positive value of $\Delta\Delta G_{PROT}$ indicates that the mutation thermodynamically destabilizes the thermolysin cleavage site.

2.2 Computational analyses

2.2.1 Evolutionary conservation analysis

We used GEMME (Laine et al., 2019) to evaluate evolutionary distances from the WT NQO1 sequence (Uniprot ID: P15559 — isoform 1). We used HHblits (Remmert et al., 2011; Steinegger et al., 2019) to generate a multiple sequence alignment

(MSA) using UniClust30 as sequence database and an E-value threshold of 10^{-10}. The resulting MSA contained 1,602 sequences and was refined using two additional filters: first, all the columns that were not present in the WT NQO1 sequence were removed; second, all the sequences with more than the 50% of gaps were removed. Application of these two filters yielded 1,414 sequences in the MSA. The GEMME package was run using default parameters. For each position, a median score was evaluated using all the available substitutions.

2.2.2 Thermodynamic stability predictions

Changes in thermodynamic stability (ΔΔG) were calculated using the crystal structure (Faig et al., 2000) (PDB 1D4A) and the Rosetta package (GitHub SHA1 c7009b3115c22daa9efe2805d9d1ebba08426a54) with the Cartesian ΔΔG protocol (Park et al., 2016; Frenz et al., 2020). The values obtained from Rosetta in internal Energy Unit were divided by 2.9 to bring them on to a scale corresponding to kcal·mol^{-1} (Park et al., 2016). Median scores were evaluated for each position using all the available substitutions.

We used DSSP (Kabsch and Sander, 1983) to calculate the solvent accessible surface area (SASA) when identifying interface residues in NQO1. Interface residues were defined as those residues for a difference larger than 0.2 was detected between SASA calculations based on the dimer and monomer structure.

3 Results

3.1 Saturation mutagenesis by computational methods

We first used the predictive ability of evolutionary conservation analysis combined with thermodynamic stability calculations to classify all possible variants (i.e. saturation mutagenesis) in NQO1 based on their effects on the protein function and stability (Cagiada et al., 2021). For evolutionary conservation studies, we used GEMME (Laine et al., 2019) which provides a score (ΔE) for all possible single amino acid change variants of NQO1 (Supplementary Figure S1A). ΔE represents the evolutionary distance of a variant from the WT NQO1 sequence, and ΔE has been shown to be a useful predictor of the deleterious effects on function and stability of the given substitution. We used Rosetta (Park et al., 2016) to predict variant effects on thermodynamic stability (ΔΔG) using a crystal structure of NQO1 (Faig et al., 2000) as input (Supplementary Figure S1B) and subsequently calculated the median ΔΔG for all variants at each position. We performed ΔΔG evaluations using both the monomeric and dimeric structure of NQO1 to separate effects on overall stability and effects on dimerization. Specifically, we calculated ΔΔG from the monomer (Supplementary Figures S1B,S2A) to predict the change in thermodynamic stability relative to wild type of each variant. We also performed similar ΔΔG calculations using the dimer structure as input (Supplementary Figures S1C,S2B), introducing each missense variant in both chains (i.e., treating this as a homodimer) and used the resulting values to compare with experiments. Based on these two calculations, we also evaluated the ΔΔG of dimerization as the difference between the two Rosetta runs (Supplementary Figures S2C,D) to highlight which residues are involved in stabilizing the dimer and thus also those variants that might weaken dimer formation. Then, we evaluated the difference in the SASA between the dimeric and monomeric residues of NQO1 (Supplementary Figure S2C) and we classified 33 of them as interface residues. We found that for 20 of these interface residues the median ΔΔG of dimerization was >1 kcal mol^{-1}. Of these 20, 15 were stable upon mutation in the monomeric form (median ΔΔG <2 kcal mol^{-1}) and a subset of seven display a median ΔΔG of dimerization >2 kcal mol^{-1}.

Then, we combined the evolutionary conservation scores and stability calculations based on the monomeric protein for the 5,187 variants of NQO1 and plotted the results in a two-dimensional histogram (Figure 1A). We used cutoff values of 2 kcal mol^{-1} for Rosetta ΔΔG values and -3 for GEMME ΔE scores as thresholds for all the variants in order to separate them based on their effects (Luo et al., 2019). To ease analyses and interpretations, we associated each of the four defined regions with a color (Cagiada et al., 2021). 'WT-like' variants represent 48% of the available NQO1 variants (shown in green). 20% of NQO1 substitutions show high ΔΔG and high evolutionary distances and are classified as "Total-loss." These variants have substitutions that are unlikely in the evolutionary analysis (ΔE < −3) and lead to decreased stability (ΔΔG >2 kcal mol^{-1}); they thus likely compromise protein function *via* loss of protein stability (colored in red). Variants with high negative ΔE and low ΔΔG belong to the "Stable but inactive" class (colored in blue). This class contains 24% of the variants and represent those for which the evolutionary and stability analysis suggests that the protein function has been compromised, but not for stability reasons. Lastly, the remaining 8% of the variants show low stability and low evolutionary distance, and were classified as "Unstable but active" (colored in yellow). Having predicted the effects of all possible missense variants, we performed a similar classification of amino acid positions, assigning the most common variant class to each position (Figures 1C,D) and found 48% of the total positions classified as "WT-like," 25% as "-Total-loss," 22% as "Stable but inactive" and 5% as "Unstable but active."

In addition, we used the data from the dimer analysis to evaluate the number of residues involved in the stabilization of the dimer form. We found that 14 residues at the interface changed their classification to "Total-loss" if ΔΔG was evaluated using the dimer structure. Of these 14, 9 were classified as "Stable but inactive," while 5 were classified as

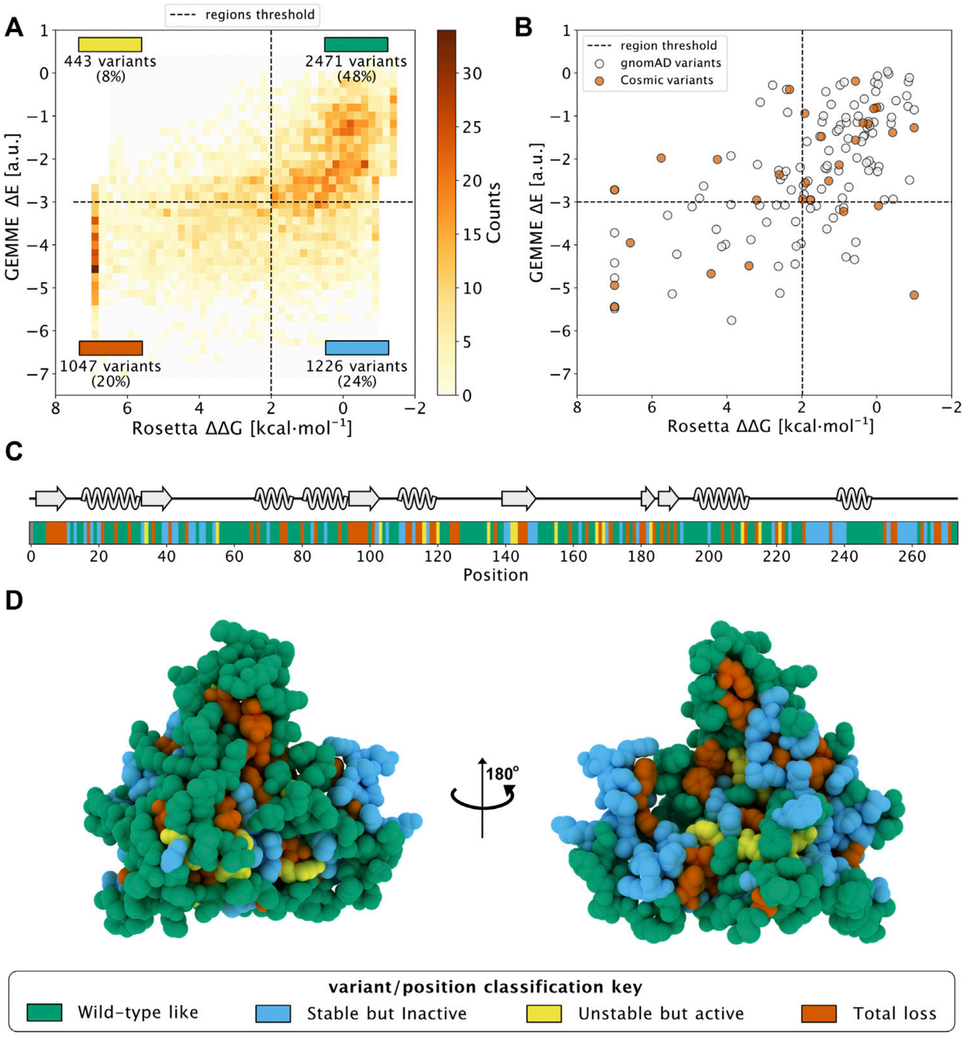


FIGURE 1
Saturation mutagenesis of NQO1 based on computational methods. **(A)** Two-dimensional histogram which combines the data from the evolutionary conservation analysis (ΔE, *y*-axis) with the thermodynamic stability (ΔΔG, *x*-axis) data from Rosetta on the NQQ1 monomer. The variants are categorized in one of the four regions, which are delimited by dashed lines. The fraction of variants, class and colour assigned to each region are indicated. The four classes of variants/positions are reported at the bottom of the figure: "WT-like" (Green), "Low stability, active" (yellow), "Stable but inactive" (blue) and "Total-loss" (red). **(B)** The scores of gnomAD (grey) and COSMIC (orange) variants in the 2D histogram. **(C)** The positional colour categories assigned using the most common colour of the position in the sequence together with the secondary structure. **(D)** The positional classification mapped to the protein crystal structure (PDB: 1D4A) (Faig et al., 2000).

WT-like using monomeric ΔΔG data. Thus, many residues at the interface appear to be conserved during evolution to preserve the stability of the dimeric form of NQO1.

Having analyzed all possible missense variants, we next looked at the results for a subset of variants that have been found in the human population. Specifically, we looked at variants that are found in the COSMIC (COSMIC v.92; https://cancer.sanger.ac.uk/cosmic) or gnomAD (gnomAD v.2.1.1; https://gnomad.broadinstitute.org/) databases, and did not find clear differences between these two sets (Figure 1B). In particular, we found variants in both sets that would be predicted as functional and others for which stability and/or conservation analyses predict loss-of-function (LoF). This result is in line with the notion that both databases may contain both potentially pathogenic as well as benign variants.

3.2 Selection of NQO1 variants to be experimentally characterized

After studying the NQO1 variants computationally, we next examined a set of the variants using a series of different

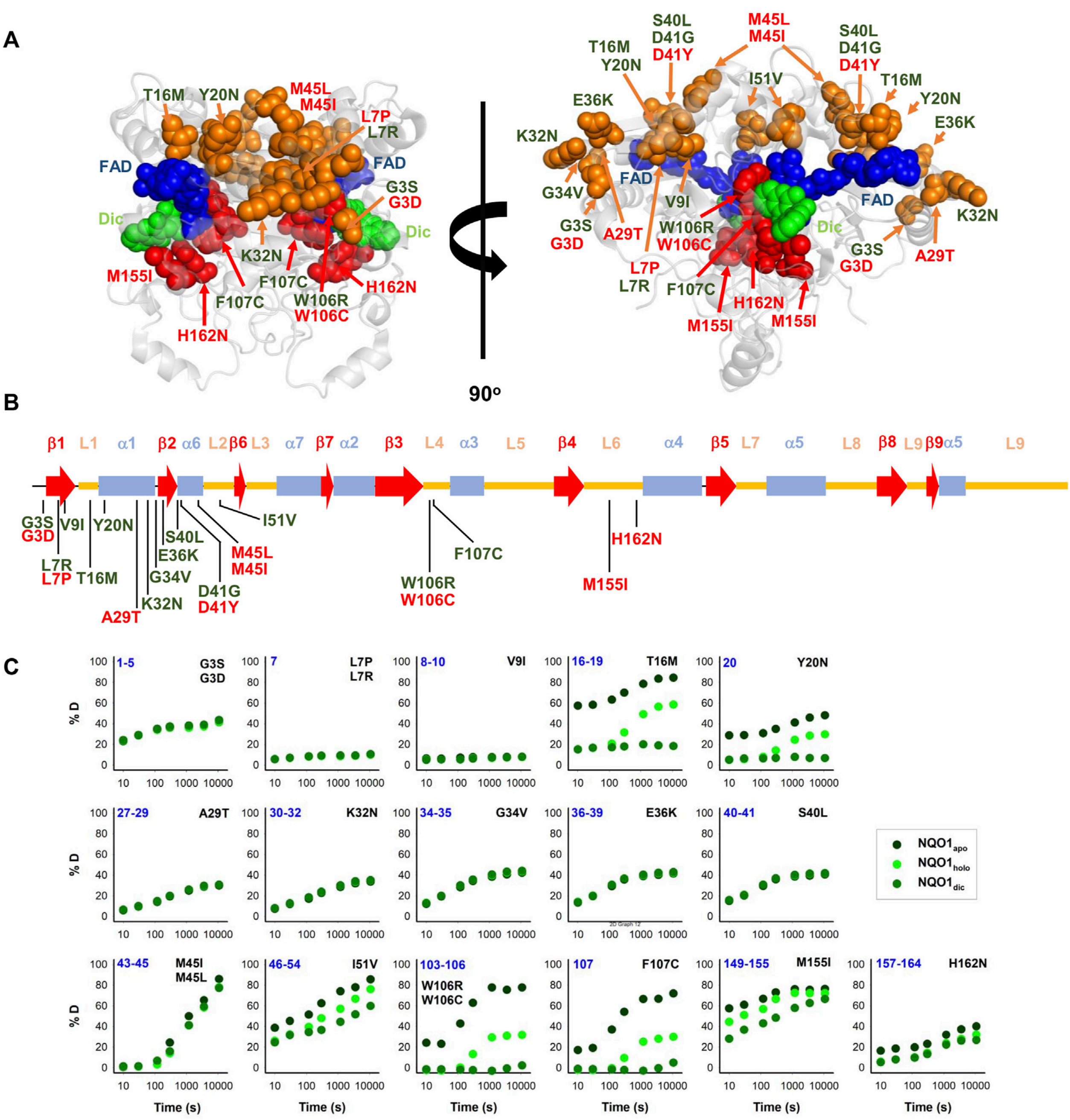


FIGURE 2
Structural features and local stability of the substituted residues and the variants characterized experimentally in this work. **(A)** The residues mutated were mapped onto the structure of NQO1 (PDB code 2F1O) (Asher et al., 2006). Residues are depicted as spheres, and the colour code indicates substitutions located in the 1–51 region (orange) or the active site (red). Variants were labelled in red (from COSMIC) or in green (from gnomAD). FAD and Dic are shown as dot representations in cyan and blue, respectively. **(B)** Location of substitutions in the sequence of NQO1 regarding secondary structure elements (from (Faig et al., 2000)). Substitutions were labelled in red (from COSMIC) or in green (from gnomAD). **(C)** HDX of segments containing the 22 variants experimentally investigated in this work. The segments are labelled in blue. The colour code corresponds to HDX for NQO1$_{apo}$, NQO1$_{holo}$ and NQO1$_{dic}$ states. HDX data are from (Vankova et al., 2019).

experiments. In this study, we have thus extended our previous work on 8 naturally-occurring variants in NQO1 (Pacheco-García et al., 2020) to a set of 22 variants (Table 1; Figure 2). Overall, this set included thirteen variants found in the gnomAD database and nine variants found in the COSMIC database. Seventeen of these variants clustered in the N-terminal part of the protein (residues 1–51), whereas five were located in the segment comprising residues 106–162 (in close proximity to the active site). Nine variants affected residues buried inside the protein structure (with less than 10% of SASA), whereas the rest are at positions that are partially or highly solvent-exposed (Table 1). The chemical nature of the changes introduced by the substitutions is also quite diverse, and the substitutions are located in different elements of secondary structure (Figure 2B). Based on our computational analysis the 22 variants represent well the heterogeneous scale of effects on NQO1 function and stability. Indeed, of the 22 variants selected 14 are classified by the computational models as "WT-like," 4 as "Total-loss" and 4 as "stable but inactive."

Results from a recent hydrogen/deuterium exchange (HDX) study on WT NQO1 (Vankova et al., 2019) enables us to evaluate the local stability of the protein segments in which these residues are found as well as the effect of FAD and dicoumarol binding (two ligands of functional and stability relevance) (Figure 2C). The L7P, L7R and V9I substitutions are located in regions with high stability that do not change upon binding of FAD or dicoumarol (Dic; a competitive inhibitor of NADH). Variants G3S, G3D, A29T, K32N, G34V, E36K, S40L and H162N are located in regions with intermediate HDX stability and whose local stability is hardly sensitive to ligand binding. Nevertheless, these positions could still, in principle, affect protein folding, stability, or solubility and indirectly affect the binding of the substrates. M45L and M45I are found in a region with low stability and not responsive to ligand binding. Y20N is located in a region with intermediate stability and where HDX shows a response to ligand binding; the remaining six variants (T16M, I51V, W106R, W106C, F107C and M155I) are in regions with low stability and also their local stability respond directly to ligand binding. This last group of variants may thus have a greater potential to affect enzymatic activity [preventing either the formation of the "stable" holo-protein, a precatalytic state, or the formation of a catalytically relevant state, the Dic state, with FAD and the inhibitor Dic bound (Anoz-Carbonell et al., 2020)]. Although this suggestion is simple, it must be noted that regions of either high or low local stability may play roles in enzyme functional and allosteric response, and that local effects can propagate far from the perturbed site (due to ligand binding or amino acid substitutions) (Luque and Freire, 2000; Luque et al., 2002; Naganathan, 2019). Therefore, we next performed an experimental characterization of variant effects on protein stability and function and compared the results with our computational analyses.

3.3 Variant effects on the expression levels, solubility and stability of NQO1

We first experimentally analysed the effect of these 22 NQO1 variants on the expression levels and solubility of the protein (upon expression in *E.coli*) as well as their effects on thermal stability (Supplementary Table S1; Figures 3, 4).

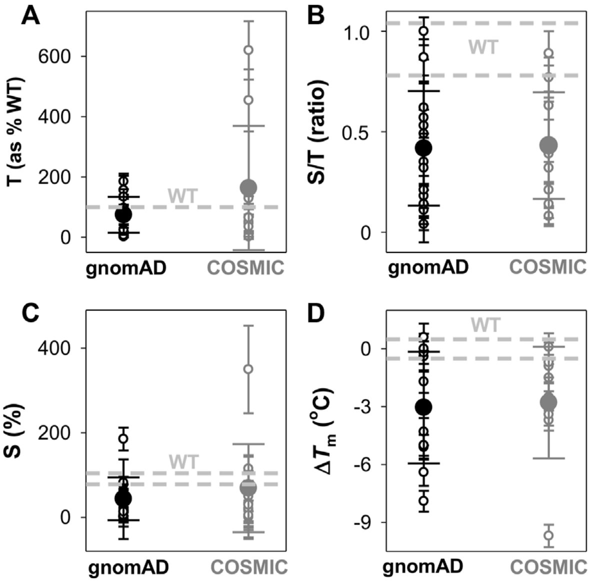


FIGURE 3
Overall effects of gnomAD and COSMIC variations on the solubility/aggregation propensity and thermal stability of the NQO1 protein. **(A–C)** Expression analyses of solubility/aggregation propensity of NQO1 variants in *E.coli* at 37°C. The total amount of NQO1 protein [T, **(A)**] as well as the ratio of soluble/total (S/T) protein **(B)** were determined from induced cells sonicated before (T) and after (S) centrifugation at 21,000 *g*. The levels of NQO1 protein were determined by western-blotting (Supplementary Figure S3). The amount of soluble protein (S) is calculated as the product of T and S/T and shown in **(C)**. Data are the mean ± s.d. from at least three independent expression experiments for each variant; **(D)** Thermal stability of NQO1 variants as holo-proteins. ΔT_m values correspond to the difference between a given variant and the WT protein. Data are the mean ± s.d. from three-six technical replicas. Small circles indicate the effects of individual variants and large circles (and corresponding errors) are those for each data set (mean ± s.d.). For reference, the values corresponding to WT NQO1 are shown in light grey. Variants are grouped in the gnomAD and COSMIC sets.

The analysis of the total (T) expression of the variants vs. WT NQO1 at 37°C (Figures 3A,B; Supplementary Figure S3; Supplementary Table S1) revealed that some variants (G3S, G3D, L7P, and V9I) showed higher total expression levels, in agreement with our previous report (Pacheco-García et al., 2020). This is likely the result of codon optimization used in the mutagenesis. Most of the variants showed relatively high expression levels, ranging from 25% to full WT levels, indicating that these variants mildly to moderately reduced total expression levels. The L7R, G34V, S40L, D41G, and D41Y variants showed extremely low expression levels, thus preventing further biophysical characterization. Interestingly, overall, no substantial differences were observed between the average effect of the gnomAD and COSMIC sets of variants.

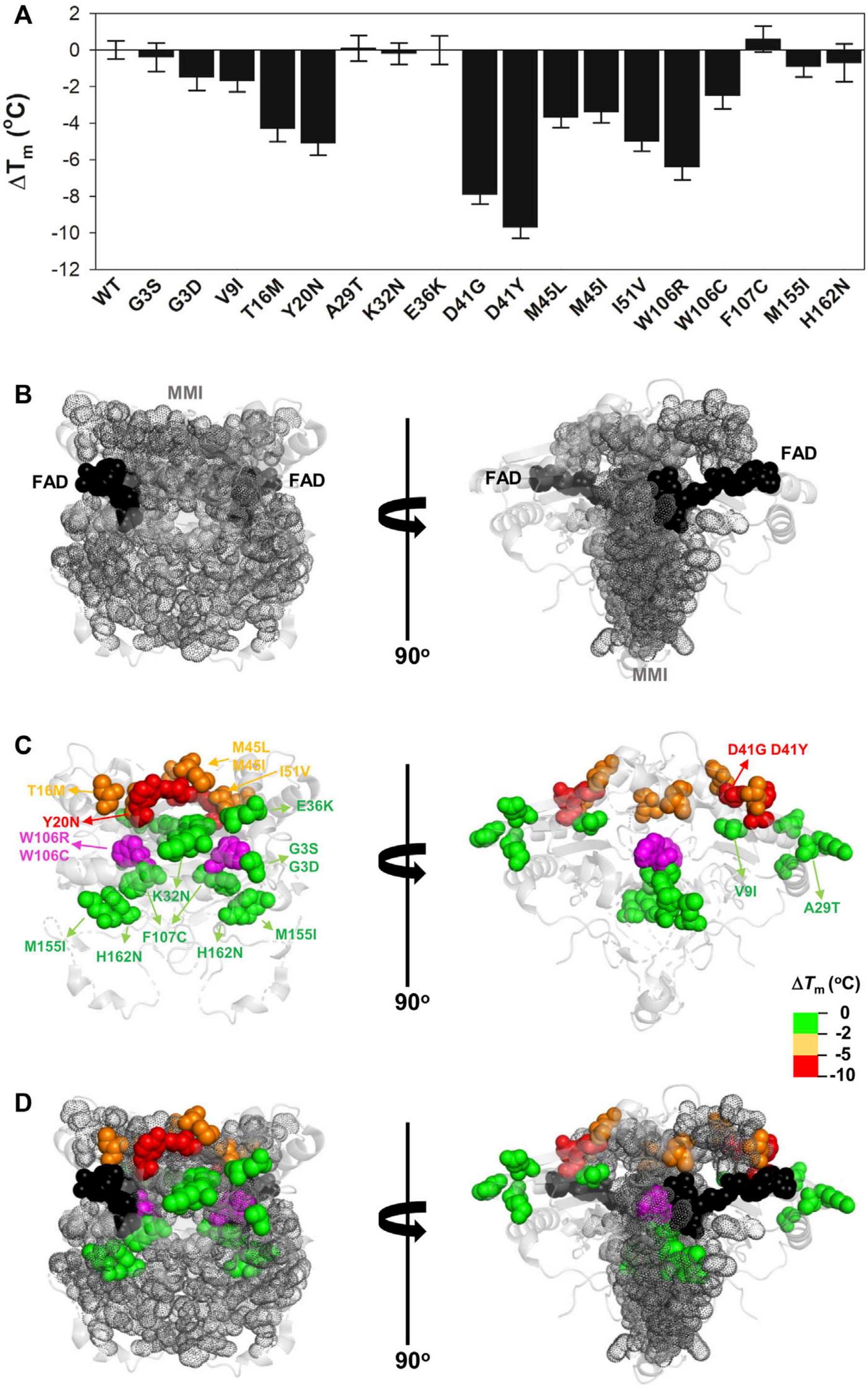


FIGURE 4
Variant effects on thermal stability related to their location near the MMI or the bound FAD. **(A)** Experimental ΔT_m values for individual variants; **(B–D)** Structural location of the FAD (black spheres) and the MMI (grey dots) **(B)** and mutated residues (colour scale according to their destabilizing effect) **(C)**. **(D)** shows an overlay of **(B,C)**. Note that two views rotated 90° are shown. The structural model used for display was PDB code 2F1O (Asher et al., 2006). The residue W106 is highlighted in magenta due to the widely different effects of the W106R/W106C substitutions.

We then determined the fraction of NQO1 existing as soluble protein (S/T ratios; Figure 3B; Supplementary Table S1). WT NQO1 showed a ratio of ~0.9 (Supplementary Table S1; Figure 3B). Again, although some variants showed much lower S/T ratios than WT NQO1, most of them showed values between 0.2 and the WT ratio. Only five variants showed lower S/T ratios than 0.15 (L7P, L7R, S40L, F107C and M155I). Expression under milder conditions (25°C) did not allow for purification of enough protein of the L7P, L7R, S40L, and G34V variants for further biophysical characterization.

We used the product of total expression levels and S/T ratios (i.e. the S values) as a proxy to evaluate the overall effect of amino acid substitutions on NQO1 solubility/aggregation propensity when expressed at 37°C (Supplementary Table S1; Figure 3C). Nine substitutions reduced the solubility below 20% of the WT protein (L7R, G34V, S40L, D41G, D41Y, M45I, W106R, F107C, and M155I).

Next, we determined the thermal stability of the remaining eighteen variants as holo-proteins (i.e. those that were expressed well as soluble proteins and were stable during purification) (Figures 3D, 4, Supplementary Table S1). Nine variants showed a thermal stability close that of the WT protein ($\Delta T_m \leq 2$°C; variants G3S, G3D, V9I, A29T, K32N, E36K, F107C, M155I, and H162N), whereas five variants decreased thermal stability by 2–5°C (variants T16M, M45L, M45I, I51V, and W106C) and four decreased the stability by 5–10°C (Y20N, D41G, D41Y, and W106R) (Supplementary Table S1). Most of the variants that destabilized the holo-protein by more than 2°C are found in the MMI or close to the FAD bound (Figure 4). Here, we note that the reported T_m and ΔT_m values are "apparent" values that cannot be regarded as reporting effects on thermodynamic stability since thermal unfolding is irreversible and kinetically-controlled (Pey et al., 2004). The W106R and W106C variants show different effects, highlighting the importance of both the location and the chemical nature of the change. The effects of the two mutations at W106 were intriguing. Both substitutions are non-conservative changes at a residue in the active site with low solvent exposure and a low structural stability with strong ligand-dependent responsiveness based on HDX studies (Vankova et al., 2019). However, their effects are very different, with W106R causing a much larger decrease in stability than W106C.

To end this section, the observed effects pinpoint that some variants in both the COSMIC and gnomAD databases decrease solubility and thermal stability of NQO1, and overall the two groups do so to similar extents (Figure 3); this observation was also seen in the computational predictions (Figure 1).

3.4 FAD binding and the stability of the FAD binding site

Sixteen out of the eighteen variants that yielded good levels of soluble proteins were prepared as apo-proteins to determine their affinity for FAD by titrations monitored by tryptophan fluorescence (Figures 5A, 6; Supplementary Figure S4; Supplementary Table S2). The D41Y and D41G variants were too unstable to obtain apo-proteins in sufficient amounts and quality.

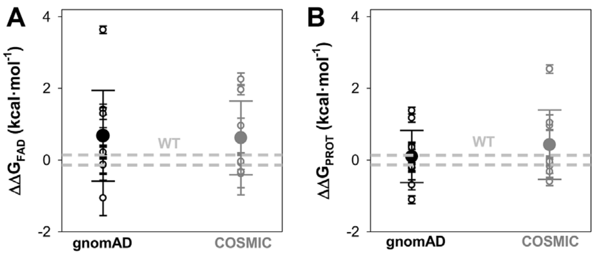


FIGURE 5
Overall effects of gnomAD and COSMIC mutations on the FAD binding affinity and the local stability of the FAD binding site (FBS). **(A)** Changes in apparent FAD binding free energies ($\Delta\Delta G_{FAD}$) calculated from the difference of binding affinity between WT and a given variant from titrations. Errors are those from linear propagation; **(B)** Local stability of the TCS (next to the FBS) from proteolysis kinetics. Changes in local stability ($\Delta\Delta G_{PROT}$) calculated from the difference of the second-order rate constant between WT and a given variant. Errors are those from linear propagation; For reference, the values corresponding to WT NQO1 are shown in light grey. Variants are grouped in the gnomAD and COSMIC sets.

The G3S, G3D, V9I, K32N, E36K, M45L, M45I, W106C, and F107C variants showed less than a 0.5 kcal mol^{-1} increase in FAD binding free energy (corresponding to less than a 2.5-fold increase in K_d). The Y20N and A29T variants showed a moderate decrease in binding affinity (between 3 and 5-fold higher K_d; thus, a change in FAD binding free energy of 0.5–1.0 kcal mol^{-1}). Five variants (T16M, I51V, W106R, M155I and H162N) markedly decreased the binding affinity for FAD (10–500-fold increase in K_d; between 1 and 3.7 kcal mol^{-1} decrease in binding free energy). Inspection of a structural model of NQO1 allows us to rationalize the effect of these substitutions due to their proximity to the bound FAD, with some exceptions. For instance, the W106R, W106C and F107C substitutions are in proximity to the FAD molecule, but have widely different effects (from ca. 500-fold lower affinity in W106R, to WT-like affinity for W106C and even *higher* affinity than WT for F107C). These results show that NQO1 responds to natural variations very differently even at the same site (i.e. two highly non-conservative variants at the site W106).

Mutational effects on FAD binding affinity ($\Delta\Delta G_{FAD}$) and proteolysis rates ($\Delta\Delta G_{PROT}$) with thermolysin have previously been shown to correlate well (Medina-Carmona et al., 2017a; Medina-Carmona et al., 2019b; Pacheco-García et al., 2020). The site for initial cleavage by thermolysin (TCS) is generally between Ser72–Val73, close to the FAD binding site (Medina-Carmona et al., 2016). All the variants investigated showed proteolysis

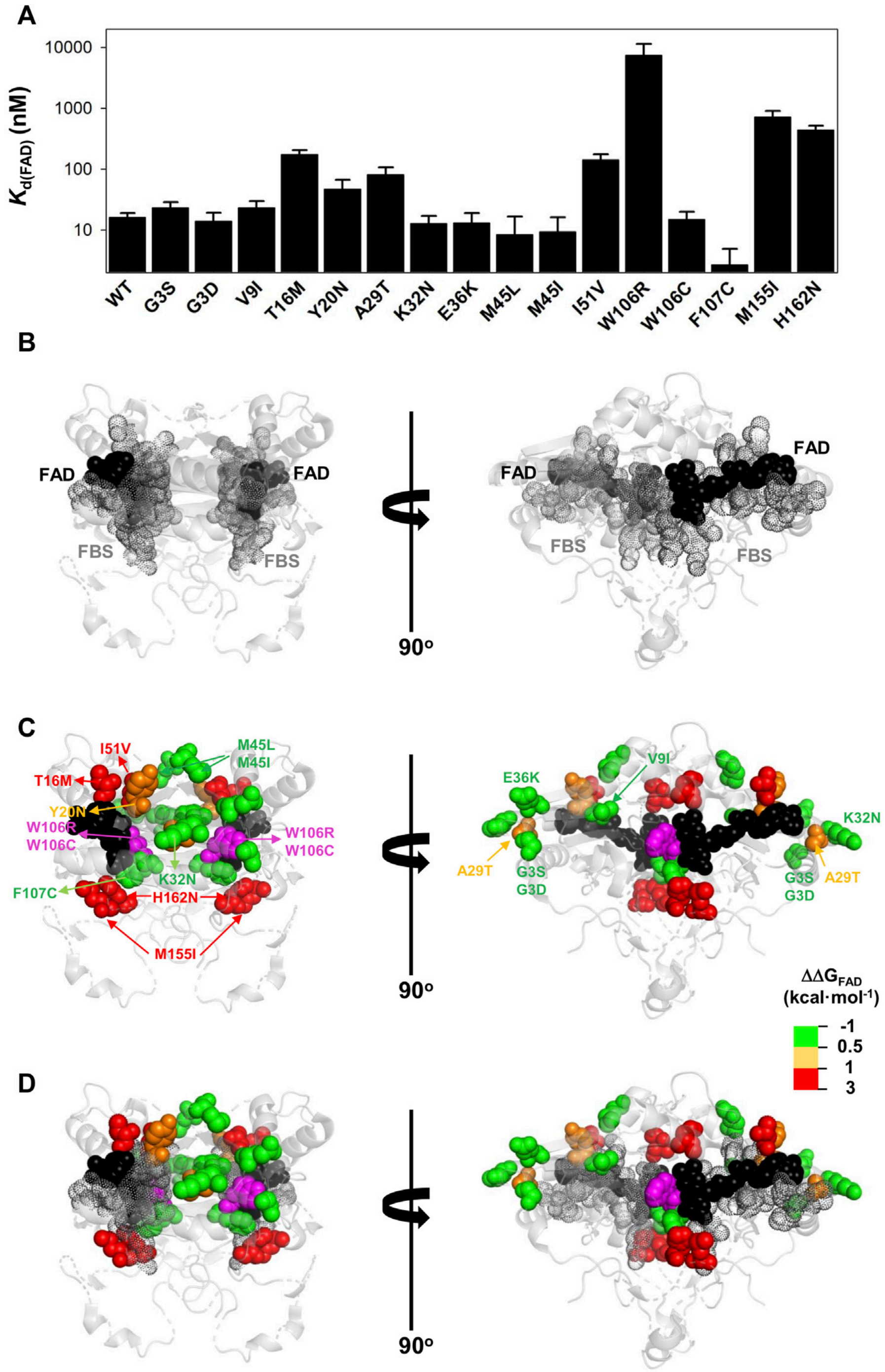


FIGURE 6
Variant effects on the FAD binding affinity. **(A)** FAD binding affinity of each variant was determined by titrations of apo-proteins. At least two independent experiments were carried out for each variant and fitted using a single-type-of-independent binding sites to obtain K_d values (note the logarithmic scale of the *y*-axis). Errors are those fittings. These K_d values were used to calculate the binding free energy difference ($\Delta\Delta G_{FAD}$) between a given variant and the WT protein (note that a positive value indicates lower affinity). **(B–D)** Structural location of the FAD (black spheres) and the FBS (grey dots) **(B)** and mutated residues (colour scale according to their destabilizing effect on FAD binding) **(C)**. The residue W106 is highlighted in magenta due to the widely different effects of the W106R/W106C substitutions. **(D)** shows an overlay of **(B,C)**. Note that two views rotated 90° are shown. The structural model used for display was PDB code 2F1O (Asher et al., 2006).

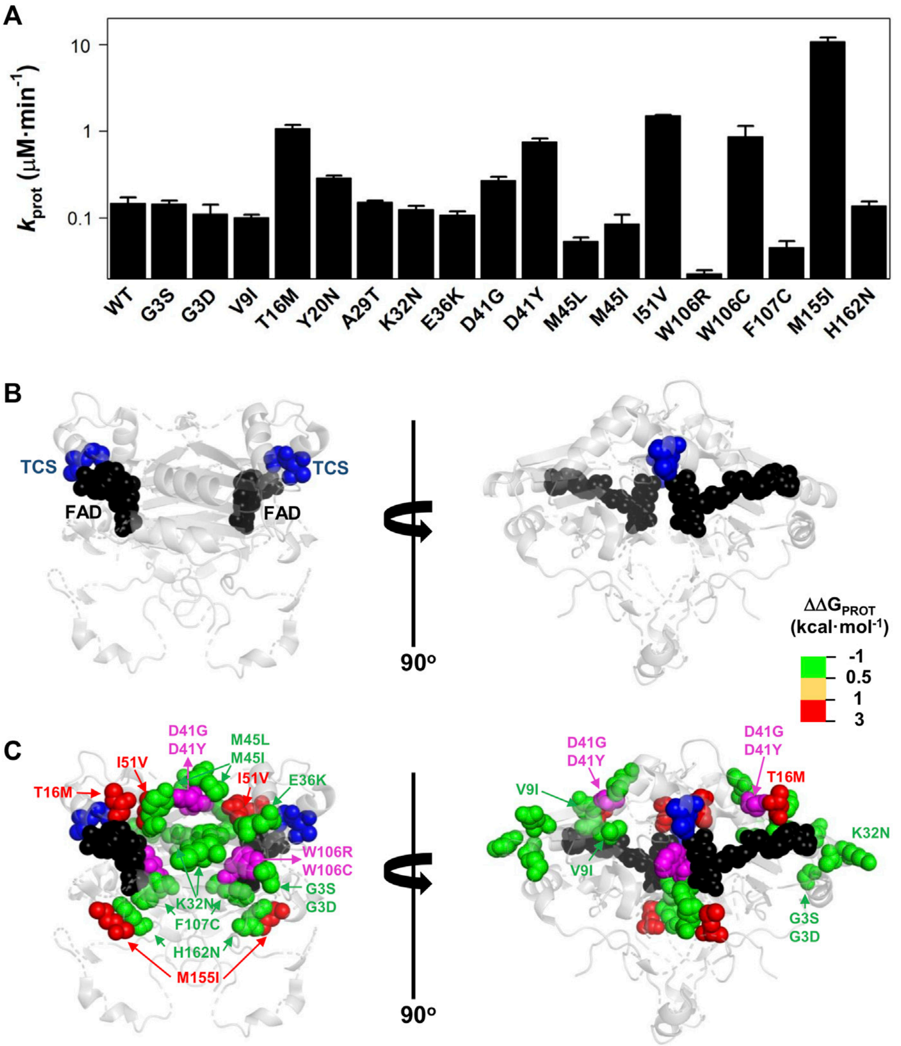


FIGURE 7
Variant effects on the local stability of the FBS from proteolysis. **(A)** Second-order rate constants for proteolysis of NQO1 variants (Supplementary Figure S5C). Errors are those fittings. These k_{PROT} values were used to calculate the TCS local energy free difference ($\Delta\Delta G_{PROT}$) between a given variant and the WT protein (note that a positive value indicates lower affinity). **(B,C)** Structural location of the FAD (black spheres) and the TCS (blue spheres) **(B)**. In **(C)**, mutated residues (colour scale according to their destabilizing effect on FAD binding) **(C)** are overlayed with TCS and FAD. The residues D41 and W106 are highlighted in magenta due to the widely different effects of the D41G/D41Y and W106R/W106C substitutions. Note that two views rotated 90° are shown. The structural model used for display was PDB code 2F1O (Asher et al., 2006).

patterns that were similar to that of WT NQO1 (Supplementary Figure S5). The previously observed correlation between $\Delta\Delta G_{FAD}$ and $\Delta\Delta G_{PROT}$ holds for the 16 variants for which both FAD binding affinity and protease sensitivity can be compared (Supplementary Figure S6). The T16M, I51V, W106C, and M155I mutations destabilized locally the TCS by 1–2.5 kcal mol^{-1} and the residues affected by these substitutions are in general close to the TCS (with the exception of M155I, the most destabilizing mutation) (Figure 7). Again, the results for the W106R/C variants were very different: both affected the local stability by ~ 1 kcal mol^{-1}, but with opposite signs (Supplementary Table S2).

Overall, the negative impact on FAD binding affinity and the stability of the FAD binding site in the holo-state was similar between variants from COSMIC and gnomAD sets (Figure 5).

3.5 Comparison of experimental analyses and computational predictions

We then proceeded to compare the experimental data to each other and to computational predictions. To ease comparison between the calculated ΔΔG values and thermal melting measurements, we first estimated $\Delta\Delta G_{melting}$ from the ΔT_m

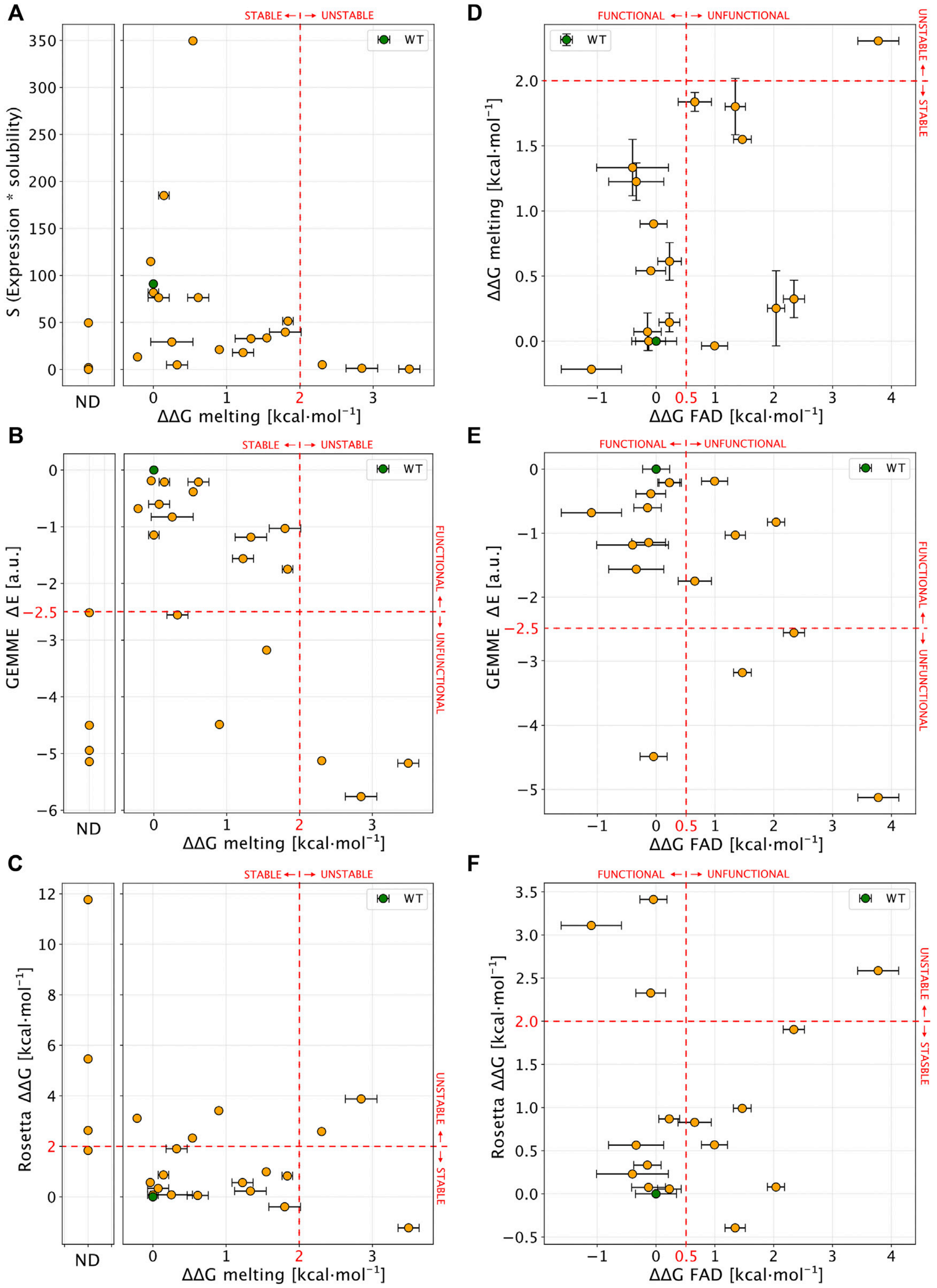


FIGURE 8
Comparison of experimental results with computational scores. **(A)** Scatter plot of $\Delta\Delta G_{melting}$ (*x*-axis) and S values (expression * solubility). Not-determined variants (ND) in thermal stability experiment are reported in a separate plot. **(B,C)** show a comparison between $\Delta\Delta G_{melting}$ and computational predictors. **(D)** Correlation between $\Delta\Delta G_{melting}$ and $\Delta\Delta G_{FAD}$ for NQO1 variants detected in both the experiments. **(E,F)** Comparison of $\Delta\Delta G_{FAD}$ with the computational results. Red lines, if present, show the boundary of experimental and computational classes for each comparison. In each panel errors are reported as a black bar on every single variant, if present.

values using an empirical relationship (Watson et al., 2018), again noting that these are not strictly experimental thermodynamic values as unfolding is not reversible either by temperature (Pey et al., 2004) or chemical denaturants (Figure S7). We first compared the experimental values of $\Delta\Delta G_{melting}$ to the levels of soluble protein (S values, Figure 8A). We found that unstable variants (those with $\Delta\Delta G_{melting} > 2$ kcal mol^{-1} or not amenable for purification, not-determined or ND) mostly showed level of S close to zero (<5) except for L7P. Stable variants (here defined as $\Delta\Delta G_{melting} < 2$ kcal mol^{-1}) instead showed a wide range of S values (76 ± 85%; mean ± s.d.).

We next compared $\Delta\Delta G_{melting}$ with the computational scores (Figures 8B,C). Overall, we found a good agreement for most of the unstable and Not-Determined (ND) variants, which showed $\Delta\Delta G$ >2.0 kcal mol^{-1} and evolutionary distance, $\Delta E < -3$ indicating predicted loss of stability and function. The only exception was D41Y which displayed a stabilizing behaviour in Rosetta $\Delta\Delta G$ predictions. Experimentally stable variants ($\Delta\Delta G_{melting} < 2$ kcal mol^{-1}) displayed low evolutionary distances ($\Delta E > -2.5$ kcal mol^{-1}) except for W106C and T16M. This observation for T16M supports the notion that detrimental effects on protein function may not be connected to thermodynamic destabilization for this variant (Pacheco-García et al., 2020).

We then compared $\Delta\Delta G_{FAD}$ with the other experimental and computational observables (Figures 8D–F). For most (15 out of 16) of the variants where $\Delta\Delta G_{FAD}$ could be measured, the $\Delta\Delta G_{melting}$ was <2 kcal mol^{-1}, indicating stable variants which was also confirmed by Rosetta $\Delta\Delta G$ calculations (13 out of 16 substitutions). Seven of the fifteen variants showed a $\Delta\Delta G_{FAD} > 0.5$ kcal mol^{-1} indicating loss of function. Of these, three variants were captured by evolutionary conservation analysis ($\Delta E < -2.5$ kcal mol^{-1}).

To summarize, for 14 out of 22 tested variants (G3S L7P L7D V9I T16M Y20N K32N G34V E36K D41G M45L M45I W106R, M155I) the predictions from the computational protocols match the experimental results, in terms of variant effects on protein stability and function.

While the results are overall encouraging, there remains differences between computation and experiments for the effects of some mutations (eight out of twenty-two; G3D, A29T, S40L, D41Y, I51V, W106C, F107C, and H162N). For five of these (G3D, S40L, D41Y, W106C, and F107C) it appears that there is a difference between the stability prediction by Rosetta and experiments (noting again that the latter are not equilibrium measurements). For the partially exposed S40L and D41Y the reason for the discrepancy is perhaps related to specific interactions made by these two residues whose effects are not captured by the Rosetta calculations. Both W106C and F107C involve substituting aromatic residues with a cysteine, suggesting problems with evaluating such substitutions. In other two cases (G3D, A29T, and H162N) the GEMME scores did not capture properly the variant effects, possibly because some specific interactions in human NQO1 may not be present in other homologs of NQO1 and thus, not captured by the evolutionary analysis. Lastly, for I51V the behaviour is opposite from both computational predictions.

