Article

Insights into protein structural, physicochemical, and functional consequences of missense variants in 1,330 disease-associated human genes

Sumaiya Iqbal[1,2,3], Jakob B. Jespersen[4,5+], Eduardo Perez-Palma[6,+], Patrick May[7,+], David Hoksza[7,8], Henrike O. Heyne[1,2,3,9], Shehab S. Ahmed[10], Zaara T. Rifat[10], M. Sohel Rahman[10], Kasper Lage[1,4], Aarno Palotie[1,2,9], Jeffrey R. Cottrell[1], Florence F. Wagner[1,11], Mark J. Daly[1,2,3,9], Arthur J. Campbell[1,11,*], Dennis Lal[1,6,12,13,*]

[1]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[4]Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

[5]Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark

[6]Cologne Center for Genomics, University of Cologne, Cologne, Germany

[7]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[8]Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[9]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00100 Helsinki, Finland

[10]Computer Science and Engineering, Bangladesh University of Engineering and Technology, ECE Building, West Palashi, Dhaka-1205, Bangladesh

[11]Center for the Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[12]Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, USA

[13]Genomic Medicine Institute, Lerner Research Institute Cleveland Clinic, US

*corresponding authors (dlal@broadinstitute.org; arthurc@broadinstitute.org)

[+]these authors contributed equally to this work

Abstract

Inference of the structural and functional consequences of amino acid-altering missense variants is challenging and not yet scalable. Clinical and research applications of the colossal number of identified missense variants is thus limited. Here we describe the aggregation and analysis of large-scale genomic variation and structural biology data for 1,330 disease-associated genes. Comparing the burden of 40 structural, physicochemical, and functional protein features of altered amino acids with 3-dimensional coordinates, we found 18 and 14 features that are associated with pathogenic and population missense variants, respectively. Separate analyses of variants from 24 protein functional classes revealed novel function-dependent vulnerable features. We then devised a quantitative spectrum, identifying variants with higher pathogenic variant-associated features. Finally, we developed a web resource (MISCAST; http://miscast.broadinstitute.org/) for interactive analysis of variants on linear and tertiary protein structures. The biological impact of missense variants available through the webtool will assist researchers in hypothesizing variant pathogenicity and disease trajectories.

Recent advances in technologies have ushered in an era of rapid and cost-effective genome and exome sequencing. Genetic screening is now frequently applied in clinical practice, especially for the diagnosis of rare diseases and cancer, producing a plethora of benign and disease-associated genetic variants[1,2]. A large fraction of these clinically-derived variants has been captured in databases such as OMIM[3], HGMD[4], and ClinVar[5]. Simultaneously, millions of likely-benign or mildly disease-influencing variants from the general population have been cataloged and are accessible through the Exome Aggregation Consortium (ExAC)[6] and in the genome Aggregation Database (gnomAD)[7]. Missense variation causes an amino acid substitution upon a single nucleotide variation in the protein-coding region of the genome. Missense variants can be either benign or disease-causing, and both types coexist in almost every human disease gene[6]. This underscores that functional interpretation of missense variants is particularly difficult, and some specific amino acid residues and their positions are essential for protein function and thus are vulnerable to substitutions. Exome sequencing studies can identify such protein-altering missense variants; however, the missing knowledge about the consequence of missense variants on protein function severely limits the interpretation of clinical genetic screening, and our understanding of how a missense variant is implicated in a disease phenotype.

Current algorithms for variant interpretation incorporate numerous annotations to assess the pathogenicity of missense variants, including the conservation of sequences through evolution[8-10], human population variation information[11], and protein structural information[12,13]. With the latest explosion of genetic variant data, machine learning algorithms can learn and predict variant pathogenicity with reasonable accuracy[14,15]. However, variant pathogenicity predictions, combined from multiple *in silico* tools, represent only one piece of supporting evidence (referred to as PP3[16]) in the variant interpretation guidelines as proposed by American College of Medical Genetics and Genomics (ACMG). This is primarily because of the similar underlying biases in these algorithms' prediction outputs[16], resulting from the use of similar

variant sets. Further, the prediction scores do not provide biologically interpretable annotations that can be translated into possible variant-induced changes of underlying protein function.

Studying the effect of disease-associated missense variation on the protein structure and function represents a growing research area. A damaging consequence of amino acid substitution has been found to be associated with the structural property and localization of the amino acid on protein structure[17-19]. Cluster analyses of patient variants and structural conservation analyses have shown that functionally essential amino acids and regions are not randomly distributed[20-22]. In separate research studies, missense variants were found to disrupt protein function by perturbing protein-protein interactions[23], modifying functional residues[24], destabilizing the entire protein fold[25,26], or by substituting amino acid residues in protein cores[27,28] and in protein-protein interfaces[29-31]. However, characterization of variants using a comprehensive set of protein features including physicochemical, functional, and 3D-structural features, for example, secondary structure types and protein-protein interactions, has not been performed using a common dataset and single, unified pipeline. Such a thorough characterization of disease-causing and benign missense variants would serve as a valuable resource for translational genetics, providing many structural and functional insights into the impact of an amino acid substitution.

Functionally essential protein regions are shared among proteins performing similar molecular function. There is no study that provides a quantitative comparison of pathogenic and benign missense-variant-associated protein features across different protein functional classes. The current ACMG guidelines list the presence of variants in a mutational hot spot in less well-characterized protein regions or in a critical functional domain that is depleted in population variants as a moderate criterion (referred to as PM1)[16] for variant pathogenicity assessment. However, domain information may not be the only protein-based feature that can be used as a determinant of variant pathogenicity. Developing an inclusive annotation of structural, chemical,

and functional features for missense variants would be a powerful resource for the ACMG community.

In this study, we investigate the features of amino acid residues altered by missense variants on protein structure at scale. First, we built an automated pipeline to map variants from genomic location to protein 3D structure location. Using the framework, we mapped >500,000 missense variants in 5,850 genes onto >43,000 protein 3D structures from the protein data bank (PDB)[32]. The variant set was comprised of population missense variants from the gnomAD database[7] and pathogenic missense variants from the ClinVar[5] (pathogenic and likely-pathogenic) and HGMD[4] (disease mutation) databases. Second, we annotated the amino acid positions of these genes with a comprehensive set of seven protein features with forty feature subtypes. Third, we evaluated the burden of population and pathogenic missense variants in 1,330 disease-associated genes on these forty features. To do so relative to protein function, we carried out separate analyses with variants in genes grouped into 24 protein classes based on their molecular function[33]. We validated the effectiveness of identified pathogenic and population missense variant-associated features on an independent variant set. Finally, we generated the Missense variant to proteIn StruCture Analysis web SuiTe (MISCAST; http://miscast.broadinstitute.org/). MISCAST contains feature annotations for each amino acid position of the 1,330 genes alongside a quantitative comparison with protein class-specific pathogenic and population missense-variant-associated features. MISCAST provides the most comprehensive catalog of protein-based features for missense variants and to date the only resource that gives protein class-specific characterization of missense variants.

Results

Description of dataset

In this study, we characterized missense variants with 3D coordinates in at least one protein structure using seven protein features of amino acid residues: (1) 3-class secondary structure, (2) 8-class secondary structure, (3) residue exposure level, (4) physicochemical property, (5) protein-protein interaction, (6) post-translational modification, and (7) UniProt-based functional features. These seven features further have forty subtypes (description of features in **Methods** and **Supplementary Note**). We performed the characterization jointly for all variants and independently for variants in groups of genes with similar molecular function. Initial collection of solved human protein structures from Protein Data Bank[32] (accessed September 2017) resulted in 43,805 tertiary structures for 5,850 genes. For 5,724 genes, we found variants from the gnomAD database[6] (hereafter referred to as population variants). For 1,466 and 1,673 genes, we obtained pathogenic and likely-pathogenic variants from ClinVar[6] and disease mutations from HGMD[6] databases, respectively (hereafter referred to as pathogenic variants). In total, we aggregated 1,485,579, 16,570 and 47,036 variants (unique amino acid alterations) from gnomAD[7], ClinVar[5] and HGMD[4], respectively. Out of the available variants, 33% (count = 496,869), 49.1% (count = 8,137), and 65.3% (count = 30,730) variants from gnomAD, ClinVar, and HGMD, respectively, were mappable to at least one protein structure from PDB. Population variants were mappable for 4,987 and pathogenic variants for 1,330 human genes. We restricted the data set to 1,330 genes with both population and pathogenic variants to perform an unbiased comparative analysis. The final dataset was comprised of 164,915 population variants and 32,923 pathogenic missense variants across 1,330 genes (**Supplementary Table 1**), which were grouped into 24 protein classes (**Supplementary Table 2**, protein class ascertainment in **Methods**) to perform protein class-specific analysis. This missense variant and gene dataset are referred to as Disease-Associated Genes with Structure (DAGS1330) in this

study. The workflow of this study is briefly presented in **Fig. 1** and the overview of the dataset is shown in **Table 1**.

