International Journal of
Molecular Sciences

MDPI

Communication

Molecular Origins of the Mendelian Rare Diseases Reviewed by Orpha.net: A Structural Bioinformatics Investigation

Anna Visibelli [1], Rebecca Finetti [1], Neri Niccolai [1,2,*], Ottavia Spiga [1,3] and Annalisa Santucci [1,3]

[1] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, 53100 Siena, Italy; anna.visibelli2@unisi.it (A.V.); rebecca.finetti@student.unisi.it (R.F.); ottavia.spiga@unisi.it (O.S.); annalisa.santucci@unisi.it (A.S.)

[2] Le Ricerche del BarLume Free Association, Ville di Corsano, 53014 Monteroni d'Arbia, Italy

[3] Industry 4.0 Competence Center ARTES 4.0, Viale Rinaldo Piaggio, 56025 Pontedera, Italy

* Correspondence: niccolai@unisi.it; Tel.: +39-338-465-3683

Abstract: The study of rare diseases is important not only for the individuals affected but also for the advancement of medical knowledge and a deeper understanding of human biology and genetics. The wide repertoire of structural information now available from reliable and accurate prediction methods provides the opportunity to investigate the molecular origins of most of the rare diseases reviewed in the Orpha.net database. Thus, it has been possible to analyze the topology of the pathogenic missense variants found in the 2515 proteins involved in Mendelian rare diseases (MRDs), which form the database for our structural bioinformatics study. The amino acid substitutions responsible for MRDs showed different mutation site distributions at different three-dimensional protein depths. We then highlighted the depth-dependent effects of pathogenic variants for the 20,061 pathogenic variants that are present in our database. The results of this structural bioinformatics investigation are relevant, as they provide additional clues to mitigate the damage caused by MRD.

Keywords: rare diseases; missense pathogenic variants; protein structure; databank analysis; structural bioinformatics

check for updates

Citation: Visibelli, A.; Finetti, R.; Niccolai, N.; Spiga, O.; Santucci, A. Molecular Origins of the Mendelian Rare Diseases Reviewed by Orpha.net: A Structural Bioinformatics Investigation. *Int. J. Mol. Sci.* **2024**, *25*, 6953. https://doi.org/10.3390/ijms25136953

Academic Editor: Alexandre G. De Brevern

Received: 28 March 2024
Revised: 11 June 2024
Accepted: 21 June 2024
Published: 25 June 2024

1. Introduction

Rare diseases (RDs) are defined by the World Health Organization as affecting fewer than 65 per 100,000 people, a characteristic that is mainly responsible for the lack of knowledge, expertise, and, therefore, effective treatments. Today, RD is emerging as a public health priority, and an increasing number of international networks are active to increase its visibility at the global level and to expand and share research, medical, and social care strategies. The fact that more than 70% of RDs are of genetic origin [1], and, therefore, the same DNA mutation is present in each cell type, means that a wide variety of effects occur in the affected human body. As a result, Mendelian diseases (MRD) are almost impossible to cure, although there are approaches to treat or manage some of the associated signs and symptoms [2]. If the molecular origins of MRD can be attributed to missense variants, and thus to well-defined changes at the protein level, we could, in principle, explore the correlations between the protein structural changes that occur, and the abnormal functions observed to develop rational therapeutic strategies. It is interesting to note that missense pathogenic variants are very common, as they occur in about half of the items that are present in the ClinVar genomic variant database [3]. The assignment of protein mutation sites for genomic missense variants to surface, core, or interaction regions has been proposed [4] when structural information is available from the Protein Data Bank [5]. However, despite the large number of known protein sequences, the limited number of experimentally resolved protein structures is a significant barrier to studying the structure–function correlation of proteins involved in MRD. Artificial intelligence (AI) has recently partially overcome the problem of limited structural information for investigating the effects of molecular changes on protein function [6]. Millions of reliably calculated protein structure models are currently available in the freely accessible AlphaFold database (https://alphafold.ebi.ac.uk/, (accessed on 14 October 2023)), providing broad coverage of the entire content of UniProtKB, the standard repository for protein sequences and annotations [7]. In addition, AI has recently developed AlphaMissense, a new powerful tool for predicting pathogenicity scores for all observed missense genomic variants [8]. Thus, structural bioinformatics can operate efficiently to provide powerful shortcuts for suggesting the protein basis of pathologies at the atomic level and possible remedies for MRD, as we describe in the present report with the implementation of a procedure to scan the database provided by Orpha.net, which correlates each MRD point with the corresponding mutated gene [9]. Orpha.net is therefore a suitable starting point for a structural investigation routine, which we have named Orphanetta (Orpha.net topological analysis). Orphanetta provides a general network of molecule-based information about the structural features of MRD pathogenic variants. It can therefore be a powerful tool to guide the search for new potential treatments of any pathology, present in the Orpha.net database, having structurally defined missense mutation sites.