4 Discussion

With advances in sequencing technologies, we are uncovering the large genetic variability in the *human genome*. To exploit the availability of this information at the clinical level, we must be able to establish genotype-phenotype correlations accurately and at a large-scale. Although detailed characterization of mutational effects is obviously useful, it is difficult to perform this at such scale (many genes, many variants). However, we may use experimental characterization on a more modest scale to test the performance of current predictive tools in the hope that we can improve them. In this work, we have carried out such an exercise for the human NQO1 protein. The rationale for selecting this system is three-fold: 1) human NQO1 is a multifunctional protein and mutational effects may affect these functions through complex mechanisms (Pacheco-Garcia et al., 2022a; Pacheco-Garcia et al., 2022b). Therefore, contrasting experimental characterization and computational predictions can be challenging for current predictive tools and may help to improve them; 2) Altered NQO1 functionality is associated with increased risk of developing cancer and neurological disorders (Salido et al., 2022). Indeed, the presence of a highly deleterious polymorphic variant in NQO1 is associated with increased cancer risk and affects multiple protein functions through allosteric effects (Lajin and Alachkar, 2013; Pacheco-Garcia et al., 2022b); 3) Over a hundred of missense variants in human NQO1 have been found in human population (i.e. the gnomAD database) or in cancer cell lines as somatic mutations (i.e. the COSMIC database). However, the impact of these mutations on NQO1 multifunctionality and their potential role in cancer development are largely unexplored.

Theoretical advances and new methodologies in the fields of sequence evolution and structure predictions allow us to perform large-scale *in silico* mutagenesis studies on target proteins. Although the current state-of-art algorithms are often [but not always (Frazer et al., 2021)] less accurate at predicting pathogenicity compared to detailed experimental testing, they provide a fast and effective way to predict LoF and sometimes to generate mechanistic hypotheses regarding which properties a variant might affect (Stein et al., 2019; Cagiada et al., 2021).

Here, we first performed *in silico* saturation mutagenesis of WT NQO1, predicting the changes in thermodynamic stability ($\Delta\Delta G$) and evolutionary conservation (ΔE) for 5,187 variants. We combined the two scores to perform a global analysis on how the NQO1 function may be perturbed. Approximately 44% of variants are predicted to cause LoF, with 45% of these drastically affecting the protein stability. This analysis enabled us to obtain an overview on the possible biologically relevant positions and variants. Indeed, although we know that the ability of computational tools to assign biological functions and predict overall pathogenicity is rapidly increasing, we are still at a point where computational methods may not predict LoF perfectly, and often do not shed much light on the mechanisms of action. This might in particular hold for proteins like NQO1 where multiple biological functions are present, and where some of which may differ between orthologues.

We then used the information provided from the *in silico* saturation mutagenesis to select 22 naturally-occurring variants with a diverse range of predicted effects on protein stability and function to be experimentally tested. We selected nine mutations found in COSMIC and thirteen from gnomAD. Of these variants, 36% severely affected protein foldability and solubility (upon expression in *E.coli*) or reduced conformational stability (at least a 5°C decrease in T_m). A quarter of the variants had severely affected FAD binding (a 5-fold decrease in affinity, i.e. a 1 kcal mol^{-1} of binding free energy penalization). For 64% of the variants, experimental characterization and computational predictions agreed in the variant effects on protein stability and function, whereas the remaining 36% of the mutations might be explained by limitations known for the tools used in the prediction process. Although, at this point, this level of agreement is reasonable, it also pinpoints the necessity of improving these predictive tools.

COSMIC mutations are in general somatic (actually, 86% of the COSMIC mutations of NQO1 are labelled as *confirmed* in this database; accessed by 17 August 2022) and likely come from samples that underwent many mutational events in different genes. Thus, the identification of a mutation in the COSMIC database does not imply that this mutation is a driver mutation [here we may define a driver mutation as a mutation with the ability to drive tumorigenesis and confer selective advantages in a tumor cell and a somatic tissue (Martínez-Jiménez et al., 2020)]. Mutations in the gnomAD database belong to heterogeneous groups (many different sequencing projects, some of them case-control studies), and likely reflect genetic variability in the *germline* and in general presence or absence in gnomAD is not sufficient to assign a label as pathogenic or benign. When we examine the NQO1 variants investigated in this work found in the gnomAD database (v.2.1), allele frequencies are overall comparable in *control* vs. *all* samples (Supplementary Table S3). This suggests that there is no strong bias towards *case* samples, and thus the allele frequencies in gnomAD may represent well their presence in a *healthy* population. The presence of these mutations in the germline may predispose somatic cells towards a new mutational event in the WT allele [as occurs in familial cancer cases (Martínez-Jiménez et al., 2020)], thus largely decreasing the NQO1 activity and function.

Our combined experimental and computational analyses provide information on the potential LoF character and the mechanisms by which the variants may exert their effects (protein stability and/or function). Due to its role in the antioxidant defense and stabilization of oncosuppressor proteins, it is likely that NQO1 play a role in cancer development. Homozygous NQO1 knock-out mice revealed cancer-associated phenotypic traits when exposed to chemical or radiological insults (Radjendirane et al., 1998; Long et al., 2000; Iskander et al., 2005; Iskander et al., 2008). Thus, the presence of LoF variants in NQO1 and increased cancer risk may resemble a recessive inheritance (Lajin and Alachkar, 2013). The p.P187S polymorphism (with an allele frequency of ~0.25, Supplementary Table S3) dramatically decreases the intracellular stability of NQO1 thus preventing its interaction with oncosuppressors, reducing enzyme activity and affecting almost the entire structure of NQO1 (Pacheco-Garcia et al., 2021). Noteworthy, it only associates with cancer in homozygotes (Lajin and Alachkar, 2013). Due to the low frequency of most gnomAD NQO1 variants, their presence would be rare even in compound heterozygotes. In fact, 98% of the homozygous samples containing NQO1 missense variations correspond to homozygotes for P187S. However, an additional (*somatic*) mutational event in a WT/P187S genetic background (about 25% of human population) may substantially enhance the LOF phenotype in this common genetic background.

To conclude, we present a test of predictive tools against the experimental characterization of large set of naturally-occurring mutations on NQO1 stability and function. Further steps will be taken to provide a wider perspective on the multifunctionality of NQO1 (i.e. intracellular degradation and stability, high-resolution structural stability in different ligation states, enzyme function and cooperativity, interaction with protein partners, allosteric communication of mutational effects) and the relationships between the genetic diversity of NQO1 in human population and its link with individual propensity towards disease development.

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

AP conceived the project and supervised the experimental work; JP-G and KT-M carried out expression, purification and characterization of proteins; MC carried out computational analysis; ES contributed to selection of variants and contributed with reagents; KL-L supervised the computational work; KL-L and AP received funding; M.C, KL-L and AP wrote the original draft; All authors contributed to and approved the final version of the manuscript.

Funding

This work was supported by the ERDF/Spanish Ministry of Science, Innovation and Universities—State Research Agency (Grant Number. RTI 2018-096246-B-I00), Consejería de Economía, Conocimiento, Empresas y Universidad, Junta de Andalucía (Grant Number. P18-RT-2413), ERDF/Counseling of Economic transformation, Industry, Knowledge and Universities (Grant B-BIO-84-UGR20) and Comunidad Valenciana (Grant Number. CIAICO/2021/135). This work is a contribution from the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (to KL-L; NNF18OC0033950).

Acknowledgments

MC and KL-L acknowledge access to computational resources the Biocomputing Core Facility at the Department of Biology, University of Copenhagen.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1063620/full#supplementary-material

References

Abildgaard, A. B., Stein, A., Nielsen, S. V., Schultz-Knudsen, K., Papaleo, E., Shrikhande, A., et al. (2019). Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. *Elife* 8, e49138. doi:10.7554/eLife.49138

Anoz-Carbonell, E., Timson, D. J., Pey, A. L., and Medina, M. (2020). The catalytic cycle of the antioxidant and cancer-associated human NQO1 enzyme: Hydride transfer, conformational dynamics and functional cooperativity. *Antioxidants* 9, 772. doi:10.3390/antiox9090772

Arnedo-Pac, C., Lopez-Bigas, N., and Muiños, F. (2022). Predicting disease variants using biodiversity and machine learning. *Nat. Biotechnol.* 40, 27–28. doi:10.1038/s41587-021-01187-w

Asher, G., Dym, O., Tsvetkov, P., Adler, J., and Shaul, Y. (2006). The crystal structure of NAD(P)H quinone oxidoreductase 1 in complex with its potent inhibitor dicoumarol. *Biochemistry* 45, 6372–6378. doi:10.1021/bi0600087

Asher, G., Tsvetkov, P., Kahana, C., and Shaul, Y. (2005). A mechanism of ubiquitin-independent proteasomal degradation of the tumor suppressors p53 and p73. *Genes Dev.* 19, 316–321. doi:10.1101/gad.319905

Beaver, S. K., Mesa-Torres, N., Pey, A. L., and Timson, D. J. (2019). NQO1: A target for the treatment of cancer and neurological diseases, and a model to understand loss of function disease mechanisms. *Biochim. Biophys. Acta. Proteins Proteom.* 1867, 663–676. doi:10.1016/j.bbapap.2019.05.002

Ben-Nissan, G., and Sharon, M. (2014). Regulating the 20S proteasome ubiquitin-independent degradation pathway. *Biomolecules* 4, 862–884. doi:10.3390/biom4030862

Cagiada, M., Johansson, K. E., Valanciute, A., Nielsen, S. V., Hartmann-Petersen, R., Yang, J. J., et al. (2021). Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol. Biol. Evol.* 38, 3235–3246. doi:10.1093/molbev/msab095

di Francesco, A., di Germanio, C., Panda, A. C., Huynh, P., Peaden, R., Navas-Enamorado, I., et al. (2016). Novel RNA-binding activity of NQO1 promotes SERPINA1 mRNA translation. *Free Radic. Biol. Med.* 99, 225–233. doi:10.1016/j.freeradbiomed.2016.08.005

Faig, M., Bianchet, M. A., Talalay, P., Chen, S., Winski, S., Ross, D., et al. (2000). Structures of recombinant human and mouse NAD(P)H:quinone oxidoreductases: Species comparison and structural changes with substrate binding and release. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3177–3182. doi:10.1073/pnas.050585797

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., et al. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature* 599, 91–95. doi:10.1038/s41586-021-04043-8

Frenz, B., Lewis, S. M., King, I., DiMaio, F., Park, H., and Song, Y. (2020). Prediction of protein mutational free energy: Benchmark and sampling improvements increase classification accuracy. *Front. Bioeng. Biotechnol.* 8, 558247. doi:10.3389/fbioe.2020.558247

Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A., and Lindorff-Larsen, K. (2022). Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* 38, 110207. doi:10.1016/j.celrep.2021.110207

Iskander, K., Barrios, R. J., and Jaiswal, A. K. (2008). Disruption of NAD(P)H:quinone oxidoreductase 1 gene in mice leads to radiation-induced myeloproliferative disease. *Cancer Res.* 68, 7915–7922. doi:10.1158/0008-5472.CAN-08-0766

Iskander, K., Gaikwad, A., Paquet, M., Long, D. J., Brayton, C., Barrios, R., et al. (2005). Lower induction of p53 and decreased apoptosis in NQO1-null mice lead to increased sensitivity to chemical-induced skin carcinogenesis. *Cancer Res.* 65, 2054–2058. doi:10.1158/0008-5472.CAN-04-3157

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

Katsonis, P., Wilhelm, K., Williams, A., and Lichtarge, O. (2022). Genome interpretation using *in silico* predictors of variant impact. *Hum. Genet.* 141, 1549–1577. doi:10.1007/s00439-022-02457-6

Laine, E., Karami, Y., and Carbone, A. (2019). Gemme: A simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.* 36, 2604–2619. doi:10.1093/molbev/msz179

Lajin, B., and Alachkar, A. (2013). The NQO1 polymorphism C609T (Pro187Ser) and cancer susceptibility: A comprehensive meta-analysis. *Br. J. Cancer* 109, 1325–1337. doi:10.1038/bjc.2013.357

Li, R., Bianchet, M. A., Talalayt, P., and Amzel, L. M. (1995). The three-dimensional structure of NAD(P)H:quinone reductase, a flavoprotein involved in cancer chemoprotection and chemotherapy: Mechanism of the two-electron reduction. *Proc. Natl. Acad. Sci. U. S. A.* 92, 8846–8850. doi:10.1073/pnas.92.19.8846

Lienhart, W. D., Gudipati, V., Uhl, M. K., Binter, A., Pulido, S. A., Saf, R., et al. (2014). Collapse of the native structure caused by a single amino acid exchange in human NAD(P)H:Quinone oxidoreductase. *FEBS J.* 281, 4691–4704. doi:10.1111/febs.12975

Long, D. J., Waikel, R. L., Wang, X. J., Perlaky, L., Roop, D. R., and Jaiswal, A. K. (2000). NAD(P)H:quinone oxidoreductase 1 deficiency increases susceptibility to benzo(a)pyrene-induced mouse skin carcinogenesis. *Cancer Res.* 60, 5913–5915.

Luo, S., Kang, S. S., Wang, Z. H., Liu, X., Day, J. X., Wu, Z., et al. (2019). Akt phosphorylates NQO1 and triggers its degradation, abolishing its antioxidative activities in Parkinson's disease. *J. Neurosci.* 39, 7291–7305. doi:10.1523/JNEUROSCI.0625-19.2019

Luque, I., and Freire, E. (2000). Structural stability of binding sites: Consequences for binding affinity and allosteric effects. *Proteins* 4, 63–71. doi:10.1002/1097-0134(2000)41:4+<63::aid-prot60>3.3.co;2-y

Luque, I., Leavitt, S. A., and Freire, E. (2002). The linkage between protein folding and functional cooperativity: Two sides of the same coin? *Annu. Rev. Biophys. Biomol. Struct.* 31, 235–256. doi:10.1146/annurev.biophys.31.082901.134215

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* 20, 555–572. doi:10.1038/s41568-020-0290-x

Martínez-Limón, A., Alriquet, M., Lang, W. H., Calloni, G., Wittig, I., and Vabulas, R. M. (2016). Recognition of enzymes lacking bound cofactor by Protein quality control. *Proc. Natl. Acad. Sci. U. S. A.* 113, 12156–12161. doi:10.1073/pnas.1611994113

Martínez-Limón, A., Calloni, G., Ernst, R., and Vabulas, R. M. (2020). Flavin dependency undermines proteome stability, lipid metabolism and cellular proliferation during vitamin B2 deficiency. *Cell Death Dis.* 11, 725. doi:10.1038/s41419-020-02929-5

McInnes, G., Sharo, A. G., Koleske, M. L., Brown, J. E. H., Norstad, M., Adhikari, A. N., et al. (2021). Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.* 108, 535–548. doi:10.1016/j.ajhg.2021.03.003

Medina-Carmona, E., Betancor-Fernández, I., Santos, J., Mesa-Torres, N., Grottelli, S., Batlle, C., et al. (2019). Insight into the specificity and severity of pathogenic mechanisms associated with missense mutations through experimental and structural perturbation analyses. *Hum. Mol. Genet.* 28, 1–15. doi:10.1093/hmg/ddy323

Medina-Carmona, E., Fuchs, J. E., Gavira, J. A., Mesa-Torres, N., Neira, J. L., Salido, E., et al. (2017). Enhanced vulnerability of human proteins towards disease-associated inactivation through divergent evolution. *Hum. Mol. Genet.* 26, 3531–3544. doi:10.1093/hmg/ddx238

Medina-Carmona, E., Neira, J. L., Salido, E., Fuchs, J. E., Palomino-Morales, R., Timson, D. J., et al. (2017). Site-to-site interdomain communication may mediate different loss-of-function mechanisms in a cancer-associated NQO1 polymorphism. *Sci. Rep.* 7, 44532. doi:10.1038/srep44532

Medina-Carmona, E., Palomino-Morales, R. J., Fuchs, J. E., Padín-Gonzalez, E., Mesa-Torres, N., Salido, E., et al. (2016). Erratum: Conformational dynamics is key to understanding loss-of-function of NQO1 cancer-associated polymorphisms and its correction by pharmacological ligands. *Sci. Rep.* 6 (1), 21939. doi:10.1038/srep21939

Medina-Carmona, E., Rizzuti, B., Martín-Escolano, R., Pacheco-García, J. L., Mesa-Torres, N., Neira, J. L., et al. (2019). Phosphorylation compromises FAD binding and intracellular stability of wild-type and cancer-associated NQO1: Insights into flavo-proteome stability. *Int. J. Biol. Macromol.* 125, 1275–1288. doi:10.1016/j.ijbiomac.2018.09.108

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

Oh, E. T., Kim, J. W., Kim, J. M., Kim, S. J., Lee, J. S., Hong, S. S., et al. (2016). NQO1 inhibits proteasome-mediated degradation of HIF-1α. *Nat. Commun.* 7, 13593. doi:10.1038/ncomms13593

Pacheco-Garcia, J. L., Anoz-Carbonell, E., Loginov, D. S., Vankova, P., Salido, E., Man, P., et al. (2022). Different phenotypic outcome due to site-specific phosphorylation in the cancer-associated NQO1 enzyme studied by phosphomimetic mutations. *Arch. Biochem. Biophys.* 729, 109392. doi:10.1016/j.abb.2022.109392

Pacheco-Garcia, J. L., Anoz-Carbonell, E., Vankova, P., Kannan, A., Palomino-Morales, R., Mesa-Torres, N., et al. (2021). Structural basis of the pleiotropic and specific phenotypic consequences of missense mutations in the multifunctional NAD(P)H:quinone oxidoreductase 1 and their pharmacological rescue. *Redox Biol.* 46, 102112. doi:10.1016/j.redox.2021.102112

Pacheco-García, J. L., Cano-Muñoz, M., Sánchez-Ramos, I., Salido, E., and Pey, A. L. (2020). Naturally-occurring rare mutations cause mild to catastrophic effects in the multifunctional and cancer-associated NQO1 protein. *J. Pers. Med.* 10, E207–E231. doi:10.3390/jpm10040207

Pacheco-Garcia, J. L., Loginov, D. S., Anoz-Carbonell, E., Vankova, P., Palomino-Morales, R., Salido, E., et al. (2022). Allosteric communication in the multifunctional and redox NQO1 protein studied by cavity-making mutations. *Antioxidants* 11, 1110. doi:10.3390/antiox11061110

Park, C., and Marqusee, S. (2004). Probing the high energy states in proteins by proteolysis. *J. Mol. Biol.* 343, 1467–1476. doi:10.1016/j.jmb.2004.08.085

Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., et al. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212. doi:10.1021/acs.jctc.6b00819

Pey, A. L. (2018). Biophysical and functional perturbation analyses at cancer-associated P187 and K240 sites of the multifunctional NADP(H):quinone oxidoreductase 1. *Int. J. Biol. Macromol.* 118, 1912–1923. doi:10.1016/j.ijbiomac.2018.07.051

Pey, A. L., Megarity, C. F., and Timson, D. J. (2004). FAD binding overcomes defects in activity and stability displayed by cancer-associated variants of human NQO1. *Biochim. Biophys. Acta* 1842, 2163–2173. doi:10.1016/j.bbadis.2014.08.011

Radjendirane, V., Joseph, P., Lee, Y. H., Kimura, S., Klein-Szanto, A. J., Gonzalez, F. J., et al. (1998). Disruption of the DT diaphorase (NQO1) gene in mice leads to increased menadione toxicity. *J. Biol. Chem.* 273, 7382–7389. doi:10.1074/jbc.273.13.7382

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175. doi:10.1038/nmeth.1818

Ross, D., and Siegel, D. (2018). NQO1 in protection against oxidative stress. *Curr. Opin. Toxicol.* 7, 67–72. doi:10.1016/j.cotox.2017.10.005

Salido, E., Timson, D. J., Betancor-Fernández, I., Palomino-Morales, R., Anoz-Carbonell, E., Pacheco-García, J. L., et al. (2022). Targeting HIF-1α function in cancer through the chaperone action of NQO1: Implications of genetic diversity of NQO1. *J. Pers. Med.* 12, 747. doi:10.3390/jpm12050747

Siegel, D., Bersie, S., Harris, P., di Francesco, A., Armstrong, M., Reisdorph, N., et al. (2021). A redox-mediated conformational change in NQO1 controls binding to microtubules and α-tubulin acetylation. *Redox Biol.* 39, 101840. doi:10.1016/j.redox.2020.101840

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, 575–588. doi:10.1016/j.tibs.2019.01.003

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinforma.* 20, 473. doi:10.1186/s12859-019-3019-7

Vankova, P., Salido, E., Timson, D. J., Man, P., and Pey, A. L. (2019). A dynamic core in human NQO1 controls the functional and stability effects of ligand binding and their communication across the enzyme dimer. *Biomolecules* 9, 728. doi:10.3390/biom9110728

Watson, M. D., Monroe, J., and Raleigh, D. P. (2018). Size-dependent relationships between protein stability and thermal unfolding temperature have important implications for analysis of protein energetics and high-throughput assays of protein-ligand interactions. *J. Phys. Chem. B* 122, 5278–5285. doi:10.1021/acs.jpcb.7b05684

Xu, Q., Tang, Q., Katsonis, P., Lichtarge, O., Jones, D., Bovo, S., et al. (2017). Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. *Hum. Mutat.* 38, 1123–1131. doi:10.1002/humu.23222

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fmolb.2022.1075570

OPEN ACCESS

EDITED BY
Laurent Roberto Chiarelli,
University of Pavia, Italy

REVIEWED BY
Julie Carol Mitchell,
Oak Ridge National Laboratory (DOE),
United States
Debsindhu Bhowmik,
Oak Ridge National Laboratory (DOE),
United States

*CORRESPONDENCE
Piero Fariselli,
✉ piero.fariselli@unito.it

SPECIALTY SECTION
This article was submitted to Protein Biochemistry for Basic and Applied Sciences, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 20 October 2022
ACCEPTED 15 December 2022
PUBLISHED 05 January 2023

CITATION
Benevenuta S, Birolo G, Sanavia T, Capriotti E and Fariselli P (2023), Challenges in predicting stabilizing variations: An exploration. *Front. Mol. Biosci.* 9:1075570.
doi: 10.3389/fmolb.2022.1075570

Challenges in predicting stabilizing variations: An exploration

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

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

An open challenge of computational and experimental biology is understanding the impact of non-synonymous DNA variations on protein function and, subsequently, human health. The effects of these variants on protein stability can be measured as the difference in the free energy of unfolding ($\Delta\Delta G$) between the mutated structure of the protein and its wild-type form. Throughout the years, bioinformaticians have developed a wide variety of tools and approaches to predict the $\Delta\Delta G$. Although the performance of these tools is highly variable, overall they are less accurate in predicting $\Delta\Delta G$ stabilizing variations rather than the destabilizing ones. Here, we analyze the possible reasons for this difference by focusing on the relationship between experimentally-measured $\Delta\Delta G$ and seven protein properties on three widely-used datasets (S2648, VariBench, Ssym) and a recently introduced one (S669). These properties include protein structural information, different physical properties and statistical potentials. We found that two highly used input features, i.e., hydrophobicity and the Blosum62 substitution matrix, show a performance close to random choice when trying to separate stabilizing variants from either neutral or destabilizing ones. We then speculate that, since destabilizing variations are the most abundant class in the available datasets, the overall performance of the methods is higher when including features that improve the prediction for the destabilizing variants at the expense of the stabilizing ones. These findings highlight the need of designing predictive methods able to exploit also input features highly correlated with the stabilizing variants. New tools should also be tested on a not-artificially balanced dataset, reporting the performance on all the three classes (i.e., stabilizing, neutral and destabilizing variants) and not only the overall results.

KEYWORDS

protein stability, single-point mutation, stability predictors, machine learning, stabilizing variants

1 Introduction

Non-synonymous DNA variations can affect the stability of the protein structure, jeopardising its function with potential pathogenic outcomes (Casadio et al., 2011; Yue et al., 2005; Hartl, 2017; Martelli et al., 2016; Cheng et al., 2008; Compiani and Capriotti, 2013; Birolo et al., 2021). For this reason, the impact of these variations on the protein structure and its stability is a widely studied problem, even though its *in silico* prediction is still challenging for bioinformaticians.

The effects of non-synonymous variants on the protein stability are usually expressed as the difference in the Gibbs free energy of unfolding ($\Delta\Delta G$), measured in kcal/mol and defined as the difference between the unfolding free energy of the mutated structure (M) of the protein and its wild-type form (W):

$$\Delta\Delta G_{WM} = \Delta G_M - \Delta G_W. \tag{1}$$

With this notation, destabilizing variants are associated to a negative $\Delta\Delta G$, while this value is positive for the stabilizing ones. For those mutations showing $\Delta\Delta G$ values close to zero, the significant experimental uncertainties (Montanucci et al., 2019b; Benevenuta and Fariselli, 2019) make their $\Delta\Delta G$ signs less reliable. At the same time, these $\Delta\Delta G$ values indicate a minimal variation of folding stability of the variants. To account for this issue, we consider an intermediate class of variants named "neutral", whose $\Delta\Delta G$ values are in the range between -0.5 and 0.5 kcal/mol. The choice of 0.5 kcal/mol is based on the average experimental error, as reported in Capriotti et al. (2008).

Over the years, several computational tools have been developed to predict the $\Delta\Delta G$, combining machine learning methods, statistical potential energy, physico-chemical properties, sequence features and evolutionary information (Benevenuta et al., 2021; Pancotti et al., 2021; Montanucci et al., 2019a; Pires et al., 2014b; Worth et al., 2011; Samaga et al., 2021; Pires et al., 2014a; Rodrigues et al., 2018; Rodrigues et al., 2021; Schymkowitz et al., 2005; Li et al., 2021; Cheng et al., 2006; Kellogg et al., 2011; Capriotti et al., 2005; Li et al., 2020; Chen et al., 2020; Dehouck et al., 2011; Laimer et al., 2016; Savojardo et al., 2016; Savojardo et al., 2019). As highlighted in Sanavia et al. (2020), most of these tools still provide over-optimistic performance due to sequence similarity between the proteins used in the training and test sets. To provide a more realistic estimate of their performance, we recently compared the predictive ability of 18 popular tools on a novel manually-curated dataset in Pancotti et al. (2022). This dataset, named S669, was extracted from ThermoMutDB (Xavier et al., 2021) and it contains only variants belonging to proteins having less than 25% sequence identity with those of S2648 (Pires et al. (2016)) and VariBench (Nair and Vihinen (2013)), two datasets on which most of the state-of-the-art methods were trained. Our analysis underlined that, across all the methods, the performance is more accurate in predicting destabilizing variants rather than the stabilizing ones.

As possible solutions to this issue, methods had accounted for unbalanced training datasets by "artificially balancing" it or by exploiting the antisimmetry property, which is a relationship imposed on $\Delta\Delta G$ values by thermodynamics. Specifically, given the wild-type W and the mutated M protein structures, the folding free energy $\Delta\Delta G_{WM}$ from W to M is equal and has the opposite sign of the folding free energy $\Delta\Delta G_{MW}$ from M to W, considering identical experimental conditions:

$$\Delta\Delta G_{WM} = \Delta G_M - \Delta G_W = -(\Delta G_W - \Delta G_M) = -\Delta\Delta G_{MW}. \quad (2)$$

Using this property, any unbalanced training dataset can be "artificially balanced" by introducing the reverse variants, which are substitutions created from the experimentally-measured variants, henceforth named "direct". Considering the mutation from W to M as the "direct" variant, its "reverse" is simply defined as:

$$M \rightarrow W, \text{with } \Delta\Delta G_{MW} = -\Delta\Delta G_{WM}. \quad (3)$$

By artificially balancing the training dataset or by enforcing the antisymmetric property in the model itself, a method should theoretically be able to predict stabilizing variations with the same accuracy as the destabilizing ones. This statement has been already verified on S669 when considering both direct and reverse variants (Pancotti et al., 2022). Here we showed that, when considering only the direct variants, the performance is highly unbalanced with the direct stabilizing variants, which were predicted much worse than the direct destabilizing ones. Investigating the reasons for this discrepancy is the main objective of the present work.

For most methods, this imbalance could be due to any reason concerning the tool architecture or to the training phase. It is difficult, in these cases, to isolate where the problem lies. On the other hand, for an untrained method such as DDGun3D Montanucci et al. (2019a), whose prediction is a linear combination of its features, this issue can only arise from the features themselves.

For these reasons, we decided to study the impact on the predictions of the following properties of residue substitutions: the difference in hydrophobicity and volume, the logarithm of the conservation ratio, the Blosum62 evolutionary score, the relative solvent accessibility and the Skolnick and Bastolla-Vendruscolo potentials. These are the most common features considered by the state-of-the-art methods and they include the DDGun3D input features.

In this study, we showed that some of these commonly-used features are only useful to predict destabilizing variants and unhelpful for the stabilizing ones. Our findings highlight an intrinsic difference between these two classes, suggesting the importance of using different properties for the stabilizing variants in order to develop methods with more consistent performance between variant classes.

2 Materials and methods

2.1 Structural information, physical properties and statistical potentials

We considered seven different features that include the DDGun3D inputs and two more properties (conservation and volume of the amino acids involved). All these features are not specific only to DDGun3D, but they are commonly employed by $\Delta\Delta G$ predictors. We analyzed:

1) two physical properties:

- **Difference in hydrophobicity**: the difference in hydrophobic regions between wild-type and mutant residues according to the Kyte-Doolittle scale (Kyte and Doolittle, 1982);
- **Volume difference**: the volume difference between the wild-type and the mutated residue measured in $\mathring{A}^3$ (Zamyatnin, 1972).

2) three features based on conservation and structural information:

- **Logarithm of the conservation ratio**, defined as $\log(\frac{CONS_W+\epsilon}{CONS_M+\epsilon})$, with $\epsilon = 0.01$ to avoid invalid results when any of the conservation frequencies were equal to zero. If the mutation changes an amino acid into a less conserved one, the logarithm value is < 0;
- **Blosum evolutionary score**: the difference between the wild-type and mutant residues in the Blosum62 substitution matrix, $B(W, M)$ (Henikoff and Henikoff, 1992);
- **Relative Solvent Accessibility**: a measure of the extent of burial or exposure of the residue in the 3D protein structure, ranging from 0 for completely buried to one for completely exposed. It was computed through the DSSP program (Kabsch and Sander, 1983; Touw et al., 2015).

TABLE 1 Datasets composition. The variants are grouped according to their $\Delta\Delta G$ values into three classes: destabilizing ($\Delta\Delta G \leq -0.5$ kcal/mol), neutral ($|\Delta\Delta G| <$ 0.5 kcal/mol) and stabilizing ($\Delta\Delta G \geq 0.5$ kcal/mol). The corresponding percentages are reported into brackets.

	Destabilizing	Neutral	Stabilizing
S2648	1,597 (60%)	755 (29%)	295 (11%)
S669	387 (58%)	195 (29%)	85 (13%)
Ssym	225 (33%)	234 (34%)	225 (33%)
VariBench	800 (56%)	426 (30%)	194 (14%)
S4428	2,461 (55%)	1,311 (30%)	656 (15%)

3) two features based on the statistical potentials:

- **Skolnick potential**: the difference in the interaction energy (measured through the Skolnick et al. (1997) statistical potential) between the wild-type and substituted residues with their sequence neighbours within a 2-residue window

$$\sum_{i=-2}^{i=2}(sk(W,a_i)-sk(M,a_i));$$

- **Bastolla-Vendruscolo potential**: the difference in the interaction energy, measured as the Bastolla statistical potential (Bastolla et al., 2001) between the wild-type and mutant residues with its structural neighbours,

$$\sum_{i\in I}(bv(W,a_i)-bv(M,a_i));$$

where I is the set of amino acid residues in the structural neighbourhood of radius $5\mathring{A}$ around the substituted position.

2.2 Datasets

We divided the analysis in two parts. Firstly, we studied the abilities to predict the stabilizing variations of 18 protein stability predictors on:

- **S669** (Pancotti et al., 2022) a recent manually-curated dataset extracted from ThermoMutDB (Xavier et al., 2021) whose variants belong to proteins having less than 25% sequence identity with those of S2648 and VariBench.

Secondly, we studied the correlation and the predictive ability of each different feature on a dataset that we named **S4428**, given by the combination of S669 and:

- **Ssym** (Pucci et al., 2018) which provides 684 balanced (i.e., half direct and half reverse) variations;
- **S2648** (Dehouck et al., 2011) and **VariBench** (Nair and Vihinen, 2013), two of the most used datasets, both extracted from Protherm (Kumar et al., 2006) database. They contain respectively 2,648 and 1,420 manually curated variants with experimentally measured $\Delta\Delta G$ values.

After merging all the datasets, we excluded 19 variants from the analysis due to errors in their 3D neighbors. The composition of all the datasets, their intersection and the distribution of their $\Delta\Delta$Gs are reported in Table 1, Figures 1, 2, respectively.

2.3 The unbalanced predictions of the state-of-the arts methods on never-before-seen variants

As shown in Figure 1, Ssym, S2648 and VariBench, which are the datasets most commonly used to train and test predictive methods, share a large number of variants. On the other hand, S669 has no variants in common with the other three datasets and its variants lie in proteins with less than 25% of sequence identity with the proteins in the other three datasets.

Since these characteristics are required for a proper test set, we will only assess $\Delta\Delta G$ prediction performance on the S669 dataset and not on the much larger S4428. The methods' performance in the different variant classes were evaluated by Pearson's correlation coefficient (ρ) and root mean square error (RMSE) between the experimental and the predicted $\Delta\Delta G$ (Table 2), defined as:

$$\rho = \frac{Cov\left(\Delta\Delta G^{exp}, \Delta\Delta G^{pred}\right)}{\sigma_{\Delta\Delta G^{exp}}\ \sigma_{\Delta\Delta G^{pred}}} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\Delta\Delta G_i^{exp} - \Delta\Delta G_i^{pred}\right)^2}{N}}. \tag{5}$$

where Cov is the covariance, σ is the standard deviation and N is the number of variants.

2.4 Assessing the impact of the input features in the prediction of destabilizing, neutral and stabilizing variants

In order to assess the relevance of the seven features listed in Section 2.1, we computed the Pearson's correlation coefficient (ρ) between each feature and the experimental $\Delta\Delta$Gs in S4428 (Table 3). To remove the bias in the ρ values on the whole dataset caused by the under-representation of the stabilizing and neutral variants with respect to the destabilizing ones (Table 1), we randomly under-sampled the available data to balance the classes. Specifically, we generated 100 random balanced subsets of 984 variants extracted from the original dataset. For each class we selected 328 elements, half the

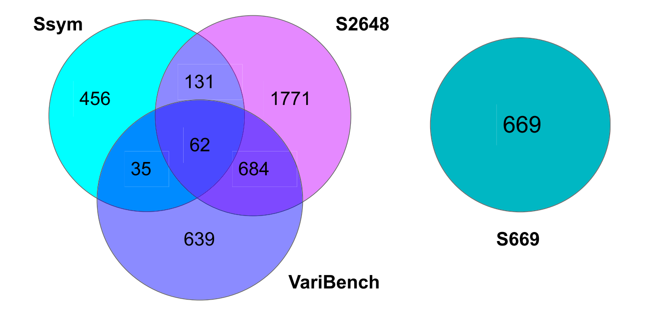


FIGURE 1
Venn diagrams showing the number of shared variants among the Ssym, VariBench, S2648 and S669 datasets.

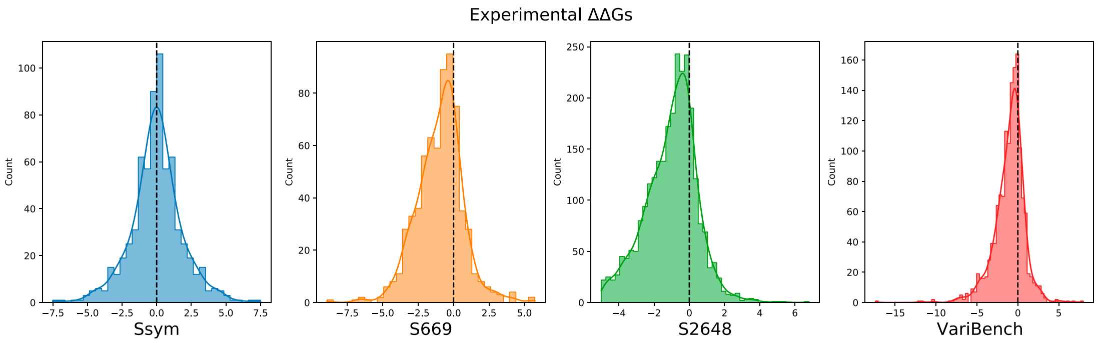


FIGURE 2
Distribution of the experimental ΔΔ*G* values in the Ssym, S669, S2648 and VariBench datasets.

TABLE 2 Pearson's correlations and root mean square error (RMSE) between the experimental and estimated ΔΔGs. The ΔΔGs are predicted by 18 commonly used protein stability prediction tools on the direct variants of the S669 dataset. The correlations and RMSE are calculated on each class separately ("Destabilizing"-"Neutral"-"Stabilizing"), on the whole dataset ("Total") and only on the destabilizing and stabilizing variants, excluding the neutral ("Non-neutral").

	Pearson/RMSE				
Dataset	Total	Destabilizing	Neutral	Stabilizing	Non-neutral
MAESTRO	0.50/1.44	0.42/1.46	−0.01/0.84	0.28/2.26	0.48/1.63
ACDC-NN	0.46/1.49	0.34/1.60	0.09/0.69	0.09/2.14	0.44/1.71
INPS3D	0.43/1.50	0.35/1.40	0.03/1.00	0.02/2.55	0.42/1.67
DDGun3D	0.43/1.60	0.32/1.69	0.13/0.94	0.13/2.22	0.41/1.80
INPS-Seq	0.43/1.52	0.26/1.56	0.10/0.92	0.13/2.25	0.42/1.70
ACDC-NN-Seq	0.42/1.53	0.28/1.64	0.08/0.76	0.07/2.18	0.40/1.75
PremPS	0.41/1.51	0.43/1.48	−0.02/0.84	−0.08/2.53	0.04/1.72
PopMusic	0.41/1.51	0.37/1.40	0.11/0.96	0.09/2.67	0.39/1.69
DUET	0.41/1.52	0.34/1.48	0.02/0.89	0.10/2.54	0.38/1.72
Dynamut	0.41/1.60	0.32/1.81	0.16/0.66	0.29/2.00	0.40/1.85
SDM	0.41/1.67	0.33/1.81	0.16/1.01	0.09/2.14	0.40/1.88
DDGun	0.40/1.75	0.25/1.75	0.12/1.29	0.11/2.46	0.39/1.90
SAAFEC-Seq	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
mCSM	0.36/1.54	0.30/1.42	−0.01/0.96	0.06/2.73	0.33/1.73
I-Mutant3.0	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
I-Mutant3.0-Seq	0.34/1.56	0.23/1.53	0.05/0.92	0.21/2.53	0.33/1.75
MuPro	0.25/1.61	0.19/1.45	0.08/1.08	−0.01/2.84	0.24/1.78
FoldX	0.21/2.32	0.20/2.25	0.01/2.28	0.17/2.66	0.24/2.33

size of the smallest class of variants on S4428. For each feature, the average Pearson's correlation coefficient and its standard deviation are reported in Table 3 (column "Total balanced").

We also evaluated the impact of the seven features in terms of their discriminative power using the Receiver Operating Characteristic (ROC) curve and its Area-Under-the-Curve (AUC-ROC) metric. We removed the size effect by under-sampling the variants 100 times in order to have 100 subsets of 656 variants for each pair of classes. On these under-sampled datasets we used each feature to calculate three different ROC curves separating three pairs of classes: Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing. Here, the assumption is that the higher the AUC scores, the better the separation of the two classes by the variable and, therefore, the more informative the variable is. The results of this analysis are shown in Table 4 and in the Supplementary Materials. The ROC curves and AUC scores of the logarithm of the conservation ratio, the volume difference and the difference in hydrophobicity were calculated using the values with opposite signs to help the interpretation. In Table 4, we also

TABLE 3 Pearson correlation coefficient (ρ) between the experimental ΔΔGs and the seven considered features in the S4428 dataset. On the column "Total", we have the correlation in the whole unbalanced dataset, while on the column "Total Balanced" we reported the average Pearson's correlation coefficient and its standard deviation calculated in 100 random subsets with the same number of variants for each class (Stabilizing-Neutral-Destabilizing). The last three columns report the correlation on each class separately, considering all the possible variants.

	Pearson's correlation coefficient				
	Total	Total balanced	Destabilizing	Neutral	Stabilizing
Accessibility	0.22	0.08 ± 0.02	0.28	0.02	−0.21
Bastolla-Vendruscolo potential	0.45	0.45 ± 0.03	0.36	0.13	0.18
Blosum evolutionary score	0.12	0.02 ± 0.02	0.26	−0.02	−0.18
Difference in hydrophobicity	−0.21	−0.18 ± 0.03	−0.19	−0.04	−0.04
Volume difference	−0.32	−0.34 ± 0.03	−0.22	−0.12	−0.15
Logarithm of the conservation ratio	−0.37	−0.37 ± 0.02	−0.27	−0.11	−0.21
Skolnick potential	0.36	0.35 ± 0.03	0.29	0.07	0.13

TABLE 4 Ability of each feature to separate the different classes and differences in distributions. We extracted one hundred subsets of size 328 from each class and used the variable score to calculate three different ROC curves: Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing. The assumption, here, is that the higher the AUC-ROC, the better the variable can separate between the two classes and the more informative it is. For each pair of classes we also computed the Mann-Whitney-Wilcoxon test two-sided to establish the statistical significance of the differences in the distributions.

AUC-ROC scores and *p*-values						
	Destabilizing-neutral		Destabilizing-stabilizing		Neutral-stabilizing	
	AUC ± std	*p-values*	*AUC ± std*	*p-values*	*AUC ± std*	*p-values*
Accessibility	0.69 ± 0.02	1.98e-79	0.55 ± 0.02	1.43e-04	0.36 ± 0.01	1.18e-23
Bastolla-Vendruscolo potential	0.66 ± 0.02	1.40e-58	0.75 ± 0.01	8.50e-84	0.63 ± 0.02	1.26e-22
Blosum evolutionary score	0.58 ± 0.02	5.61e-17	0.50 ± 0.02	7.08e-01	0.42 ± 0.02	1.21e-09
Difference in hydrophobicity	0.59 ± 0.02	9.41e-22	0.60 ± 0.02	2.28e-16	0.51 ± 0.02	4.22e-01
volume difference	0.60 ± 0.02	6.90e-24	0.72 ± 0.02	1.24e-63	0.64 ± 0.02	1.02e-22
Logarithm of the ratio of the conservation	0.66 ± 0.02	1.00e-58	0.75 ± 0.01	1.32e-84	0.63 ± 0.02	1.29e-20
Skolnick potential	0.64 ± 0.02	2.56e-41	0.71 ± 0.02	4.27e-63	0.60 ± 0.02	4.97e-14

reported the *p*-values of the two-sided Mann-Whitney Wilcoxon test computed between the distributions of the scores for each pair of classes.

2.5 Training a predictor with the reduced set of features

To evaluate the combined effect of the Blosum evolutionary score, difference in hydrophobicity and accessibility on the predictions of the different classes, we trained a Random Forest regressor on 100 random balanced subsets of S4428 (excluding S669) and then tested its performance on S669 (Table 5). We used three different sets of features: one with all the seven variables ("full"), a reduced set with all the variables except for Blosum evolutionary score and difference in hydrophobicity ("reduced") and one where we also excluded the accessibility ("red-no-acc"). Therefore, we tested to which extent the removal of the evolutionary-based component, the hydrophobicity and the accessibility affects DDGun3D predictions. The original DDGun3D score was computed as:

$$\mathrm{DDGun3D_{full}} = \left(0.20 \cdot S_{Bl} + 0.29 \cdot S_{Sk} + 0.18 \cdot S_{Hp} + 0.33 \cdot S_{BV}\right]\Big) \cdot (1.1 - acc), \tag{6}$$

where S_{Bl}, S_{Sk}, S_{BV}, S_{Hp}, acc are, respectively, the components related to the Blosum evolutionary score, the Skolnick and the Bastolla-Vendruscolo potentials, the difference in hydrophobicity and the accessibility, with the same coefficient and definition as in the original paper Montanucci et al. (2019a). We defined two "reduced" scores by dropping the corresponding components:

$$\begin{aligned}\mathrm{DDGun3D_{reduced}} &= (0.29 \cdot S_{Sk} + 0.33 \cdot S_{BV}]) \cdot (1.1 - acc),\\ \mathrm{DDGun3D_{red-no-acc}} &= (0.29 \cdot S_{Sk} + 0.33 \cdot S_{BV}]).\end{aligned} \tag{7}$$

3 Results

We divided the results into four sections: the first shows the methods' performance on the different variant classes on S669, a blind testing set, the second explores the correlation of the seven features

TABLE 5 Predictions with different sets of features. We tested to which extent the removal of the Blosum component, the hydrophobicity and the accessibility affects the performance of a Random Forest regressor and DDGun3D on S669. We used three different sets of features: one with all the seven variables, one with all the variables except for Blosum evolutionary score and the difference in hydrophobicity ("reduced") and one where we also excluded the accessibility ("red-no-acc"). Results are reported in terms of Pearson correlation coefficient and RMSE.

	Pearson/RMSE		
	Total	Destabilizing	Stabilizing
DDGun3D$_{\text{full}}$	0.43/1.6	0.32/1.69	0.13/2.22
DDGun3D$_{\text{reduced}}$	0.37/1.69	0.29/1.96	0.17/1.96
DDGun3D$_{\text{red-no-acc}}$	0.34/1.75	0.26/2.02	0.15/2.0
Random Forest$_{\text{full}}$	0.42/1.56	0.28/1.70	0.15/2.09
Random Forest$_{\text{reduced}}$	0.40/1.58	0.27/1.72	0.16/2.15
Random Forest$_{\text{red-no-acc}}$	0.4/1.58	0.26/1.71	0.19/2.13

with the experimental $\Delta\Delta G$ values, the third section investigates their discriminative power and the fourth section analyze their combined impact on the predictions.

3.1 The unbalanced predictions among stabilizing, neutral and destabilizing variants

In a previous study (Pancotti et al., 2022) we assessed the performance of 18 $\Delta\Delta G$ prediction methods on the S669 dataset, a test dataset with no intersection with the methods' training sets and whose variants are in proteins with less than 25% of sequence identity with those in the training sets. We used both direct and reverse variants (see Introduction) to show that, when dealing with never-before-seen variants, those methods that do not respect the antisymmetry property have a worse performance on stabilizing variants, while the antisymmetric ones performed consistently well on both classes.

In this study, we focused uniquely on direct variants and we showed that the difference in the non-antisymmetric methods' performance between destabilizing and stabilizing variants is even bigger when we do not consider the reverse variants. In addition, even antisymmetric methods showed a highly uneven performance between the classes (Table 2). The correlations on the whole dataset and on the non-neutral variants (destabilizing and stabilizing) were relatively good, with most methods having $\rho \geq 0.4$, showing that the predictions captured the general trend of the experimental $\Delta\Delta G$s. Most methods also predicted relatively well the destabilizing variants, with a $\rho \geq 0.3$. None of the methods, however, was able to predict the stabilizing variants with a good correlation. The highest Pearson's correlation on the stabilizing variants was ≤ 0.3 and the lowest RMSE was ≥ 2.

This result confirms the difficulty of current methods in adequately recognizing the direct-stabilizing variants.

3.2 Relevance of the seven features: Correlation on S4428

We assessed the relevance of the seven features of interest by computing their Pearson's correlation coefficient (ρ) with the experimental $\Delta\Delta G$s considering the S4428 dataset (Table 3). When considering the whole dataset, all the features showed absolute correlations ranging from $\rho = 0.12$, observed with the Blosum evolutionary score, to $\rho = 0.45$, reached by the Bastolla-Vendruscolo potential (Table 3, column "Total").

However, when the different classes were balanced by randomly sampling 100 balanced subsets of ~1,000 variants (column = "Total balanced"), the correlation dropped significantly for the Blosum evolutionary score and for the accessibility, since these two features are correlated with the destabilizing variants ($\rho = 0.26$ and $\rho = 0.28$, respectively) but anti-correlated with the stabilizing ones ($\rho = -0.18$ and $\rho = -0.21$, respectively). The reason for the anti-correlation of the Blosum score with stabilizing variants and the correlation with the destabilizing ones is the symmetry of the score, since $B(W, M) = B\ (M, W)$. On the other hand, for the accessibility this happens because, in general, the more one amino acid is buried, the greater impact its mutation has on the stability of the protein (either stabilizing or destabilizing). Therefore, this feature should be used to modulate the impact of the mutation, not to predict the sign.

The anti-correlation observed between the experimental $\Delta\Delta G$ and the logarithm of the conservation ratio is coherent with the assumption that the substitution of an amino acid with a less conserved one ($\log\left(\frac{CONS_W}{CONS_M}\right) > 0$) will likely have a disruptive effect, while substituting it with a more conserved one ($\log\left(\frac{CONS_W}{CONS_M}\right) < 0$) will likely have a stabilizing effect. The volume difference and the hydrophobicity between the two amino acids involved were also anti-correlated with the experimental $\Delta\Delta G$, meaning that a big absolute volume difference is associated with a greater effect, with a positive difference (i.e., mutated amino acid larger than the wild-type) being disruptive and a negative difference (i.e., wild-type greater than the mutated) being stabilizing. For hydrophobicity, while there seems to be a general trend (Figure 3, $\rho = -0.21$ on the whole dataset), neutral and stabilizing variants showed little correlations ($\rho = -0.04$ for both).