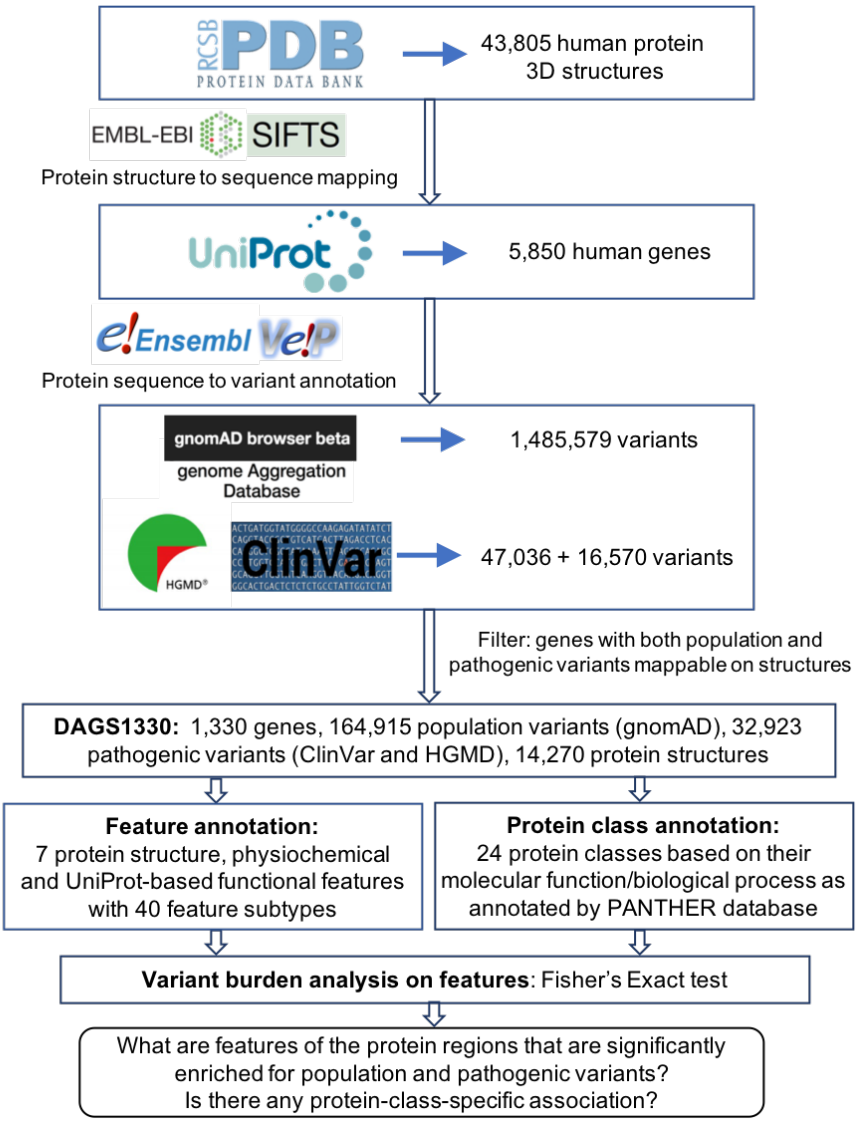


Fig. 1 | Flowchart of the study. Genetic variant to protein structure mapping pipeline, feature annotation and statistical analysis.

Table 1. Number of genes, genetic variants/amino acid alterations, protein structures in the dataset.

	Population variants	Pathogenic variants	
	gnomAD missense	ClinVar missense (likely/-pathogenic)	HGMD missense (disease mutation)
Genes with protein 3D structure	5,724	1,466	1,673
Number of amino acid alterations in the genes with protein 3D structure	1,485,579	16,570	47,036
Number of amino acid alterations mapped on the structure	496,869	8,137	30,730
Genes with variants mapped on the structure	4,897	1,330[*,+]	
Number of unique 3D structures on which the variants were mapped	29,870	14,270[*]	
Number of amino acid alterations in DAGS1330 genes that were mapped on the structure	164,915[+]	32,923[*,+]	

[*] combined (unique) counts from HGMD and ClinVar databases.

[+] Genes and variants that are considered in the final statistical analysis, referred to as Disease Associated Genes with Structure (DAGS1330).

Feature patterns of population and pathogenic missense variants on protein 3D structure across all protein classes.

Our understanding of which protein features are more likely to be disease relevant is still in its infancy. To identify structural, physiochemical, and functional protein features that are associated with pathogenic and population missense variants, we performed burden analysis with seven features including in total 40 subtypes (ascertainment and definition of features in **Methods** and **Supplementary Note**) across DAGS1330 genes. Fisher's Exact test was performed for individual feature subtypes of the population and pathogenic variants, quantifying the fold-enrichment and significance of the association. Notably, out of the seven features, physiochemical properties of amino acid and UniProt-based functional features could be curated from the protein sequence only. However, our dataset comprised of the missense variants that

could be mapped on the protein 3D structure only, thus, the identified features depict the characteristics of variants on protein structure. All seven protein features showed a deviating distribution of population and pathogenic missense variants for at least one feature subtype. The direction of the association was heterogeneous across feature subtypes (**Fig. 2**). In total, 18 out of 40 (45%) protein feature subtypes were significantly enriched for pathogenic variants. The UniProt-based functional protein features had the highest number of feature subtypes (5/6) that were enriched for pathogenic variants (**Fig. 2**). In contrast, 14 out of 40 (35%) feature subtypes showed significant depletion of pathogenic variants and enrichment of population variants. The physiochemical properties showed enrichment of population variants in six out of eight subtypes. Eight protein feature subtypes (20%) did not show any specific association to pathogenic or population variants.

A wide range of effect sizes was observed for the enrichment of pathogenic variants across the 18 protein feature subtypes with odds ratio (OR) values ranging from 1.21 to 21.75 (**Fig. 2**). The top three features associated with pathogenic variants were intermolecular disulfide bonds (OR = 21.75), SUMOylation sites (OR = 5.75), and O.GlcNAc sites (OR = 5.55). We further investigated for a possible bias in our analysis due to the wide range of population to pathogenic variant ratio per gene (0.02 – 156.0). When the analyses were repeated using only the identical number of variants per gene (20,552 pathogenic and 20,552 population variants in 588 genes, details in **Supplementary Note**), the patterns of population and pathogenic missense-variant-associated features remained consistent for 35 out of 40 features, considering the direction of association (**Supplementary Fig. 1**). The major deviations compared to the full DAGS1330 dataset were the associations of salt-bridge ionic bond and hydrogen bond sites with pathogenic variants which were observed the opposite way for the full dataset (details in **Supplementary Note**).

Protein feature and subtypes	Pathogenic yes/no	Population yes/no	OR	p-value (cut-off: 5.0e-05)
3-class secondary structure				
beta-strand/sheet	7605/25318	32537/132378	1.22	9.56e-43 (*)
helix	11510/21413	59593/105322	0.95	4.96e-05 (*)
coil	13808/19115	72785/92130	0.91	2.27e-13 (*)
8-class secondary structure				
beta-strand	414/32509	1733/163182	1.20	1.21e-03 (-)
beta-sheet	7191/25732	30804/134111	1.22	2.37e-39 (*)
three10-helix	998/31925	5736/159179	0.87	3.64e-05 (*)
alpha-helix	10248/22675	52995/111920	0.95	3.45e-04 (-)
pi-helix	264/32659	862/164053	1.54	4.12e-09 (*)
coil:loop	7472/25451	37528/127387	1.00	8.18e-01 (-)
coil:bend	3158/29765	16087/148828	0.98	3.70e-01 (-)
coil:turn	3178/29745	19170/145745	0.81	9.14e-26 (*)
Residue exposure level				
core: <5%	12859/20064	37517/127398	2.18	< 1.00e-100 (*)
buried: [5,25%)	10711/22212	46927/117988	1.21	3.56e-49 (*)
medium-buried: [25,50%)	6558/26365	50254/114661	0.57	< 1.00e-100 (*)
medium-exposed: [50,75%)	2392/30531	25748/139167	0.42	< 1.00e-100 (*)
exposed: >75%	403/32520	4469/160446	0.44	1.45e-66 (*)
Physiochemical properties of amino acids				
hydrophobic	11236/21687	60328/104587	0.90	2.23e-17 (*)
aliphatic	8328/24595	50665/114250	0.76	4.35e-88 (*)
aromatic	2908/30015	9663/155252	1.56	1.59e-83 (*)
polar	14618/18305	82293/82622	0.80	2.31e-74 (*)
positively_charged	6247/26676	31607/133308	0.99	4.25e-01 (-)
negatively_charged	3581/29342	19178/145737	0.93	8.98e-05 (-)
neutral	4790/28133	31508/133407	0.72	1.25e-88 (*)
special	7069/25854	22294/142621	1.75	< 1.00e-100 (*)
Protein-protein interactions				
disulfide	3879/613427	264/908236	20.00	< 1.00e-100 (*)
saltbridge	2946/614360	7061/901439	0.61	< 1.00e-100 (*)
hydrogen	45664/571642	69228/839272	0.97	3.06e-07 (*)
nonbonded	537359/79947	689036/219464	2.14	< 1.00e-100 (*)
Post-translational modifications				
acetylation	1812/31111	7483/157432	1.23	1.28e-13 (*)
methylation	164/32759	788/164127	1.04	6.31e-01 (-)
o-linked n-acetylglucosamine	143/32780	125/164790	5.75	7.19e-43 (*)
phosphorylation	7231/25692	31686/133229	1.18	9.01e-30 (*)
sumoylation	273/32650	248/164667	5.55	5.15e-78 (*)
ubiquitination	1990/30933	12525/152390	0.78	7.65e-24 (*)
UniProt-based functional features				
functional site	2198/30725	5950/158965	1.91	< 1.00e-100 (*)
functional/binding region	1499/31424	2341/162574	3.31	< 1.00e-100 (*)
sequence motif/region	5553/27370	21404/143511	1.36	5.79e-75 (*)
molecular processing associated region	276/32647	1385/163530	1.00	1.00e+00 (-)
modular Domain	23590/9333	70727/94188	3.37	< 1.00e-100 (*)
modified Residues	912/32011	1249/163666	3.73	< 1.00e-100 (*)

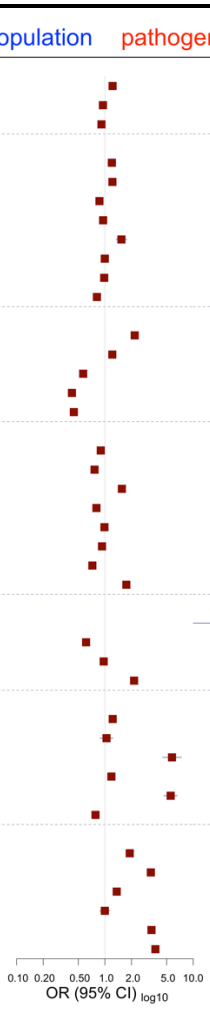


Fig. 2 | Burden of pathogenic and population variants in protein features, taking 32,923 pathogenic and 164,915 population variants in 1,330 genes. The plot shows results of the two-tailed Fisher's Exact test for 40 feature subtypes. The y-axis outlines the feature names, followed by the feature counts with and without (yes/no) pathogenic and population variants, the odds ratio (OR) and the significance (p-value). The squares (brown) show the OR and the bars (blue) show the 95% confidence interval on the x-axis. OR=1 indicates the neutral value (no enrichment or depletion) while the OR > 1.0 (and < 1.0) indicates enrichment of pathogenic variants (and population variants). If the association is significant (p-value < pcut-off = 5.0e-05), the corresponding p-value is followed by (*), otherwise (-). For protein-protein interaction types, the feature counts correspond to all bond-annotations available for the amino acid residues with pathogenic and population variants. For the rest of the features, the counts correspond to the number of amino acid residues.