2. Results

The Orphanetta workflow is summarized in Figure 1.

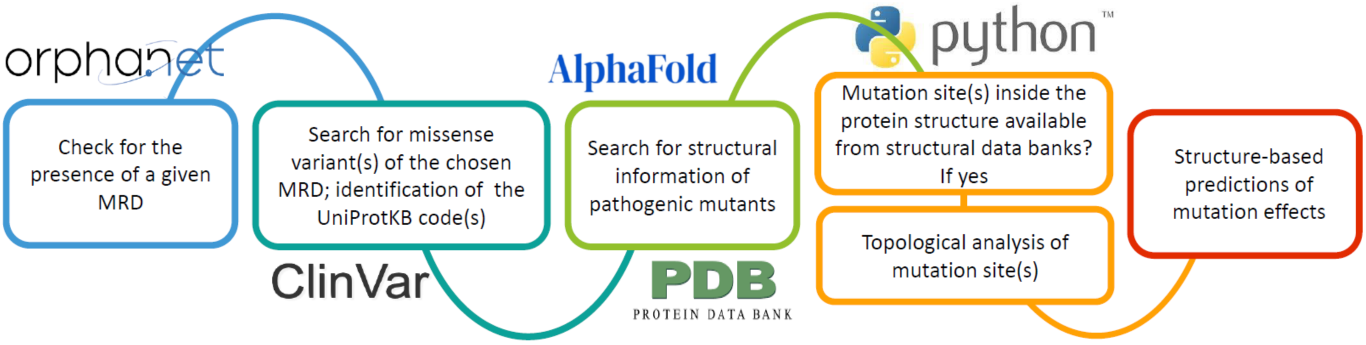

orphanet
Check for the presence of a given MRD
Search for missense variant(s) of the chosen MRD; identification of the UniProtKB code(s)
ClinVar
AlphaFold
Search for structural information of pathogenic mutants
PDB
PROTEIN DATA BANK
python™
Mutation site(s) inside the protein structure available from structural data banks? If yes
Topological analysis of mutation site(s)
Structure-based predictions of mutation effects

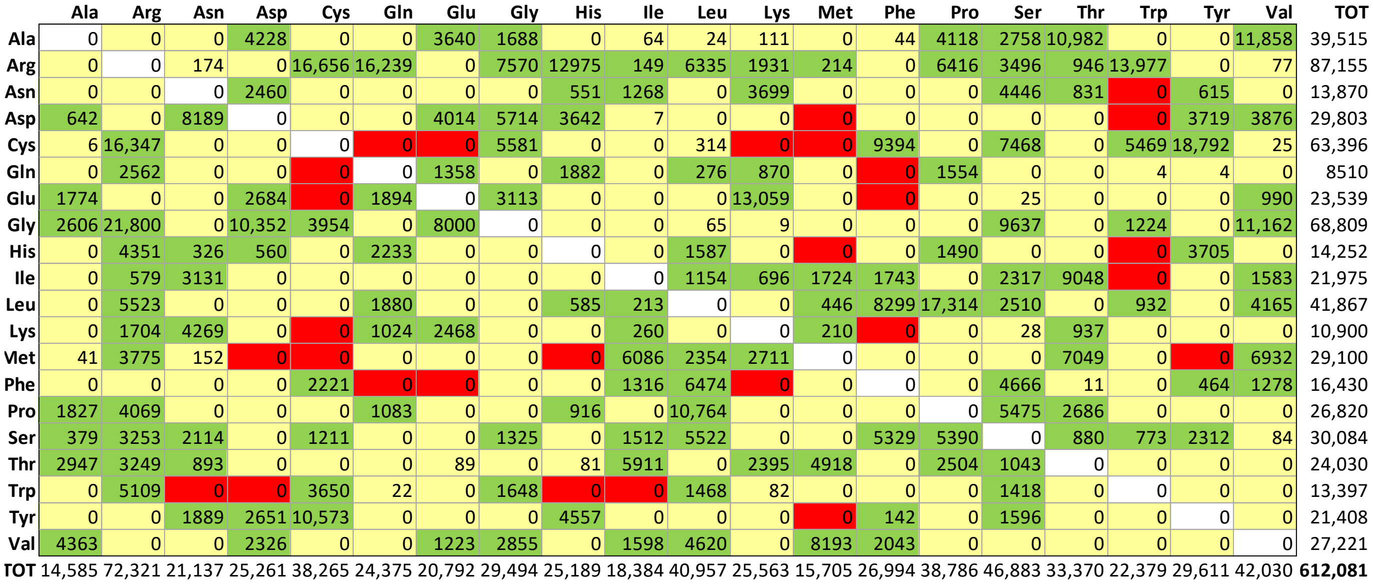
Figure 1. The Orphanetta workflow.