3.3 Discriminative power of the seven features: AUC-ROC on S4428

We also evaluated the discriminative power of the seven features using the AUC-ROC metric. We assumed that a specific feature is informative and useful for the prediction tools if it is able to separate the three pairs of classes (Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing). The higher the AUC-ROC scores for these separations, the more informative the feature is.

To remove the size effect, we under-sampled the variants 100 times in order to have 100 subsets of 656 variants for each pair of classes. In addition, due to their anti-correlation with the $\Delta\Delta G$s, the ROC curves of the logarithm of the conservation ratio, the volume difference and the difference in hydrophobicity were calculated using the values with opposite signs to respect the monotonic assumption that higher AUC-ROC means higher discriminative power, making them more immediately interpretable to the reader. In this way, instead of having a score of e.g.,: 0.3, the score would become 1–0.3 = 0.7. All the distributions of the features and the statistical differences between the classes computed using the two-sided Mann-Whitney Wilcoxon test are displayed in Figure 3 and in Table 4. The average AUC scores with their standard deviations are reported in Table 4, while the ROC curves are displayed in the Supplementary Materials.

The boxplots clearly show that none of the features perfectly separates all the three classes (Figure 3).

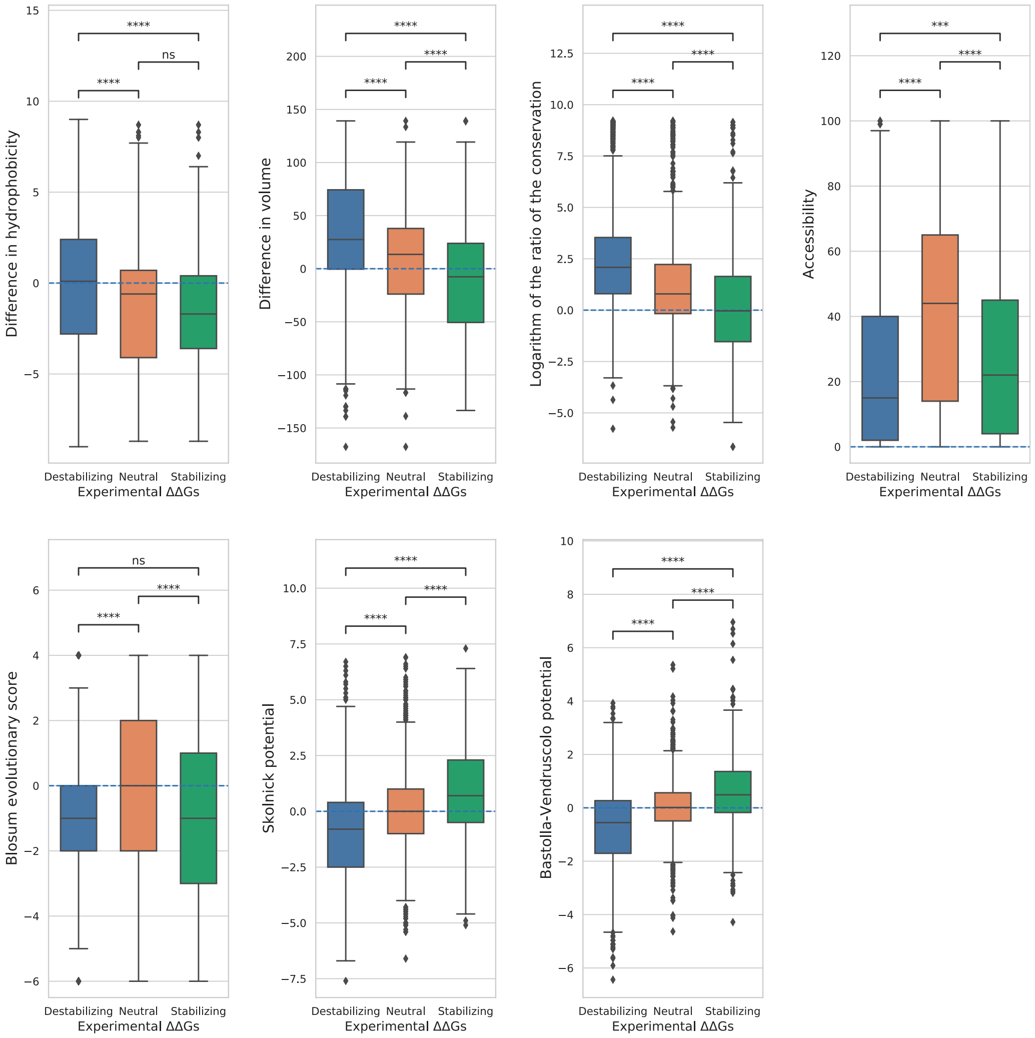


FIGURE 3
Distributions of the features. Boxplots showing the distributions of the features on the three classes. The variations are considered neutral if $\Delta\Delta G \in [-0.5, 0.5]$, stabilizing if $\Delta\Delta G < -0.5$ and destabilizing if $\Delta\Delta G > 0.5$. For each pair of classes we computed the Mann-Whitney-Wilcoxon test two-sided to establish the difference in the distributions. The p-values are reported here in a compact way: "ns" - $p > 0.05$, * - $0.01 < p \leq 0.05$, ** - $1.0e{-}03 < p \leq 0.01$, *** - $1.0e{-}04 < p \leq 1.0e{-}03$, **** - $p \leq 1.0e{-}04$, the actual values of the p-values are in Tab.4.

Overall, we found that the features with the highest absolute Pearson's correlation coefficients on the balanced datasets (Table 3, "Total balanced") were the best at separating between destabilizing and stabilizing variants, while those with a poor correlation also showed a poor discriminative power. The AUC scores "Destabilizing-Stabilizing" of the Bastolla-Vendruscolo potential, the Skolnick potential, the logarithm of the conservation ratio and the volume difference were all greater than 0.7, reflecting their high correlations on the balanced dataset. On the other hand, the accessibility and the difference in hydrophobicity, which are characterized by a low correlation on the balanced datasets (0.08 and −0.18, respectively), also showed low AUCs when separating destabilizing variants from stabilizing (0.55 and 0.6, respectively). The Blosum evolutionary score, which was uncorrelated on the balanced dataset, showed also random behaviour (AUC = 0.5) in separating the destabilizing from the stabilizing variants and the two distributions were not significantly different (p-value of Mann-Whitney test = 0.71). A non-significant p-value (p = 0.4) was also observed for the difference in hydrophobicity between stabilizing and neutral variants. This feature, while being able to slightly separate the destabilizing variants from the other two classes (AUC = 0.59 and AUC = 0.6), was not able to separate stabilizing from neutral variants (AUC = 0.51).

The accessibility separated fairly well the neutral variants from the other two classes, but not the destabilizing from the stabilizing (Figure 3), suggesting that it should only be used as a modulator of the $\Delta\Delta G$ absolute value. For the other best performing features, the two potentials and the conservation showed sightly higher similarity between neutral and stabilizing variants than between neutral and destabilizing, while the volume difference showed an opposite trend (Figure 3).

3.4 Improving the predictions on the stabilizing variants

We evaluated the effects of including or excluding the Blosum evolutionary score, the difference in hydrophobicity and the accessibility in a Random Forest predictor and in DDGun3D. The aim of this analysis was not to outperform existing methods, but to analyze how the combination of these variable affects the predictions on the different classes.

Table 5 shows the results obtained by these two predictors when using three possible sets of features: "all variables", "reduced", "red-no-acc". "Reduced" includes all variables except for Blosum evolutionary score and difference in hydrophobicity, while "red-no-acc" also excludes the accessibility.

The results show that, excluding the Blosum evolutionary score and difference in hydrophobicity, the correlation decreases on the destabilizing class for both methods: ρ = 0.32 for DDGun3D and ρ = 0.28 for Random Forest with "all variables", compared to ρ = 0.29 and 0.27 with the "reduced" set. The same behaviour in the two predictors was observed when the "red-no-acc" set was considered. Given the high unbalance in S669 towards destabilizing mutations, with 387 variants being destabilizing and only 85 being stabilizing, the decreasing performance on the destabilizing variants affects the overall correlation on the whole dataset too.

Removing the Blosum evolutionary score and the difference in hydrophobicity, however, increases the correlation on the stabilizing class, as expected. In addition, the correlation increases when we also remove the accessibility from the Random Forest predictor's features (ρ = 0.19 vs. ρ = 0.15), but not when DDGun3D is used (ρ = 0.15 vs. ρ = 0.13). However, it is worth noticing that, while in the Random Forest predictor the accessibility is used as any other feature, in DDGun3D it is used as modulator of the $\Delta\Delta G$ (see Eqs 6, 7). Indeed, using the Random Forest predictor, the additional removal of the accessibility improves the performance with respect to the removal of only the Blosum evolutionary score and the difference in hydrophobicity (ρ = 0.19 vs. ρ = 0.16), while for DDGun3D this additional removal negatively affects the correlation (ρ = 0.15 vs. ρ = 0.17).

4 Discussion and conclusion

Our study is based on the observation that none of the tools available to predict the $\Delta\Delta G$ is very accurate on the stabilizing variants and, in general, the predictions are skewed towards neutral and destabilizing variations. The existing datasets for $\Delta\Delta G$ prediction share a large number of variants and are strongly unbalanced towards the destabilizing variants (Table 1; Figure 2). For this reason, choosing features that favour only the most abundant class can lead to a good overall performance at the cost of penalizing the prediction of the less abundant class.

To better understand the weakness in correctly predicting the stabilizing variants, we evaluated seven properties commonly used by the computational $\Delta\Delta G$ predictors. We considered two features based on physical properties (hydrophobicity and volume), three structural information and conservation-based features (Blosum evolutionary score, conservation and relative solvent accessibility) and two statistical potentials-based features (Skolnick potential and Bastolla-Vendruscolo potential). For each of them, we analyzed the ability to predict the experimental $\Delta\Delta G$ by computing the Pearson's correlation coefficient (ρ) between them and the experimental $\Delta\Delta G$s of three of the most used dataset (S2648, VariBench and Ssym) and a recently-released one (S669), all combined in the S4428 dataset. We also computed the AUC-ROCs for each variant to judge how these features separate the different classes of variations. The results showed that the volume difference, the logarithm of the conservation ratio, and the statistical potentials are better than random in each possible separation, i.e., destabilizing vs neutral, destabilizing vs stabilizing and neutral vs stabilizing variants. On the other hand, the difference in hydrophobicity, the Blosum evolutionary score and the accessibility likely showed random results in at least one of the separations. Although hydrophobicity difference has an anti-correlation trend in the balanced dataset (ρ = −0.18), it cannot separate the neutral variants from the stabilizing ones (AUC-ROC = 0.51 ± 0.02), and it has marginal ability to separate destabilizing from neutrals (AUC-ROC = 0.59 ± 0.02). Thus, hydrophobicity difference alone can contribute to entangling stabilizing and neutral variants at prediction time. In turn, this may lead to enforce the role of the destabilizing variants during the method training.

Due to the symmetry of the Blosum evolutionary score, it is not surprising that the AUC-ROC in separating destabilizing and stabilizing variants was 0.5 ± 0.02. However, the score also failed to separate the other two classes from the neutral ones (AUC-ROCs = 0.58/0.42 ± 0.02), which could lead to predictions skewed towards the neutral class.

The accessibility is actually fairly good at separating the neutral from the other two classes (AUC-ROCs = 0.69/0.36 ± 0.02), but not the destabilizing from the stabilizing (AUC-ROC = 0.55 ± 0.02). Accessibility is then a good marker of the general impact of a variation (being them stabilizing or destabilizing) since its absolute value of ρ on the stabilizing variants ($|\rho|$ = 0.21) is one of the highest among the considered features.

Among the seven features, the logarithm of the conservation ratio best described the stabilizing variants (ρ = −0.21), consolidating the known important role of the conservation in the impact of the genetic variants. Moreover, the logarithm of the conservation ratio and the volume difference showed the smallest drop in correlation between destabilizing and stabilizing variants (Tab.3). The other two best performing features, the Bastolla-Vendruscolo potential and the Skolnick potential, showed a bigger drop in correlation, even though the Bastolla-Vendruscolo potential by itself reached a correlation of ρ = 0.45 on the full dataset, which is not that far off from the results that many predictors get using way more features and information.

We tested the impact of removing Blosum62, hydrophobicity and accessibility (the three worst features at separating the stabilizing variants from the rest) on prediction by examining two models: DDGun3D, a mostly linear $\Delta\Delta G$ predictor with limited interaction between features and a Random Forest, a fully non-linear generic method. Independently from the predictor used, dropping out Blosum62 and hydrophobicity improves the performance on stabilizing variants at the cost of a loss on destabilizing ones.

This is not surprising for a linear model where each feature has a fixed and independent effect on all variants. On the other hand, a more powerful method like Random Forest, that can model interactions and varying non-linear effects, could be able to use these features (Blosum62 and hydrophobicity) selectively only for the non-stabilizing variants for which they are informative. However, our tests suggest that, in practice, including them can be harmful for stabilizing variant prediction regardless of the model.

On the other hand, the behaviour of the accessibility is more complex since the less accessible residues tend to have a higher impact than the accessible ones. In some way, accessibility can modulate the stability effect as, without it, the performance on the stabilizing variants improves in the Random Forest model but worsens in DDGun3D when it is not used as a shrinking effect (Table 5). This result also agrees with previous studies that led to the tool PopMusic2 (Dehouck et al., 2011).

Our results suggest not incorporating the substitution matrix score (e.g., Blosum62) and the difference in hydrophobicity in future predictive computational tools, while using the difference in accessibility only to modulate the impact of the variant. Furthermore, we suggest using the features that better separate the stabilizing from the neutral variants (such as the Bastolla-Vendruscolo potential, the logarithm of the conservation ratio and the volume difference) to avoid the compression of the predictions towards the neutral values. Finally, new tools should also be tested on not-artificially balanced datasets, reporting the performance specifically for each variant class.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://academic.oup.com/bib/article/23/2/bbab555/6502552#supplementary-data.

Author contributions

TS, PF, EC, and GB conceptualized the research. SB and GB prepared the data. SB ran the analyses and drafted the manuscript. All the authors contributed to the final 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 (PRIN201744NR8S).

Acknowledgments

We thank the EU projects GenoMed4All (H2020 RIA n.101017549) and BRAINTEASER (H2020 RIA n.101017598). We also thank PNRR M4C2 HPC—1.4 “CENTRI NAZIONALI”- Spoke 8 for fellowship support.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1075570/full#supplementary-material

References

Bastolla, U., Farwer, J., Knapp, E. W., and Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins Struct. Funct. Bioinforma.* 44, 79–96. doi:10.1002/prot.1075

Benevenuta, S., and Fariselli, P. (2019). On the upper bounds of the real-valued predictions. *Bioinform Biol. Insights* 13, 1177932219871263. doi:10.1177/1177932219871263

Benevenuta, S., Pancotti, C., Fariselli, P., Birolo, G., and Sanavia, T. (2021). An antisymmetric neural network to predict free energy changes in protein variants. *J. Phys. D Appl. Phys.* 54, 245403. doi:10.1088/1361-6463/abedfb

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

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

Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinforma.* 9, S6–S9. doi:10.1186/1471-2105-9-S2-S6

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, 1161–1170. doi:10.1002/humu.21555

Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S., and Li, M. (2020). Premps: Predicting the impact of missense mutations on protein stability. *PLoS Comput. Biol.* 16, e1008543. doi:10.1371/journal.pcbi.1008543

Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinforma.* 62, 1125–1132. doi:10.1002/prot.20810

Cheng, T. M., Lu, Y.-E., Vendruscolo, M., Liò, P., Blundell, T. L., et al. (2008). Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput. Biol.* 4, e1000135. doi:10.1371/journal.pcbi.1000135

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

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 Bioinforma.* 12, 151. doi:10.1186/1471-2105-12-151

Hartl, F. U. (2017). Protein misfolding diseases. *Annu. Rev. Biochem.* 86, 21–26. doi:10.1146/annurev-biochem-061516-044518

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. doi:10.1073/pnas.89.22.10915

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

Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.* 79, 830–838. doi:10.1002/prot.22921

Kumar, M. D., 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

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi:10.1016/0022-2836(82)90515-0

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

Li, B., Yang, Y. T., Capra, J. A., and Gerstein, M. B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks. *PLOS Comput. Biol.* 16, e1008291. doi:10.1371/journal.pcbi.1008291

Li, G., Panday, S. K., and Alexov, E. (2021). Saafec-seq: A sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability. *Int. J. Mol. Sci.* 22, 606. doi:10.3390/ijms22020606

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 genomics* 17, 397–247. doi:10.1186/s12864-016-2726-y

Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., and Fariselli, P. (2019a). Ddgun: An untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinforma.* 20, 335. doi:10.1186/s12859-019-2923-1

Montanucci, L., Martelli, P. L., Ben-Tal, N., and Fariselli, P. (2019b). A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* 35, 1513–1517. doi:10.1093/bioinformatics/bty880

Nair, P. S., and Vihinen, M. (2013). V ari b ench: A benchmark database for variations. *Hum. Mutat.* 34, 42–49. doi:10.1002/humu.22204

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., et al. (2022). Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Briefings Bioinforma.* 23, Bbab555. doi:10.1093/bib/bbab555

Pancotti, C., Benevenuta, S., Repetto, V., Birolo, G., Capriotti, E., Sanavia, T., et al. (2021). A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes* 12, 911. doi:10.3390/genes12060911

Pires, D. E., 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. doi:10.1093/nar/gku411

Pires, D. E., 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. E., Chen, J., Blundell, T. L., and Ascher, D. B. (2016). *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.* 6, 19848. doi:10.1038/srep19848

Pucci, F., Bernaerts, K. V., Kwasigroch, J. M., and Rooman, M. (2018). Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* 34, 3659–3665. doi:10.1093/bioinformatics/bty348

Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* 30, 60–69. doi:10.1002/pro.3942

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, W350–W355. doi:10.1093/nar/gky300

Samaga, Y. B., Raghunathan, S., and Priyakumar, U. D. (2021). Scones: Self-consistent neural network for protein stability prediction upon mutation. *J. Phys. Chem. B* 125, 10657–10671. doi:10.1021/acs.jpcb.1c04913

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, 2542–2544. doi:10.1093/bioinformatics/btw192

Savojardo, C., Martelli, P. L., Casadio, R., and Fariselli, P. (2019). On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform* 22, 601–603. doi:10.1093/bib/bbz168

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

Skolnick, J., Godzik, A., Jaroszewski, L., Kolinski, A., and Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct? *Protein Sci.* 6, 676–688. doi:10.1002/pro.5560060317

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., et al. (2015). A series of pdb-related databanks for everyday needs. *Nucleic acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028

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

Xavier, J. S., Nguyen, T. B., Karmarkar, M., Portelli, S., Rezende, P. M., Velloso, J. P. L., et al. (2021). ThermoMutDB: A thermodynamic database for missense mutations. *Nucleic Acids Res.* 49, D475–D479. doi:10.1093/nar/gkaa925

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353, 459–473. doi:10.1016/j.jmb.2005.08.020

Zamyatnin, A. (1972). Protein volume in solution. *Prog. biophysics Mol. Biol.* 24, 107–123. doi:10.1016/0079-6107(72)90005-3

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fmolb.2022.1100992

Calmodulin variants associated with congenital arrhythmia impair selectivity for ryanodine receptors

Giuditta Dal Cortivo[†], Valerio Marino[†], Silvia Bianconi and Daniele Dell'Orco*

Department of Neurosciences, Biomedicine and Movement Sciences, Section of Biological Chemistry, University of Verona, Verona, Italy

OPEN ACCESS

EDITED BY
Yong Wang,
Zhejiang University, China

REVIEWED BY
Razvan L. Cornea,
University of Minnesota Twin Cities,
United States
Zhiguang Yuchi,
Tianjin University, China

*CORRESPONDENCE
Daniele Dell'Orco,
✉ daniele.dellorco@univr.it

[†]These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 17 November 2022
ACCEPTED 12 December 2022
PUBLISHED 05 January 2023

CITATION
Dal Cortivo G, Marino V, Bianconi S and Dell'Orco D (2023), Calmodulin variants associated with congenital arrhythmia impair selectivity for ryanodine receptors.
Front. Mol. Biosci. 9:1100992.
doi: 10.3389/fmolb.2022.1100992

Among its many molecular targets, the ubiquitous calcium sensor protein calmodulin (CaM) recognizes and regulates the activity of ryanodine receptors type 1 (RyR1) and 2 (RyR2), mainly expressed in skeletal and cardiac muscle, respectively. Such regulation is essential to achieve controlled contraction of muscle cells. To unravel the molecular mechanisms underlying the target recognition process, we conducted a comprehensive biophysical investigation of the interaction between two calmodulin variants associated with congenital arrhythmia, namely N97I and Q135P, and a highly conserved calmodulin-binding region in RyR1 and RyR2. The structural, thermodynamic, and kinetic properties of protein-peptide interactions were assessed together with an in-depth structural and topological investigation based on molecular dynamics simulations. This integrated approach allowed us to identify amino acids that are crucial in mediating allosteric processes, which enable high selectivity in molecular target recognition. Our results suggest that the ability of calmodulin to discriminate between RyR1 an RyR2 targets depends on kinetic discrimination and robust allosteric communication between Ca^{2+}-binding sites (EF1-EF3 and EF3-EF4 pairs), which is perturbed in both N97I and Q135P arrhythmia-associated variants.

KEYWORDS

point mutation, kinetics, kinetic discrimination, molecular dynamics, protein structure network, protein-protein recognition, free-energy

1 Introduction

Calmodulin is a widely expressed calcium sensor protein capable of binding up to four Ca^{2+} ions with micromolar affinity, thereby acquiring conformations that allow specific recognition of more than 300 molecular targets (Stevens, 1983). The two globular domains of CaM are individually composed by a pair of helix-loop-helix EF-hand motifs, each of which has a different affinity for Ca^{2+}. The N-terminal domain, consisting of the EF1 and EF2 motifs, shows lower affinity for Ca^{2+} (K_D ~10 μM) than the C-terminal domain (K_D ~1 μM), consisting of the EF3 and EF4 motifs (Figure 1A). Ca^{2+} ion binding occurs with positive cooperativity within each domain, but in the absence of a molecular target, no interdomain cooperativity is observed (Linse et al., 1991). Indeed, binding of a target to CaM induces substantial changes in the apparent affinity and cooperativity for Ca^{2+} binding (Newman et al., 2008; Theoharis et al., 2008; Valeyev et al., 2008), which is closely related to the high structural plasticity of CaM (Astegno et al., 2014). Recognition of a wide variety of molecular targets gives CaM the ability to regulate diverse biochemical processes, such as transmembrane ion transport, cell mobility and proliferation, apoptosis, cytoskeleton remodeling, and protein folding (Chin and Means, 2000; Carafoli, 2002). This versatility is the result of both CaM's very high sensitivity in detecting even minute changes in intracellular Ca^{2+} concentration and remarkable selectivity in target activation.

Ryanodine receptors (RyR), sentinels of massive intracellular Ca^{2+} stores contained in the sarcoplasmic reticulum, release Ca^{2+} into the cytosol in response to sarcolemmal depolarization, thereby facilitating mobilization of the myofilaments and enabling cell contraction in cardiac and skeletal muscle cells. To achieve cell relaxation, Ca^{2+} must be rapidly resequestered or extruded from the cytosol (Capes et al., 2011). CaM is one of the proteins that precisely regulate the activity of RyR1 and RyR2 receptors, which are predominantly expressed in skeletal and cardiac muscle, respectively (Capes et al., 2011). Because of its trace-level expression in the heart, the contribution of RyR1 to cardiac function is currently unknown, while that of RyR2 is much better understood. In fact, although RyR2 is also strongly expressed in neurons and in visceral and arterial smooth muscle, it is in cardiomyocytes that the essential action of CaM in channel regulation has been accurately established. Even though RyR2 can bind Ca^{2+}, it is precisely the interaction with CaM that allows fine-tuning of channel inhibition, which is strictly dependent on the concentration of free Ca^{2+} (Sorensen et al., 2013; Van Petegem, 2015). This enables the proper dynamics of Ca^{2+}-induced Ca^{2+} release (CICR) mechanism, a key step in the regulation of the excitation-contraction process in cardiomyocytes (Sorensen et al., 2013).

Dysregulation of Ca^{2+} release *via* RyRs is associated with life-threatening diseases in both skeletal and cardiac muscle (Capes et al., 2011). This became particularly evident 10 years ago, when two missense mutations in CaM were found to be associated with heart failure and sudden cardiac death due to Catecholaminergic Polymorphic Ventricular Tachycardia (CPVT) and long QT syndrome (LQTS) (Nyegaard et al., 2012). To date, the number of point mutations in any of the three genes (*CALM1-3*) encoding CaM in humans and associated with heart diseases has risen to 17 (Jensen et al., 2018). On the other hand, structural knowledge has been accumulating on the CaM-RyR2 interaction (Sondergaard et al., 2017; Brohus et al., 2019; Chi et al., 2019; Sondergaard et al., 2019; Holt et al., 2020; Sondergaard et al., 2020). Unprecedented insight into the functional properties of RyR2 and its regulation by CaM has come from cryogenic-electron microscopy (cryo-EM) (Peng et al., 2016; Chi et al., 2019; Gong et al., 2019), which shed light on the three previously identified CaM-binding domains (CaMBD) shared by RyR receptors, which could interact individually or in groups with CaM lobes to regulate channel function (Lau et al., 2014). Cryo-EM revealed a complex CaM-RyR2 recognition mechanism, in which apo- and Ca^{2+}-bound CaM bind to distinct but overlapping RyR2 sites, and demonstrated that Ca^{2+}-bound CaM is one of many possible regulators competing for RyR2 gating (Gong et al., 2019). Among the possible CaM-RyR binding interfaces, a common region was identified by structural investigations, which corresponds to the K3614-L3644 stretch in RyR1 and the R3581-L3611 stretch in RyR2 (Lau et al., 2014). This finding raises the question of why an interaction based on an extremely similar region of the two targets (3 different amino acids out of 31; Figure 1B) results in a severe and life-threatening phenotype only in the case of the interaction with RyR2, which is specific to cardiac muscle cells. In a more general context, it is important to identify the molecular determinants that allow CaM to discriminate among different targets, even when selective recognition is based on a few different amino acids.

To unravel the molecular mechanisms underlying the selective perturbation of the recognition between CaM and RyR1 or RyR2 in arrhythmia-associated conditions, we focused on a CaM binding target region comprising 31 amino acids with a very similar sequence in the two RyRs (known as CaMBD2). We conducted a comprehensive biophysical investigation of the structural, kinetic, and thermodynamic properties underlying the association between peptides comprising the analogous RyR1/RyR2 regions and three CaM variants, namely the wild type protein and the N97I and Q135P variants (Figure 1A), which are associated with LQTS and LQTS/CPVT, respectively (Jensen et al., 2018). Integrating spectroscopic investigation with molecular dynamics simulations and protein structure network analysis, we found that despite the very high sequence similarity of the two RyR1/2 interacting regions, these disease-associated CaM mutations alter CaM selectivity for the specific RyR channel.

2 Materials and methods

2.1 Plasmid and peptide preparation

The gene encoding for human wild type (WT) calmodulin (Uniprot entry: P0DP23) was cloned into a pET24 (+) vector (resistance to kanamycin); cloning, codon optimization, point mutations and sequence check was done by Genscript. WT and

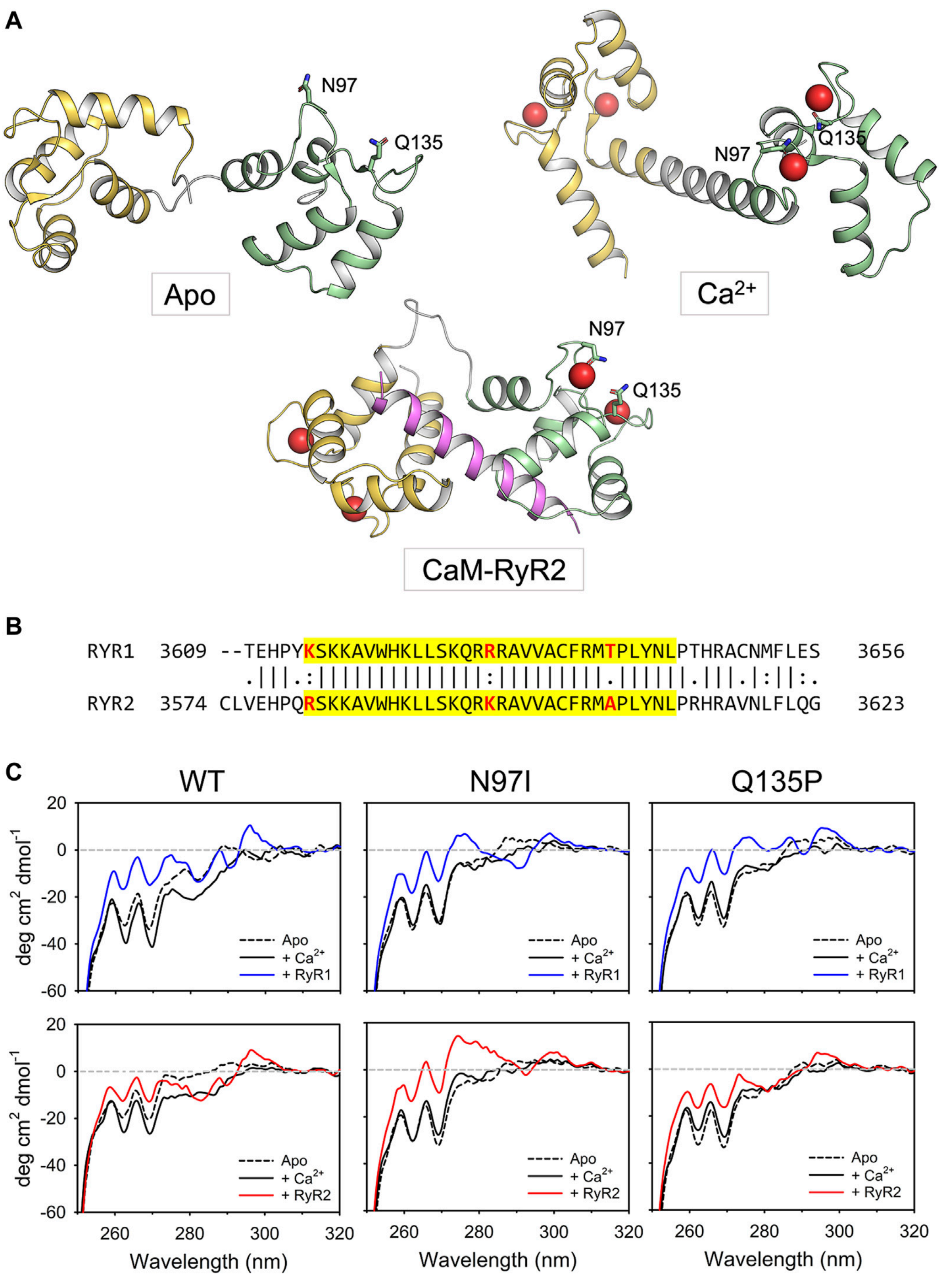


FIGURE 1
Tertiary and quaternary structure of CaM variants in complex with RyR1/2 peptides. **(A)** Cartoon representation of CaM in three different conformational states, namely apo (PDB: 1DMO (Zhang et al., 1995)), Ca^{2+}-bound (PDB: 1CLL (Chattopadhyaya et al., 1992)) and complexed with the RyR2 peptide used in this study (PDB: 6Y4O (Holt et al., 2020)). The N-terminal domain is colored in yellow, the linker region in grey, while the C-terminal domain in green. The RyR2 peptide is represented in magenta. Ca^{2+} ions are represented as red spheres while the side chains of N97 and Q135 are represented in sticks with C atoms colored according to the structural region, O atoms in red and N atoms in blue. **(B)** Pairwise sequence alignment of the Calmodulin Binding Domain-2 of human RyR1 (Uniprot entry P21817) and RyR2 (Uniprot entry Q92736). The sequence relative to the two RyR peptides employed in this study is highlighted in yellow, the residues that are not identical in such region are represented in bold and colored in red. **(C)** Near-UV CD spectra (250–320 nm) of 50 μM CaM were collected in the presence of 500 μM EGTA (black dashed line) and after sequential additions of 1 mM Ca^{2+} (black solid line, 500 μM free Ca^{2+}) and 100 μM RyR peptides (blue solid line for RyR1, red solid line for RyR2). The spectrum of sole buffer was considered as blank and subtracted; each curve represents the average of five accumulations. Temperature was set at 25°C and signal was normalized to protein concentration.

disease-related CaM variants, namely N97I and Q135P, were expressed as His-tag proteins, as in (Dal Cortivo et al., 2022). The nomenclature used in this paper is referred to the mature protein that lacks the Met in position 1. RyR1 and RyR2 peptides, encompassing regions K3614-L3644 (KSKKAVWHKLLSKQRRRAVVACFRMTPLYNL) and R3581-L3611 (RSKKAVWHKLLSKQRKRAVVACFRMAPLYNL) of the human RyR1 and RyR2 channels respectively, were purchased from Genscript as lyophilized powder (purity >95%, assessed *via* mass spectrometry and HPLC). From now on, RyR1 and RyR2 refer to the two respective peptides.

2.2 Protein expression and purification

The expression and purification of CaM variants were performed as described previously (Dal Cortivo et al., 2022). Briefly, after the transformation of *E. coli* BL21-DE3 strain by thermal shock, cells were grown at 37°C until the OD_{600} reached 0.6. After the expression induction (1 mM IPTG) cells were grown for 4 h at 37°C. Cells were then lysed using lysozyme and DNAse and, after centrifugation, the soluble fraction was loaded onto a His-trap FF crude column (GE Healthcare). After a one-step elution, the His-tagged CaM was dialyzed to allow the cleavage of His-tag at 25°C, overnight, using Tobacco Etch Virus protease (TEV, Promega) in a ratio TEV: CaM of 1U: 450 µg. Reverse immobilized metal affinity chromatography was performed, allowing the collection of the tag-free protein in the flow-through. Purified proteins were then quantified *via* Bradford assay, using a calibration line specific for human CaM (Alphalyze), aliquoted, flash frozen in liquid N_2 and stored at −80°C until use.

2.3 Circular dichroism spectroscopy

Secondary and tertiary structure of CaM variants were assessed using a Jasco-710 spectropolarimeter supplied with a Peltier-type cell-holder as previously detailed (Dal Cortivo et al., 2022). Briefly, far UV (200–250 nm) circular dichroism (CD) spectra were collected both using 10 µM of sole CaM or in co-presence of 20 µM RyR peptides. Spectra of isolated CaM variants and CaM-RyR1/2 complexes were measured in the absence of cations (apo conditions; 300 µM EGTA) and after the addition of saturating Ca^{2+} (600 µM, 300 µM free Ca^{2+}) using a 0.1 cm path length quartz cuvette. Near UV CD spectra of 50 µM CaM were collected in the presence of 500 µM EGTA (apo) and after sequential additions of 1 mM Ca^{2+} (500 µM free Ca^{2+}) and 100 µM RyRs using 1 cm path length quartz cuvette. Spectra were normalized based on protein concentration measured by Bradford assay and molecular weight of the formed complex. Each spectrum represents the average of five accumulations and the spectrum of the working buffer (20 mM Tris pH 7.5, 150 mM KCl, 1 mM DTT) was considered as reference and subtracted. Temperature and time response were set at 25°C and 4 s, respectively.

2.4 Limited proteolysis

Susceptibility of CaM variants to proteolysis was assessed incubating 26.4 µM of protein with 0.4 µM Trypsin (Sigma), i.e. at a protein:trypsin ratio 66:1, in the presence of 2 mM EGTA or Ca^{2+}. After 10 min incubation at 25°C (conditions previously optimized to maximize differences in proteolytic patterns (Dal Cortivo et al., 2022)), reactions were blocked adding sample buffer 4x and boiling for 10 min. Each reaction product, together with the untreated sample, was loaded on a 15% SDS-PAGE run at 200 V for 45–50 min and Coomassie blue-stained.

2.5 Fluorescence titrations

The apparent affinity of CaM variants for RyR1 and RyR2 was assessed following the intrinsic fluorescence of the only Trp residue present in both peptides. Briefly, 1 µM of each peptide was incubated with increasing concentrations of CaM (0–4 µM) using 20 mM Tris pH 7.5, 150 mM KCl, 1 mM DTT and 100 µM Ca^{2+} as working buffer. The fraction of peptide bound (fb) to CaM (Astegno et al., 2016) as function of the protein concentration, and was calculated as follows:

$$fb = \frac{y - y_0}{y_{max} - y_0}$$

where y_0 and y_{max} are the wavelengths at which the isolated peptide and the saturated complex showed their maximum fluorescence intensity, respectively. Each titration was repeated in triplicate. Each dataset was fitted to a one-site saturation ligand binding function to obtain individual K_D values for each replica.

2.6 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) measurements were performed as previously elucidated (Dal Cortivo et al., 2022). Briefly, the MicroCal PEAQ instrument was set to 25°C to perform titrations of RyR1 and RyR2 peptide (125 µM loaded in the syringe) with 10 µM WT-, N97I- and Q135P-CaM in 20 mM Tris pH 7.5, 150 mM KCl, 1 mM DTT, 5 mM Ca^{2+} working buffer. Thirty 1-µL injections were performed, with 150 s delay between each injection and setting the stirring to 750 rpm; each titration was performed at least in triplicate. Dilution effect was considered as blank and subtracted; it was measured by titrating the peptide with sole buffer. Data were fitted to a “one set of sites” model to obtain the number of binding sites

(N), the dissociation constant (K_D) and the enthalpy change (ΔH). These parameters were then used to calculate ΔG, the entropy change (ΔS) and $\Delta\Delta G$ as follows:

$$\Delta G = RTlnK_D = \Delta H - T\Delta S$$
$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$$

2.7 Surface plasmon resonance

Surface plasmon resonance (SPR) experiments were performed using a SensiQ Pioneer instrument and His-Cap sensor chips (from FortèBio and Sartorius). Following a previously optimized protocol (Dal Cortivo et al., 2019; Dal Cortivo et al., 2022) 1000 RU (1 RU = 1 pg mm^{-2}) of each His-CaM variant were immobilized on the surface of the sensor chip using 20 mM Tris pH 7.5, 150 mM KCl, 0.005% Tween 20 as working buffer. Immobilized His-CaM variants were washed overnight using a flowrate of 5 µL/min to remove unbound proteins. DTT (100 µM) and Ca^{2+} (5 mM) were freshly added before the beginning of titration experiments. Titrations were performed by injecting increasing concentrations of RyR1 and RyR2 (ranging from 250 nM to 3–4 µM) for 60 s and following the dissociation for 300 s using a flowrate of 20 µL/min. Data were fitted using a 1:1 Langmuir model. The dissociation phase was considered first to obtain k^{off} values that were used to calculate the k^{on} in the same binding process, according to a pseudo-first-order kinetic scheme. Each titration was repeated at least in triplicate.

2.8 Molecular modeling of CaM-RyR peptide complexes

The molecular modeling of Ca^{2+}-loaded human CaM complexed to RyR1 and RyR2 peptides was performed with the BioLuminate interface of Maestro chemical simulation suite (v. 12.5.139, Schroedinger) starting from the high-resolution (1.84 Å) X-ray structure of CaM-RyR2 peptide (residues S3582-M3605) complex, with Protein Data Bank entry: 6Y4O (Holt et al., 2020). RyR2 residues K3581, T3606, and P3607 were modelled by BioLuminate to maximize the sequence coverage of the peptide used for *in vitro* experiments. The CaM-RyR2 peptide complex structure was then subjected to the *Protein preparation* pipeline provided by BioLuminate, previously detailed in (Dal Cortivo et al., 2022), which consisted of: i) assignment of bond orders (comprising 0-order bonds to ions) according to the Chemical Components Dictionary database; ii) addition of H atoms; iii) modeling of the missing loop (residues K77-D80) using *Prime*; iv) sampling of the orientation of water molecules; v) calculation of the protonation states of ionizable residues at neutral pH (7.5) using PROPKA; vi) H-bond optimization. Arrhythmia-associated CaM variants N97I and Q135P were introduced by BioLuminate *Mutate residue* tool after selecing the highest-ranked non-clashing rotamer. Analogously, modeling of RyR1 peptide (residues R3614-P3640) was achieved by introducing three mutations (underlined) in RyR2 peptide (residues K3581-P3607), resulting in the following sequence: RSKKAVWHKLLSKQRKRAVVACFRMAP.

2.9 Molecular dynamics simulations and *in silico* analysis of CaM-RyR peptide complexes' stability

CaM-RyR1/2 peptide complexes were subjected to all-atom molecular dynamics (MD) simulations on GROMACS (v. 2020.3) simulation package (Abraham et al., 2015), using CHARMM36 m as forcefield (Huang et al., 2017). Protein complexes were placed in the center of a dodecahedral simulation box containing ~36,000 atoms, their charge was neutralized with 150 mM KCl, then structures underwent energy minimization and serial 2-ns equilibration in NVT and NPT ensembles, as previously elucidated (Marino et al., 2015). Four 300-ns simulations at 1 atm constant pressure and 310 K constant temperature were performed for each combination of variant and peptide. The convergence and consistency of the conformational space sampled by the simulations was assessed by Principal Component Analysis as described in (Borsatto et al., 2019; Marino and Dell'Orco, 2019). The diagonalization of the covariance matrix calculated on the coordinates of α-carbons (Cα) allows to identify the directions (eigenvectors) of the largest collective motions (eigenvalues) of the simulated systems. Thus, the consistency of the conformational space sampled by each replica was assessed by projecting the frames from the individual 300-ns trajectories onto the first two principal components of the concatenated trajectory. Such projections were classified using Linear Discriminant Analysis, a method that reduces the dimensional space, maximizes intercluster distances, and minimizes intracluster distances. Moreover, the first 20 principal components of each replica and the concatenated trajectories were considered the essential subspace and compared using RMSIP metric, which is defined as follows:

$$RMSIP = \sqrt{\frac{1}{S}\sum_{n=1,m=1}^{S}\left(v_n^i \cdot v_m^j\right)^2}$$

where S is the number of principal components,v_n^i and v_m^j identify the n^{th} and m^{th} principal components of trajectories i and j.

The Root-Mean Square Fluctuation (RMSF) of Cα and Ca^{2+}-ions in each CaM EF-hand (indicative of the structural flexibility of the protein and the apparent affinity for Ca^{2+}, respectively), was calculated with respect to the average structure of the complex by means of GROMACS' *gmx rmsf* function.

The final frames of each of the four 300-ns trajectories of WT CaM-RyR1 and WT CaM-RyR2 complexes were subjected to the *Residue scanning* analysis (provided by BioLuminate) to evaluate the effects of arrhythmia-associated mutations on the Gibbs free energy of protein-peptide folding ($\Delta\Delta G_f^{app}$) and binding ($\Delta\Delta G_b^{app}$), with respect to the WT. This tool computes the mutation-specific thermodynamic cycle based on the Molecular Mechanics/Generalized Born and Surface Area continuum solvation (MM/GBSA). Although this method does not include the contribution of conformational changes, it provides apparent free energy variations ($\Delta\Delta G_{app}$, in kcal/mol) which can reliably estimate mutation-associated differences in stability and affinity, rather than precise thermodynamic measurables.

2.10 Protein structure network analysis

All-atom MD simulations provide dynamical structural information, which was condensed in a static Protein Structure Network (PSN) using PyInteraph (Tiberti et al., 2014) with the default parameters for angles and distances cut-off. In detail, H-bonds, electrostatic and hydrophobic interactions were translated into distance and angle constraints between two residues, and the percentage of the trajectory in which such constraints were satisfied was calculated by PyInteraph. Then, if such percentage was higher than the persistence threshold (pT), which is based on the size of the largest hydrophobic cluster, the interaction was translated into an edge between the nodes (residues) of the PSN, otherwise the interaction was considered transient and thus discarded. For further details see Refs (Marino and Dell'Orco, 2016; Marino and Dell'Orco, 2019).

The topology of the PSN was analyzed by means of the degree centrality (number of edges reaching a node) to identify the residues involved in the highest number of interactions (hubs) and thus entitled to preserve protein folding and mediating inter/intramolecular communication of state-specific structural information. Differences in hub degree were calculated with respect to the degree of the same residue in the WT, hubs were defined as nodes with at least six edges in at least one of the tested cases.

Regardless of their structural position, intra/intermolecular signaling between two residues of any binding event or conformational change occurs through persistent non-bonded interactions. Therefore, we monitored the differences in intramolecular communication between EF-hands (identified by their representative bidentate Ca^{2+} coordinating residues E31 (EF1), E67 (EF2), E104 (EF3) and E140 (EF4)) using the previously defined Communication Robustness index (Marino and Dell'Orco, 2016), which is calculated as follows:

$$CR(XY) = \frac{nXY \cdot pT}{l}$$

where nXY is the number of shortest paths (of length l) between residues X and Y and pT is the persistence threshold used to filter out transient non-bonded interactions.

3 Results

3.1 Effects of Ca^{2+} and RyR1/2 peptides on secondary and tertiary structure of CaM variants

Like any other calcium sensor protein, CaM is expected to change its secondary and tertiary structure upon binding to Ca^{2+}. Furthermore, interaction of CaM with an essentially unstructured target peptide can induce a specific folding of the latter, as observed in previous studies (Astegno et al., 2016; Dal Cortivo et al., 2022). To assess the structural consequence of Ca^{2+} binding to the three CaM variants in the presence and in the absence of RyR1 and RyR2 peptides we used circular dichroism (CD) spectroscopy.

Near UV CD spectra (Figure 1C) provide information as to the microenvironment of aromatic residues, which are essentially located in the hydrophobic core of a folded protein, and thus represent a fingerprint of protein tertiary structure. CaM lacks tryptophan (W) residues, therefore the apo-to Ca^{2+}-bound transition of the isolated protein was characterized by minor spectral changes in the phenylalanine (F) and tyrosine (Y) bands. Addition of Ca^{2+} led to appreciable spectral changes in the Y region for WT and N97I variants, while the change was less pronounced for the Q135P variant (Figure 1C). Interestingly, addition of RyR1 (blue line) or RyR2 peptides (red line) significantly increased the dichroic signal in the F bands, which became less negative for all three variants. A small positive band in the W region was now visible for all three variants, which could be attributed to the burial of the peptide's W residue upon interaction with CaM. Conversely, the Y region in the presence of RyR1/2 peptides showed a similar pattern in the case of WT CaM, which was significantly perturbed in the mutants' spectra (Figure 1C). Indeed, a fully positive band was observed for N97I in the presence of RyR2 while interaction with the same peptide led to essentially unperturbed spectrum in the case of Q135P, at odds with the effect observed with RyR1, which significantly affected the Y band in both CaM variants. Since both RyR1 and RyR2 peptides possess one Y residue located far from the structural interface (Holt et al., 2020), this effect seems to be related to a slight conformational change occurring at the level of protein tertiary structure rather than at the level of the protein-peptide assembly.