Feature patterns of protein class-specific population and pathogenic missense variants on protein 3D structure.

Proteins are heterogenous in function and structure. We next explored if the features of pathogenic and population variants obtained from the full DAGS1330 dataset were consistent across different protein classes, having heterogenous molecular functions (see **Methods** for protein class ascertainment, **Supplementary Table 2** for protein class-specific genes and variant counts, **Supplementary Fig. 2 – 8** for protein class-specific Fisher's Exact test outputs, and **Table 2** for the summary of the results). Through the protein functional class-specific analyses, we aim to capture both shared and unique function-specific features of pathogenic and population variants on protein structure.

We observed that pathogenic variants are enriched in β-strand/sheet residues for 20 out 24 of protein classes (**Supplementary Fig. 2a** and **3b**). This observation is consistent with the joint analysis with full DAGS1330 dataset (**Fig. 2**). In contrast, cell junction proteins displayed a deviating characteristic with enrichment of population variants in β-strand/sheet (OR = 0.37, p = 1.40e-13, **Supplementary Fig. 2a**), specifically, β-sheet (OR = 0.35, p = 3.91e-14, in **Supplementary Fig. 3b**). Moreover, pathogenic variants of cell adhesion molecules were significantly frequent in both stable β-sheet residues (OR = 1.72, p = 7.52e-25, **Supplementary Fig. 3b**) and flexible coils, specifically, random loops (OR = 1.28, p = 3.67e-06, **Supplementary Fig. 3f**), and bends (OR = 1.35, p = 2.19e-05, **Supplementary Fig. 3g**). The enrichment of pathogenic variants followed a negative gradient of exposure levels of amino acids (**Fig. 2**) and was consistent for the majority of the protein classes (**Supplementary Fig. 4**). Amino acid residues in the interior of the 3D structure with < 25% relative accessible surface area (RSA) were enriched with pathogenic variants (**Supplementary Fig. 4a-b**), whereas the residues with ≥ 25% RSA were enriched with population variants (**Supplementary Fig. 4c-e**). Among

physiochemical property-based groups of amino acids, pathogenic variants were observed significantly more frequent on aromatic (tryptophan, tyrosine, phenylalanine) amino acids in 23 out of 24 protein classes, as well as on special amino acids (cysteine, proline, glycine) in all 24 protein classes (**Supplementary Fig. 5c** and **5h**). Specifically, the cell adhesion molecules, transcription factors, phosphatases, and signaling molecule proteins carried greater than 2-fold enrichments of pathogenic variants changing an aromatic amino acid. At positions of special amino acids, calcium-binding and extracellular matrix proteins showed enrichments of pathogenic variants with OR > 3.0 ($p < 1.0e-100$).

The inter-chain disulfide bond had the strongest enrichment in the joint analysis (OR = 21.75, $p < 1.0e-100$, **Fig. 2**). While such an association has been demonstrated [27], our protein class-specific analysis revealed that the enrichment was primarily directed by pathogenic variants from three protein classes alone (**Supplementary Fig. 6a**). In contrast, the non-bonded van der Waals interaction sites were enriched with pathogenic variants from all protein classes except phosphatases and nucleic acid binding proteins (**Supplementary Fig. 6d**). The pathogenic variants in kinases, nucleic acid binding and cytoskeletal proteins were found more frequent on salt-bridge and hydrogen bond sites (**Supplementary Fig. 6c-d**), which are protein class-specific features as these sites were found depleted of pathogenic variants for the full set (**Fig. 2**). Heterogeneous patterns of association between missense variants and post-translational modification (PTM) types were observed across different protein classes (**Supplementary Fig. 7**). The amino acid residues near phosphorylation sites were found to be enriched with pathogenic variants in joint analysis (**Fig. 2**); however, a differential pattern was observed for hydrolases (OR = 0.80, $p = 7.29e-11$) and cell adhesion molecules (OR = 0.58, $p = 3.81e-15$) with significant depletion of pathogenic variants around phosphorylation sites (**Supplementary Fig. 7d**). The cell junction proteins pathogenic variants showed an enrichment around methylation sites (OR = 9.10, $p = 9.29e-11$, **Supplementary Fig. 7b**), whereas there was no significant association observed for the full dataset. For nine out of twenty-four protein

classes, we observed significant enrichment of population variants around ubiquitination sites; however, the trend was different for transcription factors (OR = 1.44, p = 1.77e-06) and extracellular matrix proteins (OR = 2.05, p = 2.51e-05), with an enrichment of pathogenic variants around ubiquitination sites (**Supplementary Fig. 7f**). Modular domains and modified residues were identified as shared functional features for a majority of the protein classes, twenty-three and twenty-two, respectively, out of twenty-four classes (**Supplementary Fig. 8**). The pathogenic variants of transporters, kinases, transcription factors, and nucleic acid binding proteins were observed to be associated with functional sites, binding regions and sequence motifs. The protein class-specific analysis also showed that only the transfer/carrier protein (OR = 8.90, p = 3.68e-09) and enzyme modulator (OR = 3.09, p = 6.20e-10) carry enrichment of pathogenic variants on protein regions involved in molecular processing and signaling (**Supplementary Fig. 8d**); however, we observed no association with this feature collectively for the full variant set (**Fig. 2**).

Table 2. Protein-class-specific pathogenic and population missense variant-associated features. Here, the columns are for seven different protein features where each cell outlines the feature subtypes that had significant burden of pathogenic and population variants (two-tailed Fisher's Exact test, p < pcut-off, 5.0e-05). The '✗' indicates that no subtype of the corresponding protein feature was significantly associated with any variant type.

Pathogenic missense variant-associated features							
Protein class	3-class secondary structure	8-class secondary structure	Residue exposure levels	Physio chemical properties of amino acid	Protein-protein Interactions	Post-translational modification types	UniProt-based functional features
All classes	β-strand/sheet	β-sheet, π-helix	core, buried	aromatic, special	disulfide, nonbonded Van der Waals	acetylation, O.GlcNAc, phosphorylation, SUMOylating	functional site, functional/Binding region, sequence motif/region, modular domain, modified residue
Calcium-Binding Protein	✗	✗	core, buried	special	nonbonded Van der Waals	✗	functional/binding region, modular domain
Cell Adhesion Molecule	β-strand/sheet, coils	β-sheet, coil-loop, coil-bend	core, buried	aromatic, special	nonbonded Van der Waals	✗	functional site, modular domain, modified residue
Cell Junction Protein	helices	α-helix, π-helix	core	✗	nonbonded Van der Waals	methylation	functional site, modular domain
Chaperone	✗	✗	✗	✗	salt-bridge, nonbonded Van der Waals	✗	sequence motif/region, modular domain
Cytoskeletal Protein	✗	✗	core	✗	salt-bridge, hydrogen, nonbonded	acetylation, phosphorylation, SUMOylating	function/Binding region, modular domain

					Van der Waals		
Defense/ Immunity Protein	coils	✗	core	aromatic, special	✗	✗	sequence motif/region, modified residue, modular domain
Enzyme Modulator	✗	✗	core	aromatic, special	salt-bridge, nonbonded Van der Waals	✗	functional site, functional/Binding region, modular domain, modified residue, molecular processing-associated region
Extracellular Matrix Protein	✗	✗	core, buried	special	hydrogen, nonbonded Van der Waals	acetylation, SUMOylating, ubiquitination	functional site, sequence motif/region, modified residues
Hydrolase	β-strand/sheet	β-sheet, β-strand, coil-bend	core	aromatic, special	disulfide, salt-bridge, nonbonded Van der Waals	✗	functional site, modified residue, modular domain
Isomerase	β-strand/sheet	β-sheet	core	✗	salt-bridge, nonbonded Van der Waals	✗	modular domain
Kinase	helices	α-helix	core, buried	aromatic, special	salt-bridge, hydrogen, nonbonded Van der Waals	phosphorylation, SUMOylating	functional site, functional/Binding region, sequence motif/region, modified residue, modular domain