As of 10 October 2023, the Orpha.net database listed 4338 genes associated with MRD, which were the starting point for our investigation. ClinVar incorporated all the latter genes, adding the relevant information on the molecular consequence of their reported variants. Thus, we have delineated 3145 missense-mutated proteins involved in MRD that represent the target of our structural analysis.

2.1. Deriving Structural Information from the Orpha.net Database

Among all the 3145 proteins that were involved in MRD, several imply impossible double or triple nucleotide codon changes, as we observed in their amino acid replacement matrix, see Figure 2. Then, as we have previously performed in the general case of all the ClinVar missense variants [10], we performed a preliminary removal of the latter anomalous items. Thus, for our structural Bioinformatics analysis, we have obtained a final dataset containing 2797 proteins undergoing MRD pathogenic variants.

Based on the corresponding UniProtKB accession codes, we searched for the presence of each of these 2797 proteins in the Protein Data Bank [5]. As this search yielded only 1738 non-redundant experimentally resolved protein structures, we investigated for the presence of the additional structural information in the database of AlphaFold predicted structures [6]. We have considered only those files that ensured reliable AlphaFold models, i.e., the ones having pLDDT scores higher than 0.8. Accordingly, we have analyzed the structural features of 2515 missense-mutated proteins that form the complete repertoire of our structural Bioinformatics investigation. Multiple MRD pathogenic variants are associated with the latter proteins and the complete list of the structurally defined 20,061 pathogenic variants is given in Table S1.



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	TOT
Ala	0	0	0	4228	0	0	3640	1688	0	64	24	111	0	44	4118	2758	10,982	0	0	11,858	39,515
Arg	0	0	174	0	16,656	16,239	0	7570	12975	149	6335	1931	214	0	6416	3496	946	13,977	0	77	87,155
Asn	0	0	0	2460	0	0	0	0	551	1268	0	3699	0	0	0	4446	831	0	615	0	13,870
Asp	642	0	8189	0	0	0	4014	5714	3642	7	0	0	0	0	0	0	0	0	3719	3876	29,803
Cys	6	16,347	0	0	0	0	0	5581	0	0	314	0	0	9394	0	7468	0	5469	18,792	25	63,396
Gln	0	2562	0	0	0	0	1358	0	1882	0	276	870	0	0	1554	0	0	4	4	0	8510
Glu	1774	0	0	2684	0	1894	0	3113	0	0	0	13,059	0	0	0	25	0	0	0	990	23,539
Gly	2606	21,800	0	10,352	3954	0	8000	0	0	0	65	9	0	0	0	9637	0	1224	0	11,162	68,809
His	0	4351	326	560	0	2233	0	0	0	0	1587	0	0	0	1490	0	0	0	3705	0	14,252
Ile	0	579	3131	0	0	0	0	0	0	0	1154	696	1724	1743	0	2317	9048	0	0	1583	21,975
Leu	0	5523	0	0	0	1880	0	0	585	213	0	0	446	8299	17,314	2510	0	932	0	4165	41,867
Lys	0	1704	4269	0	0	1024	2468	0	0	260	0	0	210	0	0	28	937	0	0	0	10,900
Met	41	3775	152	0	0	0	0	0	0	6086	2354	2711	0	0	0	0	7049	0	0	6932	29,100
Phe	0	0	0	0	2221	0	0	0	0	1316	6474	0	0	0	0	4666	11	0	464	1278	16,430
Pro	1827	4069	0	0	0	1083	0	0	916	0	10,764	0	0	0	0	5475	2686	0	0	0	26,820
Ser	379	3253	2114	0	1211	0	0	1325	0	1512	5522	0	0	5329	5390	0	880	773	2312	84	30,084
Thr	2947	3249	893	0	0	0	89	0	81	5911	0	2395	4918	0	2504	1043	0	0	0	0	24,030
Trp	0	5109	0	0	3650	22	0	1648	0	0	1468	82	0	0	0	1418	0	0	0	0	13,397
Tyr	0	0	1889	2651	10,573	0	0	0	4557	0	0	0	0	142	0	1596	0	0	0	0	21,408
Val	4363	0	0	2326	0	0	1223	2855	0	1598	4620	0	8193	2043	0	0	0	0	0	0	27,221
TOT	14,585	72,321	21,137	25,261	38,265	24,375	20,792	29,494	25,189	18,384	40,957	25,563	15,705	26,994	38,786	46,883	33,370	22,379	29,611	42,030	**612,081**

Figure 2. Amino acid distributions of missense pathogenic variants reviewed by Opha.net dataset. Rows describe how each of the natural amino acids has been replaced by column residues. Colors refer to the number of codon nucleotides involved in mutations: green, yellow, and red indicate, respectively, one-, two-, and three-nucleotide changes.