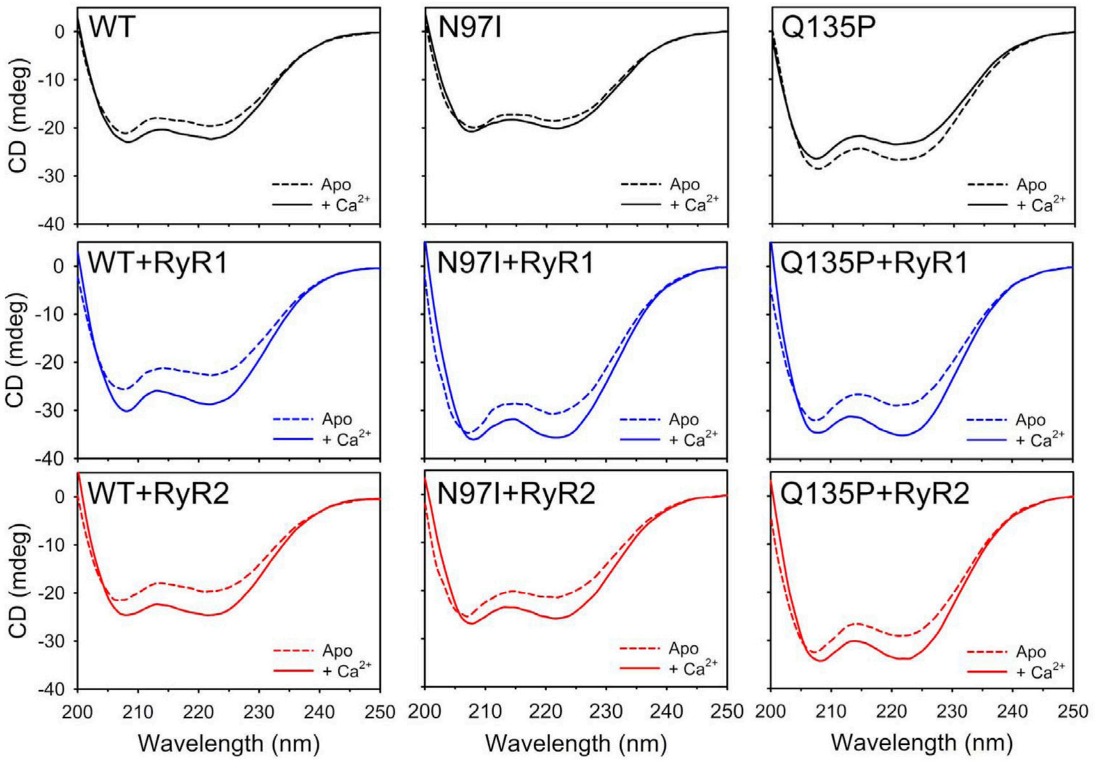


FIGURE 2
Investigation of secondary structure of CaM variants and their RyR1/2 peptide complexes. Far-UV CD spectra (200–250 nm) of 10 μM CaM variants alone (top panels, black), and incubated with 20 μM RyR1 (center panels, blue), or 20 μM RyR2 (bottom panels, red), were collected in the presence of 300 μM EGTA (dashed lines) and after the addition of 600 μM Ca^{2+} (300 μM free Ca^{2+}, solid lines). The spectrum of sole buffer was considered as blank and subtracted; each curve represents the average of five accumulations. Temperature was set at 25°C.

CD spectroscopy in the far UV region was used to monitor changes in protein secondary structure upon Ca^{2+}- and peptide-binding (Figure 2), which in many calcium sensor proteins is accompanied by an increase of the dichroic signal attributed to an increased α-helix content or to the achievement of a more compact structure (Astegno et al., 2014; Marino et al., 2015). Spectral shape was found to depend on the CaM variant (Supplementary Table S1). As observed in our recent study (Dal Cortivo et al., 2022) the two minima at 208 and 222 nm typical of a mainly α-class protein shifted to more negative values upon addition of Ca^{2+} for WT and N97I CaM ($\Delta\theta_{222}/\theta_{222}$ = 18% and 8.6%, respectively) at odds with Q135P which, similarly to other variants carrying point mutations in the EF4 motif, decreased the intensity of the spectrum upon Ca^{2+}-binding ($\Delta\theta_{222}/\theta_{222}$ = −11.6%), suggesting a loss of α-helix content and/or a less compact structure (Figure 2). However, addition of either RyR1 or RyR2 peptides dampened such differences, as all three variants displayed an increase of the dichroism signal (Figure 2, Supplementary Table S1) and a transition towards a coiled-coil conformation as indicated by the $\theta_{222}/\theta_{208}$ values approaching unity in the presence of Ca^{2+} ($\theta_{222}/\theta_{208}$ ranging from 0.96 to 1.02, Supplementary Table S1) as well as a more similar $\Delta\theta_{222}/\theta_{222}$ value (ranging from 16.7% to 22.3%), indicative of the binding-induced folding of the peptide into an α-helix secondary structure.

Digestion with trypsin of isolated CaM variants (Figure 3) confirmed the stabilizing role of Ca^{2+} binding, as judged by the presence of undigested bands for all three variants, which was more pronounced in the presence of Ca^{2+} than in the apo form (Figure 3A,B). Moreover, proteolytic patterns suggested that Q135P and N97I CaM are less stable than the WT in the absence of Ca^{2+} as shown by the higher intensity of the bands at low molecular mass. Moreover, Q135P appeared to be the less stable variant, regardless of the presence of Ca^{2+}.

Overall, these data suggest that the N97I and Q135P substitutions induced small but detectable conformational changes in the secondary and tertiary structure of CaM, and the binding of either RyR1 or RyR2 peptide occurred in any case, leading to similar quaternary structures.

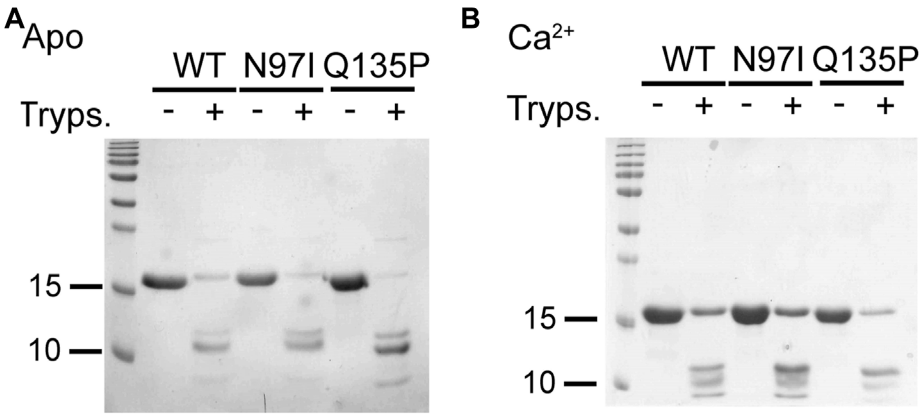


FIGURE 3
Limited proteolysis of CaM variants. CaM variants (26.6 μM) were incubated with .4 μM trypsin (ratio 1:66 trypsin: CaM) in the presence of 2 mM EGTA **(A)** or Ca^{2+} **(B)**. Reactions were run for 10 min at 25°C and then blocked by boiling samples for 10 min at 96°C. Proteolytic fragments were loaded on a 15% SDS-PAGE and Coomassie blue-stained. For each investigated condition, the untreated protein was loaded as a control.

3.2 Affinity for RyR1 and RyR2 peptides of CaM variants assessed by fluorescence spectroscopy and isothermal titration calorimetry

We took advantage of the lack of W residues in CaM and the presence of a single W residue in both RyR1 and RyR2 peptides to monitor the protein-peptide interaction by fluorescence spectroscopy, titrating a fixed amount of peptide with increasing amounts of CaM and evaluating the fraction of the peptide-bound protein in each condition (Figure 4A). This permitted an estimation of the apparent dissociation constant K_D, reported in Table 1, which was used to quantify the affinity of each protein variant for the peptide. Interestingly, WT CaM showed a 2-fold higher affinity for RyR2 than RyR1 (${K_D}^{RyR2}$ = 110 nM vs ${K_D}^{RyR1}$ = 256 nM, respectively; $p < 0.01$), while the two pathogenetic variants showed individually very similar affinities for each peptide (${K_D}^{RyR1}$ = 198 nM vs ${K_D}^{RyR2}$ = 199 nM for N97I; ${K_D}^{RyR1}$ = 269 nM vs ${K_D}^{RyR2}$ = 317 nM for Q135P). The Q135P variant showed a more scattered behavior in the three replicates of RyR2 titrations compared to other variants (Figure 4B). It is worth noting that both pathogenetic variants resulted in significantly higher K_D values when titrated against RyR2 peptides compared to the WT (${K_D}^{WT}$ = 110 nM vs ${K_D}^{N97I}$ = 199 nM, p = 0.013 and ${K_D}^{Q137P}$ = 317 nM, p = 0.056; Figure 4B and Table 1).

Overall, fluorescence spectroscopy measurements suggested that, besides a significantly decreased affinity for the RyR2 target compared to the WT, both N97I and Q135P pathogenetic variants apparently lose the capability to discriminate between RyR1 and RyR2 targets, a characteristic that was clearly observed for WT CaM.

To probe the binding process by an essentially independent approach, we used isothermal titration calorimetry (ITC). In these experiments, either RyR1 or RyR2 peptides were titrated with a fixed amount of CaM variants (Figure 5). In line with our recent observations of other CaM variants titrated with RyR2 (Dal Cortivo et al., 2022), an exothermic binding process characterized by a 1:1 protein-peptide stoichiometry was detected in each case. Although the binding process was enthalpy-driven in each case, changes observed in the steepness of the transition (Figure 5A) indicated a mutation-specific affinity for each variant (Table 2). Interestingly, the binding of both RyR1 and RyR2 to Q135P CaM was entropically unfavored (10.14 kcal/mol vs 6.55 kcal/mol entropic contributions to ΔG, respectively). The affinity of WT CaM for RyR1 and RyR2 peptides was very similar (12.31 nM vs 8.62 nM, respectively, p = 0.34) while, interestingly, N97I showed higher affinity for RyR2 compared to RyR1 (${K_D}^{RyR1}$ = 19.93 nM; ${K_D}^{RyR2}$ = 12.32 nM; p = 0.04). The Q135P pathogenetic variant showed again a more scattered interaction pattern with both peptides compared to other variants (Figure 5B) and the binding affinity was significantly reduced compared to the WT (${K_D}^{RyR1}$ = 62.03 nM, p = 0.016; ${K_D}^{RyR2}$ = 83.45 nM, p = 0.008). ITC experiments thus essentially confirmed the results of fluorescence titrations, despite being less sensitive in detecting specific changes in affinity.

3.3 Mutation-specific effects on CaM-RyR1/2 interaction kinetics

Using a recently optimized procedure (Vallone et al., 2018; Dal Cortivo et al., 2019; Dal Cortivo et al., 2022), we immobilized CaM variants on the surface of a sensor chip through homogeneous histidine tag-mediated coupling. This allowed us to monitor the kinetics of the interaction with RyR1 and RyR2 peptides by surface plasmon resonance

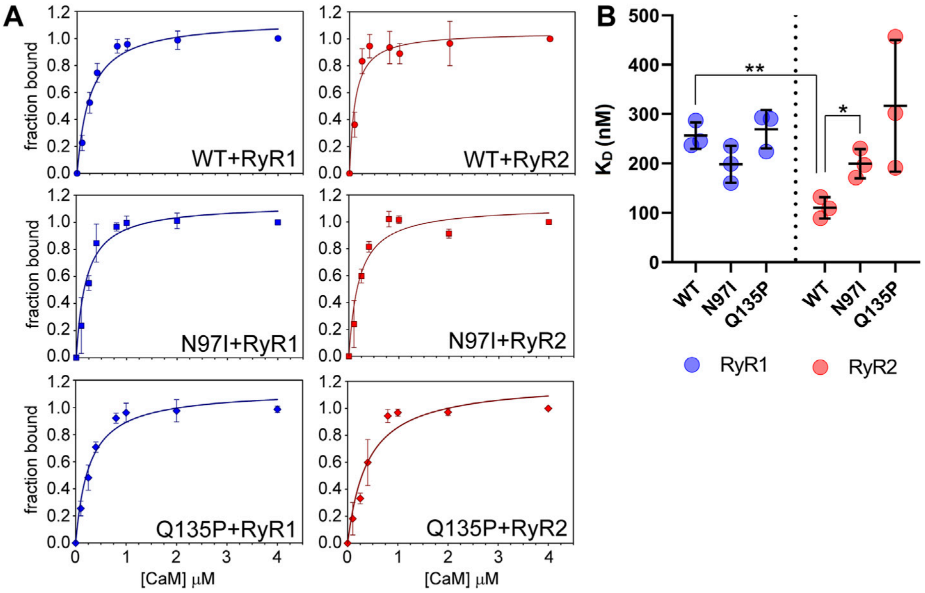


FIGURE 4
CaM affinity for RyR1 and RyR2 peptides assessed by fluorescence spectroscopy. **(A)** One micromolar RyR1 (left column) or RyR2 (right column) peptide was incubated with increasing amounts of WT (top row), N97I (middle row) and Q135P (bottom row) CaM in the presence of 100 µM Ca^{2+}. Data are reported as a function of the peptide fraction bound to CaM (see Materials and Methods for details). Curves report the mean ± std of each point obtained in three technical replicas. Representative fitting to one-site saturation ligand binding curve is superposed to each titration set. **(B)** Scatter plot of replicates reporting the K_D values calculated from fitting procedures in each individual dataset using a one-site binding model. Stars represent *t*-test statistical significance: * *p*-value ≤ 0.05, ** *p*-value ≤ 0.01.

TABLE 1 Apparent affinities of CaM-RyR1/2 complexes assessed by fluorescence spectroscopy.

Variant	K_D^{RyR1} (nM)	K_D^{RyR2} (nM)
WT	256 ± 27	110 ± 22
N97I	198 ± 38	199 ± 30
Q135P	269 ± 39	317 ± 134

under Ca^{2+} saturating conditions. Increasing amounts of peptides were injected for 60 s on the chip where similar levels of CaM variants were previously immobilized to follow the association process, while the dissociation phase was monitored by flowing only running buffer for 300 s (Figure 6A). All sensorgrams could be fitted to a 1:1: Langmuir model compatible with a pseudo-first order association kinetics and a single exponential dissociation kinetics (Figure 6A, blue lines (RyR1) and red lines (RyR2)). Individual association (k^{on}) and dissociation (k^{off}) rate constants are reported in each panel of Figure 6A.

Kinetic analysis revealed an unexpected pattern. On one hand, when comparing the kinetics of CaM interaction with RyR1 and RyR2 peptides it was striking how the association and the dissociation process were unaffected by the presence or absence of point mutation, as the differences between k^{on} and k^{off} values for each RyR peptide were not statistically significant (Figure 6A, *p*-value ranging from 0.05 to 0.89 in all comparisons). However, it should be noted that upon interaction with RyR1 and RyR2, each pathogenic variant showed significantly different association and dissociation kinetics from WT (Figure 6B). Indeed, both N97I and Q135P CaM associated with RyR1, respectively, 19-fold faster (p = 0.027) and 12-fold faster (p = 0.029) than the WT. The same pattern was observed upon interaction with RyR2, where N97I CaM associated with the peptide 13-fold ($p < 0.001$) faster than the WT, while Q135P associated 6-fold faster (p = 0.0045) (Figure 6B). Differences in the dissociation rates of CaM-RyR complexes were even more significant. Both mutants dissociated faster than the WT from RyR1 and RyR2, the most significant differences being observed for the latter peptide ($k^{off}_{N97I}/k^{off}_{WT} = 2.5$, $p < 0.00001$; $k^{off}_{Q135P}/k^{off}_{WT} = 2.9$, $p < 0.0001$). In conclusion, kinetics of CaM-peptide recognition was found to be mutation-dependent, suggesting that selectivity for target is also based on kinetic discrimination.

3.4 Mutation-specific effects on the structural and dynamic properties of the CaM-RyR1/2 complexes monitored by MD simulations and free-energy calculations

To gain atomistic insights on the molecular properties detected for the arrhythmia-associated CaM variants, we ran extensive 1.2 µs MD simulations for each combination of

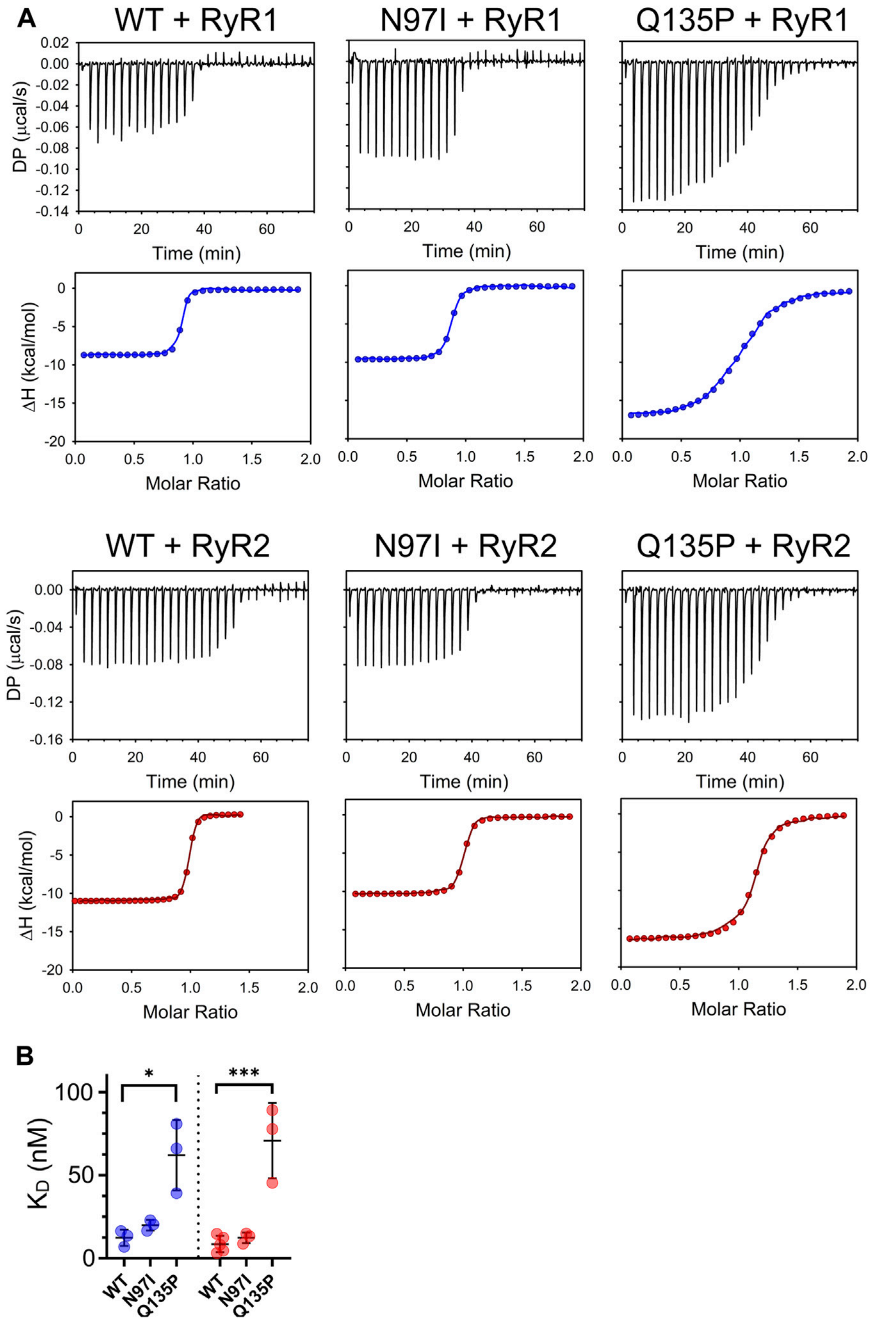


FIGURE 5
Thermodynamics of CaM-RyR1/2 peptide interaction assessed by isothermal titration calorimetry. **(A)** Examples of ITC titration curves obtained for each CaM variant upon interaction with RyR1 or RyR2 peptides. Measurements were performed at 25°C using 20 mM Tris pH 7.5, 150 mM KCl, 5 mM Ca^{2+} as working buffer and setting stirring at 750 rpm. Each titration consisted in thirty 1-µL injections of 125 µM RyR1 or RyR2 (into the titrant syringe) with 10 µM CaM variants. **(B)** Scatter plot of replicates summarizing the K_D values calculated from the fitting using a one-site binding model (see Materials and Methods). Stars represent the *p*-values: *$p \leq 0.05$, ***$p \leq 0.001$. Data for WT CaM titration with RyR2 are from (Dal Cortivo et al., 2022).

CaM variants complexed with RyR peptides. The convergence and consistency of such simulations was assessed by analyzing the conformational space sampled by each replica using metrics derived by principal component analysis (PCA) of Cα motions, as described in Methods section. The overlap of the projection of the

TABLE 2 Thermodynamics of CaM-RyR1/2 peptide association assessed by isothermal titration calorimetry.

Variant (n)	K_D (nM)	ΔH (kcal/mol)	ΔG (kcal/mol)	-TΔS (kcal/mol)	ΔΔG (kcal/mol)
RyR1					
WT (3)	12.31 ± 4.87	−9.94 ± 1.23	−10.78	−0.84	—
N97I (3)	19.93 ± 3.23	−9.58 ± 0.16	−10.50	−0.92	0.28
Q135P (4)	62.03 ± 21.13	−19.97 ± 4.37	−9.83	10.14	0.95
RyR2					
WT[a] (5)	8.62 ± 2.20	−11.28 ± 0.62	−11.01	0.27	—
N97I (3)	12.32 ± 3.18	−10.02 ± 0.13	−10.78	−0.76	0.23
Q135P (3)	83.45 ± 7.99	−16.20 ± 0.14	−9.65	6.55	1.36

n represents the number of independent titrations performed for each variant/peptide combination

[a]These data are taken from (Dal Cortivo et al., 2022).

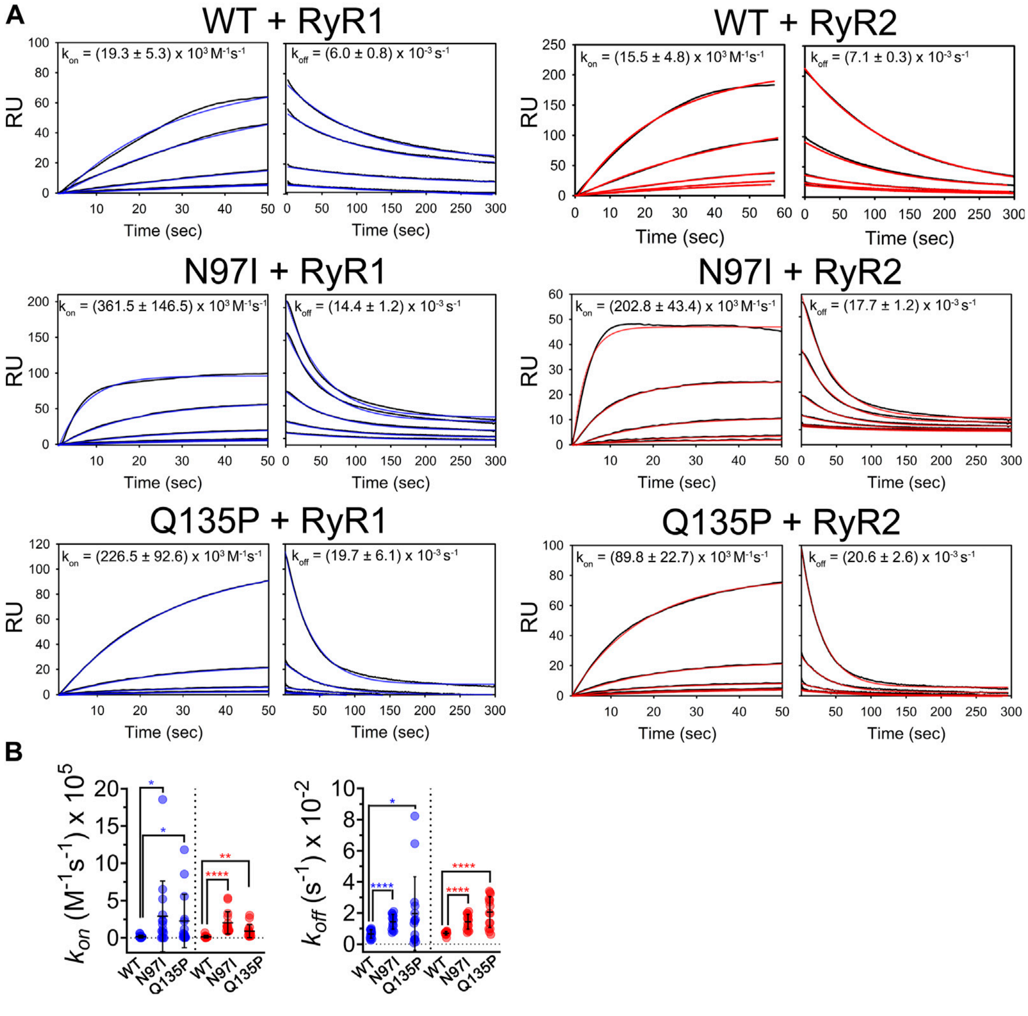


FIGURE 6
Kinetics of CaM-RyR1/2 peptide interaction investigated by surface plasmon resonance. **(A)** Sensorgrams collected by flowing different amounts of RyR1 and RyR2 (125 nM—2 μM) on immobilized His-CaM variants using 20 mM Tris pH 7.5, 150 mM KCl, 0.005% Tween 20, 5 mM Ca^{2+}, 100 μM DTT as a running buffer. Association and dissociation phases were followed for 60 s and 300 s, respectively. Experimental curves (black solid lines) are shown together with theoretical curves (red or blue solid lines) according to a 1:1 Langmuir binding model; fitting for association and dissociation phases led to the rate constants (k^{on} and k^{off}) reported in each panel (mean ± s.e.m. of 8-20 independent binding curves). **(B)** Scatter plot of replicates and statistical analysis comparing rate constants for WT and each pathogenic CaM variant. Stars represent *p*-values: **p* ≤ 0.05, ***p* ≤ 0.01, *****p* ≤ 0.0001.

conformations sampled by the trajectories onto their first two principal components (Supplementary Figure S1), as well as of the density plots defined by the LDA classifier (Supplementary Figure S2) indicate that most conformations were accessible from different replicas, thus implying that all the replicas of each simulation were consistent. The same conclusion could be drawn from the comparison of the RMSIP (Supplementary Figure S3), which in all cases was found to be higher than 0.782, confirming that the replicas of MD simulations were reproducible and consistent, and therefore could be concatenated for further analyses.

The Cα-Root-Mean Square Fluctuation (RMSF) is a convenient index to evaluate the flexibility of the backbone of proteins or protein complexes, as it represents the root-mean square deviation of Cα atoms from the average structure mediated over the simulated timeframe. RMSF profiles suggest that CaM variants complexed to RyR1 peptide are very similar to the WT in terms of flexibility, with the largest differences observed in the N-terminal lobe (Figure 7, top panel) in particular for the Q135P substitution. The flexibility of the C-terminal domain was virtually unaffected by the mutations, which is surprising, as both mutations affect residues located at the C-terminal domain. A similar situation was observed with the CaM-RyR2 peptide complex, where a significantly higher flexibility was detected in the N-terminal lobe for both variants. Interestingly, the effects of the mutations on the flexibility of the backbone were significantly larger in this case, and, at odds with the CaM-RyR1 complex, the N97I variant exhibited the most prominent destabilization of the N-lobe. In contrast, the flexibility of both RyR1 and RyR2 peptides within each complex was essentially unaltered by the presence of pathogenetic variants, except for a lower flexibility shown by N97I CaM in complex with RyR1 (Figure 7, bottom panels).

The RMSF index can also be computed on Ca^{2+} ions to evaluate the tendency of cations to dislocate from the optimal bound geometry following mutations in the coordinating groups. Potential changes in cation binding affinity indeed reflect on the RMSF index, as a tight interaction of Ca^{2+} ions with its coordinating groups essentially leads to smaller fluctuations, while a loss of affinity would result in increased Ca^{2+} mobility around the optimal position (Marino et al., 2018). As to the CaM-RyR1 complex, both CaM variants exhibited no differences in Ca^{2+}-coordination in EF3 and EF4 compared with WT CaM, even though substitutions were localized in these motifs (Supplementary Table S2). On the other hand, Ca^{2+}-binding to EF1 and EF2 was significantly affected by the variants, suggesting an allosteric enhancement of the Ca^{2+}-affinity in the case of the N97I mutant, at odds with the behavior displayed by the Q135P mutation (Supplementary Table S2). Overall, the Ca^{2+}-RMSF profiles of CaM-RyR2 complexes suggested that Ca^{2+} may be significantly more loosely bound (ΔRMSF ~0.2 Å) with respect to CaM-RyR1 complexes in all simulated cases (Supplementary Table S2). While differences were almost negligible for the EF3-bound Ca^{2+} ion, fluctuations in EF2 and EF4 were significantly larger, suggesting a local, as well as a long-range negative effect of the mutations on Ca^{2+}-affinity. Interestingly, in the presence of the Q135P substitution, also the Ca^{2+}-ion in EF1 exhibited higher RMSF values compared to the WT, pointing again towards an allosteric rearrangement of the Ca^{2+}-binding motif.

To dissect the effects of point mutations on the energetics of protein-peptide folding *versus* binding processes, we used a simplified thermodynamic cycle starting from the conformations sampled by MD simulations. This allowed us to estimate the apparent relative free energy of folding ($\Delta\Delta G^{f}_{app}$) and binding ($\Delta\Delta G^{b}_{app}$) for each variant/peptide (see Methods). The analysis suggested a stabilizing effect on the protein-peptide complex for the N97I substitution, both for RyR1 and RyR2 peptides ($\Delta\Delta G^{f}_{app}$ = −24.10 kcal/mol and -5.58 kcal/mol, respectively, Supplementary Table S3). An opposite trend was detected for the Q135P substitution, which was predicted to significantly destabilize both complexes ($\Delta\Delta G^{f}_{app}$ = 44.92 kcal/mol and 41.48 kcal/mol, respectively, Supplementary Table S3). Noteworthy, CaM variants carrying the N97I or Q135P substitutions displayed both a moderately increased apparent affinity for RyR1 ($\Delta\Delta G^{b}_{app}$ = −0.47 kcal/mol and −0.15 kcal/mol, respectively), and moderately decreased apparent affinity for RyR2 ($\Delta\Delta G^{b}_{app}$ = 0.13 kcal/mol and 0.53 kcal/mol), partly in line with fluorescence titration data. Taken together, modeling results suggest that CaM pathogenic variants perturb both the affinity for RyR peptides and the stability of the protein-peptide complex, but the specific effect appears to be mutation-dependent.

3.5 The topology of the protein-peptide structure network is altered by arrhythmia-associated mutations

The Protein Structure Network (PSN) defined by persistent non-bonded interactions between residues that occur during exhaustive MD simulations enables the study of the structural dynamics and the allosteric properties of Ca^{2+}-sensor proteins (Marino and Dell'Orco, 2016; Marino and Dell'Orco, 2019). In addition, comparison of variant-specific PSNs provides information on the rewiring of those PSNs due to the presence of the mutation and thus the effects on the network topology.

Analysis of the variant-specific and target-specific PSN topology allows the identification of residues (hubs) involved in more than four persistent non-bonded interactions (hub degree) and thus responsible for preservation of protein structure,

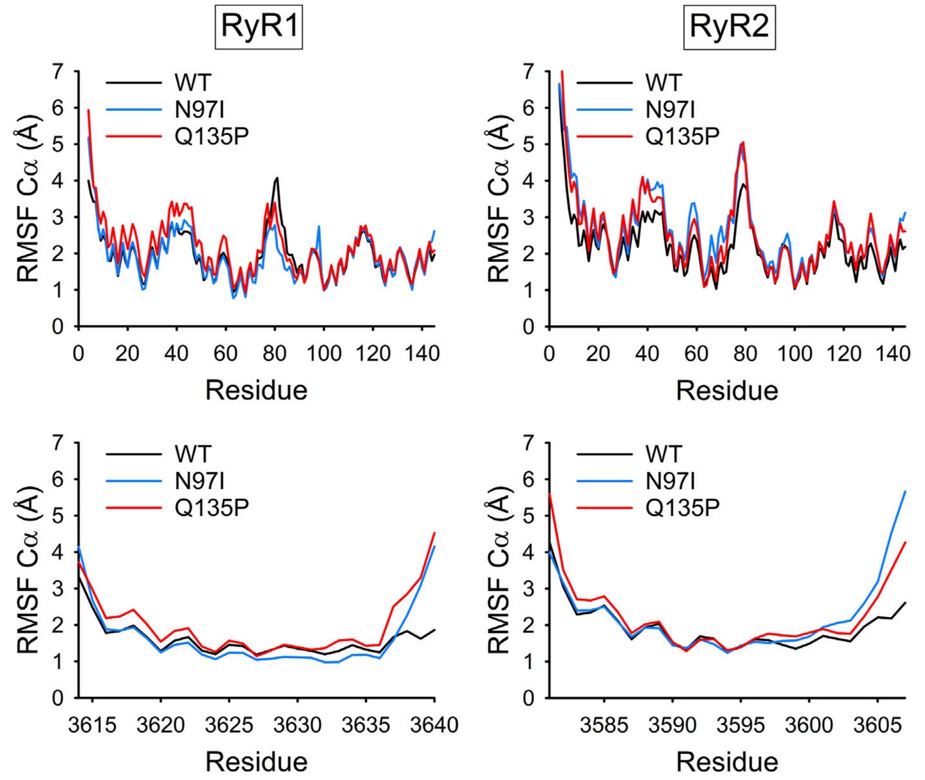


FIGURE 7
Backbone flexibility of CaM-RyR1/2 complexes. Cα-Root-Mean Square Fluctuation (RMSF) of CaM (top panels) and RyR1 or RyR2 peptides (bottom panels) calculated over 1.2 μs MD simulations of the respective complex (WT: black, N97I: blue, and Q135P: red).

dynamics, and intra/intermolecular communications. Mutations affecting hub residues are more likely to destabilize the entire network than those affecting "peripheral" (i.e., less central) residues, involved in three or fewer interactions. Analysis of the change in connectivity of hub residues with degree ≥6 (Figure 8, top) suggested that in the presence of RyR1 the N97I variant slightly decreased the overall connectivity of these high-degree hubs (ΔDegree = −2), with residues M51, F89, and L105 losing one interaction and residue M36 gaining one interaction compared with WT. On the other hand, the Q135P substitution significantly reduced the overall connectivity of the hubs (ΔDegree = −9), with only F19 showing an additional interaction compared to the WT. As for the CaM-RyR2 complex, both N97I and Q135P variants surprisingly increased the overall degree of the hubs (ΔDegree = 14 and 19, respectively), in clear contrast to the effects exerted on the CaM-RyR1 complex (Figure 8, top).

Considering only the hub residues belonging to RyR peptides, no change in overall connectivity in the CaM-RyR1 complex could be appreciated for both variants, with only two residues showing a ΔDegree = ± 1 in both cases (Figure 8, bottom). In contrast, in the CaM-RyR2 complex both variants exhibited a general increase in hub connectivity (ΔDegree = 6 for N97I and 9 for Q135P), particularly in those located at the C-terminal of the peptide, namely R3595, V3599, F3603, A3606.

3.6 Unlike WT CaM, pathogenic variants are unable to discriminate between binding to RyR1 and RyR2

The allosteric properties of CaM should depend closely on its Ca^{2+}-binding characteristics and are expected to reflect the existence of connecting routes of non-bonded interactions between its EF-hand Ca^{2+}-binding motifs. We therefore searched for the existence of such persistent communication paths over the time frame of MD simulations. To this end, the intramolecular communication between the four EF-hands of CaM, represented by the 12th residues (glutamate) of the Ca^{2+}-binding loops was assessed by the Communication Robustness (CR) index, which takes into account the number and the length of the shortest paths between residues (see Methods).

In the CaM-RyR1 peptide complexes, all variants showed a robust communication between EF1-EF2, EF1-EF4, and EF2-EF4 (Figure 9A), and no robust communication between EF1-EF3 and EF2-EF3 (CR threshold for significant robustness was

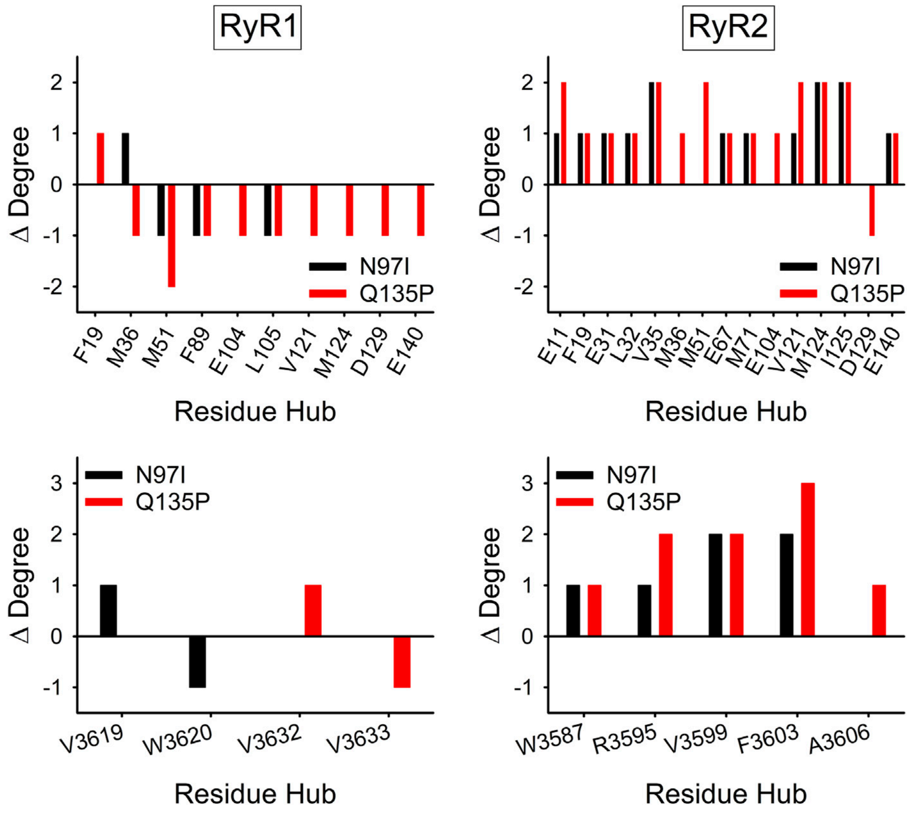


FIGURE 8
Effects of CaM variants on the connectivity of hub residues of CaM (top panels) and RyR1/2 peptides (bottom panels). Hubs were defined as residues with degree ≥6 in the Protein Structure Network (PSN) of at least one variant. ΔDegree is calculated as the difference in hub degree between the variant (N97I: black, Q135P: red) and the WT.

set at 0.1). Interestingly, the only significant difference between WT and pathogenic CaM variants was observed in the EF3-EF4 communication, which was absent in the WT (CR < 0.1) but very robust in the case of both variants (CR > 0.3). When the same analysis was performed for CaM-RyR2 peptide complex, a pattern of higher robustness in communication between EF-hand pairs was found for both pathogenic variants, which showed higher CR indexes in all cases except for EF1-EF3 communication. While qualitatively in line with the increased connectivity observed for CaM-RyR2 hubs in the presence of pathogenic mutations (Figure 9B), this analysis revealed another important feature in terms of specificity in target recognition. Indeed, WT CaM was the only variant able to discriminate RyR1 and RyR2 targets through the appearance of allosteric communication between EF1-EF3 and EF3-EF4 (Figure 9B), as the communication between this specific EF-hand pairs was very significant (CR ~ 0.4) only when in complex with RyR2, whereas it was essentially lost when binding to RyR1 (CR < 0.2, Figure 9B). In contrast, in the presence of either RyR peptide, both N97I and Q135P CaM variants showed non-robust communication between EF1-EF3, and a highly robust (CR > 0.3) communication between EF3-EF4, thus losing any ability to discriminate between targets.

4 Discussion

The redundancy of genes encoding CaM in the human genome, its high conservation among the reigns, and its promiscuity in terms of molecular partners all suggest that missense mutations in the genes encoding CaM be deleterious and probably incompatible with life. However, the finding that 17 CaM-missense mutations are associated with lethal arrhythmia and lead to no other apparent phenotype suggests the existence of a specific effect of disease-associated CaM variants on a target that is intimately connected with pathology. In other words, CaM variants associated with LQTS and CPVT must recognize RyR channels in skeletal (RyR1) and cardiac (RyR2) muscle cells differently from WT CaM. Several lines of evidence suggest that the CaM mutants analyzed in this study (N97I and Q135P) cause arrhythmia through dysregulation of RyR2 function. Indeed, the N97I substitution results in an altered interaction with the IQ

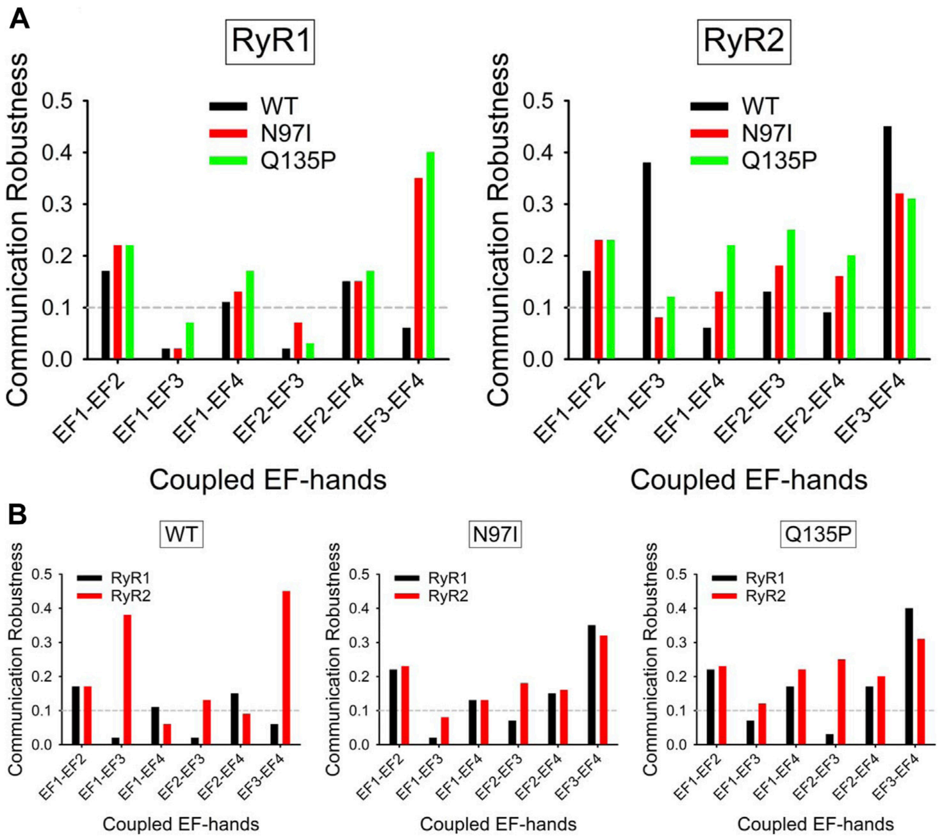


FIGURE 9
Robustness of intramolecular communication between EF-hands in CaM-RyR1/2 complexes. **(A)** Intramolecular communication among EF-hands in the CaM-RyR1 (left) and CaM-RyR2 (right) complexes. Communication robustness between EF-hands (EF1 to EF4, represented by their bidentate Glu residues) in CaM WT (black), N97I (red), and Q135P (green) variants in complex with RyR peptides. **(B)** Effects of specific RyR peptides on the intramolecular communication among EF-hands in CaM-RyR complexes. Communication robustness between EF-hands (EF1 to EF4, represented by their bidentate Glu residues) in CaM variants in complex with RyR1 (black) and RyR2 (red) peptides.

domain of the cardiac voltage-gated calcium channel (Wang et al., 2018), but on the other hand, another substitution at the same locus, i.e. N97S, was found to affect RyR2-mediated calcium release, specifically increasing it compared to WT CaM (Sondergaard et al., 2015). A direct perturbation of the interaction with RyR2 is therefore expected for the N97I variant. Similarly, the Q135P substitution was found to affect binding to the IQ domain of Cav1.2 (Wang et al., 2020), but also the activation threshold of RyR2, thus likely promoting spontaneous Ca^{2+} release in cardiomyocytes during diastoles (Sondergaard et al., 2019). The involvement of these CaM variants in RyR2 dysregulation thus appears substantial.

Our structural analysis clearly showed that the N97I and Q135P substitutions lead to small but appreciable alterations of protein secondary and tertiary structure. Such effects are substantially in line with previous observations by NMR and X-ray crystallography that individual mutations perturb the conformation of CaM and its interaction with the target peptide, in a mutation-specific manner (Wang et al., 2018; Dal Cortivo et al., 2019). A common fingerprint linking arrhythmia-associated variants N53I in EF2 (Holt et al., 2020) with D95V/H in EF3 and D131V/H/E in EF4 (Dal Cortivo et al., 2022) seems to be the fact that the single mutation can perturb to various extent protein tertiary structure, but all variants alter the protein intramolecular dynamics, affecting and destabilizing the N-terminal domain, which has been suggested to permit the functional recognition of RyR2 target *via* allosteric interactions (Westerlund and Delemotte, 2018; Dal Cortivo et al., 2022). Clearly, understanding the ability of CaM variants to discriminate between specific targets requires a complementary approach to the structural analysis, and requires full characterization of the binding process. Considering the very high sequence similarity between RyR1 and RyR2 CaM-binding domains investigated in this study, we performed an exhaustive characterization using several techniques.

The W residue in the RyR2 (and RyR1) peptide mimicking the channel region interacting with CaM has been shown to be essential for the molecular recognition both at low and high Ca^{2+} levels (Brohus et al., 2019). Titrations experiments based on intrinsic W fluorescence showed that WT CaM binds RyR2 with double affinity compared to RyR1, which is quite surprising considering the minor difference in the peptide sequences, and indicates a high target selectivity. On the other hand, sensitivity of ITC experiments did not allow to distinguish the affinity of WT CaM for RyR1 and RyR2 peptides. Care should be taken when interpreting K_D values obtained by ITC. It has previously been shown for calcium sensor proteins, which undergo significant conformational changes upon Ca^{2+} and target binding, that the apparent K_D value obtained by ITC may not represent the true dissociation constant. In fact, in ITC titrations, an enthalpic factor related to conformational changes is inevitably added to the enthalpy change of the binding process (Dell'Orco et al., 2010; Dell'Orco et al., 2012) and cannot be distinguished from the pure ΔH of binding. Since RyR peptides fold upon binding, it is thus not surprising that ITC does not fully discriminate between different RyRs. Our ITC results are also qualitatively in line with previous ones obtained by (Lau et al., 2014), who measured a 46 nM affinity for RyR1 and 47 nM affinity for RyR2. The slight differences can be attributed to different experimental conditions, including shorter RyR1 and RyR2 peptides lacking the -LYNL sequence at the C-terminal and different buffer used in (Lau et al., 2014). ITC patterns in our study and the emerging thermodynamics are also similar to those obtained for N53I (Holt et al., 2020) and F141L (Sondergaard et al., 2017) arrhythmogenic CaM variants binding to RyR2 peptides and with the results obtained very recently by us with the D95V/H, D129V, and D131H/E arrhythmogenic variants (Dal Cortivo et al., 2022).

While useful to reveal intrinsic perturbations in the CaM-peptide binding thermodynamics associated with pathogenic mutations, ITC experiments are evidently insufficient to fully describe the recognition between CaM and very similar targets, and point to the necessity of going beyond characterization of the equilibrium to assess the determinants of target discrimination. Kinetic characterization of the interaction between CaM variants and RyR1 and RyR2 peptides showed that both N97I and Q135P pathogenic variants associated with RyR1 and RyR2 peptides significantly faster than the WT, however their dissociation from the RyR2 peptide was more than two-fold faster than WT CaM, suggesting the involvement of kinetic discrimination in target selectivity.

A deeper insight into the recognition process emerged from MD simulations and analysis of RMSF profiles, which suggested that Ca^{2+}-binding affinities in specific EF-hands can be modulated by the pathogenic variants even when mutations are located far away from the binding sites. This would be possible if allosteric mechanisms mediated the structural communication between EF-hands, not only within the same domain, but also between different domains. It is relevant that both N97I and Q135P substitutions, located respectively in EF3 and EF4, reflect in a perturbation of the flexibility of EF2 and EF1 Ca^{2+} binding loops in the presence of RyR1 and RyR2 peptide, and specifically, increase the flexibility in this latter case. This finding is especially interesting considering previous results that suggest that the functional recognition between CaM and its RyR target involves allosteric interactions initiated by the N-terminal lobe of CaM (Dal Cortivo et al., 2022), in spite of its lower affinity for Ca^{2+}. It has been postulated that long-range electrostatic interactions of amino acids at the N-terminal domain are responsible for initiating CaM binding to the target, while short-range hydrophobic interactions in the C-terminal lobe may account for selectivity (Westerlund and Delemotte, 2018). Our simulations indeed support this view.

Interesting topological properties, with deep implications for target selectivity emerged from PSN analysis. Our data indeed indicate as a peculiarity of the pathogenic variants the significantly increased connectivity in the PSN formed by the complex of CaM with RyR2, which is the target channel in pathologic conditions. This increased connectivity evidently extends beyond the intramolecular communication and reaches protein-target intermolecular contacts, thus inducing a complete rewiring of the network in the co-presence of pathogenic mutations and cell-specific target. Finally, topological analysis of the structural dynamics that emerged from MD simulations suggests that the ability of WT CaM to discriminate highly similar RyR1 and RyR2 targets is based on the robust allosteric communication between EF1-EF3 and EF3-EF4 pairs in CaM-RyR2 recognition, the former of which is lost in both N97I and Q135P variants.