Ligase	β-strand/sheet	β-sheet	core	aromatic, special	nonbonded Van der Waals	phosphorylation, SUMOylating	function/Binding region, modular domain
Lyase	✗	✗	core	✗	nonbonded Van der Waals	✗	modular domain
Membrane Traffic Protein	β-strand/sheet	β-sheet	✗	✗	nonbonded Van der Waals	✗	sequence motif/region, modified residue, modular domain
Nucleic Acid Binding	helices	α-helix	core, buried	aromatic, special	Salt-bridge, hydrogen	phosphorylation	functional site, functional/Binding region, sequence motif/region, modular domain
Oxidoreductase	β-strand/sheet	β-sheet	core, buried	aromatic, special	nonbonded Van der Waals	SUMOylating	functional site, functional/Binding region, modular domain
Phosphatase	✗	✗	core	aromatic	✗	acetylation, phosphorylation, SUMOylating	sequence motif/region, modular domain
Protease	β-strand/sheet	β-sheet, π-helix	core, buried	aromatic, special	disulfide, salt-bridge, hydrogen, nonbonded Van der Waals	✗	modified residue, modular domain
Receptor	β-strand/sheet	β-sheet	core, buried	aromatic, special	nonbonded Van der Waals	✗	functional site, sequence motif/region, modified residue, modular domain

Signaling Molecule	β-strand/sheet	β-sheet	core, buried	aromatic, special	disulfide, nonbonded Van der Waals	✗	modified residue, modular domain
Transcription Factor	helices	α-helix	core, buried	aromatic, special	nonbonded Van der Waals	phosphorylation, ubiquitination	functional site, functional/Binding region, sequence motif/region, modular domain
Transfer/Carrier Protein	✗	π-helix	core, buried	special	nonbonded Van der Waals	acetylation, O.GlcNAc, phosphorylation	sequence motif/region, modular domain, modified residue, molecular processing-associated region
Transferase	β-strand/sheet	β-sheet	core, buried	aromatic, special	salt-bridge, hydrogen, nonbonded Van der Waals	phosphorylation, SUMOylating	functional site, functional/Binding region, modified residue, modular domain
Transporter	helices	α-helix	core	special	hydrogen, nonbonded Van der Waals	acetylation, O.GlcNAc, phosphorylation	functional site, functional/Binding region, sequence motif/region, modified residue, modular domain

Population missense variant-associated features

	3-class secondary structure types	8-class secondary structure types	Residue exposure levels	Physiochemical properties	Interaction types	Post-translational modification types	Uniprot-based functional features

All classes	helices, coils	3_{10}-helix, coil-turn	medium-buried, medium-exposed, exposed	hydrophobic, aliphatic, polar, neutral	salt-bridge, hydrogen	ubiquitination	✗
Calcium-Binding Protein	✗	✗	medium-buried, medium-exposed, exposed	aliphatic, polar, positively-charged, neutral	hydrogen	acetylation, ubiquitination	✗
Cell Adhesion Molecule	helices	α-helix	medium-buried, medium-exposed	aliphatic, polar, positively-charged, neutral	hydrogen	acetylation, phosphorylation, ubiquitination	functional/Binding region, sequence motif/region
Cell Junction Protein	β-strand/sheet	β-sheet	medium-buried, medium-exposed	✗	✗	✗	sequence motif/region
Chaperone	✗	✗	✗	✗	✗	✗	✗
Cytoskeletal Protein	✗	✗	medium-exposed	neutral	✗	✗	✗
Defense/Immunity protein	helices	α-helix	medium-exposed	hydrophobic, aliphatic, polar, positively-charged, neutral	✗	✗	✗
Enzyme Modulator	✗	✗	medium-buried, medium-exposed, exposed	polar, negatively-charged, neutral	✗	ubiquitination	✗

Extracellular Matrix Protein	✗	✗	medium-buried, medium-exposed	hydrophobic, aliphatic, polar, positively-charged, neutral	✗	✗	molecular processing-associated region
Hydrolase	helices	α-helix, coil-turn	medium-buried, medium-exposed, exposed	aliphatic, polar, neutral	✗	methylation, phosphorylation, ubiquitination	sequence motif/region, molecular processing
Isomerase	✗	✗	medium-buried, medium-exposed	✗	✗	acetylation	✗
Kinase	coils	coil-turn	medium-buried, medium-exposed, exposed	polar, neutral	✗	✗	molecular processing-associated region
Ligase	helices	α-helix	medium-buried, medium-exposed	✗	salt-bridge	ubiquitination	✗
Lyase	✗	✗	medium-buried, medium-exposed	✗	salt-bridge, hydrogen	✗	sequence motif/region
Membrane Traffic Protein	✗	✗	✗	✗	✗	✗	✗
Nucleic Acid Binding	coils	✗	medium-buried, medium-exposed, exposed	aliphatic, polar, neutral	nonbonded Van der Waals	✗	✗

Oxidoreductase	✗	✗	medium-buried, medium-exposed	aliphatic, polar, neutral	hydrogen	✗	✗
Phosphatase	✗	coil-turn	medium-buried, medium-exposed, exposed	neutral	hydrogen, nonbonded Van der Waals	ubiquitination	✗
Protease	helices	α-helix	medium-buried, medium-exposed	aliphatic, polar, positively-charged, neutral	✗	ubiquitination	molecular processing-associated region
Receptor	helices	coil-turn	medium-buried, medium-exposed, exposed	aliphatic, polar, positively-charged, neutral	hydrogen	✗	✗
Signaling Molecule	helices	3_{10}-helix, α-helix, coil-turn	medium-buried, medium-exposed, exposed	aliphatic, polar, positively-charged, negatively-charged, neutral	salt-bridge, hydrogen	ubiquitination	sequence motif/region
Transcription Factor	coils	coil-loop, coil-bend	medium-buried, medium-exposed, exposed	polar, negatively-charged, neutral	✗	✗	✗
Transfer/Carrier Protein	✗	✗	medium-exposed, exposed	aliphatic, neutral	salt-bridge, hydrogen	✗	functional site

Transferase	coils	coil-turn	medium-buried, medium-exposed, exposed	aliphatic, polar, neutral	✗	ubiquitination	molecular processing-associated region
Transporter	coils	coil-bend	medium-buried, medium-exposed	aliphatic	salt-bridge	✗	molecular processing-associated region

Validation of MISCAST annotation

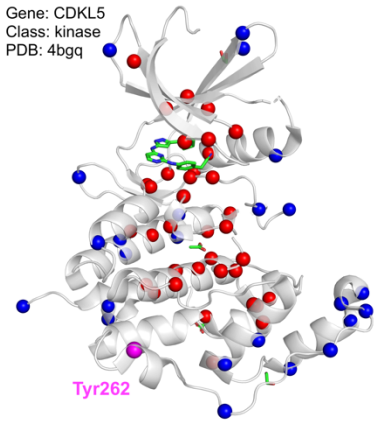
We developed an online resource for exploration of the generated data named Missense variant to proteIn StruCture Analysis web SuiTe (MISCAST; http://miscast.broadinstitute.org/). MISCAST includes interactive visualization of features and variants on linear and 3D protein structure (**Fig. 3a**), amino-acid-wise feature annotations (**Fig. 3b**), and protein class-specific features of pathogenic and population variants (**Fig. 3d**). In addition, through MISCAST, the users can obtain amino-acid-wise variant summary reports that can assist in rationalizing the pathogenicity of variants according to features of the reference amino acid. We expect that amino acid substitutions that have a deleterious impact on protein structure and function have greater numbers of pathogenic variant-associated protein features (found significant in the association analysis after Bonferroni correction). Vice versa, benign population variants are presumed to have greater numbers of population variant-associated features. To test this hypothesis, we devised a single spectrum per amino acid defined by the difference of the number of pathogenic and population variant-associated features, termed as MISCAST feature difference spectrum (MISCAST-fdspec). The positive values in MISCAST-fdspec (an example is shown in **Fig. 3c**) indicate comparatively higher pathogenic variant-associated features whereas the negative values represent relatively higher population variant-associated features. The

MISCAST-fdspec score per amino acid of a gene can be computed taking all class-specific or protein class-specific pathogenic and population variant-associated features.

An independent validation set, comprised of 18,094 pathogenic and likely-pathogenic variants (referred as PLP) and 9,222 benign and likely-benign (referred as BLB) variants (see Methods for validation set preparation) in 1,286 genes, showed differential distribution of PLP and BLB variants in all class-specific MISCAST-fdspec bands (**Fig. 4a**) with a 3.6-fold enrichment of PLP variants ($p < 2.2e-16$) in the positive bands. Separate evaluations of independent variants in 249 nucleic acid binding proteins and 208 transferases also show opposite distribution of PLP and BLB variants along protein class-specific MISCAST-fdspec, resulting in 2.43-fold (**Fig. 4b**, $p = 2.43e-37$) and 2.92-fold (**Fig. 4c**, $p = 1.96e-30$) enrichment of PLP variants in the positive bands, respectively. Finally, as another independent approach to show the utility of our newly generated protein feature annotation, we evaluated readouts from high-throughput mutagenesis for BRCA1[34] and PTEN[35] (details in Supplementary Note). Missense variants leading to loss-of-function of BRCA1 were 5.14-fold more enriched ($p = 1.13e-22$) in positive bands of nucleic acid binding protein class-specific MISCAST-fdspec (**Supplementary Fig. 9a**). We also show that functionally-verified damaging missense variants in PTEN are 6-fold enriched ($p = 3.58e-40$) in positive MISCAST-fdspec bands, specific to the phosphatase protein-class (**Supplementary Fig. 9b**).