2.2. Topological Assignments of Protein Mutants Responsible for MRD

The effect of amino acid replacements on protein evolution has been considered since the times when protein structural information was limited to a handful of experimentally resolved examples [11–13]. Nowadays, the wealth of the available protein structures, both experimentally obtained and high-quality predicted, allows us to discuss amino acid replacements also in terms of their 3D location. Thus, the structure-based analysis can yield powerful information for understanding the molecular mechanisms of diseases [14]; we carried out the topological analysis of mutation sites that are present in each of the 2515 proteins of our dataset to distinguish the outer and inner locations of the amino acid replacements by using POPScomp [15]. Hence, the Q(SASA) parameter has been used to assign the mutation site topology from protein cores to their most external regions [4]. Accordingly, Q(SASA) values lower than 0.15 have been considered diagnostic of inner positions of the missense mutation site, and the topology of all the variants related to MRD diseases are distributed as reported in Figure 3 and Table S1.

All the internal mutation sites related to MRD represent the very large majority of all the variants, in agreement with previous suggestions for the incidence of inner residues in general on pathogenicity [16]. Data reported in Figure 4 clearly show how the replacement profiles of amino acids are quite different in the case of buried or exposed mutation sites.

Furthermore, from inspection of the amino acid replacement matrices of the latter two groups of variants, shown in Figure 5, several features are worth a preliminary discussion. As far as fully buried mutation sites are concerned, Gly and Arg are equally abundant and much more frequent than all the other replaced amino acids, see Figure 5b, and Cys, despite its low occurrence in proteins, exhibits a very frequent involvement in pathogenicity. In the case of fully exposed mutation sites, see Figures 3 and 4, a main signal arises from the observation that very frequent substitutions occur for Met and Arg, representing, respectively, 53% (a total of 535 in the fully exposed regions) and 17% (a total of 1092 in the fully buried regions) of the total ones. The relevance of these findings in relation to MRD onset is underlined in Section 3 of the present report.

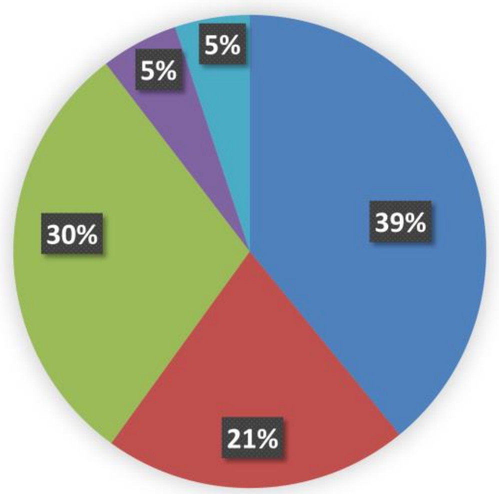


Figure 3. The sterical distribution of mutation sites in MRD proteins of our dataset. Structural data are clustered according to the POPS algorithm [4], and the Q(SASA) parameter categorizes the 20.248 mutation sites as follows: (i) 🟦 fully buried < 0.15; (ii) 🟥 internal > 0.15 and <0.24; (iii) 🟩 intermediate > 0.24 and <0.60; (iv) 🟪 external > 0.60 and <0.80; and (v) 🟦 fully solvent-exposed > 0.8.