In conclusion, we presented an in-depth characterization of the recognition between a common CaM-binding region in RyR1 and RyR2 and two arrhythmia-associated CaM variants in their Ca^{2+}-bound states. In a broader context, our analysis suggests that a complex combination of factors may influence the discrimination ability of CaM towards its many targets, which includes kinetic discrimination and specific allosteric communication between its Ca^{2+}-binding EF-hand motifs.

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

GDC designed and performed the *in vitro* experiments and analyzed the results. VM designed and performed molecular modeling and simulations and analyzed the results. SB contributed to protein expression and purification and discussion. DDO conceived and coordinated the study, contributed to analyze the data and wrote the manuscript with contributions from all the authors.

Funding

This work has received funding from the Italian Ministry of University and Research (PRIN 2017, 201744NR8S to DDO) and support from CINECA through the Italian Super Computing Resource Allocation project (ISCRA Grant: HP10CLGGAV to VM). Partial support by Fondo Sociale Europeo REACT EU—Programma Operativo Nazionale “Ricerca e innovazione” 2014–2020 (Project “Metalli pesanti e degenerazioni retiniche: geni o ambiente?”) is acknowledged.

Acknowledgments

Centro Piattaforme Tecnologiche of the University of Verona is acknowledged for providing access to the spectroscopic and computational platforms.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1100992/full#supplementary-material

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1, 19–25. doi:10.1016/j.softx.2015.06.001

Astegno, A., La Verde, V., Marino, V., Dell'orco, D., and Dominici, P. (2016). Biochemical and biophysical characterization of a plant calmodulin: Role of the N- and C-lobes in calcium binding, conformational change, and target interaction. *Biochim. Biophys. Acta* 1864, 297–307. doi:10.1016/j.bbapap.2015.12.003

Astegno, A., Maresi, E., Marino, V., Dominici, P., Pedroni, M., Piccinelli, F., et al. (2014). Structural plasticity of calmodulin on the surface of CaF2 nanoparticles preserves its biological function. *Nanoscale* 6, 15037–15047. doi:10.1039/c4nr04368e

Borsatto, A., Marino, V., Abrusci, G., Lattanzi, G., and Dell’orco, D. (2019). Effects of membrane and biological target on the structural and allosteric properties of recoverin: A computational approach. *Int. J. Mol. Sci.* 20, 5009. doi:10.3390/ijms20205009

Brohus, M., Sondergaard, M. T., Wayne Chen, S. R., Van Petegem, F., and Overgaard, M. T. (2019). Ca(2+)-dependent calmodulin binding to cardiac ryanodine receptor (RyR2) calmodulin-binding domains. *Biochem. J.* 476, 193–209. doi:10.1042/BCJ20180545

Capes, E. M., Loaiza, R., and Valdivia, H. H. (2011). Ryanodine receptors. *Skelet. Muscle* 1, 18. doi:10.1186/2044-5040-1-18

Carafoli, E. (2002). Calcium signaling: A tale for all seasons. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1115–1122. doi:10.1073/pnas.032427999

Chattopadhyaya, R., Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). Calmodulin structure refined at 1.7 A resolution. *J. Mol. Biol.* 228, 1177–1192. doi:10.1016/0022-2836(92)90324-d

Chi, X., Gong, D., Ren, K., Zhou, G., Huang, G., Lei, J., et al. (2019). Molecular basis for allosteric regulation of the type 2 ryanodine receptor channel gating by key modulators. *Proc. Natl. Acad. Sci. U. S. A.* 116, 25575–25582. doi:10.1073/pnas.1914451116

Chin, D., and Means, A. R. (2000). Calmodulin: A prototypical calcium sensor. *Trends Cell Biol.* 10, 322–328. doi:10.1016/s0962-8924(00)01800-6

Dal Cortivo, G., Barracchia, C. G., Marino, V., D’onofrio, M., and Dell’orco, D. (2022). Alterations in calmodulin-cardiac ryanodine receptor molecular recognition in congenital arrhythmias. *Cell Mol. Life Sci.* 79, 127. doi:10.1007/s00018-022-04165-w

Dal Cortivo, G., Marino, V., Iacobucci, C., Vallone, R., Arlt, C., Rehkamp, A., et al. (2019). Oligomeric state, hydrodynamic properties and target recognition of human Calcium and Integrin Binding protein 2 (CIB2). *Sci. Rep.* 9, 15058. doi:10.1038/s41598-019-51573-3

Dell’Orco, D., Muller, M., and Koch, K. W. (2010). Quantitative detection of conformational transitions in a calcium sensor protein by surface plasmon resonance. *Chem. Commun. (Camb)* 46, 7316–7318. doi:10.1039/c0cc02086a

Dell’Orco, D., Sulmann, S., Linse, S., and Koch, K. W. (2012). Dynamics of conformational Ca2+-switches in signaling networks detected by a planar plasmonic device. *Anal. Chem.* 84, 2982–2989. doi:10.1021/ac300213j

Gong, D., Chi, X., Wei, J., Zhou, G., Huang, G., Zhang, L., et al. (2019). Modulation of cardiac ryanodine receptor 2 by calmodulin. *Nature* 572, 347–351. doi:10.1038/s41586-019-1377-y

Holt, C., Hamborg, L., Lau, K., Brohus, M., Sorensen, A. B., Larsen, K. T., et al. (2020). The arrhythmogenic N53I variant subtly changes the structure and dynamics in the calmodulin N-terminal domain, altering its interaction with the cardiac ryanodine receptor. *J. Biol. Chem.* 295, 7620–7634. doi:10.1074/jbc.RA120.013430

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2017). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14, 71–73. doi:10.1038/nmeth.4067

Jensen, H. H., Brohus, M., Nyegaard, M., and Overgaard, M. T. (2018). Human calmodulin mutations. *Front. Mol. Neurosci.* 11, 396. doi:10.3389/fnmol.2018.00396

Lau, K., Chan, M. M., and Van Petegem, F. (2014). Lobe-specific calmodulin binding to different ryanodine receptor isoforms. *Biochemistry* 53, 932–946. doi:10.1021/bi401502x

Linse, S., Helmersson, A., and Forsen, S. (1991). Calcium binding to calmodulin and its globular domains. *J. Biol. Chem.* 266, 8050–8054. doi:10.1016/s0021-9258(18)92938-8

Marino, V., Dal Cortivo, G., Oppici, E., Maltese, P. E., D’esposito, F., Manara, E., et al. (2018). A novel p.(Glu111Val) missense mutation in GUCA1A associated with cone-rod dystrophy leads to impaired calcium sensing and perturbed second messenger homeostasis in photoreceptors. *Hum. Mol. Genet.* 27, 4204–4217. doi:10.1093/hmg/ddy311

Marino, V., and Dell'Orco, D. (2016). Allosteric communication pathways routed by Ca(2+)/Mg(2+) exchange in GCAP1 selectively switch target regulation modes. *Sci. Rep.* 6, 34277. doi:10.1038/srep34277

Marino, V., and Dell'Orco, D. (2019). Evolutionary-conserved allosteric properties of three neuronal calcium sensor proteins. *Front. Mol. Neurosci.* 12, 50. doi:10.3389/fnmol.2019.00050

Marino, V., Sulmann, S., Koch, K. W., and Dell'orco, D. (2015). Structural effects of Mg^{2+} on the regulatory states of three neuronal calcium sensors operating in vertebrate phototransduction. *Biochim. Biophys. Acta* 1853, 2055–2065. doi:10.1016/j.bbamcr.2014.10.026

Newman, R. A., Van Scyoc, W. S., Sorensen, B. R., Jaren, O. R., and Shea, M. A. (2008). Interdomain cooperativity of calmodulin bound to melittin preferentially increases calcium affinity of sites I and II. *Proteins* 71, 1792–1812. doi:10.1002/prot.21861

Nyegaard, M., Overgaard, M. T., Sondergaard, M. T., Vranas, M., Behr, E. R., Hildebrandt, L. L., et al. (2012). Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *Am. J. Hum. Genet.* 91, 703–712. doi:10.1016/j.ajhg.2012.08.015

Peng, W., Shen, H., Wu, J., Guo, W., Pan, X., Wang, R., et al. (2016). Structural basis for the gating mechanism of the type 2 ryanodine receptor RyR2. *Science* 354, aah5324. doi:10.1126/science.aah5324

Sondergaard, M. T., Liu, Y., Brohus, M., Guo, W., Nani, A., Carvajal, C., et al. (2019). Diminished inhibition and facilitated activation of RyR2-mediated Ca(2+) release is a common defect of arrhythmogenic calmodulin mutations. *FEBS J.* 286, 4554–4578. doi:10.1111/febs.14969

Sondergaard, M. T., Liu, Y., Guo, W., Wei, J., Wang, R., Brohus, M., et al. (2020). Role of cardiac ryanodine receptor calmodulin-binding domains in mediating the action of arrhythmogenic calmodulin N-domain mutation N54I. *FEBS J.* 287, 2256–2280. doi:10.1111/febs.15147

Sondergaard, M. T., Liu, Y., Larsen, K. T., Nani, A., Tian, X., Holt, C., et al. (2017). The arrhythmogenic calmodulin p.Phe142Leu mutation impairs C-domain Ca2+ binding but not calmodulin-dependent inhibition of the cardiac ryanodine receptor. *J. Biol. Chem.* 292, 1385–1395. doi:10.1074/jbc.M116.766253

Sondergaard, M. T., Tian, X., Liu, Y., Wang, R., Chazin, W. J., Chen, S. R., et al. (2015). Arrhythmogenic calmodulin mutations affect the activation and termination of cardiac ryanodine receptor-mediated Ca2+ release. *J. Biol. Chem.* 290, 26151–26162. doi:10.1074/jbc.M115.676627

Sorensen, A. B., Sondergaard, M. T., and Overgaard, M. T. (2013). Calmodulin in a heartbeat. *FEBS J.* 280, 5511–5532. doi:10.1111/febs.12337

Stevens, F. C. (1983). Calmodulin: An introduction. *Can. J. Biochem. Cell Biol.* 61, 906–910. doi:10.1139/o83-115

Theoharis, N. T., Sorensen, B. R., Theisen-Toupal, J., and Shea, M. A. (2008). The neuronal voltage-dependent sodium channel type II IQ motif lowers the calcium affinity of the C-domain of calmodulin. *Biochemistry* 47, 112–123. doi:10.1021/bi7013129

Tiberti, M., Invernizzi, G., Lambrughi, M., Inbar, Y., Schreiber, G., and Papaleo, E. (2014). PyInteraph: A framework for the analysis of interaction networks in structural ensembles of proteins. *J. Chem. Inf. Model* 54, 1537–1551. doi:10.1021/ci400639r

Valeyev, N. V., Bates, D. G., Heslop-Harrison, P., Postlethwaite, I., and Kotov, N. V. (2008). Elucidating the mechanisms of cooperative calcium-calmodulin interactions: A structural systems biology approach. *BMC Syst. Biol.* 2, 48. doi:10.1186/1752-0509-2-48

Vallone, R., Dal Cortivo, G., D'onofrio, M., and Dell'orco, D. (2018). Preferential binding of Mg(2+) over Ca(2+) to CIB2 triggers an allosteric switch impaired in usher syndrome type 1J. *Front. Mol. Neurosci.* 11, 274. doi:10.3389/fnmol.2018.00274

Van Petegem, F. (2015). Ryanodine receptors: Allosteric ion channel giants. *J. Mol. Biol.* 427, 31–53. doi:10.1016/j.jmb.2014.08.004

Wang, K., Brohus, M., Holt, C., Overgaard, M. T., Wimmer, R., and Van Petegem, F. (2020). Arrhythmia mutations in calmodulin can disrupt cooperativity of Ca(2+) binding and cause misfolding. *J. Physiol.* 598, 1169–1186. doi:10.1113/JP279307

Wang, K., Holt, C., Lu, J., Brohus, M., Larsen, K. T., Overgaard, M. T., et al. (2018). Arrhythmia mutations in calmodulin cause conformational changes that affect interactions with the cardiac voltage-gated calcium channel. *Proc. Natl. Acad. Sci. U. S. A.* 115, E10556–E10565. doi:10.1073/pnas.1808733115

Westerlund, A. M., and Delemotte, L. (2018). Effect of Ca2+ on the promiscuous target-protein binding of calmodulin. *PLoS Comput. Biol.* 14, e1006072. doi:10.1371/journal.pcbi.1006072

Zhang, M., Tanaka, T., and Ikura, M. (1995). Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nat. Struct. Biol.* 2, 758–767. doi:10.1038/nsb0995-758

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 10 May 2023
DOI 10.3389/fmolb.2023.1169109

Resources and tools for rare disease variant interpretation

Luana Licata[1†‡], Allegra Via[2†‡], Paola Turina[3*‡], Giulia Babbi[3‡], Silvia Benevenuta[4‡], Claudio Carta[5‡], Rita Casadio[3‡], Andrea Cicconardi[6,7], Angelo Facchiano[8‡], Piero Fariselli[4‡], Deborah Giordano[8‡], Federica Isidori[9‡], Anna Marabotti[10‡], Pier Luigi Martelli[3‡], Stefano Pascarella[2‡], Michele Pinelli[11‡], Tommaso Pippucci[9‡], Roberta Russo[11,12‡], Castrense Savojardo[3‡], Bernardina Scafuri[10‡], Lucrezia Valeriani[13] and Emidio Capriotti[3‡]

[1]Department of Biology, University of Rome Tor Vergata, Roma, Italy, [2]Department of Biochemical Sciences "A. Rossi Fanelli", University of Rome "La Sapienza", Roma, Italy, [3]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [4]Department of Medical Sciences, University of Torino, Torino, Italy, [5]National Centre for Rare Diseases, Istituto Superiore di Sanità, Roma, Italy, [6]Department of Physics, University of Genova, Genova, Italy, [7]Italiano di Tecnologia—IIT, Genova, Italy, [8]National Research Council, Institute of Food Science, Avellino, Italy, [9]Medical Genetics Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy, [10]Department of Chemistry and Biology "A. Zambelli", University of Salerno, Fisciano, SA, Italy, [11]Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Napoli, Italy, [12]CEINGE Biotecnologie Avanzate Franco Salvatore, Napoli, Italy, [13]Center for Technology and Innovation, Trieste, Italy

Collectively, rare genetic disorders affect a substantial portion of the world's population. In most cases, those affected face difficulties in receiving a clinical diagnosis and genetic characterization. The understanding of the molecular mechanisms of these diseases and the development of therapeutic treatments for patients are also challenging. However, the application of recent advancements in genome sequencing/analysis technologies and computer-aided tools for predicting phenotype-genotype associations can bring significant benefits to this field. In this review, we highlight the most relevant online resources and computational tools for genome interpretation that can enhance the diagnosis, clinical management, and development of treatments for rare disorders. Our focus is on resources for interpreting single nucleotide variants. Additionally, we present use cases for interpreting genetic variants in clinical settings and review the limitations of these results and prediction tools. Finally, we have compiled a curated set of core resources and tools for analyzing rare disease genomes. Such resources and tools can be utilized to develop standardized protocols that will enhance the accuracy and effectiveness of rare disease diagnosis.

KEYWORDS

rare disease, genetic disorder, single nucleotide variant (SNV), genome interpretation, precision medicine, genotype-phenotype association, machine learning

1 Introduction

The recent major advances in genome sequencing and analysis technology have opened the road to exome and genome sequencing (ES/GS) as a common diagnostic tool for individual patients (Turro et al., 2020; 100,000 Genomes Project Pilot Investigators et al., 2021). Especially in the field of rare genetic diseases, the use of ES/GS has brought

Check for updates

OPEN ACCESS

EDITED BY
Barbara Cellini,
University of Perugia, Italy

REVIEWED BY
Zailong Qin,
Maternal and Child Health Hospital of Guangxi Zhuang Autonomous Region, China
Andrea Mozzarelli,
University of Parma, Italy

*CORRESPONDENCE
Paola Turina,
✉ paola.turina@unibo.it

‡ELIXIR Italy Rare Disease Community

†These authors have contributed equally to this work and share first authorship

RECEIVED 18 February 2023
ACCEPTED 25 April 2023
PUBLISHED 10 May 2023

CITATION
Licata L, Via A, Turina P, Babbi G, Benevenuta S, Carta C, Casadio R, Cicconardi A, Facchiano A, Fariselli P, Giordano D, Isidori F, Marabotti A, Martelli PL, Pascarella S, Pinelli M, Pippucci T, Russo R, Savojardo C, Scafuri B, Valeriani L and Capriotti E (2023), Resources and tools for rare disease variant interpretation.
Front. Mol. Biosci. 10:1169109.
doi: 10.3389/fmolb.2023.1169109

unprecedented progress, and holds the potential for further large-scale impact in the clinical setting, allowing early diagnosis and early, precisely tuned, treatment (Pogue et al., 2018; Liu et al., 2019; Claussnitzer et al., 2020; Bonne, 2021). Presently, the definition of a rare disease (RD) varies among different regions. In Europe, it is defined as a condition affecting not more than 1 person per 2,000 in the European population (Regulation Orphan Medicinal Product, 2000). In the United States, it is defined as a condition that affects less than 200,000 people in the country (U.S. Food and Drug Administration, 2022), while in Japan it is defined as affecting fewer than 50,000 people, or one in 2,500 (Hayashi and Umeda, 2008). Collectively, RDs represent a significant burden to health and society, as their estimated prevalence is approximately 3.5%–5.9% of the worldwide population, resulting in about 30 million people affected in Europe and 300 million worldwide (Nguengang Wakap et al., 2020). Approximately 7,000 different RDs have been identified to date, even though the exact number is debated (Hartley et al., 2018; Ferreira, 2019; Haendel et al., 2020), of which an estimated 70% are genetic (with 4,418 involved genes identified so far, November 2022), whilst the remaining are the results of infections, allergies and environmental causes. Most likely, the number of involved genes is bound to increase, as rapidly increasing quantities of exomic data are analyzed in the clinic (Boycott et al., 2018; 2019). From 2010 to 2020, the diagnosis of RDs saw a remarkable increase, with 886 new RDs being identified. During this period, the total number of genes associated with RDs grew from approximately 2,400 to over 4,000, and the number of new orphan drugs approved by the US and/or the European Union rose to 438 (Monaco et al., 2022).

Due to the very status of being rare, knowledge, research, medical expertise, and therapeutic opportunities for each particular RD are often extremely limited, and geographically sparse. Along with technological advances, the public and scientific awareness has been growing, and the knowledge on RDs is going to massively benefit from large scale data collection, integration, and sharing (Hartley et al., 2020). Many international initiatives and consortia (Gainotti et al., 2018; Azzariti and Hamosh, 2020; Bonne, 2021; Baxter et al., 2022; Laurie et al., 2022; Monaco et al., 2022) aim to significantly increase the overall percentage of RD patients with a confirmed (molecular) diagnosis, estimating that thousands of RD genes and disease mechanisms still remain undiscovered (Frésard and Montgomery, 2018; Boycott et al., 2019; Hartin et al., 2020). Exome Sequencing (ES) has been the most significant technology driving progress in the discovery and diagnosis of RDs over the past decade. While some RD diagnoses may require the integration of multiple omics data (Frésard and Montgomery, 2018; Marwaha et al., 2022), it is expected that ES will continue to play a crucial role in future efforts (Boycott et al., 2019).

The sheer re-analysis of exomic data after 1–3 years updating of the major disease variants and disease-gene association databases is reported to have increased the diagnosed cases by over 10% (Wenger et al., 2017; Setty et al., 2022). Remarkably, a further improvement in the yields could be obtained by reanalysing the data in collaboration with the clinician who made the diagnosis (Basel-Salmon et al., 2019). The contribution of research laboratories has provided an additional increase, aided by the application of novel computational and analysis tools (Eldomery et al., 2017). Thus, the fundamental step in ES data processing is the interpretation of the identified variations, i.e., the estimate of their likelihood of having a causative role in contributing to the disease. Indeed, RD-affected individuals often carry multiple variations in the gene(s) associated with the disease, with only a fraction of them being actually pathogenic (Summers, 1996). Criteria for the objective classification of variants into a five-tier system (pathogenic/likely pathogenic/uncertain significance/likely benign/benign) have been provided to the biomedical community, together with scoring rules that weight each criterion used to classify the variants. In this context, computational tools have a role in supporting the evidence framework for a benign or a pathogenic assertion (Richards et al., 2015).

This paper aims to provide an updated overview of the most frequently adopted and publicly accessible online resources and computational tools for predicting genotype-phenotype associations in RDs. In the first part of this review, we focus on the main databases collecting genes and variants associated with RDs. In addition, we describe the most popular computational methods for gene and variant prioritization, showing how information derived from molecular databases and tools can improve the diagnosis of RDs in clinical settings. Finally, we discuss the central role of FAIR data sharing in boosting research and diagnosis in the field and provide future perspectives.

2 Online resources and databases for rare diseases

Large-scale sequencing efforts on healthy individuals and patients allowed the collection of large databases of genetic variants and their association with human phenotypes. Based on their content and purposes, two groups of online resources for RDs can be identified: one group includes databases that define phenotype ontologies and controlled vocabularies for the description and classification of human diseases and phenotypes; the second group includes databases collecting the frequency of variants in the human population and their relationship with genetic disorders. Here, we summarize the most popular resources for medical diagnosis, focusing specifically on those related to RDs.

2.1 Disease and phenotype classification databases and ontologies

Nowadays, different resources for the classification of RDs are available. In particular, specific ontologies based on controlled vocabularies are defined for the description of human disorders. This enables a standardized description and classification of RDs, thereby enhancing and supporting data sharing. A standardized medical terminology was defined for developing the Medical Subject Headings (MeSH), an organized collection of hierarchical trees with increasing specificity of the downstream terms (Rogers, 1963). Later, ontologies based on diagnostic terms were created. Among them, the International Classification of Diseases (ICD), which represents the healthcare classification system maintained by the World Health Organization (World Health Organization, 2019), and the Systematized Nomenclature of Medicine (SNOMED), which implements a directed acyclic graph architecture for the

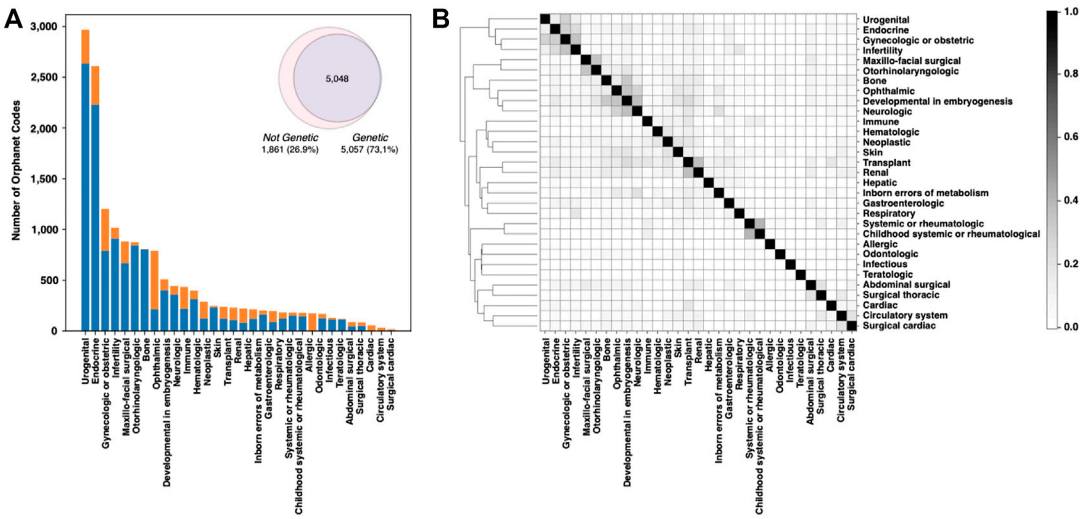


FIGURE 1
Analysis of the Orphanet database composition. **(A)** Fraction of genetic and nongenetic RDs in the different classes. **(B)** Plot showing the fraction of genes shared by each RD-class pair. Genes and Orphanet codes can be found to be associated with multiple RD classes.

automatic exploration of relationships among terms. In the 80s, the US National Library of Medicine created the Unified Medical Language System (UMLS) to harmonize the various classification systems (Bodenreider, 2004). ULMS, with its well-defined semantic relationships, is widely recognized as one of the most comprehensive resources for determining disease similarity and for the harmonization of RD data (Zhu et al., 2020). The increasing popularity of controlled vocabularies for the classification of human disorders further stimulated the creation of disease- and phenotype-oriented ontologies. The Human Phenotype Ontology (HPO) is a standardized vocabulary describing phenotypic abnormalities (Robinson et al., 2008). It is structured as a directed acyclic graph, in which a child node corresponds to a more specialized term with respect to its parent. Currently, HPO contains over 13,000 terms and over 156,000 annotations to hereditary diseases. The Disease Ontology (DO) is an open source ontology for the integration of biomedical data associated with human disease. The DO integrates concept terms from SNOMED, ICD, MeSH, and UMLS, using various semantic similarity measures (Schriml et al., 2012). The current version of DO (August 2022) collects more than 11,000 disease terms divided in 6 major classes. Mondo is the disease ontology of the Monarch Initiative (Shefchek et al., 2020) which integrates genotype-phenotype data across different species. The Mondo ontology is an open, community-driven resource which currently collects ~44,650 terms divided in three main categories (disease characteristic, disease or disorder, disease susceptibility). In Mondo, human diseases are grouped in 36 classes. Biomedical ontologies serve various purposes, such as: 1) systematizing the description of biomedical concepts for literature and clinical data recording (e.g., MeSH), 2) capturing individual clinical phenotypes, even in the absence of a recognized disease, and providing a corresponding classification in animal models (e.g., HPO), and 3) categorizing nosological entities for epidemiological and clinical management purposes (e.g., ICD). There is often overlap between these classifications, with some incorporating features from others. These ontologies of concepts are also utilized to annotate molecular data databases for the purpose of storing, analyzing, and exploring genotype-phenotype relationships.

The Online Mendelian Inheritance in Man (OMIM) database was created in the 60s by McKusick to systematically identify the relationship between disease and genetic components (Amberger et al., 2009). In November 2022, OMIM collected 7,301 phenotypes, ~85% of which were associated with at least one of the 4,674 listed genes. Focusing on the classification of rare human disorders, Orphanet is a unique resource that provides high-quality information for defining a specific nomenclature for RD (Rath et al., 2012). The description of RDs in Orphanet is based on the Orphanet Rare Disease Ontology (ORDO), a structured vocabulary capturing relationships between diseases, genes, and other relevant features. The November 2022 version of the Orphanet database collects 6,918 RDs classified in 33 major groups. All groups include genetic-caused RD, except for the “toxic effects” group. The fraction of rare genetic diseases across the remaining 32 major classes ranges from 99.6% of the *“rare inborn errors of metabolism”* to the lowest percentage of *“infectious”* diseases. The most abundant types of rare genetic disorders are those having *“developmental”* and *“neurologic”* effects (Figure 1A). Overall, RDs with genetic origin represent ~73% of the total (Figure 1A, inset). In terms of RD-associated genes, Orphanet collects more than 4,400 genes. Several of those genes are found to be associated with more than one RD class. An index of similarity (Jaccard index), based on the fraction of shared genes, has been calculated between each pair of RD groups, and is plotted in Figure 1B. The groups of “neurological” and “developmental” RD are sharing the highest number of disease-associated genes, with a Jaccard index ~0.37. The full list of the fraction of genetic RD and associated genes is reported in Supplementary Table S1.

In terms of enzymatic function, out of 5,057 genetic RDs reported in Orphanet (Supplementary Table S1), 1,596 (31.6%) are associated with enzymes, distributed among all the seven major enzyme classes (Table 1). The most represented enzyme classes are Transferases,

TABLE 1 For each enzyme class, the table lists the number of enzymes associated with Orphanet RDs, the number of the corresponding Orphanet diseases, and the number of the corresponding Reactome roots and pathways. The data were derived from DAR database (Savojardo et al., 2022) that integrates gene-disease associations reported in UniProt, Monarch, and ClinVar.

Enzyme class	Enzymes[a]	Orphanet diseases[b]	Reactome roots	Reactome pathways
All classes	1,218	1,596	27	1,098
EC 1: Oxidoreductases	186	259	20	209
EC 2: Transferases	474	738	26	799
EC 3: Hydrolases	401	611	27	592
EC 4: Lyases	63	81	15	93
EC 5: Isomerases	40	62	17	78
EC 6: Ligases	58	76	7	28
EC 7: Translocases	44	77	11	36

[a]In the distribution among classes, multiclass enzymes are counted multiple times.

[b]In the distribution among classes, diseases associated with enzymes from different classes are counted multiple times.

Hydrolases, and Oxidoreductases. Orphanet RDs can be linked to their corresponding enzyme metabolic pathways through the Reactome database, a comprehensive resource that catalogs all human metabolic reactions in 2,580 hierarchically organized pathways, with 27 main roots. Table 1 shows, for each enzyme class, the number of enzymes involved in Orphanet RDs, the number of Orphanet diseases that involve those enzymes, their Reactome roots and pathways.

The Disease And Reactome (DAR) database (Savojardo et al., 2022) provides a wealth of information on enzymes, including their relationships with Reactome pathways, molecular interactions within the pathways, and tissue expression levels as recorded in the Human Protein Atlas (Uhlén et al., 2015).

In general, the evaluation of the evidence supporting gene-disease relationships is a critical factor for an accurate diagnosis (Strande et al., 2017). To prevent mistakes in the diagnostic process, the curators of the Clinical Genome Resource (ClinGen) (Rehm et al., 2015) defined evidence-based Standard Operating Procedures for the classification of clinically relevant genes based on the presence of pathogenic variants (Section 1.3). Gene-disease relationships are classified in six groups that qualitatively describe the strength of the supporting evidence. The default class assigned to genes without any detected disease-causing variants is *"No Reported Evidence"*. Supporting evidence for gene-disease relationships is classified into four categories: *"Limited"*, *"Moderate"*, *"Strong"* and *"Definitive"*. When both supporting and conflicting evidence are present, the gene-disease relationship is classified as *"Contradictory"*. Within the ClinGen framework, the systematic review of genetic, clinical and experimental evidence, reported in databases such as OMIM and Orphanet, is used to assign one of the categories mentioned above to the reported gene-disease relationship.

2.2 Gene and protein network databases

A single gene defect is the most common origin of rare genetic diseases collected in the databases mentioned above. However, to investigate the molecular mechanisms underlying a RD, it is fundamental to understand and contextualize the resulting phenotype. At the protein level, defining the macromolecular complexes and pathways perturbed by the defective gene can be a useful strategy to understand the pathology itself and to intervene to restore the healthy phenotype.

A genetic variant can impact protein function and, depending on the central or marginal role of the mutated node inside a protein-protein interactions network, also the capability of the network to find alternative paths in the edges map. Changes in specific interactions can drastically perturb cellular networks and generate disease phenotypes (Barabasi et al., 2011; Menche et al., 2015).

Molecular interactions, mostly protein-protein interactions (PPIs), are annotated and archived, in structured formats, into several public resources. The major public databases collecting molecular interaction data can be divided into primary, predictive and meta-databases. Primary databases collect only manually curated molecular interactions, extracted from peer-reviewed journals, such as the IMEx Consortium resources (MINT (Calderone et al., 2020), IntAct (Del Toro et al., 2022), DIP (Salwinski et al., 2004), MatrixDB (Clerc et al., 2019)), and BioGRID (Oughtred et al., 2021). Meta-databases integrate data coming from primary databases, such as HiPPIE (Alanis-Lobato et al., 2017) and mentha (Calderone et al., 2013). Predictive databases use computational methods to predict PPIs (De Las Rivas and Fontanillo, 2012), such as STRING (Szklarczyk et al., 2021), IID (Pastrello et al., 2020) or ProfPPIdb (Tran et al., 2018).

In the panorama of molecular interaction resources, only the IMEx Consortium databases annotate interaction associated features, such as binding sites involved in the interaction or mutation effects (Porras et al., 2020). In particular, the IMEx mutation dataset contains annotations of experimental evidence where mutations have been shown to affect a protein interaction (~75,000 records) (IMEx Consortium Curators et al., 2019). The dataset can be used to map selected pathogenic variants to manually curated PPIs and to understand the effect of a specific variant on the interactions at protein-protein interface. Moreover, from the IntAct datasets, it is possible to download a RD specific dataset of molecular interactions extracted from literature. The dataset is enriched with experimentally proven impact of clinical mutations on interactions, and also with the non-clinical mutations which are found to impact protein functionality. So far, the dataset contains over 7,900 interactions involving about 2,500 interactors. The dataset can be visualized and filtered in the IntAct result page, or in Cytoscape (Shannon et al., 2003), using the IntAct App (Ragueneau et al., 2021).

Disease specific biological networks can also be constructed or integrated with data coming from signaling pathways databases such as Signor (Lo Surdo et al., 2023), WIKIPathway (Martens et al., 2021) or OmniPath (Türei et al., 2016). They can then be imported into Cytoscape by using resource specific CytoscapeApps (Kutmon et al., 2014; Ceccarelli et al., 2020; De Marinis et al., 2021), to gain more insight into the molecular mechanisms involved in the disease. Moreover, pathway resources such as KEGG (Kanehisa et al., 2017) and Reactome (Jassal et al., 2020) databases are very important to discover whether some disease-associated subnetworks are enriched for a particular functional pathway.

By the combination of PPI with genotype-phenotype relationships, functional similarities have been used to generate specific disease networks defining similarity across different human disorders (Goh et al., 2007; Menche et al., 2015; Buphamalai et al., 2021). Such networks have been shown to be useful for studying the biological mechanisms of diseases and for the development of gene prioritization tools (Zhang and Itan, 2019). Some examples of gene prioritization tools, specific for RDs, will be discussed in Section 2.2.

2.3 Databases of variants

The Human Genome Variation Society (HGVS) maintains comprehensive lists of databases focused on variations, from locus-specific mutation databases to SNP databases, to chromosomal variations, to other mutation databases, including nonhuman and artificial mutations. However, given the high number of resources, it is nearly impossible to perform an exhaustive description of all those that are available. We will therefore focus on selected, curated and widely used resources. None of them is specifically dedicated to RDs; however, it is possible to collect data and information on RD-associated variations.

In general, variant databases can be divided into two groups, according to whether they focus on the variant's frequency across the human populations or on their pathogenic effect.

The variant's frequency can be derived from sequencing experiments on a large set of individuals. For example, the 1,000 Genome project, started in 2008, collected and sequenced the genomes of 2,504 individuals from 26 populations worldwide, characterizing more than 88 million variants, including >99% of SNP variants with a frequency higher than 1% (1000 Genomes Project Consortium et al., 2015). The datasets and the related analyses have been freely shared with the scientific community by setting up the International Genome Sample Resource (IGSR) (Fairley et al., 2020) to ensure their future usability and accessibility. Data about these variants can be explored through the Ensembl Variation database (Hunt et al., 2018), a project aimed at automatically annotating the genomes, integrating biological data and making all information accessible via a website. Those variants were grouped into subsets, based on the origin of the individual and on the frequency of occurrence. In the same period, the UK10K project (UK10K Consortium et al., 2015) sequenced the whole genomes of about 10,000 individuals, characterizing over 24 million novel sequence variants. That information was made available via a dedicated website and via the European Genome-phenome Archive (EGA) (Freeberg et al., 2022), a resource for permanently archiving and sharing personally identified genetic, phenotypic and clinical data, obtained by biomedical research projects. Another analogous study is the "All of Us" research program, funded by NIH, sequencing 100,000 genomes from ethnic groups underrepresented in previous projects (All of Us Research Program Investigators et al., 2019). While the "All of Us" project was of broader scope, the 100,000 Genomes Project, focused on patients with an RD (161 disorders covering a broad spectrum of RDs were present) or with one among 20 different common cancer types (Turnbull et al., 2018). A pilot study, conducted on the genomes of 4,660 people, increased the diagnosis number for 25% of participants. Among them, 14% of the cases were new diagnoses based on variants found in regions usually missed in conventional, non-whole genomic tests (100,000 Genomes Project Pilot Investigators et al., 2021).

A widely used database collecting variant frequency data is the Genome Aggregation Database (gnomAD). The gnomeAD is the successor to the Exome Aggregation Consortium (ExAC), a project that was launched to aggregate and harmonize exome and genome sequencing data from a variety of large-scale sequencing projects (Karczewski et al., 2020). The National Center for Biotechnology Information (NCBI) at the NIH hosts several resources for investigating and understanding human variations. dbSNP and dbVar are two freely available databases, the former hosting a broad collection of small genetic polymorphisms (SNP, deletion/insertion polymorphisms, etc.), and the latter hosting a broad collection of large variants (>50 bp) (Lappalainen et al., 2013).

The second class of variant databases collect information about their clinical significance and their association with human disorders. To this purpose, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP), developed specific guidelines, where variants are classified into five types: *"pathogenic"*, *"likely pathogenic"*, "uncertain significance", *"likely benign"*, and *"benign"* (Richards et al., 2015). On the one hand, *"pathogenic"* and *"likely pathogenic"*, variants are classified by using multiple criteria grouped in four weighted categories: *"very strong"*, *"strong"*, *"moderate"*, and *"supporting"*. On the other hand, *"likely benign"*, and *"benign"* variants are classified using a combination of rules grouped in three weighted categories: *"stand-alone"*, *"strong"* and *"supporting"*. All previous criteria are based on eight categories of information, including among them data from population studies and computational predictions. If a variant does not meet any criteria or the evidence for benign and pathogenic is conflicting, the class assigned by default is *"uncertain significance"*. As recently shown (Tavtigian et al., 2020), the ACMG/AMP guidelines are compatible with a quantitative Bayesian formulation, whose scaling as odd of pathogenicity allows an empirical calibration of the strength of the reported evidence.

A reference database, collecting annotated genetic variants by adopting the ACMG/AMP guidelines, is ClinVar (Landrum et al., 2020) which represents one of the main sources of information for gene classification in the ClinGen database (Section 1.1). ClinVar is a freely accessible, public archive, collecting reports of variants found in patient samples, assertions about their clinical significance, and other data, including the availability of supporting evidence. ClinVar thus allows users to infer relationships between human variations and the health status of the patients. Each variation has its own accession number, and, if multiple submitted records about the same variation/condition pair are present, they are aggregated under a single accession number. The adoption of a single variant identifier allows users to review all data submitted for a single variant, regardless of the condition for which it was interpreted. In fact, ClinVar neither curates content nor modifies interpretations associated with a single record. The alleles are reported according to the HGVS standards. Focusing on protein variants, the UniProt consortium is releasing a curated file reporting a list of protein variants, grouped in 3 classes: *"Disease"*, *"Polymorphism"* and *"Unclassified"* (humsavar UniProt, 2023). In the *humsavar* file, an OMIM identifier is associated with each pathogenic variant. Alternatively, the Human Gene Mutation Database (HGMD) is a proprietary database of mutations in human genes, associated with inherited diseases, which contains both inherited and somatic mutations (Stenson et al., 2020). The GWAS Catalog, supported by a collaborative initiative between the National Human Genome Research Institute and the EMBL-EBI, is another popular freely available database of SNP-trait associations, which can be easily integrated with other resources (Sollis et al., 2023).

The collection and curation of several variant databases is supported by ELIXIR, an intergovernmental organization that brings together bioinformatics resources for life sciences from across Europe (https://elixir-europe.org/). For example, the European Variation Archive (EVA) (Cezard et al., 2022) is an open-access database of all types of genetic variations (SNP, large structural variants, observed in germline or somatic sources) from all species. Submitted variants (in Variant Call Format, VCF) are merged, normalized and annotated for functional consequences and to determine allele frequency values in a study-specific manner. Human variants are also exchanged with dbSNP and other NCBI resources. DisGeNet (Piñero et al., 2020) is another database that integrates information on human gene-disease associations and variant-disease associations from different repositories. The data are annotated with controlled vocabularies and community-driven ontologies, and several original metrics are provided to assist the prioritization of genotype–phenotype relationships. Another ELIXIR core resource collecting information about variation is Ensembl Variation (Hunt et al., 2018), a resource linked to the Ensembl Genome Browser. It stores variants found in many species (including human) and, where available, associated diseases and phenotype information. Variant data are imported from a variety of sources (e.g., dbSNP) and subjected to a quality control process. They are then classified and their consequences predicted. Moreover, variant sets are created to help people retrieve a specific group of variants from a particular dataset. For human data, the linkage disequilibrium is also calculated for each variant, by population. A list of resources and databases for RD cited in this paragraph is reported in Supplementary Table S2.

3 Tools for rare disease genome interpretation

3.1 Automatic variant calling pipelines

The analysis of next-generation sequencing (NGS) experiments requires substantial bioinformatics resources. During the last years, a variety of analytical tools have been developed for the detection of genetic variants. Such tools assist all steps of the variant calling process, including quality control and trimming, alignment to the reference genome, identification, and annotation of SNVs and short indels. Although the Genome Analysis ToolKit (GATK) Best Practices guidelines define standard procedures for setting up a variant analysis pipeline, selecting the best approach among the variety of tools for NGS data processing can still be challenging. To overcome the issue and simplify the variant calling process, several "ready-to-use" bioinformatics pipelines to process ES and GS data have been made available. Some of them include: fastq2vcf (Gao et al., 2015), SeqMule (Guo et al., 2015), ExScalibur (Bao et al., 2015), Appreci8 (Sandmann et al., 2018), JWES (Ahmed et al., 2021), OVarFlow (Bathke and Lühken, 2021) and the recent DeepVariant (Poplin et al., 2018) that integrates a deep-learning-based variant caller. Most of those pipelines integrate many variant calling tools to increase sensitivity, but they are command-line applications to be installed on local servers. Alternatively, web-based options are available, e.g., Maser (Kinjo et al., 2018), CSI NGS Portal (An et al., 2020), and the most popular Galaxy (Afgan et al., 2018). Recently, *seqr*, a web tool for the analysis of rare disease genomes, has been made available by the Broad Institute (Pais et al., 2022). They are open-source platforms that provide a user-friendly graphical interface, improving the accessibility to computation analyses of genomic data. In particular, Galaxy users can freely create custom workflows or find already existing workflows, available on Galaxy Toolshed (Blankenberg et al., 2014), which can be run on public Galaxy servers. The disadvantages of using the web-based options are the limited amount of data that can be uploaded, the CPUs time, and the limitations on some tools on the public Galaxy platforms. However, Galaxy pipelines can also be run on a local Galaxy installation, or on a paid cloud infrastructure, e.g., Amazon cloud (AWS), using CloudMan (Afgan et al., 2010). Terra is another example of a web- and cloud-based platform, providing a compute environment to run optimized pipelines on Google Cloud. Galaxy, Terra, and other analysis components are integrated in a unified environment for data analysis and management, AnVIL (Schatz et al., 2022), designed to manage and store genomics data, enable population-scale analysis, and facilitate collaborative large-scale research projects. Nevertheless, "best practices" for variant calling in clinical settings, should be considered before choosing the most appropriate sequencing strategy, and the most reliable combination of tools for read alignment/preprocessing, variant calling and filtering (Koboldt, 2020).

Furthermore, to ensure the reproducibility of complex bioinformatics analysis, different workflow languages have been used to develop specific data analysis pipelines. The NextFlow core community (Ewels et al., 2020) collected a curated set of optimized procedures for the analysis of genomic data specific for rare disease. Similar projects include Dockstore (O'Connor et al., 2017), which provides containerized tools and workflows, currently supporting 4 different languages: the Workflow Description Language (WDL), Common Workflow Language (CWL), Nextflow, and Galaxy Workflows (GWs). Moreover, several workflows accessible on Dockstore can be easily launched in web-based platforms, such as Terra. These workflow languages are designed to handle some aspects of computational workflows, such as resources, software, and execution of analysis steps. Among those, Snakemake (Köster and Rahmann, 2012) and Nextflow (Di Tommaso et al., 2017) are commonly used for developing new research pipelines, while WDL and CWL workflows are preferred for large-scale projects (Reiter et al., 2021). Recently, a specific pipeline for the analysis of rare disease genome has been made available in NextFlow (Ewels et al., 2020).

Most of the above semi-automatic pipelines help streamline the generation of variant lists (in vcf format), but lack the downstream annotation and filtering steps that are necessary to identify disease-causing variants. To this end, different data-warehousing solutions to store genomic variants, along with the relevant genomic annotations, were deployed to allow a flexible and efficient data exploration. An example is GEMINI (Paila et al., 2013) and OpenCGA that supply the platform and the analysis framework to build customized genomic databases, to efficiently store data to be queried and visualized. A list of tools for variant calling and annotation is reported in Supplementary Table S3.

3.2 Gene prioritization tools

The objective of gene prioritization is to rank a large list of potential candidate genes based on their relevance for a disease. The prioritization algorithms identify the most promising genes, as to their association to the molecular basis of a given disease and/or a specific phenotype, for defining a therapeutic and/or diagnostic procedure. From the experimental point of view, the high-throughput techniques reduced the costs for generating a high amount of information about gene mutations. On the other hand, the identification of real links between genes and diseases is still a time- and money-consuming task. Therefore, the help of computational tools to reduce the number of genes to be investigated is strongly needed. Beyond the assumption that one gene codes for one function, the possibility that defects of one gene may be related to multiple diseases is now taken into account. At the same time, more genes can be involved in a given disease. In fact, a given metabolic pathway is composed of several protein functions, a defect in any of which may result in the pathway failure. Computational tools for gene prioritization use different sources of information to rank the candidate genes. Possible features are direct experimental data on gene sequences, mutations, expression (co-expression), gene-gene and protein-protein interactions, as well as more indirect evidence as ontologies, literature, information derived by model organisms. Different types of tools may differ by the focusing level (e.g., disease-specific or not), by the applied methodology (e.g., text-mining, similarity profiling, network analysis), by the approach to select the best candidate genes (e.g., ranking or filtering into smaller subsets), by the assumptions (i.e., genes may be directly or indirectly associated with a disease), or simply by the type of experimental evidence used for the analysis. Several works list the available tools on the basis of the state-of-the-art and classification applied (Moreau and Tranchevent, 2012; Piro and Di Cunto, 2012; Gill et al., 2014; Zolotareva and Kleine, 2019; Cabrera-Andrade et al., 2020; Jacobsen et al., 2022; Yuan et al., 2022). For instance, Jacobsen et al. applied phenotype-driven methods to improve diagnostic yields for RD, and listed 16 freely available tools (Jacobsen et al., 2022). Zolotareva and Kleine listed 14 tools, classifying them according to strategies, approach types, interfaces, input, and the types of evidence sources (Zolotareva and Kleine, 2019). Smedley and Robinson compared 7 tools and summarized their features in terms of exome input, types of variants analyzed, and approach (Smedley and Robinson, 2015). Problems related to long-term maintenance of academic software are very common (Jacobsen et al., 2022) and solutions have been proposed (Rother et al., 2012). A list of tools from the cited literature is reported in Supplementary Table S4. Among all gene prioritization methods, for instance, VarElect and ToppGene are part of standard diagnostic pipelines in the clinical settings. In particular, VarElect (Stelzer et al., 2016a) is a comprehensive, phenotype-dependent, variant/gene prioritization tool, based on the GeneCards suite (Stelzer et al., 2016b). The input of VarElect is a gene list together with a free-text phenotype description, such as disease and symptom terms, which represents a useful interface for non-skilled users. The tool prioritizes the genes on the basis of scores for the associated terms, computed on the appearance frequency in the entire GeneCards knowledgebase. The latter includes also the human disease database MalaCards (Rappaport et al., 2017), the human biological pathways of Pathcards (Belinky et al., 2015), and the gene expression information in cells and tissues of LifeMap Discovery (Edgar et al., 2013), for a total of 120 sources. The results of VarElect are displayed as a table of genes with decreasing phenotype relation scores. Alternatively, the gene prioritization task can be performed by ToppGene (Chen et al., 2007; Chen et al., 2009a; Chen et al., 2009b), a suite including tools for gene list functional enrichment, candidate gene prioritization, and identification and prioritization of novel disease candidate genes in the interactome. ToppGene selects genes in the training set on the basis of their association with disease, pathway, GO term, phenotype. The test set can be given by candidate genes from linkage analysis studies, differential expression in a particular disease or phenotype, interactome knowledge. The enrichment step is based on a variety of data sources that cover 14 types of annotation. For each type of annotation of each test gene, a similarity score is generated, by comparison to the enriched terms in the training set. The prioritized gene list is ranked on the aggregated values of the 14 similarity scores.