Missense variant to proteIn StruCture Analysis web SuiTe (MISCAST)

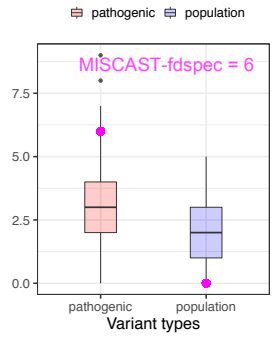
a



b **Features of amino acid residue: Tyrosine (Tyr/Y) at position 262**

1. **3-class secondary structure:** helix
2. **8-class secondary structure:** α-helix
3. **Exposure to solvent:** core (<5% relative solvent accessibility)
   - solvent accessible surface area: 3 Å^2
4. **Physiochemical property:** hydrophobic & aromatic amino acid
5. **Protein-protein interactions:** no protein complex structure available
6. **Distance to post-translational modification sites on 3D structure:**
   - **Acetylation:**
     - 4bgq:A, structure position: 33, distance to Y262: 49.912Å
   - **Phosphorylation:**
     - 4bgq:A, structure position: 262, distance to Y262: 0.0Å
7. **UniProt-based functional features:**
   - **Modular domain:** F13-F297 (Protein kinase)

c



d

Features of reference amino acid: Tyr/Y262	Pathogenic variant-associated features of kinases	Population variant-associated features of kinases
1. 3-class secondary structure: helix	helix	coils
2. 8-class secondary structure: α-helix	α-helix	coil-turn
3. Exposure to solvent: core	core, buried	core, buried
4. Physiochemical property: aromatic	aromatic, special	medium-buried, medium-/exposed
5. Protein-protein interaction: none	salt-bridge ionic, hydrogen and nonbonded Van der Waals	none
6. Post-translational modification: phosphorylation	Phosphorylation, sumoylation	none
7. UniProt-based functional features: modular domain	functional sites, functional/binding regions, sequence motif/regions, modified residues and modular domain	Molecular-processing-associated region

Fig. 3 | Overview of Missense variant to ProteIn StruCture Analysis web SuiTe (MISCAST). a, MISCAST 3D viewer with pathogenic (red) and population (blue) variants in the CDKL5 gene (Cyclin-dependent kinase-like 5) mapped on its corresponding protein structure (PDB: 4bgq). An additional reference amino acid Tyr262 is highlighted in magenta for further analysis in this overview. Features (not shown in this figure) and variants can be visualized in 1D and 3D protein structure through MISCAST and are freely downloadable for offline analysis. **b,** Structural, physicochemical and functional feature annotations for Tyr262, available in MISCAST. **c,** Quantitative comparison of protein class-specific pathogenic and population-variant-associated feature counts for Tyr262, which are 6 and 0, respectively, shown (magenta circle) on the distribution of feature counts obtained by in-training pathogenic and population variant set. Thus, for Tyr261, MISCAST-fdspec is equal to 6 – 0 = 6. A positive MISCAST-fdspec indicates that Tyr262 possesses higher variant vulnerability. **d,** Besides quantitative comparison, user can check the original features of Tyr262 that were matched with the protein function-specific features to hypothesize the molecular consequence of the Tyr262 substitution.

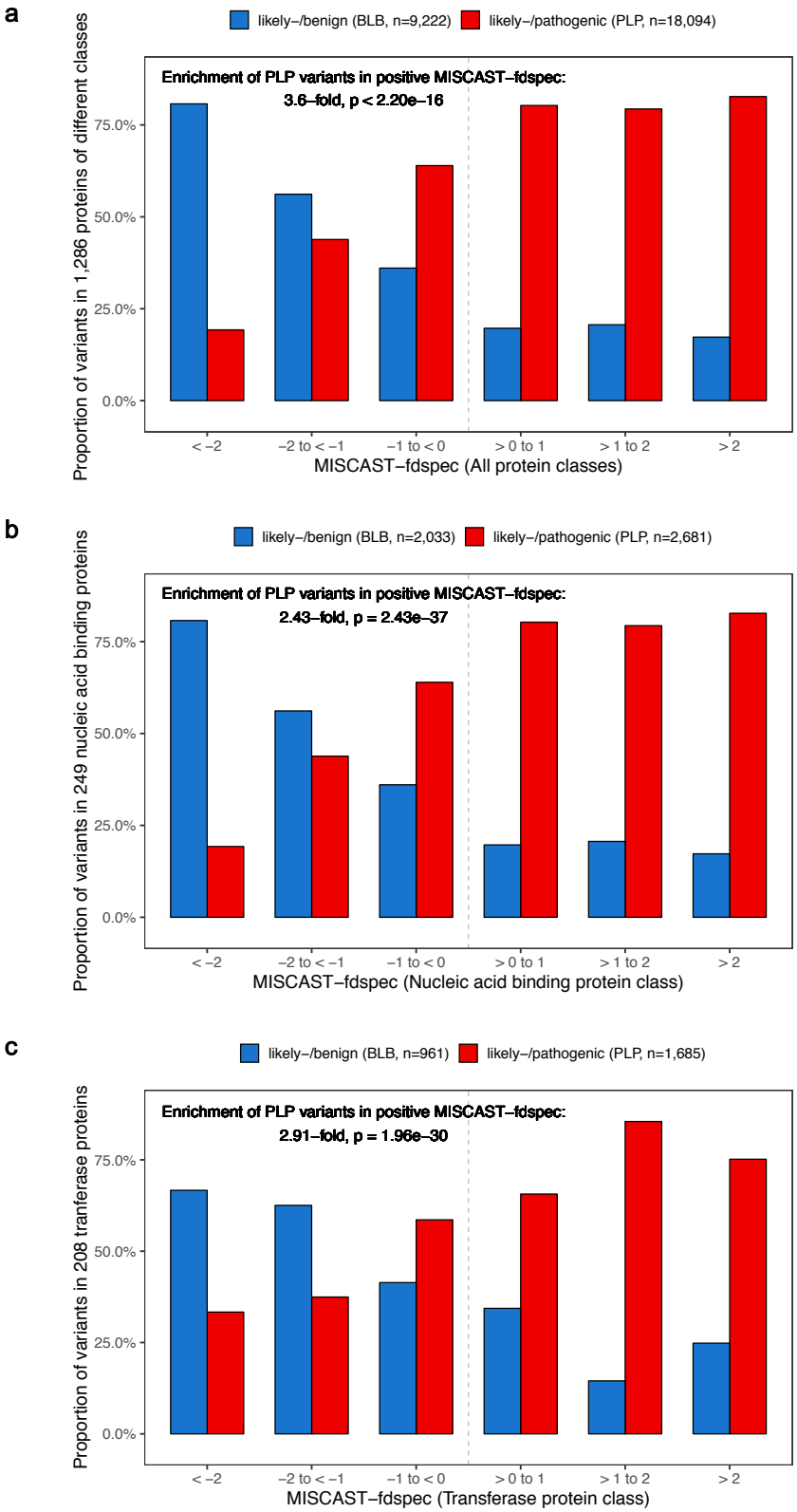

a
likely−/benign (BLB, n=9,222)
likely−/pathogenic (PLP, n=18,094)
Enrichment of PLP variants in positive MISCAST−fdspec:
3.6−fold, p < 2.20e−16
Proportion of variants in 1,286 proteins of different classes
75.0%
50.0%
25.0%
0.0%
< −2
−2 to < −1
−1 to < 0
> 0 to 1
> 1 to 2
> 2
MISCAST−fdspec (All protein classes)
b
likely−/benign (BLB, n=2,033)
likely−/pathogenic (PLP, n=2,681)
Enrichment of PLP variants in positive MISCAST−fdspec:
2.43−fold, p = 2.43e−37
Proportion of variants in 249 nucleic acid binding proteins
75.0%
50.0%
25.0%
0.0%
< −2
−2 to < −1
−1 to < 0
> 0 to 1
> 1 to 2
> 2
MISCAST−fdspec (Nucleic acid binding protein class)
c
likely−/benign (BLB, n=961)
likely−/pathogenic (PLP, n=1,685)
Enrichment of PLP variants in positive MISCAST−fdspec:
2.91−fold, p = 1.96e−30
Proportion of variants in 208 tranferase proteins
75.0%
50.0%
25.0%
0.0%
< −2
−2 to < −1
−1 to < 0
> 0 to 1
> 1 to 2
> 2
MISCAST−fdspec (Transferase protein class)

Fig | 4. Validation of MISCAST-fdspec using clinically-ascertained independent variant set. a, Proportion of pathogenic and likely-pathogenic variants from ClinVar and disease mutations from HGMD in 1,286 genes (referred as PLP), independent from the training set, and benign and likely-benign variants from ClinVar (referred as BLB) in different MISCAST-fdspec bands, quantifying the difference between the number of pathogenic- and population-variant-associated features. The most confident positive MISCAST-fdspec band (> 2) contains 4.9-times more PLP than BLB variants and that of confident negative MISCAST-fdspec band (> 2) contains 4.2-times more BLB and PLP variants. **b,** Proportion of PLP and BLB variants in 249 nucleic acid binding genes in different MISCAST-fdspec bands, computed based on nucleic acid binding protein class-specific features. **c,** Proportion of PLP and BLB variants in 208 transferase genes in different MISCAST-fdspec bands, computed based on nucleic acid binding protein class-specific features.

Discussion

Here, we investigated forty different structural, physiochemical, and functional features of amino acids altered by missense variants in 1,330 genes upon mapping of the variants onto 3D protein structures. We compared the burden of the population (gnomAD[7]) and pathogenic (combined ClinVar[5] and HGMD[36]) variants in these forty features. As expected, pathogenic and population variants showed distinct distributions for many protein features. By joint analysis on all variants in 1,330 genes, we found eighteen pathogenic- and fourteen population variant-associated protein features. Additionally, separate analyses on variants in genes of different protein functional classes resulted in novel class-specific features of pathogenic and population missense variants. We further derived a spectrum of differences between the number of pathogenic and population variant-associated features for missense variants using the study outcome, referred as MISCAST-fdspec. We expect deleterious missense variants to possess positive MISCAST-fdspec values and evaluate the method using ClinVar-ascertained and experimentally-annotated validation sets.