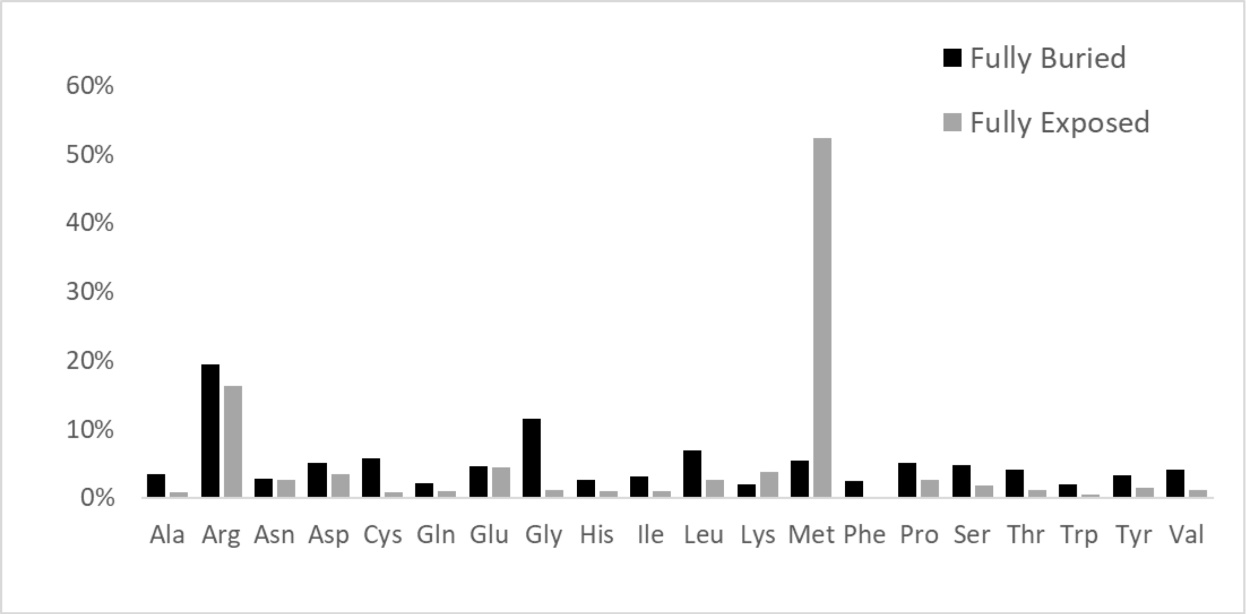
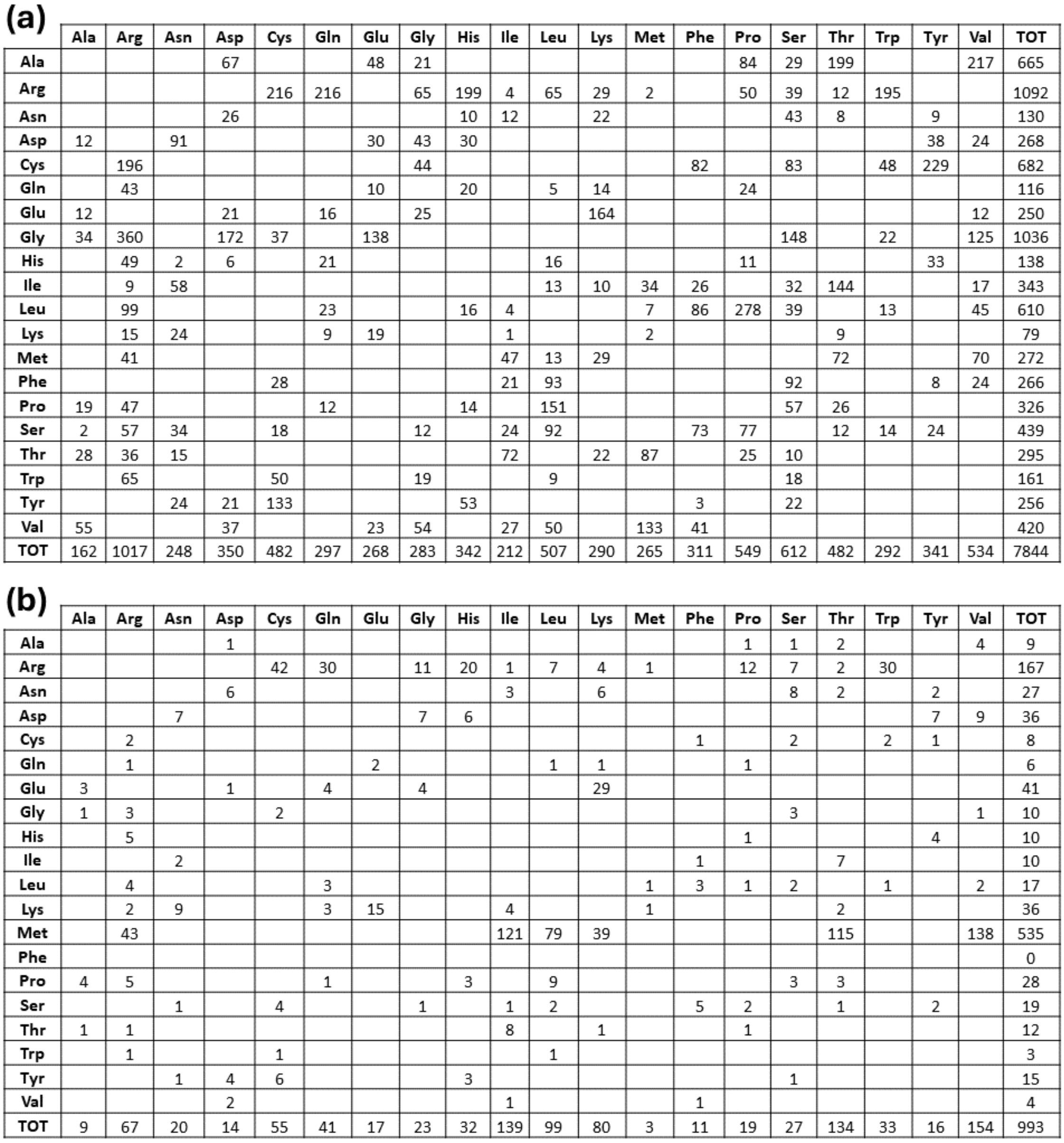


Figure 4. Profile of amino acid replacements in the structurally defined proteins that are responsible for MRD. Histograms refer to percent of amino acid variations in the fully buried and fully exposed regions, respectively, in black and grey colors.