Finally, specific algorithms for the prioritization of RD-associated genes were recently developed (Zhu et al., 2012; Liu et al., 2017; Buphamalai et al., 2021; de la Fuente et al., 2023). Among them, for instance, an algorithm was developed, based on the calculation of a vertex-similarity score between each pair of genes, that was tested on a set of ~1,600 known orphan disease-causing genes associated with 172 RDs (Zhu et al., 2012). Another method, which computes the topological similarity between genes connected in a PPI network, ranks the candidate genes combining two scores reflecting the local and global connectivity of the network (Liu et al., 2017). The success rate of this method can reach 50%–75% on a set of ~1,200 genes collected from the Orphanet database. A more comprehensive approach evaluates the impact of rare gene defects, building a multiplex network with more than 20 million gene relationships organized into 46 network layers (Buphamalai et al., 2021). The analysis of 3,771 RDs reveals distinct phenotypic modules that can be used to accurately predict RD gene candidates. A recent tool (GLOWgenes), based on 33 functional networks classified in 13 knowledge categories, was able to recover genes associated with 91 genetic diseases classified into 20 families (de la Fuente et al., 2023). When applied to 15 unsolved cases, GLOWgenes was able to identify three new genes potentially associated with syndromic inherited retinal dystrophies.

3.3 Variant interpretation methods

Variant interpretation tools are *in silico* predictive programs that can help researchers in establishing the pathogenicity of the variations identified in the gene(s) of interest. Many approaches have been developed to perform these predictions, and their number has grown very rapidly in the last years. They mainly focus on predicting the impact of a missense variation on the structure and function of the associated protein, or on predicting effects on RNA splicing.

More recently, programs addressing more general noncoding variants have also been developed (Özkan et al., 2021). Researchers and clinicians tend to use variant interpretation tools in combination, as also suggested by the ACMG/AMP guidelines (Section 1.3). Nevertheless, their concordance in asserting the variant effects (especially of the predicted benign ones) has been rather low until present. More recently, however, newly developed algorithms have shown good performance in many types of genes and mutation mechanisms. Furthermore, by using gene-specific algorithms, and by calibrating them with well-characterized sets of benign and pathogenic variants, better results may be reached, than with general use algorithms (Ghosh et al., 2017).

In the last two decades, an impressive number of methods and algorithms for single amino acid substitution (SAS) have been devised to predict the variant effect on protein structure, function and interactions, to eventually identify those involved in molecular pathogenicity. As a matter of fact, SASs represent more than 40% of the unique variants found in the Exome Aggregation Consortium (Lek et al., 2016). Those methods are obviously not specific to RDs and have a broad range of applications (Capriotti et al., 2019; Katsonis et al., 2022; Pancotti et al., 2022). A selection of the most recent methods and resources is reported in Supplementary Table S5. Many of the early methods were based on the prediction of the effect of a single mutation on the protein thermodynamic stability, as destabilization is one of the key factors in pathogenesis (Capriotti et al., 2008; Dehouck et al., 2011; Worth et al., 2011; Fariselli et al., 2015; Laimer et al., 2015; Quan et al., 2016; Savojardo et al., 2016; Yang et al., 2018; Marabotti et al., 2020; Pires et al., 2020; Montanucci et al., 2022). Subsequent efforts and developments in the field produced last-generation methods, using one of three general strategies: i) prediction of the likelihood of a missense mutation for causing pathogenic changes in a protein (Sim et al., 2012; Adzhubei et al., 2013; Carter et al., 2013; Katsonis et al., 2014; Niroula et al., 2015; Capriotti et al., 2017; Raimondi et al., 2017; Rentzsch et al., 2019; Pejaver et al., 2020; Manfredi et al., 2022; Quinodoz et al., 2022); ii) evolutionary conservation analysis of the mutated sites; iii) methods combining different strategies (Stein et al., 2019; Petrosino et al., 2021). More recently, several methods have been developed to also predict the impact of variants in noncoding regions (Rojano et al., 2019; Katsonis et al., 2022; Tabarini et al., 2022). These methods include generic tools, which predict single-nucleotide pathogenic variants across the entire genome (Quang et al., 2015; Shihab et al., 2015; Zhou and Troyanskaya, 2015; Capriotti and Fariselli, 2017; Rentzsch et al., 2019) and more specific algorithms, which predict the impact of splicing variants (Desmet et al., 2009; Cheng et al., 2019; Jaganathan et al., 2019; Rentzsch et al., 2021). In particular, splicing-affecting variants are established contributors to RD, of which they may modulate the phenotypic outcome (Li et al., 2016; Scotti and Swanson, 2016).

To assess the performance of the available variant interpretation algorithms on the variants specifically associated with RDs, we collected a dataset of SAS from ClinVar (March 2022). Such a dataset (sas-rd-202203 in Supplementary Materials) is composed of ~27,600 SAS in genes associated with rare genetic disorders from different RD classes. From RD-associated ClinVar genes, we selected 16,012 variants classified as *Pathogenic* and 11,633 *Benign*. The results of our analysis, scoring the performance of 4 state-of-the-art variant interpretation tools (CADD, FATHMM, PhD-SNP^g and VEST4), show that the selected methods reach on average 83% overall accuracy (Q2), 0.65 Matthews correlation coefficient (MCC), and >0.90 area under the ROC curve (AUC) (Supplementary Table S6). These results are in the same range of those reported in previous works, not limited to RDs (Capriotti and Fariselli, 2017; Benevenuta et al., 2021). A chromatic representation of the performance of the methods (Figure 2) reveals that VEST4 reaches the highest AUC (0.96) while FATHMM the lowest (0.83). Taking into account some possible data overfitting, we expect that the resulting measures of performance might be overestimated by no more than 2%–5% (Capriotti and Fariselli, 2017). The results of the four tools in predicting the effect of different RD classes exhibit some variation. Specifically, for the *Ophthalmic* RD class, with 7,889 variants (28.5%), the performance of the methods is slightly higher than average, reaching 85% overall accuracy, 0.69 Matthews correlation coefficient and 0.92 AUC. Conversely, the lowest performance was observed in predicting the impact of 2,152 variants associated with the *Cardiac* RD class (~7.8%), with an overall accuracy below 80%, a Matthews correlation coefficient of 0.58, and an AUC of 0.87. Although our analysis shows that state-of-the-art methods for the prioritization of causative variants in RD-associated genes result in a high-performance level, further work is needed for improving the tools' reliability, in view of the residual ~10% of misclassified variants. In this regard, it appears that an important aspect to be considered for improving the predictions reliability is the conservation level at the variation site (Capriotti and Fariselli, 2022).

3.4 Genotype/phenotype association methods

Despite the progress in our capacity to prioritize disease-causing genotypes in clinical exomes and genomes, the large number of variants that remain to be evaluated for the diagnosis-making

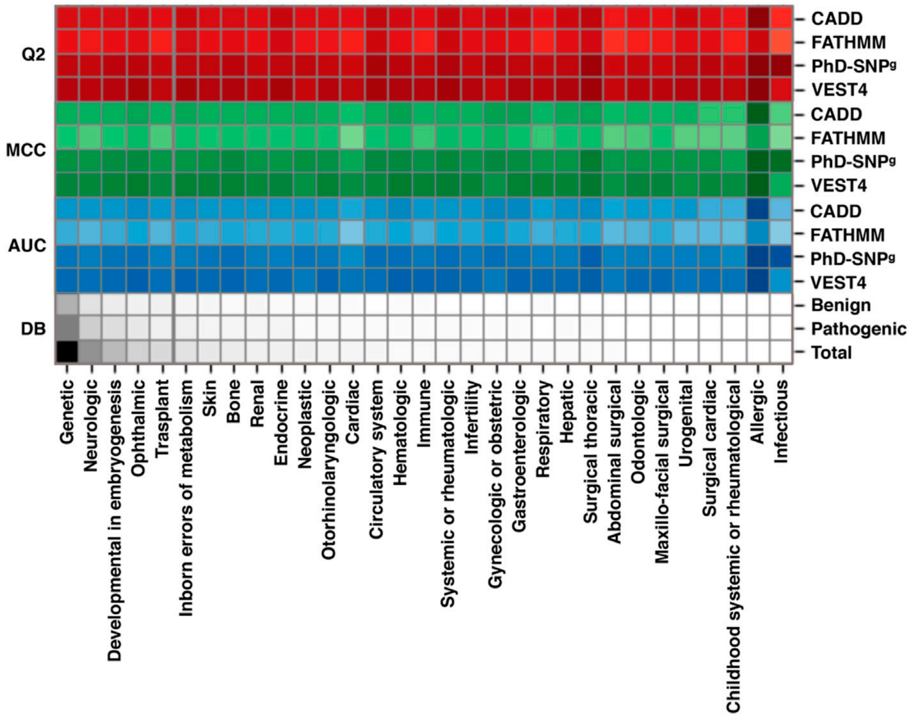


FIGURE 2
Performance of four state-of-the-art methods (CADD, FATHMM, PhD-SNP^g and VEST4) on a dataset of RD-associated variants from ClinVar database, featuring at least one annotation as Pathogenic or Benign. The scores are calculated for the different classes of RDs. All 27,648 variants (16,012 Pathogenic and 11,633 Benign) in our dataset are in the Genetic class. According to the Clinvar annotation, each variant can be classified in multiple RD groups. Performance parameters shown are: Q2, Overall Accuracy; MCC, Matthews Correlation Coefficient; AUC, Area Under the receiver operator characteristic Curve. DB indicates the fraction of each RD group in the dataset. The performance of CADD was calculated considering a Phred-like score threshold of 20. The color darkness in the drawing is proportional to the numerical values, which are reported in Supplementary Table S6. The predictions of the four methods are reported in Supplementary Materials.

process is still a challenge. Computational analysis of phenotypes, in addition to genotypes, has proven powerful to improve the standardization and automation of NGS diagnostic pipelines from raw sequences to prioritized variants. The general principle, followed by such analyses, is to compute measures of similarity between the clinical manifestation in a patient and the description of disease(s) linked to a gene. Gene and/or variant prioritization tools measure ontological similarity between a set of query terms, representing the compendium of the patient's clinical phenotypes, and the set of terms that are associated with any disease-gene (Smedley and Robinson, 2015). Algorithms underlying such tools have been developed, exploiting standardized collections of clinical terminologies, the most widely adopted of which is the Human Phenotype Ontology (HPO). The latter is used to assist clinical scientists and researchers in clustering and comparing phenotypes of patients with shared molecular background, with the aim to improve genetic diagnosis and genotype-to-phenotype correlations. Many tools that exploit the knowledge of known phenotypes of disease genes in humans and animal models have been developed. Such tools can be broadly categorized into two groups: those that take both phenotype and genotype data (VCF + HPO) as input, and those that only accept phenotype data (HPO only). These tools have been thoroughly reviewed and evaluated in a recent publication (Yuan et al., 2022). As an example, one of the earliest and most used tools is Exomiser (Robinson et al., 2014). Exomiser combines the most popular strategies for variant filtration with HPO to prioritize data in a VCF file. Despite its name, the Exomiser analysis framework is not limited to the exome but incorporates the Regulatory Mendelian Mutation (ReMM) score for relevance prediction of non-coding variations (SNVs and small InDels) (Smedley et al., 2016). In their review, Yuan et al. (2022) identified the two best performers in HPO-based gene prioritization to be LIRICAL (Robinson et al., 2020) and AMELIE (Birgmeier et al., 2020). Both of those recently published tools propose innovative and interesting analysis approaches. LIRICAL aims to overcome simple gene or variant ranking based on semantic similarity as a prioritization scheme, by introducing a likelihood-ratio test to provide an estimate of the post-test probability of candidate diagnoses. AMELIE, conversely, consists in an end-to-end machine learning approach with web interface, that finds relevant literature supporting the disease causality of genetic variants and their association with different clinical presentations. In the benchmark from Yuan et al. (2022), the two methods often resulted in quite different predictions of highly ranked causal genes, and such a complementarity suggests a possible integrative approach to further enhance the diagnostic efficiency. In a recent work, genotype/phenotype association methods were tested on a set of 4,877 molecularly diagnosed cases, affected by RDs, from the 100,000 Genome Project (Jacobsen et al., 2022). On this set, Exomiser was able to recall 82% and 92% of the diagnosed cases as

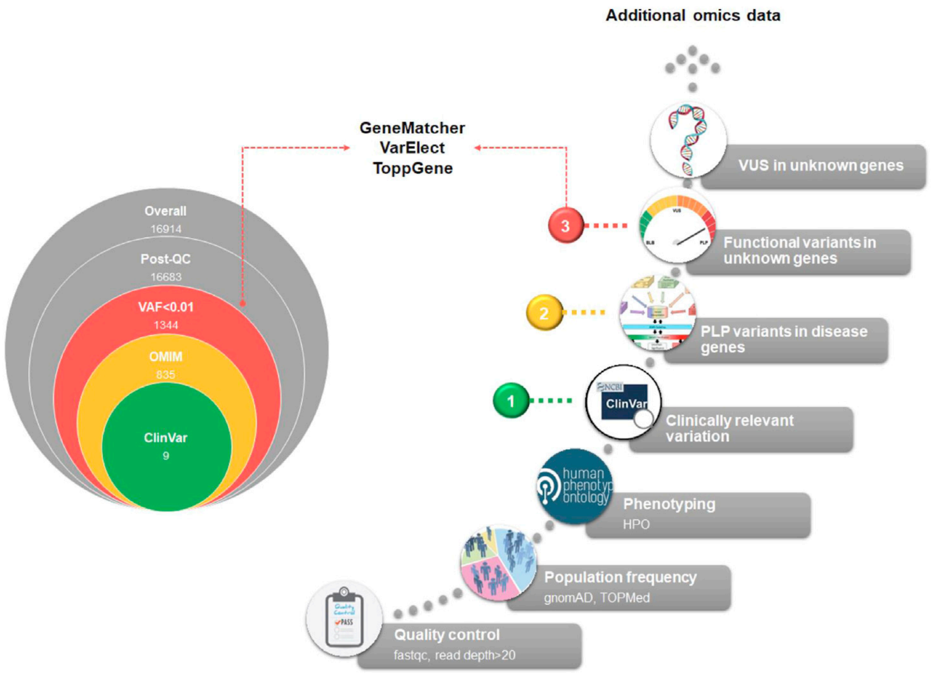


FIGURE 3
Exome analysis flowchart. A diagram of the main steps of NGS data analysis is shown. On the left, the progressive reduction by filtering in the number of likely disease-causing variants is shown, for a general patient case. The reported numbers are from a typical single patient case. On the right, the filtering process is detailed. Based on the identified variants, we can recognize three different diagnostic situations: (1, green dot) identification of P/LP variants with well-established association to RD phenotype; (2, yellow dot) identification of new P/LP variants in genes with known association to the phenotype; (3, red dot) identification of functional variants in genes with unknown association with the phenotype. A fourth case should be considered, i.e., the identification of VUS variants in genes with unknown association with the phenotype. In this case, complementing different approaches, such as short-read genome sequencing with RNA sequencing, and methyl profiling, should be considered to elucidate the molecular mechanism of the disease and improve the diagnostic yield.

the top hit, and within the top 5 scores, respectively. These positive results are going to render phenotype-genotype association tools essential for RD diagnosis in the clinical routine.

4 Use cases on rare disease genome interpretation

The diagnostic workflow of NGS genetic testing is composed of three levels of data analysis: i) quality control of raw data, ii) variant annotation, and iii) variant filtering. On the basis of the annotation level of the variants detected after the variant calling procedure, we can identify three main steps of analysis. In Figure 3 we summarized the main filtering steps, including the approximate number of unique variants that can be identified in a single subject after a clinical exome. In order to efficiently prioritize clinically relevant variations among all types of captured variants described in Figure 4, we need to adopt different analytical strategies. Several resources, including databases of genomic variation and phenotypes, population frequency data and *in silico* prediction approaches, can be used for the interpretation of each type of variant (Supplementary Table S5).

According to the variant types, distinguished by their gene location and the currently available resources for variant interpretation, we can define the three following possible cases.

4.1 Identification of pathogenic variants associated with RD phenotypes

In this case, the analysis workflow is well-defined and relatively easy. In the hypothetical case shown in Figure 4, after the application of the filters of variant allele frequency (VAF) and phenotyping (see below), known clinically relevant variants in disease genes that fit with the phenotype are selected to be reported.

The VAF is retrieved from population databases (Section 1.3), using resources that aggregate exome and genome sequencing data from a variety of large-scale sequencing projects, and make summary data available for the wider scientific community. They include sequencing data of both affected and unaffected subjects or different populations (Gudmundsson et al., 2022). In current diagnostic settings, ultra-rare and rare variants (VAF < 0.001 and VAF < 0.01, respectively), as well as private variants (not annotated in population databases) are selected. Of course, this primary filter can be modified according to the prevalence in the population of a specific disorder. Thus, in some diagnostic settings, also low-frequency variants (VAF < 0.05) can be selected (Andolfo et al., 2021).

The exact characterization of the phenotype (“*phenotyping*”) is one of the most relevant aspects of NGS genetic testing, and it is often considered a major challenge for the NGS-based genetic diagnosis. Generally, *phenotyping* is obtained using a standardized vocabulary of phenotypic abnormalities encountered

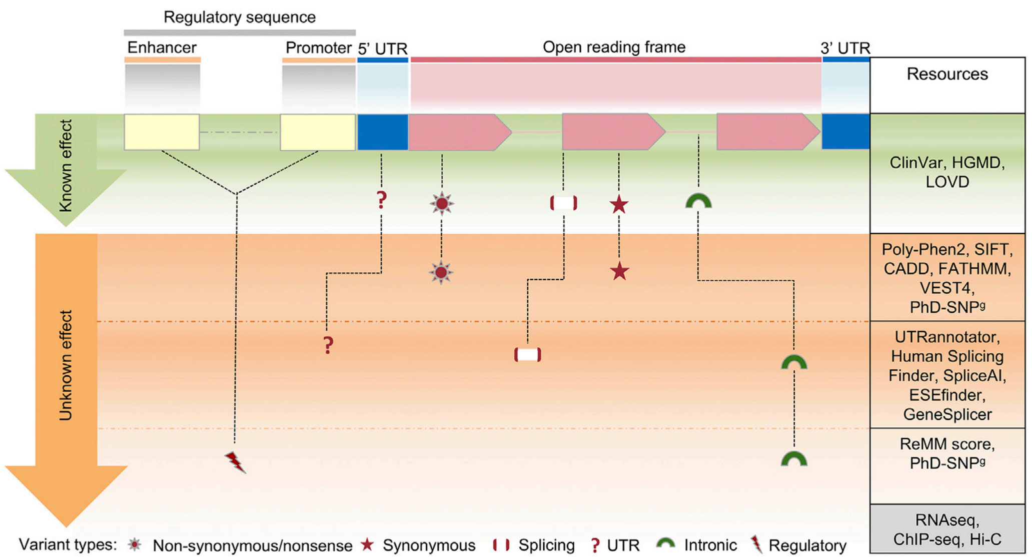


FIGURE 4
Schematic view of the clinical variant interpretation process. In a human protein-coding gene, a variant in the exons of an open reading frame can result in synonymous or nonsynonymous changes, while a variant in other areas (splice or intronic regions) can impact on splicing regulation. Changes within regulatory sequences (yellow and blue) can affect transcription and translation regulation of gene expression. On the right column, a selection of the most commonly used resources for variant interpretation is reported, distinguished by their gene location. Several methods are currently available to predict the effect of coding variants, however the interpretation of variants in deep-intronic regions or in regulatory elements is still challenging, due to the limited number of *in silico* prediction approaches. Such shortcomings can be overcome by parallel sequence analysis of the whole exome/genome together with multi-omics technologies, including RNA sequencing (transcriptome analysis), ChIP-seq (chromatin immunoprecipitation assay) and HiC (high-throughput chromosome conformation capture).

in human disease, such as that provided by HPO database (Section 2.4). Clinically relevant variants can be prioritized using public repositories reporting correlation between genetic variants and phenotypes, such as ClinVar and HGMD (Section 1.3; Supplementary Table S2).

4.2 Identification of pathogenic variants in RD associated genes

Herein, after the application of the aforementioned filters of variant frequency, phenotyping, and clinically relevant variants in disease genes, no variants are prioritized. In this case, to prioritize pathogenic/likely pathogenic (PLP) variants associated with the phenotype, ACMG/AMP guidelines for variant interpretation (Section 1.3) are used.

According to those guidelines, the pathogenicity of each variant is evaluated by gathering evidence from various sources: population data, computational and predictive data, functional data, and segregation data. Computational and predictive data are obtained by using several *in silico* prediction programs described in Section 2.3. Those tools are mainly devoted to the evaluation of the missense variants, which constitute a major set of VUS (Variant of Unknown Significance). The ACMG/AMP guidelines recommend complete concordance of predictions among all *in silico* algorithms used, without specifying the number or types of algorithms. However, many studies have provided rules for the classification of non-synonymous variants based on the integration of different prediction tools (Ghosh et al., 2017; Li et al., 2019; Nicora et al., 2022).

For phenotype characterization, the analysis of splicing variants is also relevant. The prioritization of splice site variants can be performed by web server tools, such as Human Splicing Finder (Desmet et al., 2009), MMSplice (Cheng et al., 2019), SpliceAI (Jaganathan et al., 2019) and CADD-splice (Rentzsch et al., 2021), that can highlight potential splicing-affecting variants outside the canonical splicing sites. ACMG/AMP variant classification can be achieved in such cases by using InterVar or wInterVar (Li and Wang, 2017), a web server that enables user-friendly variant interpretation with both an automated interpretation step and a manual adjustment step. Functional data that supports the pathogenic effect of newly discovered variants is not typically included in the standard diagnostic process of NGS genetic testing. Nevertheless, laboratories with an extensive experience in a specific disease area, can provide additional functional evaluation for new variants as part of their diagnostic protocols (Thouvenot et al., 2016; Ellingford et al., 2022).

Finally, segregation and allele data are fundamental to correctly assess the pathogenicity of variants. For this reason, in diagnostic settings the trio analysis, i.e., the combined genomic analysis of patient and parents, is strongly recommended (Alfares et al., 2020; French et al., 2022; Gabriel et al., 2022).

4.3 Identification of functional variants in genes with unknown RD association

Currently, the diagnostic process reaches a definitive diagnosis only in about 50% of the cases, leaving many patients with strongly-suspected genetic diseases without molecular explanations. In such cases, all variants with potential functional effects on any gene must be considered, under the hypothesis that the pathogenic role of the causative gene is still unknown. The initial filtering steps, similar to the previous scenarios, consist of removing all variants unlikely to be implicated in the disease, either because of low quality in exome or high frequency in population. Very stringent frequency thresholds are used, since it is likely that the considered disease is extremely rare. Then, the pedigree is analyzed to maintain only the variants that co-segregate with the phenotype according to any Mendelian transmission model. The variant-affected genes are prioritized to remove those that show a high grade of variability in the general population and to highlight those with a plausible biological role in the disease phenotype. The resulting set of genes with functional variants, poor population variability and biological compatibility is released in gene matching tools to search for other patients who are affected by alterations in the same genes (Section 2.2). Once a 'match' occurs, the researchers are connected through the system and can share molecular and clinical details of the patients, potentially concluding that they are both affected by the same disease, caused by the matched genes.

An example of successful gene-matching regards a 19-years-old girl seen at Federico II University Hospital, Naples, Italy. The girl was affected by a severe clinical picture, composed of complex brain malformations, extraocular muscle anomalies, severe intellectual disability, and drug-refractory epilepsy. Despite the presentation strongly suggesting an underlying genetic cause, thorough molecular and metabolic investigation failed to yield any plausible explanation. The patient was, then, enrolled in the Telethon Undiagnosed Diseases Program (TUDP) and underwent patient-parent trio ES. Variant filtering and manual revision did not find causative variants but highlighted those in four non-disease genes (PLEKHN1, NR5A2, TMEM89, DHX37). The patient's clinical and molecular descriptions were released in PhenomeCentral for gene-matching (Buske et al., 2015; Sobreira et al., 2017; Osmond et al., 2022), where only for DHX37 a consistent match with other patients with syndromic intellectual disability was found (Paine et al., 2019).

However, depending on disease type and patient selection, exome sequencing has been estimated to lead to a diagnosis in 30%–50% of rare Mendelian diseases (Frésard and Montgomery, 2018). A recent analysis shows that 14% of the recent diagnoses could be successfully performed by the combination of automatic and research approaches, looking for variants occurring in genomic regions poorly covered by exome sequencing (100,000 Genomes Project Pilot Investigators et al., 2021). Thus, the whole genome sequencing approach is becoming more relevant for the diagnosis of rare disorders (Turro et al., 2020). Accordingly, a large variety of computational approaches have been recently developed to score the impact of variants in noncoding regions (Shihab et al., 2015; Zhou and Troyanskaya, 2015; Ioannidis et al., 2016; Ionita-Laza et al., 2016; Capriotti and Fariselli, 2017; Rentzsch et al., 2019; Wells et al., 2019). In addition, for the interpretation of these potentially regulatory variants, the simultaneous and integrated use of multiple layers of omics technologies, such as whole-genome and transcriptome sequencing, is also increasingly being considered (Hasin et al., 2017; Kerr et al., 2020).

We expect that such methods will soon become the reference diagnostic tools in clinical settings. In this direction, a recent work describes approaches and discusses strategies for the diagnosis of rare and undiagnosed diseases, based on the analysis of the whole genome (Marwaha et al., 2022).

5 Data sharing and FAIRification

In the context of RDs, data sharing between institutions and across countries is crucial for maximising the potential of the generated genomic data (Saunders et al., 2019). It allows for the recruitment of larger cohorts of patients, thereby increasing statistical, and diagnostic, power. Sensitive RD patient data are collected by multiple institutions, whose registries are always difficult to aggregate. Sharing such data is essential for the development and maintenance of large databases, which are essential for federated analysis and discovery. In this context, the guiding principles of Findable, Accessible, Interoperable and Reusable (FAIR) data for humans and computers (Wilkinson et al., 2016) were developed, to ensure responsible sharing of health data and safeguarding of subjects. Since 2014, when "FAIR" acronym was first coined, and, because of their potential, FAIR principles have been widely endorsed by the RD community, the International Rare Diseases Research Consortium (IRDiRC) and the ELIXIR research infrastructure. In fact, adopting FAIR principles allows researchers and clinicians to integrate data from different resources in compliance with the restrictions of data accessibility, and thus answer questions involving multiple resources. For example, many types of genomic data, including features linked to the genomic coordinates of a reference genome, are always difficult to locate and access. A recent application of the FAIR principles to genomic data allowed the development of a track search service, which integrates metadata from various hubs, by adopting a set of recommendations for genomic data sharing (Gundersen et al., 2021). In addition, tools and pipelines developed for the analysis of genomic data, such as those described in this review, undoubtedly fall in the category of "research software", which is now considered part of FAIR by the European Commission. Indeed, FAIR principles can be applied not only to data, but to research software as well (Jiménez et al., 2017; Lamprecht et al., 2020).

A recent initiative, aiming at making FAIR ('FAIRification') 24 ERN (European Reference Networks) registries of RD patients, allowed collecting ninety-eight critical FAIRification challenges and proposing solutions to address them (Dos Santos Vieira et al., 2022). Awareness of the FAIRification challenges learned from initiatives like this one, which are strongly supported by the ELIXIR community, plays an important role in identifying solutions aimed at harmonizing RD data. Nevertheless, most resources collect unique data and there are wide differences in content, format, and language across them. This heterogeneity makes it virtually impossible to harmonize data from different resources, wasting the time and effort of data analysts and compromising any large-scale project aimed at improving RD research and supporting RD patients. It is therefore critical to put effort in the FAIRification process, both for humans and machines, so that data (including registries) can be queried in an unambiguous, global and federated way.

Inline with this need, the ELIXIR bio. tools portal (Ison et al., 2019) provides a comprehensive registry of software and databases that facilitates the search, understanding, use, and recognition of biomedical scientific resources. Among the 27,471 tools available on the portal, we identified 303 tools that are part of the *"Rare Diseases"* collection, domain, or topic and refer to a total of 165 functions described with the semantic terms of the EDAM ontology (Ison et al., 2013). After reviewing a list of 303 RD tools, we integrated them with other bio. tools methods, to develop a curated set of core resources for analyzing rare disease genomes.

The resources and tools collected in Supplementary Tables S2–S5 have been evaluated according to five criteria, related to their accessibility, update status, number of citations, and development stage (reported with "mature" tag in bio. tools). This type of evaluation, which assigns a score ranging from 1 to 5, represents a step toward the establishment of a standardized protocol for their clinical application.

6 Conclusions and future directions

Quick and accurate diagnosis are key issues for public health in general and for RDs in particular. The diagnostic delay for many RDs may at present reach up to decades (Molster et al., 2016; Heuyer et al., 2017), with an average time of about 4–5 years (Yan et al., 2020). In the journey towards diagnosis (also named the "*diagnostic odyssey*") patients may receive misdiagnosis and consequent inappropriate treatments and care. Diagnostic delay and misdiagnosis are due to many factors: RDs are infrequent, thus it is difficult to achieve a critical mass of data; data are sensitive, heterogeneous (clinical data, patient registries, variants, etc.) and usually fragmented (different communities and efforts collect data on specific RDs of interest using different formats, schemas, etc.) with poor interoperability, and a single, comprehensive repository for RDs does not exist.

In recent years, the development of new tools and resources, and the advances in data sharing practices and integrated analyses have allowed to reach an appropriate diagnosis for a sizable proportion of patients (Marwaha et al., 2022). Indeed, combining data from different sources, and using computational tools to analyze them in an integrated manner, is crucial to validate candidate variants, identify disease causative genes, perform genotype-phenotype associations, and elucidate the underlying molecular mechanism of a disease.

However, RD patients and expertise are still very scattered from each other, and knowledge and data sparsity, fragmentation, heterogeneity and poor interoperability often make integration and sharing of information extremely difficult if not impossible. Moreover, RD data are sensitive and recent technologies and practices gave rise to the further challenge of reconciling the benefits of data sharing and integration with privacy protection and ethical issues. Indeed, one of the major challenges nowadays consist in the implementation of reliable procedures for improving data sharing and the development of standardized tools and pipelines to enable reproducible research, while at the same time guaranteeing privacy rights.

To address these challenges, many international consortia have been established to create and integrate global infrastructure for RD research. At the European level, Solve-RD (solving the unsolved RDs, (Zurek et al., 2021), and RD-Connect (Lochmüller et al., 2018) enabled the creation of interdisciplinary teams to actively share and jointly analyze existing patient's data. These initiatives leverage existing computational infrastructures to share registries and standardize data among clinicians and scientists. In particular, the RD-Connect consortium promoted the development of the Genome-Phenome Analysis Platform (GPAP) (Laurie et al., 2022), and its integration with the PhenomeCentral (Osmond et al., 2022) and DECIPHER (Foreman et al., 2022). The GPAP platform facilitates the collation, discovery, sharing, and analysis of standardized genome-phenome data within a collaborative environment.

In this context, the implementation of a FAIR ecosystem of federated resources is essential for boosting research and diagnosis by decreasing RD data fragmentation and increasing data quality, with great advantages also in terms of time saving and sustainability. Although the developers and maintainers of the major RD resources and tools are already moving in the direction of FAIR data and software sharing, much still remains to be done to achieve the systematic application of FAIR principles by all players of the ecosystem, including data providers, data stewards and managers, software developers, researchers and clinicians, patients associations, research institutions, hospitals, and infrastructures. The transparent access to data and tools by the scientific community is recognized nowadays as one of the major challenges for improving RD diagnosis.

Author contributions

EC, CC, RC, AF, PF, LL, AM, PM, SP, MP, TP, CS, PT, and AV contributed to conception and design of the study. GB, SB, CS, EC, and LV performed the statistical analysis. EC wrote the first draft of the manuscript. GB, EC, CC, RC, AF, FI, LL, AM, PM, SP, MP, TP, RR, CS, PT, AV wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the following PRIN projects of the Ministero Istruzione, Università e Ricerca: "Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation" (PRIN201744NR8S), "Protein Bioinformatics for Human Health" (PRIN2017483NH8). DG is supported within the framework of CIR01_00017–"CNRBiOmics–Centro Nazionale di Ricerca in Bioinformatica per le Scienze Omiche"–Rafforzamento del capitale umano–CUP B56J20000960001.

Acknowledgments

The authors acknowledge financial support of the Rare Disease implementation study of the European Network for Life Science ELIXIR (IT-UNIBO-2022-Strategic 202201217). AM and BS also acknowledge financial support from Banca D'Italia.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1169109/full#supplementary-material

SUPPLEMENTARY TABLE S1

Analysis of the disease types in the Orphanet database.

SUPPLEMENTARY TABLE S2

Databases of ontologies, gene/protein networks and variants.

SUPPLEMENTARY TABLE S3

Tools for variant calling and annotation.

SUPPLEMENTARY TABLE S4

Resources for gene prioritization.

SUPPLEMENTARY TABLE S5

Resources and tools for variant interpretation.

SUPPLEMENTARY TABLE S6

Performance of 4 methods in the prediction of rare disease associated variants.

SUPPLEMENTARY DATA

Predictions of CADD, FATHMM, PhD-SNP^g and VEST4 on the dataset of 27,468 rare disease associated variants (sas-rd-202203) from 2,697 genes.

References

100,000 Genomes Project Pilot Investigators et al., 2021 100,000 Genomes Project Pilot InvestigatorsSmedley D., Smith K. R., Martin A., Thomas E. A., McDonagh E. M., et al. (2021). 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N. Engl. J. Med.* 385, 1868–1880. doi:10.1056/NEJMoa2035790

1000 Genomes Project Consortium et al., 2015 1000 Genomes Project ConsortiumAuton 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, 68–74. doi:10.1038/nature15393

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit7.20. doi:10.1002/0471142905.hg0720s76

Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., and Taylor, J. (2010). Galaxy CloudMan: Delivering cloud compute clusters. *BMC Bioinforma.* 11, S4. doi:10.1186/1471-2105-11-S12-S4

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi:10.1093/nar/gky379

Ahmed, Z., Renart, E. G., Mishra, D., and Zeeshan, S. (2021). JWES: A new pipeline for whole genome/exome sequence data processing, management, and gene-variant discovery, annotation, prediction, and genotyping. *FEBS Open Bio* 11, 2441–2452. doi:10.1002/2211-5463.13261

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi:10.1093/nar/gkw985

Alfares, A., Alsubaie, L., Aloraini, T., Alaskar, A., Althagafi, A., Alahmad, A., et al. (2020). What is the right sequencing approach? Solo VS extended family analysis in consanguineous populations. *BMC Med. Genomics* 13, 103. doi:10.1186/s12920-020-00743-8

All of Us Research Program InvestigatorsDenny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., et al. (2019). The "all of us" research program. *N. Engl. J. Med.* 381, 668–676. doi:10.1056/NEJMsr1809937

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

An, O., Tan, K.-T., Li, Y., Li, J., Wu, C.-S., Zhang, B., et al. (2020). CSI NGS portal: An online platform for automated NGS data analysis and sharing. *Int. J. Mol. Sci.* 21, E3828. doi:10.3390/ijms21113828

Andolfo, I., Martone, S., Rosato, B. E., Marra, R., Gambale, A., Forni, G. L., et al. (2021). Complex modes of inheritance in hereditary red blood cell disorders: A case series study of 155 patients. *Genes* 12, 958. doi:10.3390/genes12070958

Azzariti, D. R., and Hamosh, A. (2020). Genomic data sharing for novel mendelian disease gene discovery: The matchmaker exchange. *Annu. Rev. Genomics Hum. Genet.* 21, 305–326. doi:10.1146/annurev-genom-083118-014915

Bao, R., Hernandez, K., Huang, L., Kang, W., Bartom, E., Onel, K., et al. (2015). ExScalibur: A high-performance cloud-enabled suite for whole exome germline and somatic mutation identification. *PloS One* 10, e0135800. doi:10.1371/journal.pone.0135800

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi:10.1038/nrg2918

Basel-Salmon, L., Orenstein, N., Markus-Bustani, K., Ruhrman-Shahar, N., Kilim, Y., Magal, N., et al. (2019). Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 21, 1443–1451. doi:10.1038/s41436-018-0343-7

Bathke, J., and Lühken, G. (2021). OVarFlow: A resource optimized GATK 4 based open source variant calling workFlow. *BMC Bioinforma.* 22, 402. doi:10.1186/s12859-021-04317-y

Baxter, S. M., Posey, J. E., Lake, N. J., Sobreira, N., Chong, J. X., Buyske, S., et al. (2022). Centers for mendelian genomics: A decade of facilitating gene discovery. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 24, 784–797. doi:10.1016/j.gim.2021.12.005

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: Multi-source consolidation of human biological pathways. *Database J. Biol. Databases Curation* 2015, bav006. doi:10.1093/database/bav006

Benevenuta, S., Capriotti, E., and Fariselli, P. (2021). Calibrating variant-scoring methods for clinical decision making. *Bioinforma. Oxf. Engl.* 36, 5709–5711. doi:10.1093/bioinformatics/btaa943

Birgmeier, J., Haeussler, M., Deisseroth, C. A., Steinberg, E. H., Jagadeesh, K. A., Ratner, A. J., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* 12, eaau9113. doi:10.1126/scitranslmed.aau9113

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., et al. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* 15, 403. doi:10.1186/gb4161

Bodenreider, O. (2004). The unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270. doi:10.1093/nar/gkh061

Bonne, G. (2021). The Treatabolome, an emerging concept. *J. Neuromuscul. Dis.* 8, 337–339. doi:10.3233/JND-219003

Boycott, K. M., Dyment, D. A., and Innes, A. M. (2018). Unsolved recognizable patterns of human malformation: Challenges and opportunities. *Am. J. Med. Genet. C Semin. Med. Genet.* 178, 382–386. doi:10.1002/ajmg.c.31665

Boycott, K. M., Hartley, T., Biesecker, L. G., Gibbs, R. A., Innes, A. M., Riess, O., et al. (2019). A diagnosis for all rare genetic diseases: The horizon and the next Frontiers. *Cell* 177, 32–37. doi:10.1016/j.cell.2019.02.040

Buphamalai, P., Kokotovic, T., Nagy, V., and Menche, J. (2021). Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun.* 12, 6306. doi:10.1038/s41467-021-26674-1

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., et al. (2015). PhenomeCentral: A portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum. Mutat.* 36, 931–940. doi:10.1002/humu.22851

Cabrera-Andrade, A., López-Cortés, A., Jaramillo-Koupermann, G., Paz-Y-Miño, C., Pérez-Castillo, Y., Munteanu, C. R., et al. (2020). Gene prioritization through consensus strategy, enrichment methodologies analysis, and networking for osteosarcoma pathogenesis. *Int. J. Mol. Sci.* 21, E1053. doi:10.3390/ijms21031053

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10, 690–691. doi:10.1038/nmeth.2561

Calderone, A., Iannuccelli, M., Peluso, D., and Licata, L. (2020). Using the MINT database to search protein interactions. *Curr. Protoc. Bioinforma.* 69, e93. doi:10.1002/cpbi.93

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

Capriotti, E., and Fariselli, P. (2022). Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants. *Hum. Genet.* 141, 1649–1658. doi:10.1007/s00439-021-02419-4

Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinforma.* 9, S6. doi:10.1186/1471-2105-9-S2-S6

Capriotti, E., Martelli, P. L., Fariselli, P., and Casadio, R. (2017). Blind prediction of deleterious amino acid variations with SNPs&GO. *Hum. Mutat.* 38, 1064–1071. doi:10.1002/humu.23179

Capriotti, E., Ozturk, K., and Carter, H. (2019). Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 11, e1443. doi:10.1002/wsbm.1443

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 14, S3. doi:10.1186/1471-2164-14-S3-S3

Ceccarelli, F., Turei, D., Gabor, A., and Saez-Rodriguez, J. (2020). Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinforma. Oxf. Engl.* 36, 2632–2633. doi:10.1093/bioinformatics/btz968

Cezard, T., Cunningham, F., Hunt, S. E., Koylass, B., Kumar, N., Saunders, G., et al. (2022). The European variation archive: A FAIR resource of genomic variation for all species. *Nucleic Acids Res.* 50, D1216–D1220. doi:10.1093/nar/gkab960

Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinforma.* 8, 392. doi:10.1186/1471-2105-8-392

Chen, J., Aronow, B. J., and Jegga, A. G. (2009a). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinforma.* 10, 73. doi:10.1186/1471-2105-10-73

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009b). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi:10.1093/nar/gkp427

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., et al. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* 20, 48. doi:10.1186/s13059-019-1653-z

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A brief history of human disease genetics. *Nature* 577, 179–189. doi:10.1038/s41586-019-1879-7

Clerc, O., Deniaud, M., Vallet, S. D., Naba, A., Rivet, A., Perez, S., et al. (2019). MatrixDB: Integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.* 47, D376–D381. doi:10.1093/nar/gky1035

de la Fuente, L., Del Pozo-Valero, M., Perea-Romero, I., Blanco-Kelly, F., Fernández-Caballero, L., Cortón, M., et al. (2023). Prioritization of new candidate genes for rare genetic diseases by a disease-aware evaluation of heterogeneous molecular networks. *Int. J. Mol. Sci.* 24, 1661. doi:10.3390/ijms24021661

De Las Rivas, J., and Fontanillo, C. (2012). Protein-protein interaction networks: Unraveling the wiring of molecular machines within the cell. *Brief. Funct. Genomics* 11, 489–496. doi:10.1093/bfgp/els036

De Marinis, I., Lo Surdo, P., Cesareni, G., and Perfetto, L. (2021). SIGNORApp: A Cytoscape 3 application to access SIGNOR data. *Bioinforma. Oxf. Engl.* 38, 1764–1766. btab865. doi:10.1093/bioinformatics/btab865

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 Bioinforma.* 12, 151. doi:10.1186/1471-2105-12-151

Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., et al. (2022). The IntAct database: Efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* 50, D648–D653. doi:10.1093/nar/gkab1006

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human splicing finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, e67. doi:10.1093/nar/gkp215

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi:10.1038/nbt.3820

Dos Santos Vieira, B., Bernabé, C. H., Zhang, S., Abaza, H., Benis, N., Cámara, A., et al. (2022). Towards FAIRification of sensitive and fragmented rare disease patient data: Challenges and solutions in European reference network registries. *Orphanet J. Rare Dis.* 17, 436. doi:10.1186/s13023-022-02558-5

Edgar, R., Mazor, Y., Rinon, A., Blumenthal, J., Golan, Y., Buzhor, E., et al. (2013). LifeMap Discovery$^{\text{TM}}$: The embryonic development, stem cells, and regenerative medicine research portal. *PloS One* 8, e66629. doi:10.1371/journal.pone.0066629

Eldomery, M. K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J. A., Gambin, T., Stray-Pedersen, A., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med.* 9, 26. doi:10.1186/s13073-017-0412-6

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., et al. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* 14, 73. doi:10.1186/s13073-022-01073-3

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi:10.1038/s41587-020-0439-x

Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 48, D941–D947. doi:10.1093/nar/gkz836

Fariselli, P., Martelli, P. L., Savojardo, C., and Casadio, R. (2015). INPS: Predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinforma. Oxf. Engl.* 31, 2816–2821. doi:10.1093/bioinformatics/btv291

Ferreira, C. R. (2019). The burden of rare diseases. *Am. J. Med. Genet. A* 179, 885–892. doi:10.1002/ajmg.a.61124

Foreman, J., Brent, S., Perrett, D., Bevan, A. P., Hunt, S. E., Cunningham, F., et al. (2022). DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research. *Hum. Mutat.* 43, 682–697. doi:10.1002/humu.24340

Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J. I., Jene, A., et al. (2022). The European genome-phenome archive in 2021. *Nucleic Acids Res.* 50, D980–D987. doi:10.1093/nar/gkab1059

French, C. E., Dolling, H., Mégy, K., Sanchis-Juan, A., Kumar, A., Delon, I., et al. (2022). Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood. *HGG Adv.* 3, 100113. doi:10.1016/j.xhgg.2022.100113

Frésard, L., and Montgomery, S. B. (2018). Diagnosing rare diseases after the exome. *Cold Spring Harb. Mol. Case Stud.* 4, a003392. doi:10.1101/mcs.a003392

Gabriel, H., Korinth, D., Ritthaler, M., Schulte, B., Battke, F., von Kaisenberg, C., et al. (2022). Trio exome sequencing is highly relevant in prenatal diagnostics. *Prenat. Diagn.* 42, 845–851. doi:10.1002/pd.6081

Gainotti, S., Torreri, P., Wang, C. M., Reihs, R., Mueller, H., Heslop, E., et al. (2018). The RD-connect registry and biobank finder: A tool for sharing aggregated data and metadata among rare disease researchers. *Eur. J. Hum. Genet.* 26, 631–643. doi:10.1038/s41431-017-0085-z

Gao, X., Xu, J., and Starmer, J. (2015). Fastq2vcf: A concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Res. Notes* 8, 72. doi:10.1186/s13104-015-1027-x

Ghosh, R., Oak, N., and Plon, S. E. (2017). Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol.* 18, 225. doi:10.1186/s13059-017-1353-5

Gill, N., Singh, S., and Aseri, T. C. (2014). Computational disease gene prioritization: An appraisal. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 21, 456–465. doi:10.1089/cmb.2013.0158

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U A* 104, 8685–8690. doi:10.1073/pnas.0701361104

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., et al. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* 43, 1012–1030. doi:10.1002/humu.24309

Gundersen, S., Boddu, S., Capella-Gutierrez, S., Drabløs, F., Fernández, J. M., Kompova, R., et al. (2021). Recommendations for the FAIRification of genomic track metadata. *F1000Research* 10, ELIXIR–268. doi:10.12688/f1000research.28449.1

Guo, Y., Ding, X., Shen, Y., Lyon, G. J., and Wang, K. (2015). SeqMule: Automated pipeline for analysis of human exome/genome sequencing data. *Sci. Rep.* 5, 14283. doi:10.1038/srep14283

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., et al. (2020). How many rare diseases are there? *Nat. Rev. Drug Discov.* 19, 77–78. doi:10.1038/d41573-019-00180-y

Hartin, S. N., Means, J. C., Alaimo, J. T., and Younger, S. T. (2020). Expediting rare disease diagnosis: A call to bridge the gap between clinical and functional genomics. *Mol. Med. Camb. Mass* 26, 117. doi:10.1186/s10020-020-00244-5

Hartley, T., Balcı, T. B., Rojas, S. K., Eaton, A., Canada, C. R., Dyment, D. A., et al. (2018). The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. *Am. J. Med. Genet. C Semin. Med. Genet.* 178, 458–463. doi:10.1002/ajmg.c.31662

Hartley, T., Lemire, G., Kernohan, K. D., Howley, H. E., Adams, D. R., and Boycott, K. M. (2020). New diagnostic approaches for undiagnosed rare genetic diseases. *Annu. Rev. Genomics Hum. Genet.* 21, 351–372. doi:10.1146/annurev-genom-083118-015345

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1

Hayashi, S., and Umeda, T. (2008). 35 years of Japanese policy on rare diseases. *Lancet lond. Engl.* 372, 889–890. doi:10.1016/S0140-6736(08)61393-8

Heuyer, T., Pavan, S., and Vicard, C. (2017). The health and life path of rare disease patients: Results of the 2015 French barometer. *Patient Relat. Outcome Meas.* 8, 97–110. doi:10.2147/PROM.S131033

humsavar UniProt (2023). *UniProt humsavar*. Available at: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/humsavar.txt (Accessed Jan, 2023).