A large number of publicly available disease-associated missense variants is available; however, a vast majority of these variants are of uncertain significance[5], which fundamentally limits the clinical and research utility of genetic variation information. Given the complexity of variant classification, statistical approaches to objectively hypothesize the functional consequence of missense variants is pivotal[16]. MISCAST will play a key role here, as it characterizes missense variants on experimentally solved protein structures, providing function-specific features and structural insights into the effect of missense variants at the molecular level, going beyond sequence- and conservation-based methods alone. It's also worth noting that while the presence of variants in functional important protein domains has been widely utilized for variant interpretation[37,38], MISCAST further considers critical structural and physicochemical features of the amino acid that may have functional implications on the protein

structure and function upon perturbation. For reference, in our dataset, 28.3% of the pathogenic variants are not located in any protein domain according to UniProt annotation. Thus, MISCAST provides a comprehensive summary of vulnerability to structural, functional, and chemical impacts of substitutions, specifically, by reporting the original biological features to be altered along with the quantitative scoring (an illustrative overview of MISCAST in **Fig. 3**).

Our statistical analyses confirmed established association signatures between missense variants and protein features that have been found on a smaller scale and also illuminated novel protein class-specific signatures. For example, we observed an enrichment of pathogenic variants in β-strands/sheets conformation (**Fig. 2**) for all genes, thereby confirming the previous result that β-strands/sheets are more variant-intolerant and vulnerable to stabilization compared to helices upon sequence changes[39]. However, we found novel protein class-specific associations, such as pathogenic variants being significantly enriched on α-helices for five protein classes (**Supplementary Fig. 3d**). In cell junction proteins (OR = 10.54), transfer/carrier proteins (OR = 3.58), and proteases (OR = 2.71), pathogenic variants were more likely to mutate π-helix residues. The π-helical regions have been shown to undergo selection and often provide a functional advantage to proteins[40,41]; therefore, pathogenesis upon altering π-helices may be associated with perturbation of the protein functions. The protein core (Relative Solvent Area, RSA < 5%) is found to be a uniform descriptor of variant pathogenicity for all protein classes (**Supplementary Fig. 4a**), which is in line with established literature as well[27]. At the same time, for twenty-one out of twenty-four protein classes, distributions of RSA values for pathogenic missense variants showed higher maximum RSA values and accumulate a higher number of outliers (**Supplementary Fig. 10**). This result indicates the presence of pathogenic variants on the protein-protein surface[31] that may perturb essential interaction sites by altering exposed residues.

On protein structure, the special amino acids cysteine (Cys, OR = 3.84, $p < 1.00e-100$) and glycine (Gly, OR = 1.76, $p < 2.19e-154$) were found to be predominantly mutated by

pathogenic missense variation (**Supplementary Fig. 11**). The enrichment for Cys residues correlates well with the cogent association between variant pathogenicity and perturbation of covalent disulfide bond (a structure-based pathogenic-variant-associated feature), as a Cys mutation is likely to impair such a bond[27]. However, there is a protein class-specific bias in the available disulfide bond annotations[42] (**Supplementary Fig. 6a**). Specifically, 93.8% (3,642/3,879) of the disulfide bonds affected by a pathogenic variant were found in genes that encode for proteins transducing signals between cells. The other special amino acid, Gly, is the smallest amino acid having no side chain and is usually flexible. Greater than 67% of the Gly residues in our variant set is pliable coils and their substitutions by a larger amino acid is likely to have an impact on structure due to steric clashes[23]. All aromatic residues were found enriched for pathogenic variants (**Fig. 2**), with tryptophan (Trp) having the strongest association (OR = 3.27, p = 2.88e-126). Trp is the largest amino acid and it has also been shown that Trp mutations lead an increase in free energy of protein folding, making them more likely to destabilize the structure of protein[27]. Arginine (Arg) was also found to have a burden of pathogenic variants[31] (**Supplementary Fig. 11**), while the other positively-charged amino acids (lysine and histidine) were found to be associated with population variants.

Pathogenic variants were, in general, found to spatially cluster near different post-translational modification (PTM) sites[43]; however, we identified several protein class-specific novel signatures. For example, the pathogenic variants in transcription factors, extracellular matrix proteins, and structural proteins were found to be enriched near ubiquitination sites (**Supplementary Fig. 8f**), which were found depleted of pathogenic variants by joint analysis (**Fig. 2**). Kinases alone carried 15% and 56% of the pathogenic variants near phosphorylation and SUMOylation sites. Expanding the six UniProt-based functional features into twenty-five constituent features (details in **Supplementary Note**), we identified twenty-one pathogenic variant-associated features. While some of these features, such as modular domains[44], motifs, and binding sites[43], are known to be associated with pathogenic variants, our study revealed

new associations. A burden of pathogenic variants (OR = 3.9, $p < 1.0e-100$, **Supplementary Fig. 12**) on residues annotated by at least one functional feature versus no feature was also notable, showing putative association between variant pathogenicity and protein function.

It is important to note that we characterized the protein positions of missense variants that were mappable onto at least one protein 3D structure available in PDB[32], which includes structures for only one-third of all human proteins. Many of these structures have partial coverage of the protein. Therefore, our study outcomes are not exclusive of the influence that may come from accounting for the structured part of the full protein sequence. Also, we could transfer a higher proportion of pathogenic variants (~60%) compared to the population variants (~33%) onto protein structure. Out of the 1,330 disease-associated genes (>80%, n = 1,077 annotated in OMIM), >55% of the genes had at least one structure with >50% of the sequence covered in the structure, and >75% of the ClinVar and HGMD pathogenic variants could be mapped onto the structure for those genes. This observation is consistent with the positive correlation observed between the number of known variants in proteins and the number of proteins being molecularly solved and reported in the literature[45].

While performing the protein class-specific analyses, we also observed issues due to having smaller variant sets that could be mapped on the 3D structures. For one (lyase) and four (chaperons, cytoskeletal protein, defense/immunity proteins and membrane-traffic proteins) protein classes, we found three or fewer pathogenic- and population-variant-associated features (**Table 2**), respectively, after applying the stringent significance cut-off. In addition, the power of the MISCAST-fdspec metric is affected by unavailability of 3D structure information of a reference amino acid, as, in that case, the quantitative score is computed based on the physicochemical and functional features only. In future work, we plan to include high-fidelity homology models and computational models of protein structures to scale up the generated resource to the full human proteome.

Currently, a variety of in silico tools aid in the interpretation of sequence variants, primarily providing a prediction score generated using different algorithms, training cohorts, and features[16]. The three most commonly used missense variant interpretation tools[16] include PolyPhen2[8], SIFT[10] and Mutation Taster[46]. While these tools can stratify pathogenic variants from benign with a reasonable accuracy (65 – 80%)[47], progress has not been made to provide users with biologically-relevant features of the location and context of missense variants within 3D structures. This is a problem in particular for molecular scientists who need a pre-selection of variants to study protein function. In this work, besides generating a quantitative score to rationalize missense variant pathogenicity, we also provide the relevant features to support the candidate variant selection by molecular biologists (**Fig. 3**). Moreover, the gnomAD cohort has been leveraged for the first time in our work to assess genetic burden on protein features. MISCAST is also the only resource to date that gives protein functional class-specific characterization of missense variants, providing many function-specific structural insights into the effect of missense variants at the molecular level. Thus, MISCAST can serve as a powerful resource for the translation of personal genomics to precision medicine by translating genetic variant into 3D protein context: it can help to delineate variant pathogenicity, select candidate variants for functional assay, and aid in generating hypothesis for targeted drug development.

Online Methods

Protein structure and sequence collection.

The Protein Data Bank (PDB)[32] was mined on September 08, 2017 to collect all solved protein three-dimensional (3D) structures that were fully or partly (chimeric) indexed as homo sapiens. Our initial curation resulted in 43,805 protein structures, both single-chain and multiple-chain, including structures solved by six different experimental methods. Around 85% of the structures in this collection were solved by X-ray crystallography, of which over 50% of the structures in our collection are of resolution greater than 2.0 Å. The next highest proportion of the structures (~10%) were solved using nuclear magnetic resonance (NMR).

From UniProt[48], the corresponding protein identifiers and gene name for the 43,805 protein structures were collected, which resulted in 5,870 unique protein identifiers and 5,850 unique gene names. Mapping of coordinates of amino acid residues in 3D structures to linear protein sequence was derived from the SIFTS database[49].

Missense variant collection and mapping on protein structure

Protein-coding missense variants in the general population were retrieved from the genome aggregation Database (gnomAD[6]) public release 2.0.2. Consolidated files for all available exomes and genomes (coding-only subset) were downloaded in Variant Call Format[50] (VCFs) (http://gnomad.broadinstitute.org/downloads). Missense variant extraction was performed with vcftools based on the pre-annotated consequence "CSQ" field (VEP, Ensembl[51] v92). All annotations refer to the human reference genome version GRCh37.p13/hg19. Entries passing gnomAD standard quality controls (Filter = "PASS" flag) and annotated to a canonical transcript were extracted (CSQ canonical = "YES" flag). The canonical transcript is defined as the longest CCDS translation with no stop codons according to Ensembl[51]. Exome and genome missense variants calls were consolidated into one single file using in-house Perl scripts (available upon request) matching identical genomic position and annotation.