2.3. Predicting the Structural Effects of Specific Amino Acid Replacements

The mutation matrices shown in Figure 5 confirm that MRD pathogenicity arising from amino acid substitutions in the proteins of our dataset is not univocal, but it is determined by the topology of the mutation site. For instance, replacements of amino acids bearing hydrophobic bulky side chains with other ones having electric charges, in the case they occur in the protein interior, determine a disruption of the folding nucleus, and the unfolded protein undergoes a fast proteolytic digestion. Whenever the same type of event occurs in the solvent-exposed protein surface, the protein folding process is fully conserved, but this feature causes a strong change in the protein interaction pattern with its molecular environment, ranging from the inhibition of protein quaternary assemblies to changes in protein–ligand interactions. In general, the effects of an amino acid substitution can determine the (i) reduction in the structural stability, (ii) inhibition of folding nucleus formation, (iii) interference in protein–protein, protein–nucleic acids, or protein–ligand interactions. To predict the effect of amino acid replacements, they are usually grouped into four main categories, which must undergo further subdivision to account for their hydrophobicity, polarity, electric charge, and side chain size. Hence, we have considered 10 different subgroups of amino acids, as reported in Table 1.



(a)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	TOT
Ala				67			48	21							84	29	199			217	665
Arg					216	216		65	199	4	65	29	2		50	39	12	195			1092
Asn				26					10	12		22				43	8		9		130
Asp	12		91				30	43	30										38	24	268
Cys		196						44						82		83		48	229		682
Gln		43					10		20		5	14			24						116
Glu	12			21		16		25				164								12	250
Gly	34	360		172	37		138									148		22		125	1036
His		49	2	6		21					16				11				33		138
Ile		9	58								13	10	34	26		32	144			17	343
Leu		99				23			16	4			7	86	278	39		13		45	610
Lys		15	24			9	19			1			2				9				79
Met		41								47	13	29					72			70	272
Phe					28					21	93					92			8	24	266
Pro	19	47				12			14		151					57	26				326
Ser	2	57	34		18			12		24	92			73	77		12	14	24		439
Thr	28	36	15							72		22	87		25	10					295
Trp		65			50			19			9					18					161
Tyr			24	21	133				53					3		22					256
Val	55			37			23	54		27	50		133	41							420
TOT	162	1017	248	350	482	297	268	283	342	212	507	290	265	311	549	612	482	292	341	534	7844

(b)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	TOT
Ala				1											1	1	2			4	9
Arg					42	30		11	20	1	7	4	1		12	7	2	30			167
Asn				6						3		6				8	2		2		27
Asp			7					7	6										7	9	36
Cys		2												1		2		2	1		8
Gln		1					2				1	1			1						6
Glu	3			1		4		4				29									41
Gly	1	3			2											3				1	10
His		5													1				4		10
Ile			2											1			7				10
Leu		4				3							1	3	1	2		1		2	17
Lys		2	9			3	15			4			1				2				36
Met		43								121	79	39					115			138	535
Phe																					0
Pro	4	5				1			3		9					3	3				28
Ser			1		4			1		1	2			5	2		1		2		19
Thr	1	1								8		1			1						12
Trp		1			1						1										3
Tyr			1	4	6				3							1					15
Val				2						1				1							4
TOT	9	67	20	14	55	41	17	23	32	139	99	80	3	11	19	27	134	33	16	154	993

Figure 5. Amino acid replacement matrices of fully buried (**a**) and fully exposed (**b**) pathogenic variants. As in Figure 3, rows describe how each of the natural amino acids has been replaced by column residues.

Thus, possible effects of the amino acid variations in the structurally characterized proteins of our Orpha. net-derived datasets are predicted and listed in Table S1. In Table S2 are listed all the 5221 MRD reviewed from Orpha.net, whose structural origins can be tracked with the Orphanetta procedure.

Table 1. Physico–chemical properties of natural amino acids. Sizes of amino acid side chains are classified according to ref. [17].

Electrically Charged Side Chains	Polar Uncharged Side Chains	Hydrophobic Side Chains	Special Cases
Positive: Arg, His, Lys Negative: Asp, Glu	Small size: Ser, Thr Large size: Asn, Gln, Tyr	Small size: Ala, Val Medium size: Ile, Leu, Met Large size: Phe, Trp	Cys; Gly; Pro

3. Discussion

The genetic information offered by the Orpha.net database has been our starting point for collecting additional clues at a molecular level for possible MRD remediation. Our topological analysis of all the amino acid replacements that are correlated to MRD, clearly confirms what has been already observed in general [17], i.e., low solvent exposure of mutation sites is mainly correlated to the onset of genetic diseases.

Data shown in Figure 5 indicate the very high frequency of pathogenic variants found for Arg and Gly in protein cores. The presence of Arg in inner protein regions, indeed, is very critical to maintain a positive charge, when it is needed in hydrophobic environments or to bind internal water molecules [18]. Thus, MRDs frequently come from pathogenic variants of Arg, a residue that has CG (U, C, A, G) codons that are particularly unstable [16,19]. It is well known that Gly residues are largely conserved, as with their small dimensions they can play unique roles in the structure of folded proteins [20,21]. As shown in Figure 5a, the core glycines of our structural dataset are mostly replaced by charged amino acids or residues with larger side chains, in both cases perturbing the folding nucleus formation. Figure 5a also evidences how frequently Cys substitutions in the inner protein region are lethal for the folding process, as they interrupt the cysteine-bridge network, which is necessary to stabilize the correct protein structure. Upon defining fully exposed protein residues exhibiting a Q(SASA) value above 0.8, we have selected 1.002 mutation sites. They are mainly due to Met and Arg substitutions, respectively, at 53 and 17%. It is worth noting that pathogenicity always occurs whenever the mutation involves a surface-exposed Met occupying the amino terminus position. This finding is in total agreement with a recent investigation [22] that underlines how such a mutation in the signal peptide interferes with protein targeting, translocation, processing, and stability. Several very pathogenic effects can arise from replacements of Arg occupying protein surface positions, as the latter amino acid drives most of the protein interactions with other proteins, nucleic acids, and ligands. As reported in Figure 5b, there are 167 cases where Arg is changed into all the possible alternatives given by single nucleotide variations. It would be precious information to know when these Arg mutation sites occur at the interface with their molecular partners for a better understanding of the mechanisms of pathogenicity. Very likely, this task will be possible in the near future thanks to artificial intelligence procedures such as AlphaFold-Multimer [23]. Thus, from any of the MRDs listed in Table S2, we can find the structurally defined protein(s) involved in missense pathogenic variants. Then, from the UniProtKB code(s) reported in Table S1, the mutation site topology can be delineated to predict the structure and function damage caused by the amino acid replacement.

4. Materials and Methods

4.1. Dataset of Missense Variants

As of 10 October 2023, the Orpha.net database (https://www.orpha.net/, (accessed on 10 October 2023)), providing a standardized classification and coding system for all the known rare diseases, listed 4338 genes associated with MRDs. This information represents the starting point for the present investigation. ClinVar databank [3] incorporated all the latter genes, adding the relevant information on the molecular consequence of the reported variants. Thus, we have delineated 3145 pathogenic missense variants involved in MRD that represent the target of our structural analysis.

4.2. Structural Analysis

To complete the dataset of the present structural analysis, the Protein Data Bank (PDB) [5], a widely used repository for 3D structural data of biological biopolymers, was accessed to retrieve structural information for MRD pathogenic variants. In addition, AlphaFold [6] was used to obtain structural information on the proteins that were not present in the PDB. AlphaFold's algorithm analyzes the amino acid sequence of a protein to predict the distances between pairs of amino acids, which are then used to generate a 3D model of the protein structure. For each residue, AlphaFold outputs a predicted Local Distance Difference Test (pLDDT) score, to assess the reliability of specific regions within the structure of interest. In the present investigation, we have considered only predicted models possessing very high reliability with a pLDDT > 0.8. The effects of amino acid replacements are only briefly discussed in terms of the corresponding site topology. Specific analysis of the structural changes due to mutations would require molecular dynamics simulations and could be performed only for single MRD cases. The latter methodology, indeed, needing large computational resources and times, is not suited for high-throughput investigations.

4.3. Atom Depth Calculations

The POPScomp program [15] computes the Solvent Accessible Surface Area (SASA) of any structure in a suitable PDB format. Q(SASA), which represents the ratio between the SASA of a generic Xyz amino acid inserted in the protein structure and a reference value related to the SASA of the amino acid side chain in a tripeptide GlyXyzGly. Based on the Q(SASA) parameter, it has been possible to assign the mutation site topology: (i) fully buried < 0.15; (ii) internal > 0.15 and <0.24; (iii) intermediate > 0.24 and <0.60; (iv) external > 0.60 and <0.80; and (v) fully solvent-exposed > 0.8.

5. Conclusions

The Orphanetta procedure, summarized in Figure 1, by linking the genomic information provided by the Orpha.net database to the ones available from structural data banks, seems to be well suited for AI developments that can yield fast and automatic answers for deciding the priorities for genomic editing to solve or at least mitigate the effects of MRD.

Despite this potential, this study acknowledges limitations associated with public databases. ClinVar database may be limited by inconsistencies in variant interpretation and the varying quality of submissions from different sources, which have been, however, addressed over time [24]. The Orpha.net database, instead, relies on voluntary reporting from various sources, which can lead to potential gaps or inaccuracies in the information on rare diseases, affecting the reliability and comprehensiveness of this study's findings. Future research on MRDs must undoubtedly continue, including the interpretation of findings regarding the implications of variants on protein function and their contribution to MRD pathogenicity. As an example, the observation of replacements of amino acids bearing bulky side chains with glycine in protein outer regions yields pathogenic modifications of protein surface dynamics that could be restored simply using suitable ligands [10]. These efforts will be crucial in developing targeted therapies that can mitigate the effects of these molecular disruptions, ultimately improving patient outcomes.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25136953/s1.

Author Contributions: Conceptualization, N.N.; methodology, A.V. and R.F.; formal analysis, N.N., A.V. and R.F.; investigation, N.N., A.V., R.F. and O.S.; resources, A.V. and A.S.; writing—original draft preparation, N.N.; writing—review and editing, N.N., A.V., R.F., O.S. and A.S.; supervision, N.N.; project administration, O.S. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank F. Fraternali for helpful suggestions and A. Trezza for encouraging this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nguengang Wakap, S.; Lambert, D.M.; Olry, A.; Rodwell, C.; Gueydan, C.; Lanneau, V.; Murphy, D.; Le Cam, Y.; Rath, A. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *Eur. J. Hum. Genet.* **2020**, *28*, 165–173. [CrossRef] [PubMed]
2. O'Connor, T.P.; Crystal, R.G. Genetic medicines: Treatment strategies for hereditary disorders. *Nat. Rev. Genet.* **2006**, *7*, 261–276. [CrossRef] [PubMed]
3. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J.; et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **2016**, *44*, D862–D868. [CrossRef] [PubMed]
4. Laddach, A.; Ng, J.C.F.; Fraternali, F. Pathogenic missense protein variants affect different functional pathways and proteomic features than healthy population variants. *PLoS Biol.* **2021**, *19*, e3001207. [CrossRef] [PubMed]
5. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. 2000. Available online: http://www.rcsb.org/ (accessed on 12 October 2023).
6. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
7. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [CrossRef]
8. Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L.H.; Zielinski, M.; Sargeant, T.; et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **2023**, *381*, eadg7492. [CrossRef] [PubMed]
9. Pavan, S.; Rommel, K.; Marquina, M.E.M.; Höhn, S.; Lanneau, V.; Rath, A. Clinical practice guidelines for rare diseases: The orphanet database. *PLoS ONE* **2017**, *12*, e0170365. [CrossRef] [PubMed]
10. Bongini, P.; Niccolai, N.; Trezza, A.; Mangiavacchi, G.; Santucci, A.; Spiga, O.; Bianchini, M.; Gardini, S. Structural Bioinformatic Survey of Protein-Small Molecule Interfaces Delineates the Role of Glycine in Surface Pocket Formation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 1881–1886. [CrossRef]
11. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862–864. [CrossRef]
12. Epstein, C.J. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature* **1967**, *215*, 355–359. [CrossRef] [PubMed]
13. Miyata, T.; Miyazawa, S.; Yasunaga, T. Two Types of Amino Acid Substitutions in Protein Evolution. *J. Mol. Evol.* **1979**, *12*, 219–236. [CrossRef]
14. Teng, S.; Srivastava, A.K.; Schwartz, C.E.; Alexov, E.; Wang, L. Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein protein interaction. *Int. J. Comput. Biol. Drug Des.* **2010**, *3*, 334–349. [CrossRef]
15. Cavallo, L. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.* **2003**, *31*, 3364–3366. [CrossRef] [PubMed]
16. Vitkup, D.; Sander, C.; Church, G.M. The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* **2003**, *4*, R72. [CrossRef]
17. Häckel, M.; Hinz, H.-J.; Hedwig, G.R. Partial molar volumes of proteins: Amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10–90 °C. *Biophys. Chem.* **1999**, *82*, 35–50. [CrossRef] [PubMed]
18. Harms, M.J.; Schlessman, J.L.; Sue, G.R.; García-Moreno, E.B. Arginine residues at internal positions in a protein are always charged. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 18954–18959. [CrossRef]
19. Antonarakis, S.E.; Krawczak, M.; Cooper, D.N. Disease-causing mutations in the human genome. *Eur. J. Pediatr.* **2000**, *159*, S173–S178. [CrossRef]

20. Branden, C.; Tooze, J. *Introduction to Protein Structure*, 2nd ed.; Garland Science: New York, NY, USA, 1999.
21. Parrini, C.; Taddei, N.; Ramazzotti, M.; Degl'innocenti, D.; Ramponi, G.; Dobson, C.M.; Chiti, F. Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* **2005**, *13*, 1143–1151. [CrossRef]
22. Guarnizo, S.A.G.; Kellogg, M.K.; Miller, S.C.; Tikhonova, E.B.; Karamysheva, Z.N.; Karamyshev, A.L. Pathogenic signal peptide variants in the human genome. *NAR Genom. Bioinform.* **2023**, *5*, lqad093. [CrossRef]
23. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**. [CrossRef]
24. Sharo, A.G.; Zou, Y.; Adhikari, A.N.; Brenner, S.E. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med.* **2023**, *15*, 51. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.