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., et al. (2018). Ensembl variation resources. *Database J. Biol. Databases Curation* 2018, bay119. doi:10.1093/database/bay119

IMEx Consortium CuratorsDel-Toro, N., Duesbury, M., Koch, M., Perfetto, L., Shrivastava, A., et al. (2019). Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat. Commun.* 10, 10. doi:10.1038/s41467-018-07709-6

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., et al. (2016). Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, 877–885. doi:10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48, 214–220. doi:10.1038/ng.3477

Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., et al. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinforma. Oxf. Engl.* 29, 1325–1332. doi:10.1093/bioinformatics/btt113

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., et al. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol.* 20, 164. doi:10.1186/s13059-019-1772-6

Jacobsen, J. O. B., Kelly, C., Cipriani, V., Research Consortium, G. E., Mungall, C. J., Reese, J., et al. (2022). Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. *Hum. Mutat.* 43, 1071–1081. doi:10.1002/humu.24380

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548. doi:10.1016/j.cell.2018.12.015

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi:10.1093/nar/gkz1031

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., et al. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research* 6, ELIXIR-876. doi:10.12688/f1000research.11407.1

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

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

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T.-K., Lua, R. C., Wilkins, A. D., et al. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Sci. Publ. Protein Soc.* 23, 1650–1666. doi:10.1002/pro.2552

Katsonis, P., Wilhelm, K., Williams, A., and Lichtarge, O. (2022). Genome interpretation using *in silico* predictors of variant impact. *Hum. Genet.* 141, 1549–1577. doi:10.1007/s00439-022-02457-6

Kerr, K., McAneney, H., Smyth, L. J., Bailie, C., McKee, S., and McKnight, A. J. (2020). A scoping review and proposed workflow for multi-omic rare disease research. *Orphanet J. Rare Dis.* 15, 107. doi:10.1186/s13023-020-01376-x

Kinjo, S., Monma, N., Misu, S., Kitamura, N., Imoto, J., Yoshitake, K., et al. (2018). Maser: One-stop platform for NGS big data from analysis to visualization. *Database J. Biol. Databases Curation* 2018, bay027. doi:10.1093/database/bay027

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Med.* 12, 91. doi:10.1186/s13073-020-00791-w

Köster, J., and Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi:10.1093/bioinformatics/bts480

Kutmon, M., Lotia, S., Evelo, C. T., and Pico, A. R. (2014). WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research* 3, 152. doi:10.12688/f1000research.4254.2

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinforma.* 16, 116. doi:10.1186/s12859-015-0548-6

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, , R., Martin Del Pico, E., et al. (2020). Towards FAIR principles for research software. *Data Sci.* 3, 37–59. doi:10.3233/DS-190026

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

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi:10.1093/nar/gks1213

Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., et al. (2022). The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. *Hum. Mutat.* 43, 717–733. doi:10.1002/humu.24353

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, 285–291. doi:10.1038/nature19057

Li, Q., and Wang, K. (2017). InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.* 100, 267–280. doi:10.1016/j.ajhg.2017.01.004

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. doi:10.1126/science.aad9417

Li, Q., Zhao, K., Bustamante, C. D., Ma, X., and Wong, W. H. (2019). Xrare: A machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 21, 2126–2134. doi:10.1038/s41436-019-0439-8

Liu, X., Yang, Z., Lin, H., Simmons, M., and Lu, Z. (2017). DIGNiFI: Discovering causative genes for orphan diseases using protein-protein interaction networks. *BMC Syst. Biol.* 11, 23. doi:10.1186/s12918-017-0402-8

Liu, Z., Zhu, L., Roberts, R., and Tong, W. (2019). Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: Where are we? *Trends Genet. TIG* 35, 852–867. doi:10.1016/j.tig.2019.08.006

Lo Surdo, P., Iannuccelli, M., Contino, S., Castagnoli, L., Licata, L., Cesareni, G., et al. (2023). SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* 51, D631–D637. doi:10.1093/nar/gkac883

Lochmüller, H., Badowska, D. M., Thompson, R., Knoers, N. V., Aartsma-Rus, A., Gut, I., et al. (2018). RD-connect, NeurOmics and EURenOmics: Collaborative European initiative for rare diseases. *Eur. J. Hum. Genet. EJHG* 26, 778–785. doi:10.1038/s41431-018-0115-5

Manfredi, M., Savojardo, C., Martelli, P. L., and Casadio, R. (2022). E-SNPs&GO: Embedding of protein sequence and function improves the annotation of human pathogenic variants. *Bioinforma. Oxf. Engl.* 38, 5168–5174. doi:10.1093/bioinformatics/btac678

Marabotti, A., Scafuri, B., and Facchiano, A. (2020). Predicting the stability of mutant proteins by computational approaches: An overview. *Brief. Bioinform.* 22, bbaa074. doi:10.1093/bib/bbaa074

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., et al. (2021). WikiPathways: Connecting communities. *Nucleic Acids Res.* 49, D613–D621. doi:10.1093/nar/gkaa1024

Marwaha, S., Knowles, J. W., and Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. *Genome Med.* 14, 23. doi:10.1186/s13073-022-01026-w

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347, 1257601. doi:10.1126/science.1257601

Molster, C., Urwin, D., Di Pietro, L., Fookes, M., Petrie, D., van der Laan, S., et al. (2016). Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J. Rare Dis.* 11, 30. doi:10.1186/s13023-016-0409-z

Monaco, L., Zanello, G., Baynam, G., Jonker, A. H., Julkowska, D., Hartman, A. L., et al. (2022). Research on rare diseases: Ten years of progress and challenges at IRDiRC. *Nat. Rev. Drug Discov.* 21, 319–320. doi:10.1038/d41573-022-00019-z

Montanucci, L., Capriotti, E., Birolo, G., Benevenuta, S., Pancotti, C., Lal, D., et al. (2022). DDGun: An untrained predictor of protein stability changes upon amino acid variants. *Nucleic Acids Res.* 50, W222–W227. gkac325. doi:10.1093/nar/gkac325

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

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., et al. (2020). Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *Eur. J. Hum. Genet.* 28, 165–173. doi:10.1038/s41431-019-0508-0

Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R., and Magni, P. (2022). A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci. Rep.* 12, 2517. doi:10.1038/s41598-022-06547-3

Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PloS One* 10, e0117380. doi:10.1371/journal.pone.0117380

O'Connor, B. D., Yuen, D., Chung, V., Duncan, A. G., Liu, X. K., Patricia, J., et al. (2017). The Dockstore: Enabling modular, community-focused sharing of docker-based genomics tools and workflows. *F1000Research* 6, 52. doi:10.12688/f1000research.10137.1

Osmond, M., Hartley, T., Johnstone, B., Andjic, S., Girdea, M., Gillespie, M., et al. (2022). PhenomeCentral: 7 years of rare disease matchmaking. *Hum. Mutat.* 43, 674–681. doi:10.1002/humu.24348

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* 30, 187–200. doi:10.1002/pro.3978

Özkan, S., Padilla, N., Moles-Fernández, A., Diez, O., Gutiérrez-Enríquez, S., and de la Cruz, X. (2021). "Chapter 6 - the computational approach to variant interpretation: Principles, results, and applicability," in *Clinical DNA variant interpretation. Translational and applied genomics*. Editors C. Lázaro, J. Lerner-Ellis, and A. Spurdle (Academic Press), 89–119. doi:10.1016/B978-0-12-820519-8.00007-7

Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* 9, e1003153. doi:10.1371/journal.pcbi.1003153

Paine, I., Posey, J. E., Grochowski, C. M., Jhangiani, S. N., Rosenheck, S., Kleyner, R., et al. (2019). Paralog studies augment gene discovery: DDX and DHX genes. *Am. J. Hum. Genet.* 105, 302–316. doi:10.1016/j.ajhg.2019.06.001

Pais, L. S., Snow, H., Weisburd, B., Zhang, S., Baxter, S. M., DiTroia, S., et al. (2022). seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum. Mutat.* 43, 698–707. doi:10.1002/humu.24366

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., et al. (2022). Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* 23, bbab555. doi:10.1093/bib/bbab555

Pastrello, C., Kotlyar, M., and Jurisica, I. (2020). Informed use of protein-protein interaction data: A focus on the integrated interactions database (IID). *Methods Mol. Biol. Clifton N. J.* 2074, 125–134. doi:10.1007/978-1-4939-9873-9_10

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, 5918. doi:10.1038/s41467-020-19669-x

Petrosino, M., Novak, L., Pasquo, A., Chiaraluce, R., Turina, P., Capriotti, E., et al. (2021). Analysis and interpretation of the impact of missense variants in cancer. *Int. J. Mol. Sci.* 22, 5416. doi:10.3390/ijms22115416

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

Pires, D. E. V., Rodrigues, C. H. M., and Ascher, D. B. (2020). mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res.* 48, W147–W153. doi:10.1093/nar/gkaa416

Piro, R. M., and Di Cunto, F. (2012). Computational approaches to disease-gene prediction: Rationale, classification and successes. *FEBS J.* 279, 678–696. doi:10.1111/j.1742-4658.2012.08471.x

Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V., Aguiar, L. R., de Carvalho, J. L., et al. (2018). Rare genetic diseases: Update on diagnosis, treatment and online resources. *Drug Discov. Today* 23, 187–195. doi:10.1016/j.drudis.2017.11.002

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. doi:10.1038/nbt.4235

Porras, P., Barrera, E., Bridge, A., Del-Toro, N., Cesareni, G., Duesbury, M., et al. (2020). Towards a unified open access dataset of molecular interactions. *Nat. Commun.* 11, 6144. doi:10.1038/s41467-020-19942-z

Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinforma. Oxf. Engl.* 32, 2936–2946. doi:10.1093/bioinformatics/btw361

Quang, D., Chen, Y., and Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763. doi:10.1093/bioinformatics/btu703

Quinodoz, M., Peter, V. G., Cisarova, K., Royer-Bertrand, B., Stenson, P. D., Cooper, D. N., et al. (2022). Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am. J. Hum. Genet.* 109, 457–470. doi:10.1016/j.ajhg.2022.01.006

Ragueneau, E., Shrivastava, A., Morris, J. H., Del-Toro, N., Hermjakob, H., and Porras, P. (2021). IntAct App: A Cytoscape application for molecular interaction network visualization and analysis. *Bioinforma. Oxf. Engl.* 37, 3684–3685. doi:10.1093/bioinformatics/btab319

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., et al. (2017). DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* 45, W201–W206. doi:10.1093/nar/gkx390

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

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: The Orphanet approach to serve a wide range of end users. *Hum. Mutat.* 33, 803–808. doi:10.1002/humu.22078

Regulation Orphan Medicinal Product (2000). *Regulation (EC) No 141/2000 of the European parliament and of the council of 16 december 1999 on orphan medicinal products*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000R0141.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., et al. (2015). ClinGen--the clinical genome resource. *N. Engl. J. Med.* 372, 2235–2242. doi:10.1056/NEJMsr1406261

Reiter, T., Brooks, P. T., Irber, L., Joslin, S. E. K., Reid, C. M., Scott, C., et al. (2021). Streamlining data-intensive biology with workflow systems. *GigaScience* 10, giaa140. doi:10.1093/gigascience/giaa140

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi:10.1093/nar/gky1016

Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* 13, 31. doi:10.1186/s13073-021-00835-9

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 17, 405–424. doi:10.1038/gim.2015.30

Robinson, P. N., Kohler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615. doi:10.1016/j.ajhg.2008.09.017

Robinson, P. N., Kohler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C. J., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* 24, 340–348. doi:10.1101/gr.160325.113

Robinson, P. N., Ravanmehr, V., Jacobsen, J. O. B., Danis, D., Zhang, X. A., Carmody, L. C., et al. (2020). Interpretable clinical genomics with a likelihood ratio paradigm. *Am. J. Hum. Genet.* 107, 403–417. doi:10.1016/j.ajhg.2020.06.021

Rogers, F. B. (1963). Medical subject headings. *Bull. Med. Libr. Assoc.* 51, 114–116.

Rojano, E., Seoane, P., Ranea, J. A. G., and Perkins, J. R. (2019). Regulatory variants: From detection to predicting impact. *Brief. Bioinform* 20, 1639–1654. doi:10.1093/bib/bby039

Rother, K., Potrzebowski, W., Puton, T., Rother, M., Wywial, E., and Bujnicki, J. M. (2012). A toolbox for developing bioinformatics software. *Brief. Bioinform.* 13, 244–257. doi:10.1093/bib/bbr035

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi:10.1093/nar/gkh086

Sandmann, S., Karimi, M., de Graaf, A. O., Rohde, C., Göllner, S., Varghese, J., et al. (2018). appreci8: a pipeline for precise variant calling integrating 8 tools. *Bioinforma. Oxf. Engl.* 34, 4205–4212. doi:10.1093/bioinformatics/bty518

Saunders, G., Baudis, M., Becker, R., Beltran, S., Béroud, C., Birney, E., et al. (2019). Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* 20, 693–701. doi:10.1038/s41576-019-0156-9

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. *Bioinforma. Oxf. Engl.* 32, 2542–2544. doi:10.1093/bioinformatics/btw192

Savojardo, C., Baldazzi, D., Babbi, G., Martelli, P. L., and Casadio, R. (2022). Mapping human disease-associated enzymes into Reactome allows characterization of disease groups and their interactions. *Sci. Rep.* 12, 17963. doi:10.1038/s41598-022-22818-5

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., et al. (2022). Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics* 2, 100085. doi:10.1016/j.xgen.2021.100085

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., et al. (2012). Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res.* 40, D940–D946. doi:10.1093/nar/gkr972

Scotti, M. M., and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* 17, 19–32. doi:10.1038/nrg.2015.3

Setty, S. T., Scott-Boyer, M.-P., Cuppens, T., and Droit, A. (2022). New developments and possibilities in reanalysis and reinterpretation of whole exome sequencing datasets for unsolved rare diseases using machine learning approaches. *Int. J. Mol. Sci.* 23, 6792. doi:10.3390/ijms23126792

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., et al. (2020). The Monarch initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 48, D704–D715. doi:10.1093/nar/gkz997

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., et al. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536–1543. doi:10.1093/bioinformatics/btv009

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457. doi:10.1093/nar/gks539

Smedley, D., and Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* 7, 81. doi:10.1186/s13073-015-0199-2

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.* 99, 595–606. doi:10.1016/j.ajhg.2016.07.005

Sobreira, N. L. M., Arachchi, H., Buske, O. J., Chong, J. X., Hutton, B., Foreman, J., et al. (2017). Matchmaker exchange. *Curr. Protoc. Hum. Genet.* 95, 9.31.1–9.31.15. doi:10.1002/cphg.50

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., et al. (2023). The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985. doi:10.1093/nar/gkac1010

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, 575–588. doi:10.1016/j.tibs.2019.01.003

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., et al. (2016a). VarElect: The phenotype-based variation prioritizer of the GeneCards suite. *BMC Genomics* 17, 444. doi:10.1186/s12864-016-2722-2

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016b). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* 54, 1.30.1–1.30.33. doi:10.1002/cpbi.5

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., et al. (2020). The human gene mutation database (HGMD®): Optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* 139, 1197–1207. doi:10.1007/s00439-020-02199-3

Strande, N. T., Riggs, E. R., Buchanan, A. H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S. S., et al. (2017). Evaluating the clinical validity of gene-disease associations: An evidence-based framework developed by the clinical genome resource. *Am. J. Hum. Genet.* 100, 895–906. doi:10.1016/j.ajhg.2017.04.015

Summers, K. M. (1996). Relationship between genotype and phenotype in monogenic diseases: Relevance to polygenic diseases. *Hum. Mutat.* 7, 283–293. doi:10.1002/(SICI)1098-1004(1996)7:4<283::AID-HUMU1>3.0.CO;2-A

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Tabarini, N., Biagi, E., Uva, P., Iovino, E., Pippucci, T., Seri, M., et al. (2022). Exploration of tools for the interpretation of human non-coding variants. *Int. J. Mol. Sci.* 23, 12977. doi:10.3390/ijms232112977

Tavtigian, S. V., Harrison, S. M., Boucher, K. M., and Biesecker, L. G. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum. Mutat.* 41, 1734–1737. doi:10.1002/humu.24088

Thouvenot, P., Ben Yamin, B., Fourrière, L., Lescure, A., Boudier, T., Del Nery, E., et al. (2016). Functional assessment of genetic variants with outcomes adapted to clinical decision-making. *PLoS Genet.* 12, e1006096. doi:10.1371/journal.pgen.1006096

Tran, L., Hamp, T., and Rost, B. (2018). ProfPPIdb: Pairs of physical protein-protein interactions predicted for entire proteomes. *PloS One* 13, e0199988. doi:10.1371/journal.pone.0199988

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi:10.1038/nmeth.4077

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., et al. (2018). The 100 000 genomes project: Bringing whole genome sequencing to the NHS. *BMJ* 361, k1687. doi:10.1136/bmj.k1687

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., et al. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* 583, 96–102. doi:10.1038/s41586-020-2434-2

U.S. Food and Drug Administration (2022). Medical products for rare diseases and conditions. Available at: https://www.fda.gov/industry/medical-products-rare-diseases-and-conditions (Accessed Jan, 2023).

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419. doi:10.1126/science.1260419

UK10K ConsortiumWalter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90. doi:10.1038/nature14962

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., et al. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* 10, 5241. doi:10.1038/s41467-019-13212-3

Wenger, A. M., Guturu, H., Bernstein, J. A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 19, 209–214. doi:10.1038/gim.2016.88

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

World Health Organization (2019). *International classification of diseases (ICD)*. Available at: https://www.who.int/standards/classifications/classification-of-diseases (Accessed Jan, 2023).

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

Yan, X., He, S., and Dong, D. (2020). Determining how far an adult rare disease patient needs to travel for a definitive diagnosis: A cross-sectional examination of the 2018 national rare disease survey in China. *Int. J. Environ. Res. Public. Health* 17, E1757. doi:10.3390/ijerph17051757

Yang, Y., Urolagin, S., Niroula, A., Ding, X., Shen, B., and Vihinen, M. (2018). PON-tstab: Protein variant stability predictor. Importance of training data quality. *Int. J. Mol. Sci.* 19, 1009. doi:10.3390/ijms19041009

Yuan, X., Wang, J., Dai, B., Sun, Y., Zhang, K., Chen, F., et al. (2022). Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* 23, bbac019. doi:10.1093/bib/bbac019

Zhang, P., and Itan, Y. (2019). Biological network approaches and applications in rare disease studies. *Genes* 10, 797. doi:10.3390/genes10100797

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931–934. doi:10.1038/nmeth.3547

Zhu, C., Kushwaha, A., Berman, K., and Jegga, A. G. (2012). A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst. Biol.* 6, S8. doi:10.1186/1752-0509-6-S3-S8

Zhu, Q., Nguyen, D.-T., Sid, E., and Pariser, A. (2020). Leveraging the UMLS as a data standard for rare disease data normalization and harmonization. *Methods Inf. Med.* 59, 131–139. doi:10.1055/s-0040-1718940

Zolotareva, O., and Kleine, M. (2019). A survey of gene prioritization tools for mendelian and complex human diseases. *J. Integr. Bioinforma.* 16, 20180069. doi:10.1515/jib-2018-0069

Zurek, B., Ellwanger, K., Vissers, L. E. L. M., Schüle, R., Synofzik, M., Töpf, A., et al. (2021). Solve-RD: Systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet. EJHG* 29, 1325–1331. doi:10.1038/s41431-021-00859-0

Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 01 June 2023
DOI 10.3389/fmolb.2023.1122269

Check for updates

OPEN ACCESS

EDITED BY
Sofia R. Pauleta,
New University of Lisbon, Portugal

REVIEWED BY
Adam Liwo,
University of Gdansk, Poland
Ayori Mitsutake,
Meiji University, Japan

*CORRESPONDENCE
Giovanni La Penna,
✉ giovanni.lapenna@cnr.it

RECEIVED 12 December 2022
ACCEPTED 16 May 2023
PUBLISHED 01 June 2023

CITATION
Nobili G, Botticelli S, La Penna G, Morante S, Rossi G and Salina G (2023), Probing protein stability: towards a computational atomistic, reliable, affordable, and improvable model. *Front. Mol. Biosci.* 10:1122269. doi: 10.3389/fmolb.2023.1122269

Probing protein stability: towards a computational atomistic, reliable, affordable, and improvable model

Germano Nobili[1,2], Simone Botticelli[1,2], Giovanni La Penna[3,2]*, Silvia Morante[1,2,3], Giancarlo Rossi[1,2,4] and Gaetano Salina[2]

[1]Dipartimento di Fisica, Universitá di Roma Tor Vergata, Roma, Italy, [2]INFN, Sezione di Roma Tor Vergata, Roma, Italy, [3]CNR—Istituto di Chimica Dei Composti Organometallici, Firenze, Italy, [4]Museo Storico della Fisica e Centro Studi e Ricerche E. Fermi, Roma, Italy

We present an improved application of a recently proposed computational method designed to evaluate the change of free energy as a function of the average value of a suitably chosen collective variable in proteins. The method is based on a full atomistic description of the protein and its environment. The goal is to understand how the protein melting temperature changes upon single-point mutations, because the sign of the temperature variation will allow us to discriminate stabilizing vs. destabilizing mutations in protein sequences. In this refined application the method is based on altruistic well-tempered metadynamics, a variant of multiple-walkers metadynamics. The resulting metastatistics is then modulated by the maximal constrained entropy principle. The latter turns out to be especially helpful in free-energy calculations as it is able to alleviate the severe limitations of metadynamics in properly sampling folded and unfolded configurations. In this work we apply the computational strategy outlined above in the case of the bovine pancreatic trypsin inhibitor, a well-studied small protein, which is a reference for computer simulations since decades. We compute the variation of the melting temperature characterizing the folding-unfolding process between the wild-type protein and two of its single-point mutations that are seen to have opposite effect on the free energy changes. The same approach is used for free energy difference calculations between a truncated form of frataxin and a set of five of its variants. Simulation data are compared to *in vitro* experiments. In all cases the sign of the change of melting temperature is reproduced, under the further approximation of using an empirical effective mean-field to average out protein-solvent interactions.

KEYWORDS

protein stability, protein variant, free energy, unfolding, metastatistics

1 Introduction

Many proteins are characterized by a given three-dimensional structure when they are observed in a water soluble monomeric state Branden and Tooze (1999). In order to understand the way the sequence determines the structure, the effect of single point mutations has been studied since a long time Cunningham and Wells (1989). A simple way to address the sequence-structure interplay is to measure some structural parameter as a function of temperature. Circular dichroism (CD) and many other techniques are often used and in many cases the change of this structural parameter with temperature can be taken as an indicator of the melting of the protein structure Cantor and Schimmel (1980). The change of melting temperature triggered by different single point mutations is therefore a widely used measure of the change of protein stability upon a localized change of the protein sequence and large archives of such data have been collected Guerois et al. (2002); Alexov and Sternberg (2013); Forbes et al. (2016). When this information is available, it must be interpreted in terms of the reshaping of atomic interactions.

The change of protein stability upon sequence mutations has implications in many pathologies. One example is Friedreich's ataxia, an autosomal-recessive genetic condition that causes ataxia, sensory loss and cardiomyopathy worsening over time Pandolfo (2009); Klockgether (2011); Clark et al. (2018). The cause of the disease is in mutations of the gene encoding for the frataxin (FXN) protein. Depending on the specific kind of mutation, a patient may end up with an insufficient level of frataxin, a nonfunctional frataxin or frataxin, that is, not correctly localized in the mitochondria Delatycki et al. (2000); Galea et al. (2016). Frataxin variants have also a role in cancer, as expected because of the involvement of FXN and mitochondria in the control of oxidative metabolism Schulz et al. (2006). Indeed, missense variants are found in multiple human cancer tissues Petrosino et al. (2019, 2021). The example of FXN shows that even single point mutations can have significant impact in protein stability, trafficking, plasticity, interactions with local environment and mutual interactions with other macromolecules.

Many models have been proposed to relate measured changes in protein stability with the chemical nature of the protein sequence. High through-put approaches based on atomic models have been recently developed Steinbrecher et al. (2017). Many of these approaches are summarized in this Special Issue.

The method we would like to propose here aims at predicting the change of thermal stability of a protein in a monomeric water-soluble state, when its sequence is changed by a single aminoacid. The method is based on a suitable modelling of interatomic forces, i.e., it is atomistic, and includes an explicit model of the water solution. The method was initially applied to FXN Botticelli et al. (2022) and it is here refined and discussed in more detail, to achieve better computational performance and higher accuracy in prediction. In particular, we use here the well-tempered metadynamics, one of the best performing method to enhance sampling of phase space in atomistic models. A small reference protein of 58 residues is first used to assess the methods and to understand limitations and advantages.

Many initial configurations of the protein of interest are generated, assuming the protein structure representing the native state of the wild-type sequence, but with initial conditions diversified as much as possible. This is achieved by changing the protein environment, that is, in this case the water solution of NaCl. A multiple walkers metadynamics simulation is carried out Raiteri et al. (2006); Hošek et al. (2016); Hošek et al. (2017), building an external biasing potential as a function of a suitably chosen collective variable. In the range of values spanned by the collective variable folded and unfolded protein structures are sampled.

The external potential is built so as to initially unfold the protein structure. All along the metadynamics simulation time, the external biasing potential is systematically constructed and updated to uniformly sample folded and unfolded configurations. This goal is best achieved by combining multiple walkers histories into a unique trajectory Hošek et al. (2017).

The standard analysis of multiple walkers metadynamics can be performed, but limitations in predicting experimental behaviours arise because of the huge number of configurations required to achieve a good convergence and stability. We therefore decided to exploit the features of the maximal constrained entropy approach which allows to properly re-modulate the collected set of sampled configurations by imposing suitable conformational constraints. In this way we can reliably monitor the change in free energy as a function of average values of the chosen collective variables. The quantity of interest to be calculated is therefore:

$$\Delta\Delta G = \Delta G_X - \Delta G_0 = [G_X(s) - G_X(s_0)] - [G_0(s) - G_0(s_0)], \quad (1)$$

where the subscript denotes a sequence (the pedix 0 indicates a reference sequence, usually the wild-type), and the s state variable indicates the degree of structural order, where s_0 indicates the folded native state and s indicates the unfolded state.

The approach is first applied to the study of a paradigmatic protein displaying a well defined three-dimensional structure, namely, to the bovine pancreatic trypsine inhibitor (BPTI), where the effect of a set of mutations on melting temperature has been carefully investigated Yu et al. (1995). BPTI has always been a milestone for folding studies, being one of the smallest proteins (58 aminoacids) characterized by a well defined structure. Then, the same procedure is applied to frataxin, in a truncated form of 121 aminoacids, and 5 of its variants Petrosino et al. (2019).

2 Materials and methods

The computational methods used in this work are similar to those used in Ref. (Botticelli et al., 2022). In the following we emphasize the differences that characterize this work.

2.1 Metadynamics

Let $\xi(q)$ be a collective variable (CV) function of atomic positions, q. When ξ is an observable quantity, the values, s, allowed for ξ can be used to label system macrostates. The set of coordinates q labels the system microstates, each set of q yielding one of the possible values of s. If ergodicity holds, infinitely long simulations of a trajectory $q(t)$ in a given statistical ensemble would correctly sample the statistical weight of ξ. However, because of the huge number of ways in which certain values of s of ξ are encountered, compared to others, actual numerical simulations in practice only sample the maximally degenerate values of ξ. This is precisely the case where ξ is the CV associated to folding/unfolding events.

Standard statistical ensembles and more recently generalized ensembles try to address this problem by biasing the trajectory to spend more time where ξ has a low degeneracy and less time where ξ has a large degeneracy.

The sampling of configurations obtained with the biased inverse probability of ξ is called metastatistics. We will denote by $\tilde{P}(q)$ the probability of microstates encountered along the simulated trajectory and by $\tilde{P}(\xi)$ the probability of the macrostates labeled by ξ. For simplicity with a little abuse of notation we use the same name for the metastatistics probability as function of the microscopic variables, q, and to the associated metastatistics probability as function of the macroscopic collective variable, ξ.

Many methods have been proposed to sample configurations with the inverse of the estimated probability of ξ Mitsutake et al. (2001). In this work and in the previous application of the method Botticelli et al. (2022), we used the altruistic metadynamics proposed in Refs. Hošek et al. (2016); Hošek et al. (2017). The desired metastatistics is obtained from a swarm of trajectories provided by metadynamics after building a suitable external bias, which is then kept fixed when collecting configurations in the final step of the NpT simulation (see Section 2.6). We performed simulations in the statistical ensemble associated to constant temperature T and pressure p (NpT ensemble) because macromolecules forced by an external bias undergo large and fast conformational changes. When these conditions occur, solute macromolecules exert strong perturbations over the explicit solvent and ions representing their environment. To cope with steep changes of kinetic energy of water molecules and possible temporary voids around the macromolecule, the NpT ensemble is recommended.

In the framework of metadynamics, the estimated probability of the CV is expressed by means of a sum of Gaussian functions, V_G $[\xi(q)]$, related to the inverse metastatistics probability by the formula

$$\ln \tilde{P}(\xi) = \beta V_G[\xi(q)] + C, \tag{2}$$

with $\beta = 1/(k_B T)$ where T is the temperature used in the simulation, k_B the Boltzmann constant, and C a normalization constant, that is, of no relevance in the computation of thermal averages. Different methods have been proposed to build an external bias V_G $[\xi(q)]$ such that the probability distribution of ξ is flat and transitions between folded and unfolded states of a biomolecule endowed with many degrees of freedom, are equally well sampled.

In metadynamics the external potential V_G (ξ, t) acting on the system at time t is defined as:

$$V_G(\xi(q), t) = w \sum_{t'=\tau_G, 2\tau_G, \ldots} exp\left(\frac{-(\xi(q) - s(t'))^2}{2\delta^2}\right) \tag{3}$$

where $t' < t$, $s(t) = \xi(q(t))$ is the value taken by the CV at time t, w is the Gaussian height, δ is the Gaussian width τ_G is the time interval after which a new Gaussians is added.

After a sufficiently long time $V_G(s, t)$ provides an estimate of the underlying free energy F according to the formula

$$V_G(s, t) = -F(s) + C(t) \tag{4}$$

where $C(t)$ depends on time but not on the collective variables s, V_G is the external biasing potential acting on the system at time t.

Equation above states that an equilibrium quantity, like free energy, can be estimated by a non-equilibrium dynamics in which the bias potential is changed in time, as new Gaussians are successively added. In metadynamics, when all the wells in CV distribution are filled with Gaussians, the dynamics in the CV space becomes diffusive.

2.2 Well tempered metadynamics

Well tempered metadynamics is an improved approach designed to obtain a reliable estimator of the free energy Barducci et al. (2008). The weight of each Gaussian function added to the bias V_G depends on the history of V_G ($V_G(t')$). Equation 3 changes into:

$$\begin{aligned} V_G(\xi(q), t) = w \sum_{t'=\tau_G, 2\tau_G, \ldots} exp\left(\frac{-V_G(\xi(q), t')}{k_B \Delta T}\right) \\ \times\, exp\left(\frac{-(\xi(q) - s(t'))^2}{2\delta^2}\right), \end{aligned} \tag{5}$$

where $k_B \Delta T$ is approximately the energy change when a new value of ξ is visited. An exact relation between $V_G(s, t)$ and $F(s)$ can be obtained if the rate at which the bias potential is modified is suitably decreased as the simulation progresses. With well tempered metadynamics, the biasing potential converges to

$$V_G(s, t) = -\frac{\Delta T}{T + \Delta T} F(s) + C(t). \tag{6}$$

The quantity $T + \frac{\Delta T}{T}$ is called "biasing factor".

For a finite T, the probability distribution is proportional to:

$$\exp\left(\frac{-F(s)}{k_B T}\right) \exp\left(\frac{\Delta T}{T + \Delta T} \frac{F(s)}{k_B T}\right) = \exp\left(-\frac{F(s)}{k_B (T + \Delta T)}\right) \tag{7}$$

which corresponds to effectively increasing the CV sampling temperature. Thus, the effect of well tempered metadynamics is similar to that of other non-equilibrium methods, like steered molecular dynamics, but trajectories are obtained with a quasi-equilibrium procedure Bussi et al. (2018).

In well tempered metadynamics, as the simulation proceeds the width of the added Gaussian remains constant but its height decreases (see Eq. 5). The bias, which increases monotonically, eventually changes very slowly with time. At the beginning the space of CV is flooded by Gaussians of height w. With the progress of flooding, heights of newly added Gaussians decrease. This behaviour is very important in highly complex biological systems, where the bias potential should never reach any excessively large value.

In contrast with the "non tempered" metadynamics, in the well tempered metadynamics a flat CV distibution is not expected to be achieved when convergence is obtained. A simple interpretation of the fact that the distribution of the CV at long times is not flat is the following. Since the prefactor for the accumulated Gaussians depends on the value of s, Gaussians of different heights are placed in different regions of the CV space. In order to reach a stationary distribution, it is thus necessary that the system spends more time in regions where small Gaussians are added and less time in regions where large Gaussians are added. This idea can be pushed further and used to convert metadynamics in an algorithm, not designed to flatten completely (as in non-tempered metadynamics) or partially (as in well tempered metadynamics) the histogram of the CVs but rather to enforce a predefined distribution Bussi et al. (2018).

In this work we used a biasing factor of 20 (see Eq. 6), corresponding to $\Delta T = 5700\,\mathrm{K}$, in agreement with the biasing factor used in literature for similar molecular systems Hošek et al. (2017). The energy value $R\Delta T$ is of the order of a typical energy barrier of a single hydrogen bond.

2.3 The maximal constrained entropy method

The maximal constrained entropy method (MEC method, hereafter) allows, starting from $\tilde{P}(\xi)$ of Eq. 2, to obtain a better probability for thermal average calculations. This elaboration is used to correct for limitations of $\tilde{P}(\xi)$, whatever the method used for its determination is. We remind that the method consists in post-processing the biased statistics (that we indicate with metastatistics) collected by whatever method. Since in actual simulations one works with trajectories where configurations can be enumerated, we attach the microstate index γ to the configuration $\{q\}$ and we denote by $\tilde{P}_\gamma$ the probability

$$\tilde{P}_\gamma = \frac{\tilde{w}_\gamma}{\sum_\gamma \tilde{w}_\gamma}, \tag{8}$$

where $\tilde{w}_\gamma$ is the number of microstates with label γ collected in the metastatistics and $\tilde{Z} = \sum_\gamma \tilde{w}_\gamma$ is a normalization factor.

In an infinitely long (ergodic) simulation, it is unnecessary to explicitly evaluate the weights $\tilde{w}_\gamma$, as they are automatically encoded in the degeneracy of the set of collected configurations sampled along the simulated trajectory. This means that in the following equations, where the sum over γ is extended over that actually produced configurations, we should not include the factor $\tilde{P}_\gamma$. However, we leave this redundant factor to recall that we are dealing with a finite set of configurations generated by metadynamics.

In case of the "non tempered" metadynamics, the maximal constrained entropy was employed as a viable solution to compute thermal averages as a function of the average values taken by the chosen CV, in situations where metastatistics is not fully ergodic and the CV distribution does not come out flat. As mentioned, in the case of well tempered metadynamics the CV distribution is not expected to be flat, but the maximal constrained entropy method is a powerful method to "correct" the free energy by adding *ex post* further information about the system injecting extra information. In our case we use the maximal constrained entropy to introduce in the computation of the free energy the change of number of hydrogen bonds in α-helices in folding↔unfolding processes. In general the maximal constrained entropy method can be used either to improve the estimate of the free energy for a non-converging system (e.g., in a metadynamics simulation the CV distribution is not flat) or to compute the free energy by reintroducing *ex post* degrees of freedom related to the CV (like the α-helices' hydrogen bonds in the case of frataxin, see Section 3). This second use of the maximal constrained entropy method is really powerful because allows to have a reliable estimate of the free energy while keeping efficient the simulations by limiting the degrees of freedom of the system.

2.4 Estimating the free energy

The main goal of this work is to compare the change of free energy as a function of the number of hydrogen bonds (s) computed using well tempered metadynamics and maximal constrained entropy, with the results obtained in protein thermal denaturation experiments Yu et al. (1995); Petrosino et al. (2019). Both BPTI and FXN are folded in a structure where one or two α-helices lay over a small β-sheet. The experimental measurement of the free energy difference between folded and unfolded states was obtained by measuring the molar ellipticity at 222 nm, a wavelength where the contribution of α-helix to CD spectra is dominant. Besides acting on the α-helices arrangement, the protein ternary structure can also be perturbed by destroying the intra-molecular hydrogen bonds that stabilize the β-sheet. For a small protein like BPTI (58 residues), we decided to include in the CV all the hydrogen bonds that are formed in the native folded state Parkin et al. (1996). For FXN (121 residues) we took instead as a CV the number of hydrogen bonds occurring in the β-sheet formed by 4 anti-parallel β-strands, which are observed both in 1EKG and 5KZ5 structures Botticelli et al. (2022). This choice in the case of FXN was made to reduce the number of degrees of freedom of the system thus substantially decreasing the time required to sample its phase space. The use of such CV as a way to monitor the structural transitions in the protein was inspired by several previous applications of metadynamics Barducci et al. (2006).

For both proteins and variants, the biasing potential, V_G, was obtained at the end of a systematic construction (well tempered metadynamics) in which V_G is progressively built by summing over Gaussian functions of the CV. Gaussian functions (possibly scaled by the biasing factor in the case of the well tempered metadynamics) are deposited every 20 ps along the molecular dynamics (MD) simulation time.

The accumulated final biasing potential, $V_G(\xi)$, smoothly interpolated by a polynomial of fourth order, was used for the direct computation of the change in the free energy for folded to unfolded states and *vice versa*. The free energy change defined in well tempered metadynamics is given by Eq. 6:

$$F(s) - F(s_0) = -\left(\frac{T + \Delta T}{\Delta T}\left[V_G(s) - V_G(s_0)\right]\right) \tag{9}$$

with s_0 a reference state corresponding to a given value of the CV and V_G the external biasing potential determined at the end of construction. Equation 9 holds also in the NpT statistical ensemble, when the construction of the bias V_G is performed in such statistical ensemble. In this case, the Helmoltz free energy $F(s)$ is replaced by the Gibbs free energy $G(s)$. We call the latter function G free energy, hereafter, for simplicity. The G free energy extracted from well tempered metadynamics simulations (Eq. 9), was then compared with the G free energy obtained with the maximal constrained entropy method.

The accumulated statistics used in the successive maximal constrained entropy application have been obtained by collecting the system configurations along a trajectory where the biasing potential was kept fixed (i.e., not anymore updated). Within the maximal constrained entropy method, the definition of the G free energy (see La Penna et al., 2004) is given by the formula

$$G(s) = \langle H \rangle_\lambda - T\ k_B\ \bar{S}_c(s), \tag{10}$$

in which $G(s)$ is written as the combined sum of the enthalpy in the NpT ensemble, and the (informational) entropy measured by the maximal cross-entropy. In Eq. 10 $H = U + pV$ is the enthalpy of the simulated system, λ the parameter associated with the constraint, $\bar{S}_c$ the maximal cross-entropy change due to the introduction of such a constraint, k_B the Boltzmann constant, and T some effective temperature in the stability range of the system under study. The same free energy definition holds for the Helmoltz free energy F when the enthalpy H is replaced by the energy U if one is working in a NVT ensemble.

The maximal cross-entropy in Eq. 10 is described in the following. Given an estimate, $\tilde{P}_\gamma$, of the metastatistic probability, say the one provided by metadynamics, the problem of finding the least-biased expression of the probabilities P_γ, that is, nearer to $\tilde{P}_\gamma$ and satisfies the condition

$$s = \langle \xi \rangle = \sum_\gamma P_\gamma \xi_\gamma, \tag{11}$$

is solved by determining the maximum of the cross-entropy functional Attard (2000); La Penna (2003); La Penna et al. (2004).

$$S_c\left[P, \tilde{P}\right] = -\sum_\gamma P_\gamma \ln \frac{P_\gamma}{\tilde{P}_\gamma}. \tag{12}$$

under the constraint (Eq. 11). The well-known solution of this variational problem is given by the formulae:

$$P_\gamma = \frac{\tilde{P}_\gamma}{Z_\lambda} \exp\left(-\lambda \xi_\gamma\right) \tag{13}$$

$$Z_\lambda = \sum_\gamma \tilde{P}_\gamma \exp\left(-\lambda \xi_\gamma\right) \tag{14}$$

with the parameter λ the solution of the (highly non-linear) equation:

$$s = \sum_\gamma P_\gamma \xi_\gamma = \frac{1}{Z_\lambda} \sum_\gamma \tilde{P}_\gamma \exp\left(-\lambda \xi_\gamma\right) \xi_\gamma. \tag{15}$$

The quantity exp $(-\lambda\xi_\gamma)/Z_\lambda$ is called the modulation factor of the metastatistics. Owing to Eq. 15, λ is a function of s. Inserting the solution for P_γ into Eq. 12 one gets for the cross entropy at its maximum:

$$\bar{S}_c(s) = \ln Z_\lambda + \lambda\ s. \tag{16}$$

The average of H (or simply of U in NVT simulations) is obtained using equations like

$$b_\lambda = \langle B \rangle_\lambda = \frac{1}{Z_\lambda}\ \sum_\gamma \tilde{P}_\gamma \exp\left(-\lambda \xi_\gamma\right)\ B\left(q_\gamma\right), \tag{17}$$

with B either H or U and $Z_\lambda = \sum_\gamma \tilde{P}_\gamma \exp(-\lambda\xi_\gamma)$. The identification of S_c and T in Eq. 10 with, respectively, thermodynamic state function entropy S and state variable absolute temperature T, is empirical. It must be noticed that changes in thermodynamic $T\,S$ values are also reflected in the changes of $\langle H \rangle_\lambda$ as a function of λ.

The details to compute the free energy within the maximal constrained entropy method, the direct calculation of $\langle H \rangle_\lambda$ in Eq. 10 as well as the free energy error estimate is the same we used in our previous work where the "non tempered" version of the metadynamics Botticelli et al. (2022) was employed. In this work we concentrate on collecting more accurate averge quantities (well-tempered metadynamics and longer simulations) and on applying the proposed method also to a simpler protein (BPTI). We must note that the direct calculation of $\langle H \rangle_\lambda$ in Eq. 10 includes the effects of the fluctuations of U and V due to the movement of all explicit water molecules and ions included in the atomistic model of the protein environment. The fluctuations of H are huge, while the change of the average of H with s is small. This is a serious issue when using the total enthalpy like in Eq. 10. As it is customary done in these cases, we use an approximate evaluation of $\langle H \rangle_\lambda$, where H is replaced by the effective mean-field free energy $\bar{U}$ of the protein solute. The advantage of this approximation is that the energy of the system is thermally averaged over the many degrees of freedom of water molecules and ions surrounding the much smaller solute protein aggregate.

A widely used strategy for the evaluation of the effective mean-field energy of the solute protein is the so-called molecular mechanics/Poisson-Boltzmann solvent accessible approximation (MM/PBSA) Simonson et al. (2002). In this approach the mean-field energy for solute-solvent interactions is described as the sum of polar (electrostatic) and non-polar (surface) contributions. For each protein configuration Q one writes

$$\bar{U}(Q) = U_{intra}(Q) + G_{solv,np}(Q) + G_{solv,pol}(Q), \tag{18}$$

where U_{intra} is the intra-molecular part of the potential energy in the protein force-field, given by

$$U_{intra}(Q) = U_{str}(Q) + U_{bend}(Q) + U_{tors}(Q) + U_{vdw}(Q) + U_{el}(Q). \tag{19}$$

The various contributions are the stretching (U_{str}), bending (U_{bend}), and torsional (U_{tors}) terms in the potential. U_{vdw} and U_{el} are the Lennard–Jones and Coulomb interactions, respectively, computed by summing over all the pairs of atoms of the protein.

The last two terms in Eq. 18 represent solute-solvent contributions to free energy at fixed Q. Mean field energy is the energy as a function of Q once the variables associated to solvent positions and velocities are averaged for the given value of solute positions Q. The averaging is performed at the given thermodynamic state variables p and T used in the simulation of the whole system. Within this mean-field assumption, the solute and the solvent are made independent. This is a strong approximation, since the chosen collective variable contains the number of hydrogen bonds within protein groups and once a single intramolecular hydrogen bond is broken there is a large chance for the formation of hydrogen bonds with the water molecules in the protein environment where the breaking event occurs. On the other hand, this elementary change of free energy, that does not imply a wide change in protein structure, can be calculated within the MM/PBSA approximation as the sum of $G_{solv,np}$ and $G_{solv,pol}$. Therefore, under this approximation, the change of free energy $G(s)$ depends on the number of protein configurations for which a unitary change of s is allowed independently of the configuration of the protein environment. The calculation of $G_{solv,np}$ and $G_{solv,pol}$ is described in the following.

The term $G_{solv,np}$ is the contribution to the solute-solvent free energy due to the formation of a cavity of zero charge density with the shape of the solute protein and the creation of the solute-solvent interface. Introducing a charge density in the space occupied by the solute leads to the $G_{solv,pol}$ contribution. The charge density is given in terms of the point charges q_i of the atom i sitting at the point $\vec{r}_i$, where i runs over the N_a atoms of the solute molecule.

The term $G_{solv,np}$ is calculated as an empirical linear combination of the solvent accessible surface area (SASA) for each group in the solute molecule Ooi et al. (1987) according to the formula

$$G_{solv,np} = \sum_{i}^{N_a} \sigma_i SASA_i, \tag{20}$$

where the coefficients σ_i are positive or negative for hydrophobic or hydrophilic groups, respectively (see below for details). Finally the electrostatic contribution to the solute-solvent free energy, $G_{solv,pol}$, is given by the electrostatic energy required to charge the low-dielectric solute molecule of generic shape into a high-dielectric medium like a salt-water solution. The magnitude of this contribution is obtained by a numerical finite difference solution of the Poisson–Boltzmann equation Rocchia et al. (2002).

2.5 Summary of the method

We summarize the complicated computational protocol of our theoretical analysis as follows. One starts by performing MD simulations at $T = 300$ K in the presence of the biasing potential $V_G(\xi)$ built according to the well tempered metadynamics strategy. The resulting statistics is what we call metastatistics. Using the set of collected configurations, we determine the λ parameter that maximizes the cross-entropy S_c in the maximal constrained entropy method, under the constraint $\langle \xi \rangle = s$. In the case of BPTI, ξ is the number of hydrogen bonds holding together the protein α-helices and β-sheet secondary motifs. In this case, the ξ of metadynamics and that of maximal constrained entropy coincide. In the case of BPTI, differently from FXN (see below), the ξ collective variable takes into account the whole of the secondary structure as it is observed in the crystal folded structure. Therefore, s takes integer values in the range between 0 and 16. In the case of FXN, the variable ξ used in metadynamics is the number of hydrogen bonds holding together the protein β-sheet (made of 4 anti-parallel β strands). The values of s are in the range between 0 and 15. But in the maximal constrained entropy approach we extended ξ adding to it the number of hydrogen bonds in the two α-helices. For each value of s, we get a value of λ that yields the modulating weight

$$w[q(t)] = \frac{1}{Z_\lambda} \exp\{-\lambda \xi[q(t)]\}, \tag{21}$$

with q the system configuration at time t, indexing the microstate γ, along the collected metadynamics trajectory. For details see Ref. (Botticelli et al., 2022).

2.6 Simulation parameters

Apart from the fact that differently than what was done in Ref. (Botticelli et al., 2022), in this work the metastatistics is obtained as altruistic multiple-walkers well-tempered metadynamics, most of the technical details of the simulation procedure we followed to compute the expectation values of the physical quantities of interest described in Section 2 are identical to those reported for FXN in Ref. (Botticelli et al., 2022). Below we only outline the few differences.

TABLE 1 Pairs of atoms used in Eqs 16–18 of Ref. (Botticelli et al., 2022) and related label in parameter *S*. As for FXN, see Table 1 of the same publication. Residues are those of BPTI WT sequence. Mutated residues are boldface.

β_{1-2}		α_1		α_2	
N (Tyr 35)	O (Ile 18)	N (Cys 5)	O (Pro 2)	N (Met 52)	O (Ala 48)
N (Ile 18)	O (Tyr 35)	N (Leu 6)	O (**Asp 3**)	N (Arg 53)	O (Glu 49)
N (Phe 33)	O (Arg 20)	N (Glu 7)	O (**Phe 4**)	N (Thr 54)	O (Asp 50)
N (Arg 20)	O (Phe 33)			N (Thr 54)	O (Ala 51)
N (Gln 31)	O (Phe 22)			N (Cys 55)	O (Ala 51)
N (Phe 22)	O (Gln 31)				
N (Leu 29)	O (Asn 24)				
N (Asn 24)	O (Leu 29)				

TABLE 2 Short description of the atomistic models used in metadynamics simulations. The composition of each system changes only in the protein sequence for each protein (BPTI and FXN, respectively). The number of water molecules and counterions (NaCl) is the same for all the 90 walkers representing each system, and the same (= symbol) for different variants of the same protein.