We retrieved pathogenic missense variants from two sources: the ClinVar database (ClinVar)[5], February, 2018 release and The Human Gene Mutation Database[36] (HGMD®) Professional release 2017.2. ClinVar variants were downloaded directly from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/)

in table format. Missense variants were inferred through the analysis of the Human Genome Variation Society Sequence Variant Nomenclature field[45] (HGVS, e.g. p.Gly1046Arg). To increase stringency, ClinVar variants were filtered to have only “Pathogenic” and/or “Likely Pathogenic” clinical consequence. Raw HGMD files were directly filtered for missense variants, high confidence calls (hgmd_confidence = “HIGH” flag), and disease-causing states (hgmd_variantType = “DM” flag). All annotations refer to the human reference genome version GRCh37.p13/hg19. Variants annotated to non-canonical transcripts were not considered.

Out of 5,850 genes with experimentally solved structures, we had gnomAD, ClinVar and HGMD missense variants for 5,724, 1,466 and 1,673 genes, respectively. Altogether, we collected 1,485,579; 16,570, and 47,036 amino acid alterations by missense variants from gnomAD, ClinVar and HGMD, respectively. We filtered out genes for which (i) canonical isoform protein sequences were not translated from canonical transcripts and (ii) amino acid residues affected by variants were not mappable to a protein structure. Thereafter, we had 496,869 gnomAD (33.4% of available variants), 8,137 ClinVar (49.1% of the available variants) and 30,730 HGMD (65.3% of the available variants) missense variants or amino acid alterations, respectively, that could be mapped at least one protein structure. Since ClinVar and HGMD are not mutually exclusive, we took the union of both resources, which was accomplished by comparison of their respective HGVS annotations using an in-house Perl script. Finally, we had 496,869 gnomAD (referred to as population/control) and 32,923 combined ClinVar and HGMD (referred to as pathogenic/case) missense amino acid variations from 4,897 and 1,330 genes, and these variants were mapped respectively onto 29,870 and 14,270 protein structures (PDB files). For comparative association analysis between population/pathogenic variants with protein features, we have used the 1,330 genes (Supplementary Table 1) only for which we had pathogenic (count = 32,923) and population (count = 164,915) variants. We refer to this dataset as the Disease-Associated Genes with Structure (DAGS1330). The full data overview is given in **Table 1**. The pipeline for mapping variants onto the protein 3D structures is shown in **Fig. 1**.

Protein class annotation

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) database[33] was curated to classify genes by their molecular function, biological process, and protein class. The PANTHER Protein Class ontology was adapted from the PANTHER/X molecular function ontology.

We obtained the protein class annotations for 1,330 genes in the DAGS1330 dataset with population (gnomAD) and pathogenic (ClinVar and HGMD) variants mapped on the protein 3D structures. Initially, the genes were classified into 178 groups based on similar molecular function, 234 groups based on their involvement in similar biological processes, and into 206 different protein class groups. Finally, guided by the relationships defined among different protein classes (source: http://data.pantherdb.org/PANTHER13.1/ontology/Protein_Class_13.0), the genes were grouped into 24 major protein classes. After automatic annotation, we observed that 624 genes were not assigned to any specific protein class and so as to any protein major class. We then downloaded the Ensembl family description of all human genes (version 93) using BioMart (https://www.ensembl.org/biomart). After that, we manually annotated the 624 genes into those 24 major classes based on (i) the Ensemble family description, (ii) molecular function/biological process annotation previously collected from PANTHER or (iii) molecular function/biological process annotation available in UniProt as defined by the Gene Ontology consortium. **Supplementary Table 2** lists the 24 major classes and the number of genes, the number of protein 3D structures, and the count of population/pathogenic variants mapped on protein structures for each protein class. Note that each gene may not be assigned into a unique protein class group; therefore, the genes in each group are not mutually exclusive.

Feature set mining and annotation

We annotated the amino acid residues of 1,330 genes with seven different features comprising of 40 feature subtypes. In the following, we introduce the features and we discuss how we curated these features and annotated the amino acid positions in the Supplementary Note.

1. 3-class secondary structure (subtype count: 3): (1) β-sheet/strand: β-strand and β-sheet; (2) Helices 3_{10}-helix, α-helix, π-helix; (3) Coils – turn, bend, and random loops.

2. 8-class secondary structure (subtype count: 8): (1) β-strand; (2) β-sheet; (3) 3_{10}-helix; (4) α-helix; (5) π-helix; (6) coil: turn; (7) coil: bend; (8) coil: loop.
3. Residue exposure level (subtype count: 5): (1) core (RSA < 5%); (2) buried (5% ≤ RSA < 25%); (3) medium-buried (25% ≤ RSA < 50%); (4) medium-exposed (50% ≤ RSA < 75%); (5) exposed (RSA ≥ 75%).
4. Physiochemical properties of amino acids (subtype count: 8): (1) aliphatic – alanine (Ala/A), isoleucine (Ile/I), leucine (Leu/L), methionine (Met/M), valine (Val/V); (2) aromatic – phenylalanine (Phe/F), tryptophan (Trp/W), tyrosine (Tyr/Y); (3) hydrophobic – aliphatic and aromatic amino acids; (4) positively-charged – arginine (Arg/R), histidine (His/H), lysine (Lys/K); (5) negatively-charged – aspartic acid (Asp/D), glutamic acid (Glu/E); (6) neutral – asparagine (Asn/N), glutamine (Gln/Q), serine (Ser/S), threonine (Thr/T); (7) polar – positively-charged, negatively-charged, and neutral amino acids; (8) special – proline (Pro/P), glycine (Gly/G), cystine (Cys/C).
5. Protein-protein interactions (subtype count: 4): (1) disulfide bond; (2) salt-bridge ionic interaction; (3) hydrogen bond; (4) nonbonded van der Waals interaction.
6. Post-translational modifications (subtype count: 6): (1) acetylation; (2) methylation; (3) O.GlcNAc, also known as β-linked N-acetylglucosamine; (4) phosphorylation; (5) SUMOylation; (6) ubiquitination.
7. UniProt-based functional features (subtype count: 6): (1) functional site – active site, metal binding site, binding site, site; (2) functional/binding region – zinc finger, DNA binding region, nucleotide phosphate binding region, calcium binding region; (3) sequence motif/region – region, repeat, coiled coil, motif; (4) modular domain – domain, topological domain, transmembrane, intramembrane; (5) molecular processing associated region – peptide, transit peptide, signal peptide, propeptide; (6) modified residue – modified residue, lipidation, disulfide bond, cross-link, glycosylation.

Statistical analysis

For each of the 40 subtypes of seven protein features, we carried out the two-tailed Fisher Exact test on the 2x2 contingency matrix, populated with the total number of pathogenic and population variants of one specific protein feature subtype (helix within 3-class secondary structure) and the rest within a protein structure feature-group ('not helix' equals to 'β-strand/sheet and coil' within 3-class secondary structure).

From this test, we obtained an estimate of fold enrichment, referred to as odds ratio (OR), along with the 95% confidence interval (CI) that shows whether the population/patient variants are relatively enriched or depleted in residues of any specific feature type on 3D structure. A value of OR equal to 1 indicates that there is no association between a specific type of variant with a particular feature, whereas a value of OR greater than 1 (or less than 1) indicates that the pathogenic variants are enriched (or depleted) in a particular type of feature subtype. Further, we calculated p-values showing the significance of association that were compared to the $p_{cut\text{-}off}$. This analysis was also performed taking variants from all proteins and variants from the 24 different protein classes separately, and the output gives the features that are most likely to be associated to the pathogenic and population variants in general and also separately in different protein types. We defined $p_{cut\text{-}off}$ according to the Bonferroni correction for multiple testing in statistical analysis, thus $p_{cut\text{-}off}$ was set to 5.00e-05 (0.05/(40 * 25)) as we have 40 features on which we performed the association analysis taking all proteins and separately for 24 protein classes.

Implementation of the online resource, MISCAST

The Missense variant to proteIn StruCture Analysis SuiTe, MISCAST has been made publicly available through (http://miscast.broadinstitute.org/). MISCAST was developed with the Shiny framework of R studio[52]. All the two-dimensional plots shown by the online tool are based on the ggplot2[53] and plotly[54] R libraries. To visualize the variants on one-dimensional linear protein sequence, we used the multiple sequence alignment (MSA) viewer[55] (source: http://msa.biojs.net/). To visualize the protein features and missense variants simultaneously over the one- and three-dimensional protein structures in our application, we utilized ProtVista[56] and MolArt[57]. MolArt is a JavaScript plugin enabling to view sequence annotations and overlay them over available 3D experimental or predicted structures. ProtVista[56] and MolArt[57] allows the users to visually inspect the features in the context of the 1D and 3D structures, and to contrast them with the variant positions. To support advanced analysis and visualization outside of the web environment, MolArt provides the ability to export feature and variant annotated protein structure to PyMOL. Further, the missense variant and protein feature annotations are freely downloadable in tab-delimited text format from MISCAST.

Validation of MISCAST-annotations: dataset and method

An independent pathogenic and benign missense variant set was collected from ClinVar database (ClinVar)[5], February, 2019 release and The Human Gene Mutation Database[36] (HGMD®) Professional release 2018.4. The ClinVar benign and likely-benign variants (referred as BLB) was used as the benign variant set during validation, including 9,222 variants in 4,489 amino acids. In total, 34,273 amino acid positions had ClinVar-annotated pathogenic and likely-pathogenic variants and HGMD-reported disease mutations. After removing the positions included in the training set for statistical analysis, 12,004 amino acids with 18,094 independent pathogenic variants (referred as PLP) remained the validation set. The BLB and PLP validation set contained missense variants in 1,286 genes out of the 1,330 genes. High-throughput mutagenesis readouts, classifying loss-of-function or damaging substitutions from neutral substitutions in exon 13 of BRCA1 and full PTEN protein sequence were collected form literature[34,35].

MISCAST feature difference spectrum (MISCAST-fdspec; Fig 3) was calculated taking the difference of the number of pathogenic- and population-variant associated features of a variant position. Only the features that had significant associations were considered in the MISCAST-fdspec calculation. MISCAST-fdspec > 0 indicate that a reference amino acid has comparatively more pathogenic-variant-associated features and is thereof more vulnerable to substitution. On the other hand, an amino acid with MISCAST-fdspec < 0 is expected to be more tolerant of the substitution. MISCAST-fdspec can be derived with protein class-specific or all protein class-based pathogenic- and population-variant-associated features. The maximum and minimum values for a specific protein class depends on the number of the significant pathogenic- and population-associated features, respectively (**Supplementary Table 2**).

Acknowledgements

We acknowledge Costin Leu, Giulio Genovese and Jon Bloom for insightful discussions that motivated some of the analysis presented in this manuscript. We are grateful to members of the Analytic and Translational Genetics Unit of Massachusetts General Hospital the medicinal chemistry group in the Stanley Center for useful comments and conversations. This work was supported by the Stanley Center for Psychiatric Research.

Author Contributions

D.L., A.J.C. and F.F.W. conceived the research question. S.I., D.L., A.J.C. designed the study. S.I. led the study, performed data aggregation, statistical analyses and drafted the manuscript. S.I., J.B.J, E.P and P.M. collected the data. S.I., H.O.H., P.M., A.J.C., M.J.D. and D.L. contributed to analysis concepts and methods. S.I., P.M., A.J.C., M.J.D. and D.L. interpreted the results. S.I., D.H, E.P, S.S.A. and Z.T.R. developed the online resource. D.L., M.J.D., A.J.C. and P.M. supervised the study. D.L., M.J.D., A.J.C., P.M., K.L., M.S.R., H.O.H., A.P., F.F.W. and J.R.C. provided miscellaneous support to the research. All the authors revised and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).
2. Glusman, G. Clinical applications of sequencing take center stage. *Genome Biol* **14**, 303 (2013).
3. McKusick, V.A. Mendelian Inheritance in Man and Its Online Version, OMIM. *The American Journal of Human Genetics* **80**, 588-604 (2007).
4. Stenson, P.D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1-9 (2014).
5. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
6. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
7. Karczewski, K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv 391185* (2019).
8. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
9. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
10. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).
11. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res* **27**, 1715-1729 (2017).
12. Baugh, E.H. *et al.* Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic Acids Res* **44**, 2501-13 (2016).
13. Wagih, O., Reimand, J. & Bader, G.D. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat Methods* **12**, 531-3 (2015).
14. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).

15. Pejaver, V., Mooney, S.D. & Radivojac, P. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Hum Mutat* **38**, 1092-1108 (2017).
16. Richards, S. *et al.* 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**, 405-24 (2015).
17. Reimand, J., Wagih, O. & Bader, G.D. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genet* **11**, e1004919 (2015).
18. Meyer, M.J. *et al.* mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum Mutat* **37**, 447-56 (2016).
19. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**, 3719-31 (2016).
20. Sivley, R.M., Dou, X., Meiler, J., Bush, W.S. & Capra, J.A. Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am J Hum Genet* **102**, 415-426 (2018).
21. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat Genet* **48**, 117-25 (2016).
22. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A* **112**, E5486-95 (2015).
23. Petukh, M., Kucukkal, T.G. & Alexov, E. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Hum Mutat* **36**, 524-534 (2015).
24. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647-660 (2015).
25. Kucukkal, T.G., Petukh, M., Li, L. & Alexov, E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr Opin Struct Biol* **32**, 18-24 (2015).
26. Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J Mol Biol* **425**, 3919-36 (2013).
27. Gao, M., Zhou, H. & Skolnic, J. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure* **23**, 1362 - 1369 (2015).
28. Steward, R.E., MacArthur, M.W., Laskowski, R.A. & Thornton, J.M. Molecular basis of inherited diseases: a structural perspective. *Trends Genet* **19**, 505-13 (2003).

29. David, A., Razali, R., Wass, M.N. & Sternberg, M.J. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat* **33**, 359-63 (2012).
30. Nishi, H., Nakata, J. & Kinoshita, K. Distribution of single-nucleotide variants on protein-protein interaction sites and its relationship with minor allele frequency. *Protein Sci* **25**, 316-21 (2016).
31. David, A. & Sternberg, M.J. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J Mol Biol* **427**, 2886-98 (2015).
32. Berman, H.M., et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
33. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, D183-D189 (2016).
34. Findlay, G.M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217-222 (2018).
35. Mighell, T.L., Evans-Dutson, S. & O'Roak, B.J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am J Hum Genet* **102**, 943-955 (2018).
36. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-81 (2003).
37. Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* **100**, 267-280 (2017).
38. Nicora, G. *et al.* CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. *Hum Mutat* **39**, 1835-1846 (2018).
39. Abrusán, G. & Marsh, J.A. Alpha helices are more robust to mutations than beta strands. *PLoS computational biology* **12**, e1005242 (2016).
40. Fodje, M.N. & Al-Karadaghi, S. Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering, Design and Selection* **15**, 353 - 358 (2002).
41. Weaver, T.M. The pi-helix translates structure into function. *Protein Science* **9**, 201 - 206 (2000).
42. Laskowski, R.A., Jablonska, J., Pravda, L., Varekova, R.S. & Thornton, J.M. PDBsum: Structural summaries of PDB entries. *Protein Sci* **27**, 129-134 (2018).

43. Wagih, O. *et al.* A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol Syst Biol* **14**, e8430 (2018).
44. Sun, H. & Yu, G. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci Rep* **9**, 1667 (2019).
45. Curators, T.I.C. *et al.* Capturing variation impact on molecular interactions: the IMEx Consortium mutations data set. *bioRxiv 391185* (2018).
46. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).
47. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* **32**, 358-68 (2011).
48. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).
49. Velankar, S., et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* **41**, D483–D489 (2013).
50. Danecek, P., et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
51. Zerbino, D.R., et al. . Ensembl 2018. *Nucleic acids research* **46**, D754-D761 (2017).
52. (2015)., R.T. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. (2015).
53. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* (2016).
54. Inc., P.T. Collaborative data science. (2015).
55. Yachdav, G. *et al.* MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501-3503 (2016).
56. Watkins, X., Garcia, L.J., Pundir, S., Martin, M.J. & UniProt, C. ProtVista: visualization of protein sequence annotations. *Bioinformatics* **33**, 2040-2041 (2017).
57. Hoksza, D., Gawron, P., Ostaszewski, M. & Schneider, R. MolArt: a molecular structure annotation and visualization tool. *Bioinformatics* **34**, 4127-4128 (2018).

Figure Legends

Fig. 1 | Flowchart of the study. Genetic variant to protein structure mapping pipeline, feature annotation and statistical analysis.

Fig. 2 | Burden of pathogenic and population variants in protein features, taking 32,923 pathogenic and 164,915 population variants in 1,330 genes. The plot shows results of the two-tailed Fisher's Exact test for 40 feature subtypes. The y-axis outlines the feature names, followed by the feature counts with and without (yes/no) pathogenic and population variants, the odds ratio (OR) and the significance (p-value). The squares (brown) show the OR and the bars (blue) show the 95% confidence interval on the x-axis. OR=1 indicates the neutral value (no enrichment or depletion) while the OR > 1.0 (and < 1.0) indicates enrichment of pathogenic variants (and population variants). If the association is significant (p-value < pcut-off = 5.0e-05), the corresponding p-value is followed by (*), otherwise (-). For protein-protein interaction types, the feature counts correspond to all bond-annotations available for the amino acid residues with pathogenic and population variants. For the rest of the features, the counts correspond to the number of amino acid residues.

Fig. 3 | Overview of Missense variant to ProteIn StruCture Analysis web SuiTe (MISCAST). a, MISCAST 3D viewer with pathogenic (red) and population (blue) variants in the CDKL5 gene (Cyclin-dependent kinase-like 5) mapped on its corresponding protein structure (PDB: 4bgq). An additional reference amino acid Tyr262 is highlighted in magenta for further analysis in this overview. Features (not shown in this figure) and variants can be visualized in 1D and 3D protein structure through MISCAST and are freely downloadable for offline analysis. **b,** Structural, physicochemical and functional feature annotations for Tyr262, available in MISCAST. **c,** Quantitative comparison of protein class-specific pathogenic and population-variant-associated feature counts for Tyr262, which are 6 and 0, respectively, shown (magenta circle) on the distribution of feature counts obtained by in-training pathogenic and population variant set. Thus, for Tyr261, MISCAST-fdspec is equal to 6 – 0 = 6. A positive MISCAST-fdspec indicates that Tyr262 possesses higher variant vulnerability. **d,** Besides quantitative comparison, user can

check the original features of Tyr262 that were matched with the protein function-specific features to hypothesize the molecular consequence of the Tyr262 substitution.

Fig | 4. Validation of MISCAST-fdspec using clinically-ascertained independent variant set. a, Proportion of pathogenic and likely-pathogenic variants from ClinVar and disease mutations from HGMD in 1,286 genes (referred as PLP), independent from the training set, and benign and likely-benign variants from ClinVar (referred as BLB) in different MISCAST-fdspec bands, quantifying the difference between the number of pathogenic- and population-variant-associated features. The most confident positive MISCAST-fdspec band (> 2) contains 4.9-times more PLP than BLB variants and that of confident negative MISCAST-fdspec band (> 2) contains 4.2-times more BLB and PLP variants. **b,** Proportion of PLP and BLB variants in 249 nucleic acid binding genes in different MISCAST-fdspec bands, computed based on nucleic acid binding protein class-specific features. **c,** Proportion of PLP and BLB variants in 208 transferase genes in different MISCAST-fdspec bands, computed based on nucleic acid binding protein class-specific features.