System	Protein atoms	Water molecules	Na	Cl
BPTI				
BPTI $[5\text{-}55]_{\text{BPTI}}$	892	11033	21	27
D3A	890	=	=	=
F4A	890	=	=	=
FXN				
WT	1875	13926	34	26
D104G	1870	=	=	=
A107V	1881	=	=	=
F109L	1874	=	=	=
Y123S	1865	=	=	=
S161I	1883	=	=	=
W173C	1862	=	=	=
S181F	1884	=	=	=
S202F	1884	=	=	=

Table 1 provides the list of hydrogen bonds used to define the CV for the BPTI. All the hydrogen bonds contribute to the BPTI CV and are used both in well tempered metadynamics and maximal constrained entropy. For FXN only the number of hydrogen bonds in the β-sheet, β_{1-4}, is used to generate the statistics of metadynamics. However, the total number of hydrogen bonds listed in Table 1 of Ref. (Botticelli et al., 2022), including the two

TABLE 3 Short description of the simulation stages used to build the external bias $V_G(\xi)$ and to acquire the metastatistics at constant external bias. Where α and w are not indicated, the external bias is not updated. The initial bias is zero. Therefore, stages 1–3 (6 ns) are equilibration stages. The bias construction is the same for all variants of BPTI and FXN. As for the constant bias simulation stage 16–20 (10 ns) were collected for BPTI, while 16-30 (30 ns) were collected for FXN. Values of α and w are used when applying the altruistic combination of single-walker updating (2 ns) of V_G using Eq. 3 of Ref. (Hošek et al., 2017). The resulting global altruistic bias is used in the following 2-ns stage (next line). The bias after stage 15 is approximately the same for all walkers and, therefore, is made identical for all walkers by averaging over the 90 walkers.

Stage	Time length	α	w
1–4	8	-	-
5	2	0	1
6	2	1/4	1
7	2	1/2	1
8	2	3/4	1
9–15	14	1	1/2
16-end	10–30	-	-

α-helices added to the definition of CV in the successive maximal constrained entropy step. We call this an extension of the CV ξ used in metadynamics and we indicate it with ξ'. The corresponding constrained average is indicated with s'.

Table 2 reports the number of atoms of the two systems (BPTI and FXN) we have studied. In the case of BPTI, the structure of the unique folded structure available [1BPI PDB entry Parkin et al. (1996)] has been used. As for FXN, the initial configurations of the various walkers are obtained using the available crystallographic information about the native FXN protein sequence. We used two structures: the X-ray structure of the mature human frataxin [PDB 1EKG, segment 88-210 Dhe-Paganon et al. (2000)]; the structure of FXN in the mitochondrial iron-sulfur cluster assembly machine as it was determined by electron microscopy (PDB 5KZ5, chain A, segment 42-210 Gakh et al. (2016)).

The values of α and w of Eq. 3 in Ref. (Hošek et al., 2017) and used in the successive stages of the simulation are reported in Table 3. As for the construction of the biasing potential, we remark that its construction in the present work lasted 22 ns, while in our previous application it lasted 16 ns. The exchange of the bias among walkers takes place every 2 ns. At the end of stage 15 (see Table 3), i.e., after simulating each walker for a total of 30 ns using an altruistically updated bias, the external bias that will be used in stage 16 and in the following steps is not updated any more. From stage 16 to the end the final metastatistics is collected, storing configurations along the simulated trajectory every ps. The time duration of this last simulation step was 10 and 30 ns for BPTI and FXN, respectively.

3 Results

3.1 Bovine pancreatic trypsin inhibitor (BPTI)

48 single point mutations have been studied in the case of BPTI in the literature Yu et al. (1995) via alanine-scanning. This set of mutations includes all residues, with the exception of 6 Ala and 4 Cys, mutated to Ala. The reference sequence used to study the change in melting temperature is the native sequence where Cys 14, 30, 38, and 51 are mutated in Ala. This reference variant is indicated as $[5\text{-}55]_{\mathrm{BPTI}}$, to underline the presence of the residual 5–55 disulfide bridge. The sequence is used because the native sequence has 3 disulfide bridges in the folded structure and it does not unfold at $T < 100$°C. The removal of 2 disulfide bridges allows the melting at $T < 50$°C, while the protein keeps the same folded structure as the native (WT) sequence, as summarized in Ref. (Yu et al., 1995). Therefore, we could use the structure determined for the WT sequence as the initial representation of the folded state (1BPI Parkin et al. (1996) in PDB).

According to our conventions, a positive $\Delta\Delta G$ means a larger reversible work required to unfold the given variant with respect to the reference sequence. All the variants analyzed in experiments have been already studied as part of large datasets in previous works dedicated to predictions of free energy change Guerois et al. (2002); Steinbrecher et al. (2017). In our work we are interested in predicting the sign of the free energy change, which is also the sign of the change of the melting temperature T_m, ΔT_m. As paradigmatic cases we focused, among the 48 variants, on the two displaying the largest measurable change in the absolute value of ΔT_m. The mutations with the most positive and negative value of ΔT_m [see Table 1 in Ref. (Yu et al., 1995)] are D3A and F4A, respectively.

Three representative structures of the $[5\text{-}55]_{\mathrm{BPTI}}$ reference sequence of BPTI. are displayed in Figure 1 to show how the folded (left panel) and unfolded (right panel) states look like in terms of atomic configurations. Native BPTI is folded into a ternary structure with two short α-helices and a small β-sheet. The construction of the external bias, $V_G(\xi)$, perturbs the ternary structure by breaking the intramolecular hydrogen bonds.

This is why we decided to take as a collective variable ξ the sum of the number of hydrogen bonds between the two α-helices (α) and the β-sheet (β). The number of hydrogen bonds of α-helices and β-sheets in the initially folded structure (PDB 1BPI) is 8 for both secondary structures. Therefore, the values of ξ span the range between 0 and 16. Figure 1 shows in the right panel that the unfolded state is represented by a molten globule. This occurs because of the short-range nature of the collective variable we have chosen. In the specific case of BPTI the presence of the residual disulfide bridge 5-55 that seals the N-terminus with the C-terminus also pushes the protein towards this atomic arrangement.

The evolution in time of the collective variable ξ is notoriously slow, even by using well tempered metadynamics. Therefore the convergence of the external bias $V_G(\xi)$ is expected to occur after very long simulation times. This issue is illustrated in Figure 2, where the time evolution of ξ of 4 walkers among 90 is displayed. We remind that every 2 ns the bias V_G obtained by the whole set of 90 walkers is exchanged among all of the walkers during bias construction in the altruistic approach [Eq. 3 in Hošek et al. (2017)]. Furthermore, before the bias construction the 90 walkers have been separately equilibrated for 8 ns. The figure is divided in two parts. The time evolution during the 22 ns of bias construction is displayed on the lefthand side. The time evolution at constant bias, which constitutes the metastatistics used to compute the biased equilibrium averages,

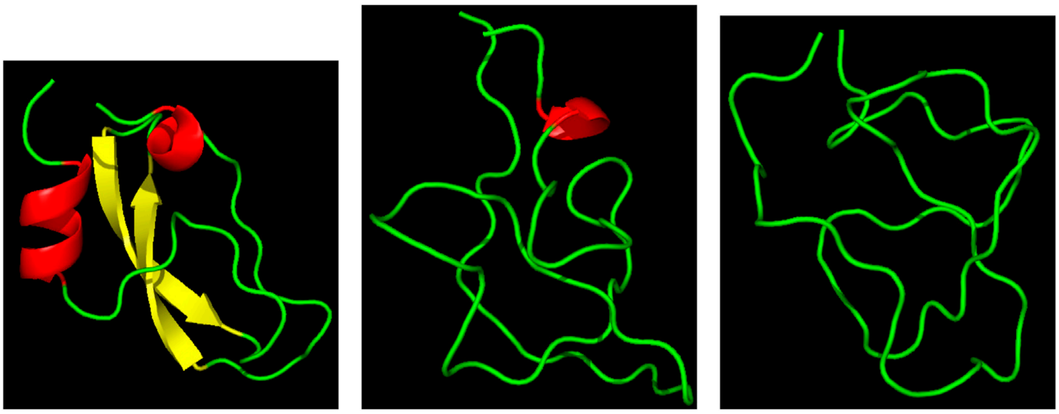


FIGURE 1
Three representative structures of $[5\text{-}55]_{\mathrm{BPTI}}$ reference sequence of BPTI. Left—ξ = 13 (folded state); middle—ξ = 4 (unfolded state); right—ξ = 4 (unfolded state). α-helices are in red; β-sheet is in yellow; the displayed ribbon interpolates the backbone atoms. The Pymol program is used for the molecular drawing Schrödinger (2015).

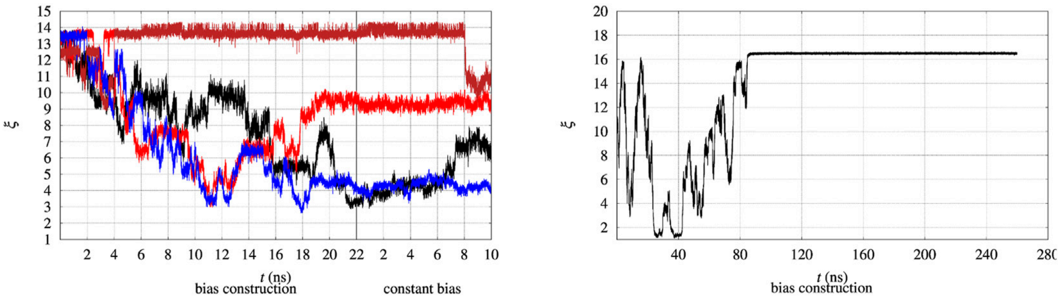


FIGURE 2
Time evolution of the collective variable ξ during the bias construction (left part) and at constant bias (right part), with the vertical line dividing the bias construction from the bias application. Left—The evolution is displayed in different colors for 4 representative walkers among 90 and for BPTI in the $[5\text{-}55]_{\mathrm{BPTI}}$ sequence. Right—The same evolution is displayed for a single walker in well-tempered metadynamics of FXN in the WT sequence.

during the last 10 ns is displayed in the righthand side. The figure clearly shows that the unfolding of the protein often occurs during bias construction, since ξ decreases from the value characterizing the folded state to values of 3–4 in 3 cases over the 4 displayed. In certain cases (red curve) the expected behaviour of a random walk of ξ in the 3–14 range is observed.

In principle metadynamics is capable of letting the system, starting from the known folded configuration, to unfold and refold. In practice this rarely occurs in affordable computational time, unless the system is sufficiently small. To illustrate the time-scale required for collecting such trajectory, the behaviour of ξ for the longer FXN chain is displayed in Figure 2 (right panel) for a single walker. In this trajectory a single-walker well-tempered metadynamics is performed for 260 ns. While the first 100 ns of the trajectory displays an ideal behaviour for metadynamics [see for instance Figure 2 in Ref. (Barducci et al., 2006)], when the bias is no more effectively updated by new Gaussian functions the system becomes frozen in a fixed configuration. This effect is expected in well-tempered metadynamics, since the height of the Gaussian functions that are added to the bias is progressively decreased by construction.

Anyway, the dynamics of ξ shows that in order to observe a proper random walk of ξ for all walkers, simulation time should have been at least 10–100 times larger. The dynamics of ξ becomes even slower when the bias is kept constant compared to bias construction (righthand side of both panels in Figure 2). This behaviour is due to the effect of noise during bias construction, occurring when new Gaussian contributions are added to V_G every 10 ps. The dynamical nature of V_G during its construction acts as a stochastic perturbation. This effect is not present when V_G is kept constant and when V_G does not change because added Gaussian heights are small.

Because of the slow dynamics of ξ, the metastatistics represents a static disorder triggered by the bias construction process.

In Figure 3 we display the behaviour of $V_G(\xi)$ during the bias construction for the BPTI reference sequence. The red arrow indicates the direction in which the iteration index of the altruistic method, occurring every 2 ns of metadynamics simulation, increases. In the process of iteration the number of Gaussian contributions to V_G keeps increasing in the region of low ξ values, while at high ξ it does not change anymore after the first 4 iterations. A full convergence of V_G is not achieved, but we notice

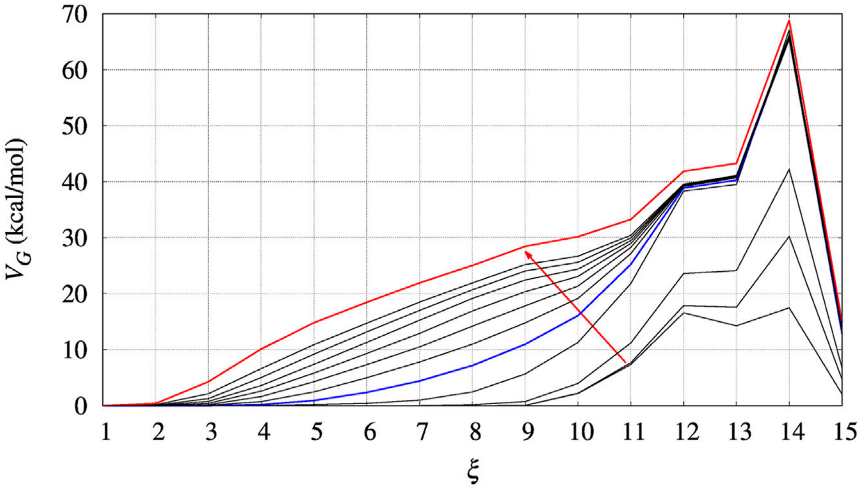


FIGURE 3
The evolution of the bias (V_G) acting on walker 1 during bias construction for BPTI, $[5\text{-}55]_{\text{BPTI}}$ sequence. Different curves are obtained combining the bias of all walkers using Eq. 3 of Ref. (Hošek et al., 2017). The red arrow indicates the increasing iteration. The red curve is that used in the final collection, while the blue curve was used to estimate the effect of a non-converged bias on the values of free energy obtained by the post-processing MEC method.

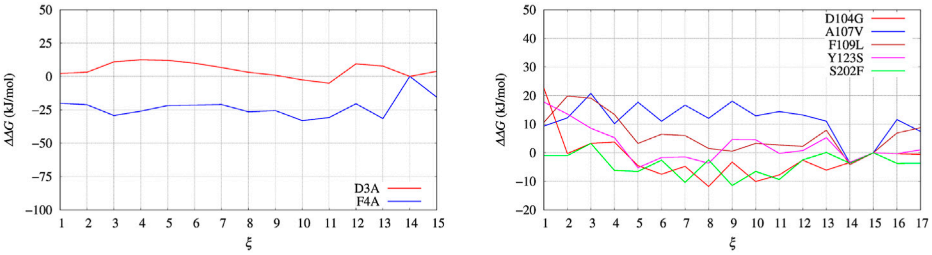


FIGURE 4
Free energy change ($\Delta\Delta G$) calculated with Eq. 9 and V_G built with well tempered metadynamics. Left panel: BPTI; Right panel: FXN.

that the change of V_G is very slow after about 10 iterations. This happens because when the protein is unfolded, many atomic configurations consistent with a low number of hydrogen bonds are possible. Then, Gaussian contributions to V_G are all added in the region of low ξ values, while no further contributions are added to the region of high ξ values.

Since the bias converges very slowly, it is worth checking the effect of choosing different bias in the calculation of interest for us, that is, $\Delta\Delta G$ as a function of the chosen collective variable for a protein variant with respet to the wild-type sequence. In Figure 3 we choose the red curve, as what we assumed as converged bias, and the blue curve, the function built after 5 iterations in the altruistic scheme (stage 9 in Table 3). The difference between results obtained with these two different choices of final bias will be described later for BPTI. We remark that the configurations used in the comparison are the results of two different 10-ns trajectories for all of the 90 walkers: one performed with the "converged" bias (stage 15 in Table 3) while the second performed with the bias of stage 9.

Times of the order of 100 ns are required to build a useful bias for each walker even for a protein of 58 residues like BPTI. This issue is illustrated by the behaviour of a single walker of the larger FXN protein (see Figure 2, right panel, discussed above). In practice such long simulation times can not be used to compare a native sequence with the usually rather large number of its variants. The method described in this article allows extracting differences in stability under point variations with computational wall-times of the order of 1 month in a high-performance computing infrastructure.

In Figure 4, left panel, we display the free energy change $\Delta\Delta G$ computed using Eq. 9, implicitly assuming that V_G has properly converged after 22 ns of multiple-walkers bias construction. In Figure 5, we also display the free energy change using the polynomial of order 4 interpolating the grid representation of the bias V_G (see Section 2.4). We remind that the polynomial interpolation is performed on each approximately converged $V_G(\xi)$ profile obtained by metadynamics. Therefore, the effect of interpolation on the free energy change $\Delta\Delta G$ as a function of sequence change can be slightly different when the difference between interpolated curves is extracted. In Figure 6 we display the comparison between the grid representation of $\Delta G = -V_G + C$ and its interpolation in all of the three sequences investigated for

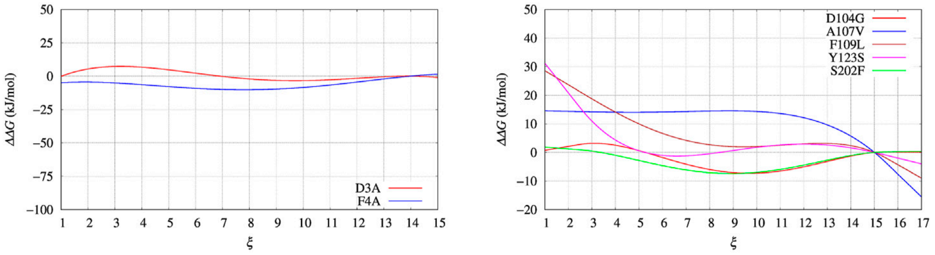


FIGURE 5
Same as for Figure 4 using the polynomial of order 4 interpolating the grid representation of the bias V_G used in Figure 4. Left panel: BPTI; Right panel: FXN.

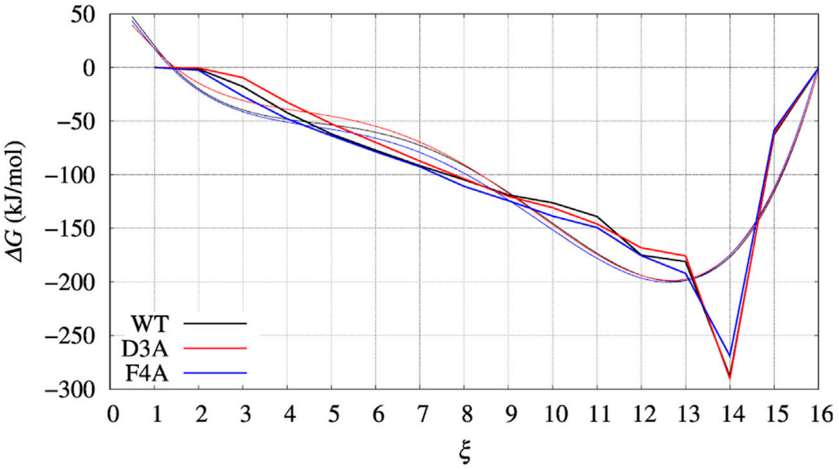


FIGURE 6
Same as for Figures 4, 5 (left panels), but comparing $\Delta G(\xi)$ for the case of BPTI variants. Color scheme is black, red, and blue for WT, D3A, and F4A variants, respectively. Thick line is the grid representation of V_G; thin line is the 4-th order polynomial interpolation of each grid (same color) in the range between 0.5 and 15.5. Out of this range the interpolation is linear, with continuous derivative at the extreme points.

BPTI. We notice that for the three variants the free energy increases by decreasing the value of the collective variable, consistently with a greater thermal stability of the protein configurations in the folded state. But, because of the smoothing of V_G exerted by the interpolation, part of the changes are lost. However, even the tiny difference between the curves $\Delta G = G(\xi)—G(\xi_0)$ is still a representation of the change in stability upon protein unfolding once the sequence is changed. In the BPTI case, we find that, consistently with experiments (see Table 4), the D3A mutation induces an increase in the stability of the protein (left panel, red curve). The opposite is true for the F4A mutation.

As discussed above, statistics severely limits the convergence of $V_G(\xi)$, since the number of Gaussian contributions to V_G giving an almost flat ξ distribution is achieved when all unfolded configurations are sampled, a condition, that is, hardly achieved even with μs-long MD simulations. Employing multiple walkers allows one to sample unfolded and folded configurations in non-infinitely long simulation but each of the walkers is not able to walk from a folded configuration towards an unfolded one and viceversa with a frequency allowing a proper sampling.

In the left panel of Figure 7 we display the distribution of the collective variable ξ along with the sampling at constant bias $V_G(\xi)$ for all the BPTI variants. The distributions that we obtain are not flat because of the technical limitations of the well tempered metadynamics method (see Section 2.2) and the limited span of sampled CV values as shown in Figure 2 and discussed above. Despite the sampling being likely insufficient to have both a converged bias and a converged distribution of ξ once a constant bias is applied, we can estimate the reversible work necessary to build a given average of ξ, s, from the biased metastatistics at our disposal. This is done using the maximal constrained entropy approach described in Ref. (Botticelli et al., 2022) and references therein.

The free energy difference between each of the two variants of BPTI D3A and F4A, and the reference sequence $[5\text{-}55]_{\mathrm{BPTI}}$ is displayed in the left panel of Figure 8 as a function of the

TABLE 4 Experimental (ΔT_m, °C), experimental $\Delta\Delta G$ Petrosino et al. (2019), and computed values of $\Delta\Delta G$ (kJ/mol) for the selected BPTI and FXN variants. Column 4: the values obtained with metadynamics. Columns 5-7: the maximal constrained entropy method is used with the effective energy for solute-solvent interactions (Eqs 18, 19). Column 5—Data published in previous article Botticelli et al. (2022); column 6—Simulation used in previous article, using the extended ξ' variable in the maximal constrained entropy method; column 7—Well tempered metadynamics, using the extended ξ' variable in the maximal constrained entropy method. Rows are reported in descending order of ΔT_m for each protein. While for BPTI the collective variable $\xi = \alpha + \beta$ is used both in metadynamics and maximal constrained entropy methods, for FXN $\xi = \beta$ is used in metadynamics and the extended variable $\xi' = \xi + \alpha = \beta + \alpha$ is used in the maximal constrained entropy method. β is the number of hydrogen bonds in the β-sheet; α is the number of hydrogen bonds in the α-helices (see Methods for details). BPTI: Unfolded state is $s = 4$; Folded state is $s = 14$ (highest peak in the distribution obtained with the meta-statistics, see Figure 7). FXN: Unfolded state is s' =21 [23 for simulation of Ref. (Botticelli et al., 2022)]; Folded state is s' =37.

Variant	ΔT_m	$\Delta\Delta G$ (exp.)	$\Delta\Delta G$ (calc.)			
		BPTI $[5\text{-}55]_{\text{BPTI}}$				
D3A	1.4	0.84	12.5	-	-	−2.5
F4A	−21.2	−12.55	−26.0	-	-	−56.2
		FXN				
D104G	3.0	0.88	2.5	20.1	16.1	58.4
S202F	−0.3	−0.67	−1.0	−7.3	14.6	2.3
A107V	−3.0	3.35	14.1	−114.2	−89.5	−11.9
F109L	−11.4	−8.74	14.1	−21.3	−32.5	−89.9
Y123S	−14.4	−20.59	4.3	−25.2	29.2	−56.4

average value s of the collective variable ξ (see Section 2). The MEC modulation is employed here and Eq. 1 is used, with X sequences identified by different colors. Since the distribution of ξ in the metastatistics displays a sharp peak in the folded state ($\xi = 14$) and a broad peak in the unfolded one (at about $\xi \sim 4$) we report in Table 4 the free energy change going from the state of average $s = 14$ (that is, the state s_0 in Eq. 1) to the state with average $s = 4$.

The direct metadynamics calculation and the maximal constrained entropy method give consistent results for the sign of $\Delta\Delta G$ in the case of BPTI. In fact, we find a slightly negative value equal to −2.5 kJ/mol for D3A at $s = 4$ (see Table 4; Figure 8, left panel). Moreover, in both cases our results are consistent with experiments. It is important to recall that among the 48 single-point mutations of $[5\text{-}55]_{\text{BPTI}}$, only 3 produces a stabilization of the protein Yu et al. (1995).

Before entering into structural details providing explanation of $\Delta\Delta G$ values, we assess what we expect to be the major source of error propagation in the MEC method. The most efficient handle to expand the sampling of atomic protein configurations is the external bias V_G, as computed by well-tempered altruistic multiple-walkers metadynamics. Therefore, we calculated $\Delta G(s)$ profiles for BPTI, which is protein small enough to easily repeat 900 ns of MD simulations, using two different choices of $V_G(\xi)$, respectively the red and blue curves in Figure 3. The profiles of $\Delta G(s)$ computed with the different choices are displayed in Figure 9, left and right panels, respectively, for red and blue curves of $V_G(\xi)$. We notice that there are a few values that are affected by the limited number of points in the statistics: $s = 2$ (left panel) and $s = 4$ (right panel). By using the bias obtained by a shorter cumulative history (the blue curve), low values of ξ (contributing to low values of average s) are rarely sampled. Apart from these limitations due to the range of sampled ξ values, the similarity in the behaviour of $G(s)$ is remarkable. In particular, we notice that the sign of $\Delta\Delta G$ (the difference between curves in each of the plots) is robust. This depends on the fact that the contribution to the calculation of ΔGs with the help of Eq. 10 depends on the energy of the populated states (with a certain value of the collective variable ξ) rather then to the number of ways the state is reached by the simulation.

Due to the collection of atomic configurations at hand and to the possibility of computing the different terms contributing to $\Delta\Delta G$, we can interpret the unusual stabilization of the D3A variant observed in experiments. The increase in unfolding free energy upon D3A mutation is partially due to the removal of the salt-bridges formed by Asp 3 that occur in the WT sequence. On the other hand, the F4A mutation reduces the steric hindrance of Phe 4 thus enhancing the chance of salt-bridges formation between the N-terminus and other protein regions. The competition between electrostatic long-range contributions and short-range interactions characterizing the hydrophobic patches can be analyzed studying the changes in the terms contributing to U.

In Table 5 the change in four terms contributing to U (see Eqs 18, 19) are reported, together with the whole change of U (last

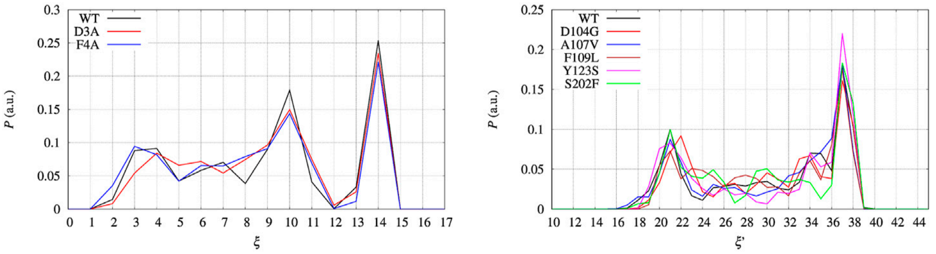


FIGURE 7
Distribution P as a function of the collective variable ξ. Left panel—BPTI, where $\xi = \beta + \alpha$, where β and α are the number of hydrogen bonds in, respectively, the β-sheet and α-helices present in the folded structure. Right panel—FXN, where $\xi = \beta$ was used to build the external bias, but the extended variable $\xi' = \beta + \alpha$ is used to represent the distribution. P is normalized as to have $\sum_i P_i = 1$, where P_i is each of the displayed values.

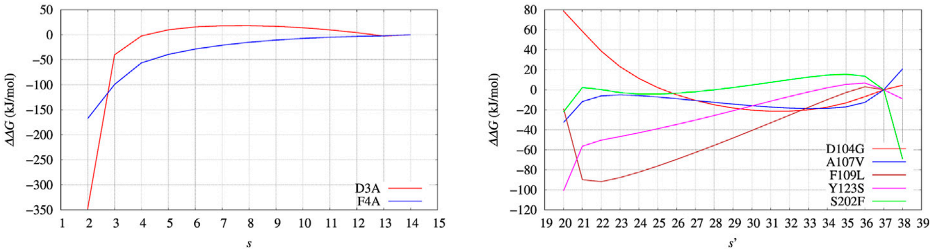


FIGURE 8
Changes of free energy variation ($\Delta\Delta G$) upon unfolding, that is, the decrease of the average number of hydrogen bonds in α helices and β-sheets, α and β, respectively. The average is $s = \langle(\alpha + \beta)\rangle$. As for FXN (right panel) the number of hydrogen bonds is calculated after using metadynamics based on $\xi = \beta$ and $\xi' = \alpha + \beta$ in the maximal constrained entropy method. Color scheme is the same as for Figure 7.

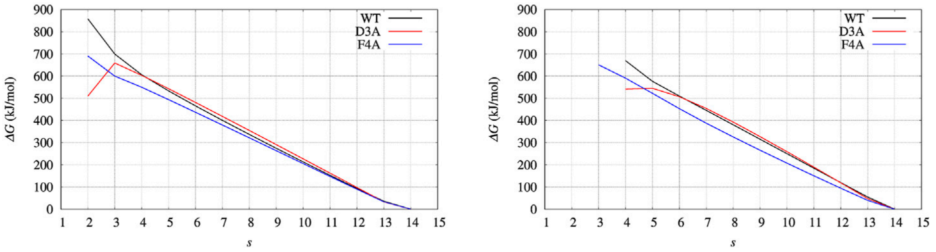


FIGURE 9
BPTI: change of free energy (ΔG) upon unfolding, that is, the decrease of the average number of hydrogen bonds in α helices and β-sheets, α and β, respectively. The average is $s = \langle(\alpha + \beta)\rangle$. Left panel is obtained with Eqs 10, 18, using the configurations collected for 10 ns with the final bias obtained (red curve in Figure 3); Right panel is obtained using the configurations collected for 10 ns with an intermediate bias, blue curve in Figure 3. Color scheme is the same as for Figure 7.

column). The latter dominates the change of G, since the contribution of the maximized cross-entropy is small compared to ΔU in Eq. 10. The significant changes in each component almost cancel each other in the sum. The smaller change in U_{vdw} in the F4A case indicates the cancellation of hydrophobic contacts between Phe 4 and the residues in the major hydrophobic core of folded BPTI when Phe 4 is replaced by Ala. The negative change of $U_{solv,np}$ for all variants indicates the release of hydrophilic sidechains into the solvent upon unfolding. This contribution tends to cancel the release of dispersive solute-solute interactions. However, the two electrostatic contributions (ΔU_{el} and $\Delta G_{solv,pol}$) span the largest range of values among the variants. Therefore, we argue that changes in the electrostatic networks become critical with respect to an almost uniform background of interactions that change upon the demolition of the hydrophobic core occurring during unfolding.

Most of the long-range salt-bridges lock the native structure into a less hydrophilic globular form, because the small size of the globule allows efficient electrostatic sealing, not allowed when the size of the globule increases. Breaking of the salt-bridges in the native form allows exposing hydrophylic groups to the solvent while the formation of the salt-bridges hides hydrophobic groups inside the globule core. Once salt-bridges are broken, that is, when the hydrogen bonds keeping the native scaffold are broken, the globular protein is allowed to expose a larger surface to the solvent, including its hydrophobic core.

In conclusion, the D3A stabilization against protein unfolding is due to the stabilization of non-native salt-bridges when the native Asp 3 is removed.

Our analysis has shown that for the small BPTI protein (58 residues) the number of configurations at constant bias we have been able to collect provides consistency between well tempered metadynamics and the maximal constrained entropy method. On the other hand, the improvement of statistics we achieve in this work, compared to our previous investigation of the FXN case, as explained in the next section, is not yet sufficient to get full consistency and robust predictions in the case of bigger proteins.

3.2 Frataxin (FXN)

The effects of single-point mutations on the unfolding process of the truncated form of FXN (residues 90-210) have been discussed in detail in Ref. (Botticelli et al., 2022). Differences of the present work compared to what was done in the previous paper are the following:

1. The well tempered metadynamics method is employed in place of a plain (constant T) metadynamics;
2. The construction of the biasing potential is made with a larger number of iterations and is, therefore, more accurate;

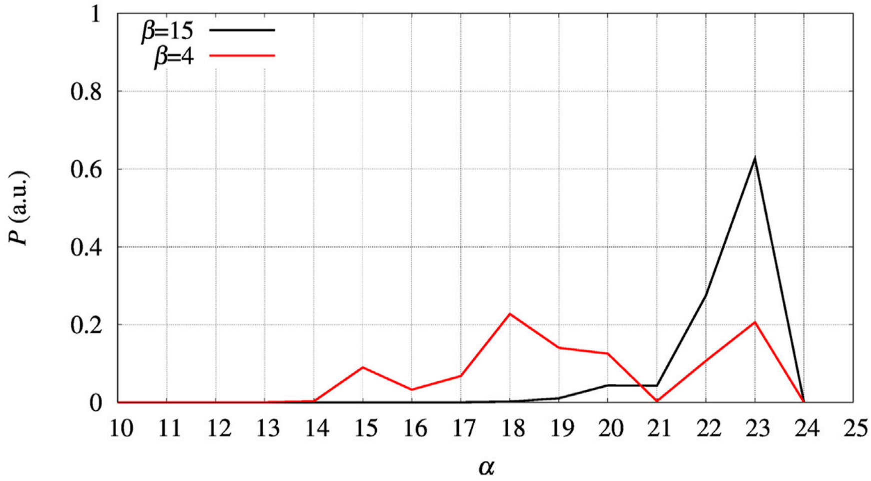


FIGURE 10
The distribution of α obtained in the 30-ns trajectories collected for the whole set of 90 walkers at constant bias $V_g(\beta)$ for FXN, WT sequence. Black curve—Distribution obtained for configurations with β = 15; red curve—Same distribution with β = 4.

TABLE 5 Energy changes (kJ/mol) starting from folded reference state, ending to unfolded state for native (WT) sequence and studied variants. Folded and unfolded states are defined as in Table 4. The energy components are those indicated in Eqs 18, 19.

Variant	ΔU_{el}	ΔU_{vdw}	$\Delta G_{solv,pol}$	$\Delta G_{solv,np}$	ΔU
			BPTI [5-55]$_{\mathrm{BPTI}}$		
WT	749.8	198.3	−264.1	−47.8	608.0
D3A	802.7	196.2	−294.4	−58.4	605.1
F4A	660.2	159.5	−170.0	−58.7	552.4
			FXN		
WT	910.8	542.7	−56.0	−127.9	1218.0
D104G	1295.0	544.9	−394.6	−133.2	1275.8
A107V	986.7	553.0	−125.1	−146.6	1206.3
F109L	788.9	583.8	−33.9	−134.3	1127.7
Y123S	754.4	563.2	37.0	−138.4	1161.3
S202F	911.1	577.0	−65.4	−135.9	1219.9

3. The maximal constrained entropy method is applied here using an extended collective variable including the number of hydrogen bonds present in the folded α-helical regions;
4. The trajectory produced at constant bias, which is used in the maximal constrained entropy method, is three times longer than in Ref. (Botticelli et al., 2022).

The change of G computed using the bias $V_G(\xi)$ as obtained out of 22 ns of bias construction is displayed in the right panels of Figures 4, 5. Again, the free energy increases upon unfolding (decrease of ξ), in agreement with what happens in similarly folded state as observed in experiments Petrosino et al. (2019) (data not shown here). However, the relative order of unfolding free energy is not well captured. Indeed, most of the variants are found to be more stable than the WT reference sequence. On the other hand, in experiments only the D104G variant among the 8 analyzed shows an increased stability of the folded state with respect to the native sequence and, therefore, a larger unfolding free energy.

An explanation of the difference between the trend showed by experiments and that predicted by direct metadynamics is in the choice of the collective variable we made to study FXN unfolding. The thermal unfolding was measured by CD at 222 nm wave-length: this means that the CD signal was mainly composed by variations in the content of α-helices. The choice of $\xi = \beta$ in metadynamics was based on the expectation that the demolition of the β-sheet would be sufficient to destabilize all the secondary motifs in the protein, including the two α-helices. This was only partially true. In Figure 10 we show the distribution of α, the number of hydrogen bonds in α-helices, in correspondence low and high values of β, 4 and 15, respectively. The distributions were computed making reference to the 30-ns long simulation at constant bias collected for the whole set of 90 walkers. The curve with $\beta = 4$ shows that α-helices are partially broken in those configurations where the β-sheet is broken. This effect is due mainly to the shortening of helix α_1 (data not shown here), which is softer than α_2 particularly in its N-terminus. Therefore, only an *a posteriori* analysis of the effect of a chosen collective variable can point to a more valid collective variable to be used in metadynamics.

The set of configurations, obtained by including all the 90 walkers simulated at constant bias, is used in the maximal constrained entropy method to overcome the above shortcoming. Results for FXN are displayed in the right panels of Figures 7, 8. In Figure 7 (right panel) we notice that the two peaks at $s' = 21$ and 37 are not due just to the choice of initial configurations (i.e. the two PDB structures used to differentiate the walkers, see Section 2). The distance in s' between the two peaks displayed in Figure 7 (right panel) is larger than the difference in α-helical values between the two PDB structures used to build the set of initial configurations, namely 37-21 compared to 23-19. Consistently with the data displayed in Figure 10, this means that the metastatistics contains configurations with a significant decrease in the number of α-helical hydrogen bonds despite the external bias forcing the unfolding being a function of the number of hydrogen bonds in the β-sheet only.

In the right panel of Figure 8 the profiles of $\Delta\Delta G$ of the 5 different variants studied with maximal constrained entropy method are compared. It is interesting to notice that the relative order of the experimental values of ΔT_m (see also Table 4) is better reproduced with the use of the augmented and updated statistics collected in this work.

The different contributions to $\Delta\Delta G$ are reported in Table 5. Again, the tendency of different contributions to compensate each other when summed is apparent. It can be noticed that, similarly to BPTI, the electrostatic contributions display a larger span among variants. In the case of D104G the value of ΔU_{el} is clearly dominant, while the opposite sign contribution of the polar solvation term is unable to compensate the effect of changes in direct electrostatic contacts. Strikingly, despite the longer accumulation of statistics and the more accurate bias construction, the reasons of the D104G stabilization can be explaind in terms of the same effects described in the previous investigation Botticelli et al. (2022). It is the removal of Asp 104 that changes the structure of the α_1 helix and the possibility of the charged residues lying in that region to form alternative salt-bridges. When α_1 helix is allowed to rotate, like in the unfolded molten globule, these interactions are not possible. However, the effect of the point mutation on the S202F variant is different as the change of dispersive interactions become significant, consistently with the introduction of a hydrophobic sidechain (Phe) in place of the small hydrophilic Ser residue. In this situation, it is possible to infer that the native-like hydrophobic core is stabilized and more work is required to destroy it and the significant change in electrostatic interactions (ΔU_{el}) is seen to positively combine with hydrophobic contributions.

Though the interactions among protein atoms and between the protein and its environment (a NaCl solution) are crudely approximated, the method is able to capture the little changes surviving when the total potential energy is computed.

4 Conclusion

In this work we refined the combination of several computational methods to predict, on the basis of fully atomistic protein models, the changes of thermal stability of proteins under single-point mutations. The method has been applied to a well-studied small protein, the bovine pancreatic trypsin inhibitor (58 residues), and to a truncated form of frataxin (121 residues). In both cases experiments were compared to computational results. The unusual effect of protein stabilization exerted by some point mutations was the special focus of this study.

We found a good agreement in the sign of representative values of $\Delta\Delta G$ upon unfolding and the sign of the shift in the melting temperature compared to experimental results. The competition between the changes in the demolition of hydrophobic cores and the changes in networks of electrostatic interactions is captured by the method. This effect was not fully analyzed in the interpretation of the unusual D3A stability in BPTI, so far.

Despite its potential, the method is computationally quite demanding, requiring extended statistical methods and, as for the collection of reliable configurations, a detailed model for atomic interactions, including explicit solvent and counterions. As discussed in the case of FXN, the direct calculation of free energy variation from the constructed bias potential is strongly affected by the choice of the collective variable in metadynamics. It was found that the maximal constrained entropy is a possible work-around to the statistical limitations of even challenging and promising methods like those based on multiple-walkers well tempered metadynamics. Numerical limitations still prevent the application to many interesting variants where the native structure becomes unstable: F33A, F22A, Y35A for BPTI; W173C for FXN. The ability of predicting the sign of the free energy change is in any case of extreme importance when the protein can adopt structures alternative to the native one.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

GN and SB made most of the reported calculations. GL designed the method and the application. GR described the method within statistical physics. SM and GS acquired the funds to perform the work. All authors equally contributed to data interpretation and manuscript writing.

Funding

This work was supported by the project PRIN-201744NR8S (MIUR, Italy) and the BIOPHYS initiative (INFN, Italy). Numerical calculations have been made possible owing to the projects IscrB_PolyFrat, IscrC_PolyBPTI, and the CINECA-INFN agreement, all providing us access to computational resources on MARCONI-100 at CINECA.

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.

References

Alexov, E., and Sternberg, M. (2013). Understanding molecular effects of naturally occurring genetic differences. *J. Mol. Biol.* 425, 3911–3913. doi:10.1016/j.jmb.2013.08.013

Attard, P. (2000). The explicit density functional and its connection with entropy maximisation. *J. Stat. Phys.* 100, 445–473. doi:10.1023/A:1018668502023

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100, 020603. doi:10.1103/PhysRevLett.100.020603

Barducci, A., Chelli, R., Procacci, P., Schettino, V., Gervasio, F. L., and Parrinello, M. (2006). Metadynamics simulation of prion protein: β-Structure stability and the early stages of misfolding. *J. Am. Chem. Soc.* 128, 2705–2710. doi:10.1021/ja057076l

Botticelli, S., La Penna, G., Nobili, G., Rossi, G., Stellato, F., and Morante, S. (2022). Modelling protein plasticity: The example of frataxin and its variants. *Molecules* 27, 1955. doi:10.3390/molecules27061955

Branden, C., and Tooze, J. (1999). *Introduction to protein structure*. London, UK: Garland Publishing Inc.

Bussi, G., Laio, A., and Tiwary, P. (2018). Metadynamics: A unified framework for accelerating rare events and sampling thermodynamics and kinetics. *(Cham Springer Int. Publ.*, 1–31. doi:10.1007/978-3-319-42913-7_49-1

Cantor, C. R., and Schimmel, P. R. (1980). *Biophysical chemistry*. San Francisco, USA: W.H. Freeman & Co.

Clark, E., Johnson, J., Dong, Y., Mercado-Ayon, E., Warren, N., Zhai, M., et al. (2018). Role of frataxin protein deficiency and metabolic dysfunction in friedreich ataxia, an autosomal recessive mitochondrial disease. *Neuronal Signal* 2, NS20180060. doi:10.1042/NS20180060

Cunningham, B. C., and Wells, J. A. (1989). High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science* 244, 1081–1085. doi:10.1126/science.2471267

Delatycki, M. B., Williamson, R., and Forrest, S. M. (2000). Friedreich ataxia: An overview. *J. Med. Genet.* 37, 1–8. doi:10.1136/jmg.37.1.1

Dhe-Paganon, S., Shigeta, R., Chi, Y.-I., Ristow, M., and Shoelson, S. E. (2000). Crystal structure of human frataxin. *J. Biol. Chem.* 275, 30753–30756. doi:10.1074/jbc.C000407200

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2016). Cosmic: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783. doi:10.1093/nar/gkw1121

Gakh, O., Ranatunga, W., Smith IV, D. Y., Ahlgren, E.-C., Al-Karadaghi, S., Thompson, J. R., et al. (2016). Architecture of the human mitochondrial iron-sulfur cluster assembly machinery. *J. Biol. Chem.* 291, 21296–21321. doi:10.1074/jbc.M116.738542

Galea, C. A., Huq, A., Lockhart, P. J., Tai, G., Corben, L. A., Yiu, E. M., et al. (2016). Compound heterozygous fxn mutations and clinical outcome in friedreich ataxia. *Ann. Neurol.* 79, 485–495. doi:10.1002/ana.24595

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387. doi:10.1016/S0022-2836(02)00442-4

Hošek, P., Kříž, P., Toulcová, D., and Spiwok, V. (2017). Multisystem altruistic metadynamics-well-tempered variant. *J. Chem. Phys.* 146, 125103. doi:10.1063/1.4978939

Hošek, P., Toulcová, D., Bortolato, A., and Spiwok, V. (2016). Altruistic metadynamics: Multisystem biased simulation. *J. Phys. Chem. B* 120, 2209–2215. doi:10.1021/acs.jpcb.6b00087

Klockgether, T. (2011). Update on degenerative ataxias. *Curr. Opin. Neurol.* 24, 339–345. doi:10.1097/WCO.0b013e32834875ba

La Penna, G. (2003). A constrained maximum entropy method in polymer statistics. *J. Chem. Phys.* 119, 8162–8174. doi:10.1063/1.1609197

La Penna, G., Morante, S., Perico, A., and Rossi, G. C. (2004). Designing generalized statistical ensembles for numerical simulations of biopolymers. *J. Chem. Phys.* 121, 10725–10741. doi:10.1063/1.1795694

Mitsutake, A., Sugita, Y., and Okamoto, Y. (2001). Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolym. Pept. Sci.)* 60, 96–123. doi:10.1002/1097-0282(2001)60:2<96::AID-BIP1007<3.0.CO;2-F

Ooi, T., Oobatake, M., Némethy, G., and Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci.* 84, 3086–3090. doi:10.1073/pnas.84.10.3086

Pandolfo, M. (2009). Friedreich ataxia: The clinical picture. *J. Neurol.* 256, 3–8. doi:10.1007/s00415-009-1002-3

Parkin, S., Rupp, B., and Hope, H. (1996). Structure of bovine pancreatic trypsin inhibitor at 125 K: Definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Cryst. D.* 52, 18–29. doi:10.1107/S0907444995008675

Petrosino, M., Novak, L., Pasquo, A., Chiaraluce, R., Turina, P., Capriotti, E., et al. (2021). Analysis and interpretation of the impact of missense variants in cancer. *Intl. J. Mol. Sci.* 22, 5416. doi:10.3390/ijms22115416

Petrosino, M., Pasquo, A., Novak, L., Toto, A., Gianni, S., Mantuano, E., et al. (2019). Characterization of human frataxin missense variants in cancer tissues. *Hum. Mutat.* 40, 1400–1413. doi:10.1002/humu.23789

Raiteri, P., Laio, A., Gervasio, F. L., Micheletti, C., and Parrinello, M. (2006). Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* 110, 3533–3539. doi:10.1021/jp054359r

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (2002). Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.* 23, 128–137. doi:10.1002/jcc.1161

Schrödinger, L. (2015). *The PyMOL molecular graphics system*. Schrödinger, LLC.Version 1.8

Schulz, T. J., Thierbach, R., Voigt, A., Drewes, G., Mietzner, B., Steinberg, P., et al. (2006). Induction of oxidative metabolism by mitochondrial frataxin inhibits cancer growth: Otto warburg revisited. *J. Biol. Chem.* 281, 977–981. doi:10.1074/jbc.M511064200

Simonson, T., Archontis, G., and Karplus, M. (2002). Free energy simulations come of age: Protein-ligand recognition. *Acc. Chem. Res.* 35, 430–437. doi:10.1021/ar010030m

Steinbrecher, T., Zhu, C., Wang, L., Abel, R., Negron, C., Pearlman, D., et al. (2017). Predicting the effect of amino acid single-point mutations on protein stability—Large-scale validation of md-based relative free energy calculations. *J. Mol. Biol.* 429, 948–963. doi:10.1016/j.jmb.2016.12.007

Yu, M.-H., Weissman, J. S., and Kim, P. S. (1995). Contribution of individual side-chains to the stability of bpti examined by alanine-scanning mutagenesis. *J. Mol. Biol.* 249, 388–397. doi:10.1006/jmbi.1995.0304

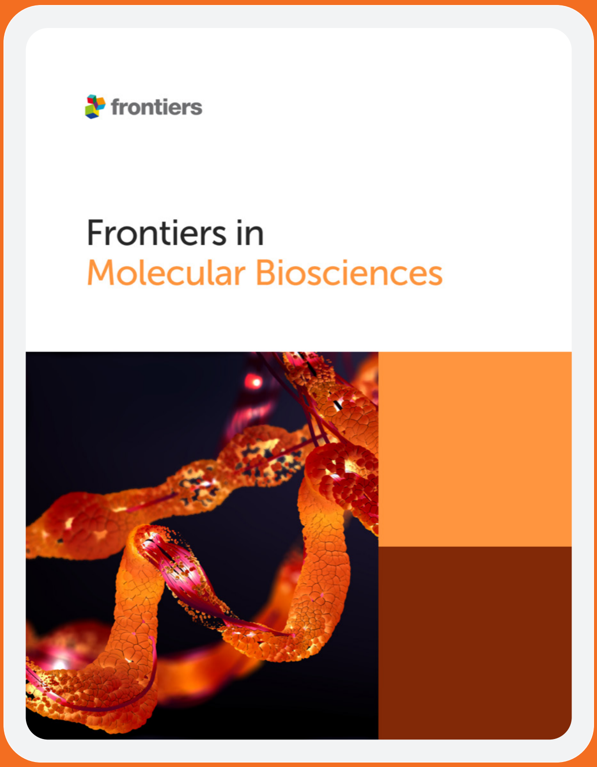
Frontiers in
Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Molecular Biosciences

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact