Interpreting the molecular mechanisms of disease variants in human membrane proteins

Johanna Katarina Sofie Tiemann[1,2], **Henrike Zschach**[2§], **Kresten Lindorff-Larsen**[1*], **Amelie Stein**[2*]

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark.; [2]Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark.

***For correspondence:** lindorff@bio.ku.dk (KL-L); amelie.stein@bio.ku.dk (AS)

Present address: [§]Center for Health Data Science, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen N., Denmark.

Abstract Next-generation sequencing of human genomes reveals millions of missense variants, some of which may lead to loss of protein function and ultimately disease. We here investigate missense variants in membrane proteins — key drivers in cell signaling and recognition. We find enrichment of pathogenic variants in the transmembrane region across 19,000 functionally classified variants in human membrane proteins. A key mechanism underlying pathogenicity in missense variants of soluble proteins has been shown to be loss of stability. We here perform structure-based estimations of changes in thermodynamic stability and evolutionary conservation analyses of 15 membrane proteins. We find evidence for loss of stability being the cause of pathogenicity in 59% of pathogenic variants, indicating that this is a driving factor also in membrane-protein-associated diseases. Our findings show how computational tools aid in gaining mechanistic insights into variant consequences for membrane proteins. To enable broader analyses of disease-related and population variants, we include variant mappings for the entire human proteome.

Introduction

Proteins carry out the majority of functions in a cell. While most proteins are robust to some sequence changes (***Soskine and Tawfik, 2010***), other single amino acid variants may render them non-functional. For nuclear and cytosolic proteins, we and others have shown that the molecular reason underlying loss of function for human pathogenic variants is often loss of protein stability (***Pey et al., 2007***; ***Yue et al., 2005***; ***Casadio et al., 2011***; ***Martelli et al., 2016***; ***Nielsen et al., 2017***; ***Abildgaard et al., 2019***; ***Gersing et al., 2021***; ***Scheller et al., 2019***; ***Clausen et al., 2020***). Proteins affected by such destabilizing variants are recognised by the cellular protein quality control system, leading to degradation and hence low levels that cause a loss-of-function phenotype (***Stein et al., 2019***). For soluble proteins, structure-based calculations of stability changes upon mutation (ΔΔG) (***Park et al., 2016***) correlate with experimental stability (***Guerois et al., 2002***; ***Kellogg et al., 2011***; ***Ó Conchúir et al., 2015***; ***Frenz et al., 2020***) as well as high-throughput abundance measurements (***Jepsen et al., 2020***; ***Cagiada et al., 2021***), allowing us to annotate variants accordingly (***Høie et al., 2022***). The loss of stability induced by such variants often leads to cellular protein degradation. This mechanistic link to degradation is not only interesting from a biophysical perspective but can also lead to development of treatments that rescue the variant from degradation (***Meng et al., 2017***;

Kampmeyer et al., 2017).

Twenty-three percent of sequences in the human proteome encode membrane proteins, including channels, transporters, enzymes, and receptors such as G-protein coupled receptors (GPCRs) (*Uhlén et al., 2015*). Located at the junction between two compartments and often exposed to small molecules in the bloodstream, membrane proteins are key in cell signaling and recognition, as well as major drug targets (*von Heijne, 2007*; *Hauser et al., 2017*). Variants in membrane proteins are associated with a number of diseases, including for example cystic fibrosis, Parkinson's, Alzheimer's and atherosclerosis (*Sanders and Nagy, 2000*; *Hamel, 2006*; *Koepsell, 2020*; *Vanier, 2010*).

Studying membrane proteins experimentally or computationally is challenging, as the proteins need to be considered in context of the lipid membrane (*Cournia et al., 2015*). Furthermore, while many soluble proteins can unfold and refold reversibly, the processes of synthesis, folding, and assembly are intrinsically linked for membrane proteins (*Hong, 2015*; *Booth and Clarke, 2010*). Therefore, large-scale and easily accessible experimental data for benchmarking computational tools are sparse. Despite recent methodological advances (*Marx and Fleming, 2021*), computational methods for membrane proteins are not as developed as those for soluble proteins. Thus, the application of computational analyses for examining a potential correlation between protein stability, cellular abundance, and function analogous to that for soluble proteins may be particularly challenging for membrane proteins.

Building on recent force field developments that make computational analysis of membrane proteins more realistic (*Alford et al., 2021*), we here set out to assess whether calculations of the change in folding free energy can be used to identify the subset of pathogenic variants that are likely caused by loss of stability. In particular, we calculate the change in folding energy between a wild-type protein and a protein variant ($\Delta\Delta G = \Delta G_{MUT} - \Delta G_{WT}$), where low $\Delta\Delta G$ correspond to substitutions that—in light of stability—appear well-tolerated, and high $\Delta\Delta G$ for variants that destabilize the protein structure. We first combined several protein annotation databases to obtain an overview of the number and types of missense variants that are found in membrane proteins. We then analyzed in more detail 15 human membrane proteins for which high-resolution structural data as well as annotations of pathogenic and benign variants were available, and calculated $\Delta\Delta G$ values for them. In addition, we used an evolutionary sequence analysis approach (*Laine et al., 2019*) to calculate a value which we term $\Delta\Delta E$ indicating the evolutionary importance of each residue. This approach has been shown to be useful in detecting detrimental variants and includes both loss of stability variants and variants that lose function due to—for example—catalytic impairment of enzymes or mistrafficking (*Frazer et al., 2021*; *Feinauer and Weigt, 2017*; *Nicoludis and Gaudet, 2018*). The combination of $\Delta\Delta G$ and $\Delta\Delta E$ has proven particularly useful for providing mechanistic insight into loss-of-function variants in soluble proteins (*Cagiada et al., 2021*; *Høie et al., 2022*). We here apply such a combined analysis to gain mechanistic insight into variant consequences in 15 selected membrane proteins.

Results

Variant annotations in human membrane proteins

We first set out to obtain an overview of the presence and properties of missense variants in human membrane proteins. We searched Uniprot (*The UniProt Consortium, 2021*) for keywords like TRANSMEM (see Methods for details) and used the results to define a list of 5,522 proteins that are thought to be embedded in the membrane (Fig. 1A). We subsequently searched the gnomAD (*Karczewski et al., 2020*) and ClinVar (*Landrum et al., 2018*) databases for missense variants in the genes encoding these proteins (see Methods for additional details). gnomAD is a database aggregating the variants observed in ca. 150,000 exome and genome sequences, and thus provides a relatively unbiased view of the variants that are present in the human population (*Karczewski et al., 2020*). ClinVar is a database containing, among other things, missense variants that have been categorized as benign, pathogenic, or *variants of uncertain significance* (VUS), the latter indicating that

Table 1. Number of variants annotated by ClinVar (as benign, pathogenic, and VUS), gnomAD and their intersection in total and for each cellular regions. VUS (ClinVar) includes conflict variants.

	all	extracellular	cytoplasmic	transmembrane	other
total	1,867,856	574,211	447,435	258,366	587,844
benign	11,063	3,489	3,544	961	3,069
benign $\bigcap$ gnomAD	10,260	3,214	3,326	908	2,812
pathogenic	8,026	2,327	2,233	1,863	1,603
patho. $\bigcap$ gnomAD	2,360	632	628	454	646
VUS	64,584	17,451	25,119	6,814	15,200
VUS $\bigcap$ gnomAD	36,700	9,907	14,693	3,574	8,526
only gnomAD	1,784,183	550,944	416,539	248,728	567,972

the pathophysiological consequences of the variant are not clear (***Landrum et al., 2018***). Almost all (98.1%) membrane proteins have at least one variant in gnomAD, and about half (44.0%) have at least one variant in ClinVar (Fig. 1A). Across the two datasets, we find $1.9 \cdot 10^6$ unique variants in 5,471 proteins. We excluded synonymous, silent, or deletion variants, which make up 0.3% (5,403 variants) from any further analysis. Nearly all (99.1%) of the non-synonymous variants are either from gnomAD or are assigned as variants of uncertain significance in ClinVar, and only 19,089 of the 1.9 million variants have an assigned status of being pathogenic or benign (Fig. 1A and Table 1), highlighting the scope of the problem of determining variant effects.

Variants that are pathogenic are expected to be depleted in the human population compared to those that are benign, and indeed we find a clear separation of the distributions of allele frequencies between the two classes (Fig. 1C). We also observe that while 92.7% of benign ClinVar variants have been observed in gnomAD, this is only true for 29.4% of pathogenic variants (Fig. 1B, Table 1). The separation in the distribution of allele frequencies between pathogenic and benign variants suggests that variants with allele frequencies $> 9.9 \cdot 10^{-5}$ are more likely to be benign than pathogenic (Fig. 1C, cut-off calculated from the receiver-operator characteristic (ROC) analysis, see methods). While the allele frequency in gnomAD appears to be a good predictor of pathogenicity (area under the curve (AUC) of 0.96; Fig. S1), we note that this result should be taken with some caution. First, many ClinVar variants are not found in gnomAD (Fig. 1B), limiting the practical utility. Second, since the presence in gnomAD might have been used to assign (lack of) pathogenicity, it is difficult to ensure that the two sets of data are independent.

We analyzed in which regions of the protein structures the ClinVar (Fig. 1D,E) and gnomAD (Table 1) variants are located. We find that most variants are found in soluble domains, though this is likely due to the fact that these regions makes up 83% of membrane proteins (Fig. 1D,E). Notably, though, we find that while the numbers of known benign and pathogenic variants are similar in the different types of soluble regions, there appears to be an almost two-fold excess of pathogenic variants compared to benign variants in the transmembrane regions (Fig. 1E, Table 1). While we cannot exclude that this enrichment is in part due to an increased focus on the transmembrane region in clinical research, we suggest—in line with previous work (***Zaucha et al., 2021***; ***Lee and Manoil, 1994***)—that this observation also reflects a decreased mutational tolerance of the transmembrane region.

To gain a better understanding of the mechanisms causing benign or pathogenic variant consequences, we mapped the variants onto known protein structures. Despite recent advances in protein structure prediction (***Varadi et al., 2021***) and analysis using computational methods (***del Alamo et al., 2022***; ***Akdel et al., 2021***), we decided to focus our work on experimentally determined structures. Specifically, we searched the protein databank (***wwPDB consortium, 2019***) for structures with at least one variant in the resolved part of the protein structure and found that 27.5% of all annotated human membrane proteins have at least some part resolved and that 15.1% of

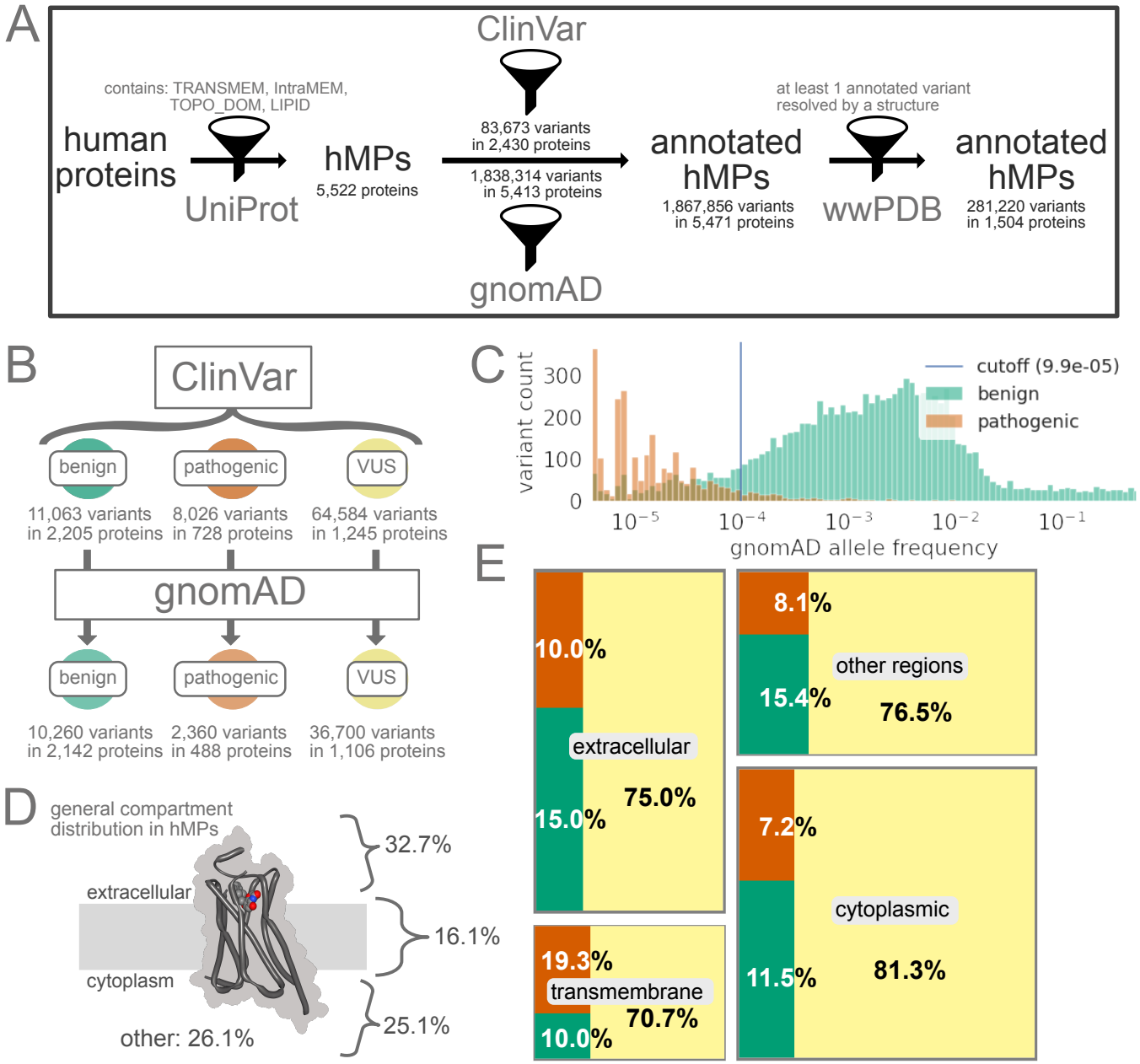

Figure 1. Overview of variants in human membrane proteins (hMPs). A. Statistics on membrane annotations, ClinVar/gnomAD annotation, and availability of high-resolution structures. B. ClinVar statistics for benign, pathogenic, and variants of unknown significance (VUS) and their coverage by gnomAD. C. gnomAD allele frequencies for all variants observed in gnomAD. Note that 71% of the pathogenic variants in ClinVar are not in gnomAD and hence missing from this analysis. D. Percentage of protein regions of different cellular elements across human membrane proteins. 'Other' includes compartments such as the lumen, mitochondrial matrix, vesicular compartments. E. Distribution of ClinVar annotations (color scheme as in B) in different regions of the proteins.

Table 2. Overview of proteins analysed in depth, including Uniprot name, protein functional class, protein length and number of amino acids within the transmembrane (TM) region, number of group A (=pathogenic), group B (=benign and/or non-rare gnomAD) and benign variants before and after filtering, the sequence depth of the multiple sequence alignment (MSA) used by GEMME, the AUC for our two predictors (ΔΔG and ΔΔE) and the MIM disease phenotype.

protein info		length		before filtering			after filtering			GEMME	AUC		MIM
name	class	all	TM	group A	group B	benign	group A	group B	benign	MSA depth	ΔΔG	ΔΔE	phenotype
NPC1	Transporter	1278	277	60	36	13	44	12	5	1486	0.69	0.84	Niemann-Pick disease
OPSD	GPCR	348	161	67	10	2	41	6	2	1183	0.79	0.69	Night blindness; Retinitis punctata albescens; Retinitis pigmentosa
GTR1	Transporter	492	261	56	9	4	42	7	2	1772	0.65	0.92	Dystonia; GLUT1 deficiency syndrome; Epilepsy
AT2A2	Transporter	1042	204	12	9	7	8	5	4	1924	0.85	1	Acrokeratosis verruciformis; Darier disease;
ACHB2	Ion channel	502	86	3	16	12	1	5	3	1789	0.8	1	Epilepsy
CXB2	Cell Junction	226	83	60	23	8	33	16	5	1351	0.42	0.84	Deafness; Bart-Pumphrey syndrome; Vohwinkel syndrome;...
S5A2	Enzyme	254	84	21	11	5	18	8	5	1177	0.63	0.9	Pseudovaginal perineoscrotal hypospadias
MC4R	GPCR	332	165	12	11	2	10	8	1	1265	0.66	0.9	Obesity
AQP2	Ion channel	271	124	10	2	1	7	1	1	1326	1	0.86	Diabetes insipidus
ACHA4	Ion channel	627	85	3	47	32	3	6	4	1107	0.33	0.94	Epilepsy; Nicotine addiction
JAGN1	Transporter	183	84	5	4	3	4	3	3	196	0.42	0.92	Neutropenia
SMO	GPCR	787	147	1	25	6	1	7	1	823	1	0.86	Curry-Jones syndrome; Pallister-Hall-like syndrome; Basal cell carcinoma
ABCG8	Transporter	673	126	1	35	11	0	15	3	86	-	-	Gallbladder disease; Sitosterolemia
ABCG5	Transporter	651	127	1	27	9	0	0	0	31	-	-	Sitosterolemia
GPT	Enzyme	408	228	7	5	1	6	2	1	1812	0.5	0.92	Congenital disorder of glycosylation; Myasthenic syndrome
FZD4	GPCR	537	206	5	13	6	2	3	2	1055	0.83	1	Exudative vitreoretinopathy; Retinopathy of prematurity
total				324	283	122	220	104	42		0.64	0.82	

the total set of variants are found in the region covered by these structures (Fig. 1A, additional data at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al). Of the 281,220 variants found in resolved regions, only 2.2% of those (6,119 variants) have been assigned as benign (18.9%) or pathogenic (50.2%) (Table S1).

Computational assessment of stability and evolution shows loss of function due to loss of stability for ca. 62% of disease variants in selected proteins

To examine the importance of changes in protein stability in membrane proteins for causing loss of function and disease, we analysed a smaller set of proteins in more detail. Specifically, we searched for proteins that had at least one pathogenic and one benign variant in the transmembrane region. As our aim was to use the Rosetta software to predict changes in thermodynamic stability, we focused on protein structures that had been determined via X-ray crystallography as Rosetta has been developed and benchmarked most extensively on such structures. These requirements narrow down the set to 16 proteins: six transporters, three ion channels, four GPCRs, two enzymes and one cell junction protein (Table 2 and additional data at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al). These 16 proteins represent different types of membrane proteins with diverse functions, structures, and involvement in different diseases. Of note, the secondary structure of these proteins is mostly α-helical (> 50%), while unstructured regions or extended strands add up to 27% (Fig. S2).

Inspired by previous analyses of soluble proteins, we investigated these membrane proteins in terms of structural stability and sequence conservation. Specifically, we developed and benchmarked a revised Rosetta protocol for stability calculations of membrane proteins (see Methods and Supplementary Material). We used this method to calculate the change in thermodynamic stability (ΔΔG) upon single amino acid substitutions. In each case, we selected a high-resolution structure (Table 2), and removed any co-crystallised molecules. We also constructed multiple sequence alignments (MSA) of each protein and used GEMME (Global Epistatic Model for predicting Mutational Effects) (***Laine et al., 2019***) to estimate the evolutionary effects of the variants. Specifically, we calculated a normalized score (ΔΔE) with $\Delta\Delta E \approx 0$ corresponding to substitutions that—in light of evolution—appear well tolerated, and $\Delta\Delta E \approx 1$ for variants that—based on the evolutionary record—are rare or absent, and expected to cause loss of function. In analyses of soluble proteins, we have previously found that a high value of ΔΔE is a good predictor for a variant to cause loss of

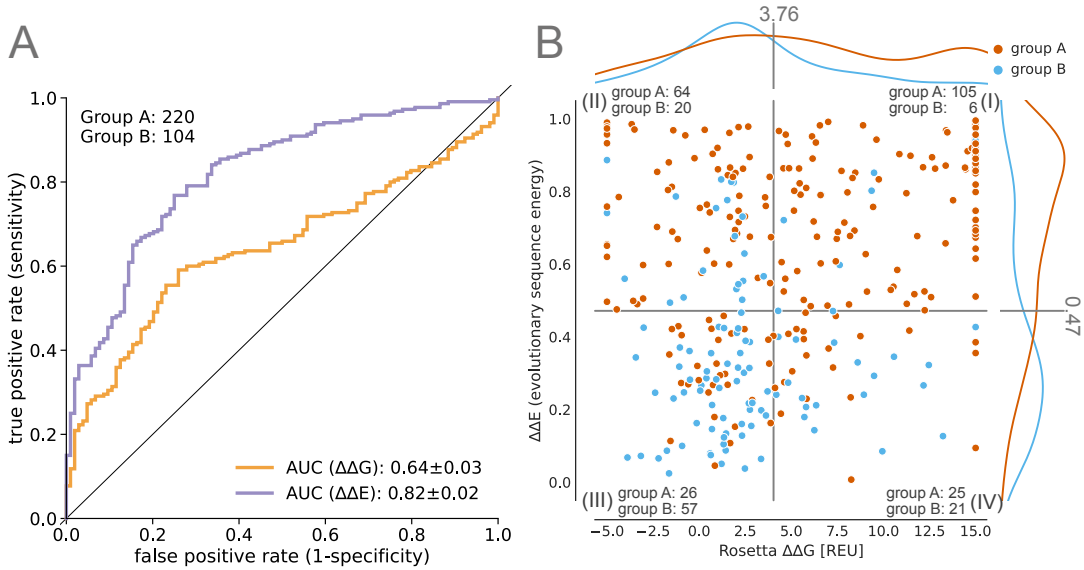

Figure 2. Correlation of ClinVar variants with computational predictors for human membrane proteins. (A) ROC curves of Rosetta ΔΔG and GEMME ΔΔE with variant counts for each group. AUC standard errors were determined by bootstrapping. (B) Distribution of *group A* (pathogenic) and *group B* (benign and/or non-rare) along ΔΔG and ΔΔE landscape and separation into quadrants by optimal cutoff obtained from ROC curves. The counts of each group are shown for the quadrants.

function and that variants with both high ΔΔE and ΔΔG are likely to cause loss of function via loss stability and cellular abundance (***Cagiada et al., 2021***).

We calculated ΔΔG and ΔΔE for all variants that have been observed in humans and where the wild-type residue was resolved using X-ray crystallography. Further, we did not analyze variants at positions where the Rosetta energy function suggested a potential incompatibility between the experimental structure and the Rosetta energy function, and variants at positions with < 50 sequences in the multiple sequence alignment (see Methods, Table 2 and Supporting Data at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al). After this quality control, we retain 220/324 pathogenic and 42/122 benign variants and lose one protein (ABCG8) as it does not have any variants left after filtering. We thus analyzed two sets of variants: A is the set of 220 variants that are described as pathogenic in ClinVar and B is the set of 104 variants that are either assigned as benign in ClinVar and/or non-rare gnomAD variants that, as discussed above, are more likely to be benign than pathogenic as their allele frequencies in gnomAD are $> 9.9 \cdot 10^{-5}$ (Fig. 1 and Fig. S1). In what follows we refer to *group B* as benign, but note that among the 104 variants in *group B* only 42 are classified as benign in ClinVar and the remainder comes from gnomAD.

To quantify how well ΔΔE and ΔΔG distinguish between the two classes of variants, *group A* (pathogenic) and *group B* (benign and/or non-rare), we constructed a ROC curve and calculated the AUC as a measure of how well each of the two scores can predict pathogenicity (Fig. 2A). We find that both scores can separate the *group A* and *group B* variants, though ΔΔE, as expected, performs better than ΔΔG (AUC 0.82 vs. 0.64). This is in line with previous observations for soluble proteins (***Nielsen et al., 2017***; ***Scheller et al., 2019***; ***Stein et al., 2019***; ***Abildgaard et al., 2019***) and the hypothesis that many, but not all, pathogenic variants are destabilized so that ΔΔG calculations can capture these pathogenic variants, but not those caused by other mechanisms for loss of function.

We further analyzed the *group A* (pathogenic) and *group B* (benign and/or non-rare) variants in terms of their ΔΔE and ΔΔG scores (Fig. 2B). To simplify the discussion, we analyze the variants in terms of whether ΔΔE and ΔΔG are low or high, with respect to cutoffs from the ROC analysis (see Methods section). This analysis separates the variants into four quadrants with only a few

variants (14.2%) falling in the quadrant of low ΔΔE and high ΔΔG (Fig. 2B (IV)), which is comparable to previous observations for soluble proteins (***Cagiada et al., 2021***; ***Høie et al., 2022***). The three remaining quadrants correspond roughly to:

(I) Variants that cause loss of function via loss of stability (high ΔΔE and high ΔΔG),
(II) Variants that cause loss of function for other reasons than loss of stability, such as substitutions at key functional sites (high ΔΔE and low ΔΔG),
(III) Variants that are expected to be tolerated both from a structural and evolutionary sequence perspective (low ΔΔE and low ΔΔG).

As expected, we find that most *group B* (benign and/or non-rare) variants have low ΔΔE (75%) and 73% of those also low ΔΔG, whereas most pathogenic variants (*group A*) have large ΔΔE (76.8%, see Fig. 2B). Among the 169 pathogenic variants that have high ΔΔE values, we find that 62.1% also have high ΔΔG values, suggesting that loss of stability plays an important role for disease in the 15 investigated membrane proteins.

Pathogenic variants in GPCRs, especially in the transmembrane region, lose function mostly by loss of stability, while this is less prominent in transporters or other protein classes

Our data set contains several members of the main membrane protein classes, namely five transporters (98 *group A*/pathogenic + 62 *group B* (benign and/or non-rare variants)), three ion channels (11 + 15 variants), four GPCRs (54 + 28 variants) and two enzymes (24 + 13 variants) (Table 2). We examined the results from our computational predictors to probe for class-specific trends. In all four classes, evolutionary conservation predictions (ΔΔE) have a high AUC (>0.8), similar to the analysis with all proteins combined (Table S2). Focusing on the transmembrane region, we find a very high AUC of 0.97 for variant classification of transporters (underlying datapoints: 44 pathogenic *group A* + 10 *group B* variants). Interestingly, we find that for GPCRs loss of stability is the main cause of pathogenicity, as indicated by an increased AUC (0.79) for ΔΔG predictions (Fig. 3A and Table S2) compared to the complete data set with 15 proteins (AUC=0.64, Fig. 2). This is even more prominent for variants located within the transmembrane region (AUC=0.81).

In the transmembrane region of GPCRs, 77.4% of the pathogenic variants have high ΔΔG values (Fig. 3B), suggesting that their pathogenicity is due to loss of stability. When separating the proteins into specific regions, namely by whether they are buried, solvent-accessible, face the membrane/lipids, and are within and outside the transmembrane regions (Fig. 3C), we see that those pathogenic variants that lose function via loss of stability are typically buried (Fig. 3C). In contrast, solvent-accessible pathogenic variants are not found to lose function due to loss of stability, and variants located in those regions are more likely to be tolerated (11.1% of *group A*/pathogenic compared to 44.4% of *group B* variants). Within the transmembrane region, most variants (90% *group A* and 97% for *group B*/benign and/or non-rare) are buried, in contact with other residues, and often facing lipids. Looking at pathogenic variants that lose function due to other reasons than loss of stability (quadrant (II)), variants in GPCRs are more often within the transmembrane (Fig. 3C) compared to all datasets (Fig. 3D), where we see a larger proportion of variants in buried sides in extracellular or intracellular environments.

Next, we focused on individual proteins and examined the location and potential mechanism behind disease variants in one GPCR and one transporter protein. We used the calculated values of ΔΔE and ΔΔG to aid in a structural analysis of the disease variants in rhodopsin (OPSD; Fig. 4A–4C) and a glucose transporter (GTR1; Fig. 4D–4F). We examined the structures of the two proteins to find the residues that interact with ligands or co-factors and searched the literature to find residues that are known to be key to function. We find that many disease variants are located at these residues, suggesting that they directly disrupt function, and some of them also decrease stability. For example, in OPSD we find a number of disease mutants at residues that interact with the retinal co-factor as well as residues in e.g. the so-called ionic lock (***Hofmann et al., 2009***) (Fig. 4A and B).

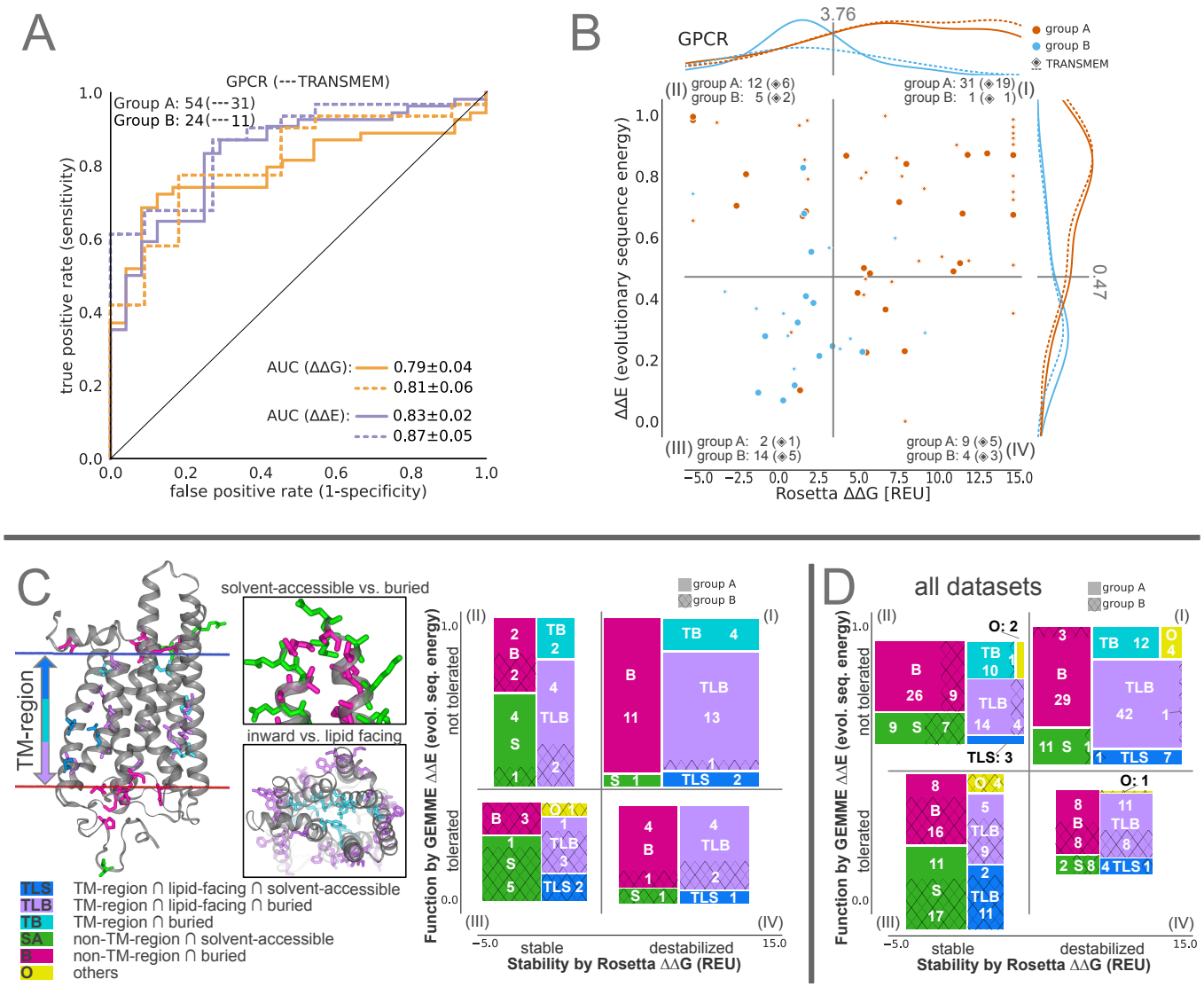

Figure 3. Analysis of GPCR variant classification performance and structural differences of variant effects. (A) ROC curve and (B) distribution of ΔΔG versus ΔΔE plotted for GPCRs for all variants and only for variants within the transmembrane region (dashed line and diamond symbol). (C) [left] Illustration of the different residue categories used in this work on OPSD, namely, whether they are inside (dark blue, turquoise, and violet) or outside of the transmembrane (TM) region (green and pink), and their orientation towards the membrane (lipid-facing: blue and violet), and whether they are solvent-accessible or buried. The structure on the left shows disease-associated variants, while the two inserts on the right illustrate solvent-accessible vs. buried, and inwards vs. lipid-facing, more generally and are not restricted to a disease relationship. Variants labeled with *other* are rare combinations, e.g., residues within the TM region that are solvent accessible but do not face the membrane, and some at the intersection between the membrane and the solvent (like lipid-facing but not placed within the TM region; see note in limitation section). For more details on how the classes were assigned, see Methods. On the right, the variant counts in the four quadrants, separated by their position in the proteins, are shown for *group A* (pathogenic, full) and *group B* (benign and/or non-rare, hashed) variants. (D) Variant counts as seen in (C) are shown summed over all 15 proteins.

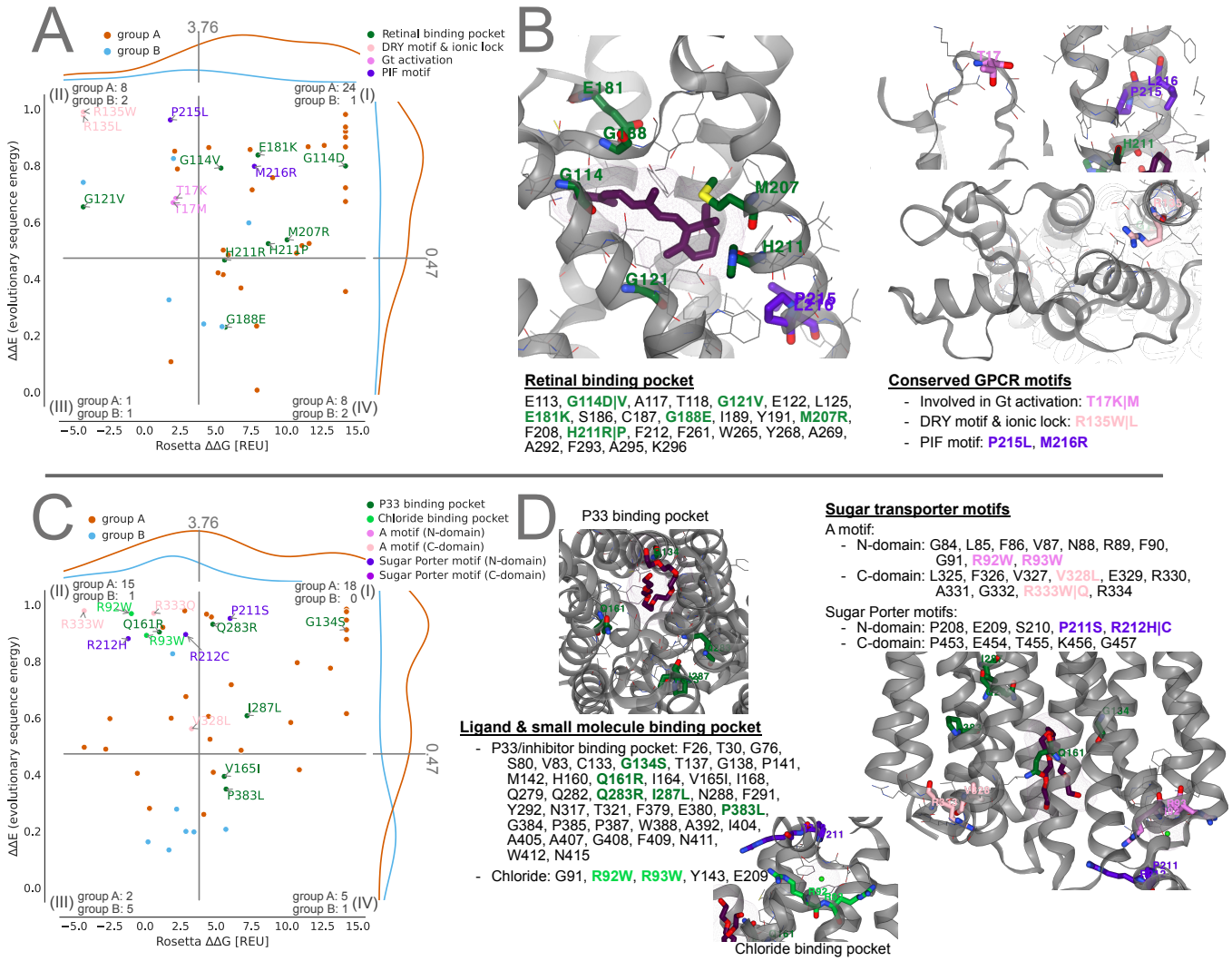

Figure 4. Mapping pathogenic variants to known motifs and binding pockets. A–C show OPSD variants which lie within the retinal binding pocket and conserved GPCR motifs. D–F show GTR1 variants within proximity of a PEG molecule (P33), matching previously identified inhibitor positions (***Custódio et al., 2021***), and the chloride binding side, along with conserved sugar transporter motifs.

Similarly, many disease variants in GTR1 are located at sites known to interact with a chloride ion that is important for function (***Custódio et al., 2021***), the sugar molecule, known inhibitors (***Kapoor et al., 2016***), or residues known to affect transport (***Custódio et al., 2021***) (Fig. 4C and D).

Looking across the two proteins (Fig. 4), most of the high-ΔΔE, low-ΔΔG disease variants are found at residues that have known functional roles. We expect such variants in quadrant (II) to lose function due to other reasons than stability (***Cagiada et al., 2021***). This also includes variation in residues in close proximity to ligands and interaction partners, which were not included in our stability calculations. Further, we find that many of the disease variants that are not located at known functional sites have both high values of ΔΔE and ΔΔG, suggesting that these variants instead disrupt the stability of the folded state.

Correlating physicochemical changes with variant effects

We examined the dataset containing all 15 proteins and the amino acid properties within the four quadrants, where quadrant (I) contains destabilized and quadrant (II) stable variants while both quadrants (I) (II) are—in light of evolution—not tolerated. Quadrants (III) and (IV) are evolutionary tolerated, but quadrant (III) contains stable and quadrant (IV) destabilized variants (see Fig. 5A and B, and methods for a more detailed quadrant definition). Across all quadrants, hydrophobic amino acids are most commonly observed (wild-type, 33%; target, 37%), which can be explained by the general preference for hydrophobic residues in membrane proteins, especially within the TM region (***von Heijne, 2007***). Almost 65% of the *group B* (benign and/or non-rare) variants located in quadrant (III) have, as expected, the same amino acid property for wild-type and target (35.1% remain hydrophobic, 21.1% charged, 8.8% polar). For the pathogenic variants that lose function due to loss of stability (quadrant (I)), we see greater changes in physicochemical properties among those substitutions (Fig. 5A). Interestingly, in quadrant (II), where variants lose function due to other reasons than stability, we see mainly hydrophobic target amino acid types (54.8%, with one third coming from charged to hydrophobic substitutions).

Inspired by the enrichment of pathogenic variants in the transmembrane region (Fig. 1E) that is enriched with hydrophobic residues (***von Heijne, 2007***), we analyze substitutions by physicochemical properties. Specifically, we calculated the median score for ΔΔG and ΔΔE for each combination of wild-type and target amino acids and arranged the amino acids by hydrophobicity (***Zhao and London, 2006***) (Fig. 5B). For variants where the target residue is more hydrophobic (e.g. Arg to Leu, Arg to Trp, or Asp to Tyr variants), we indeed see a different pattern when looking at the median stability values (ΔΔG) compared to the median ΔΔE values. These variants appear to be tolerated by protein stability, but not by evolution (Fig. 5B, dashed upper rectangle). In contrast, variants changing the residue to be less hydrophobic are indicated as not tolerated by evolution and destabilising (Fig. 5B, solid lower rectangle).

Conclusions

Here, we present an analysis of missense variants and their properties within human membrane proteins. We identified $1.9 \cdot 10^6$ unique variants in 5,471 proteins of which 99.1% are of uncertain significance, and only 19,089 have been classified as pathogenic or benign. Additionally, we see an almost 2-fold excess of pathogenic variants compared to benign variants in the transmembrane regions, which make up only 16.1% of the proteins.

We have examined the importance of changes in membrane protein stability for causing loss of function. We analysed 15 proteins and calculated the change in thermodynamic stability (ΔΔG) and evolutionary conservation (ΔΔE). Our ROC analysis shows good performance in separating benign from pathogenic variants by their sequence conservation (AUC 0.82 for ΔΔE), and we find that ca. 62% of the pathogenic variants cause loss of function via loss of stability, in line with previous findings on soluble proteins. In the 15 selected proteins, we observe that most variants have a hydrophobic wild type (33%) or target (37%) amino acid type and that almost 65% of benign and/or non-rare variants that are likely tolerated as assessed by both ΔΔE and ΔΔG do not change their

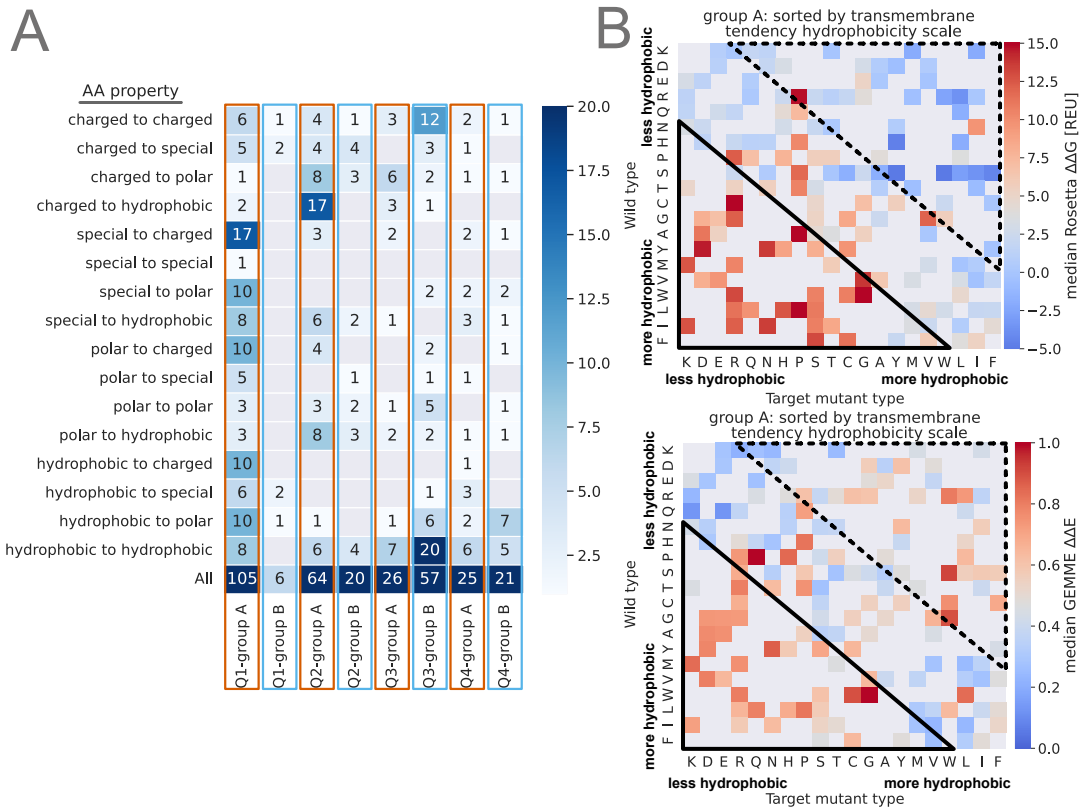

Figure 5. Variant effects depend on the chemical nature of the wild-type and target amino acids. (A) Substitution amino acid property counts for wild type and variant per quadrants and groups. Quadrant (I) contains destabilized and evolutionary not tolerated variants; quadrant (II) stable and not tolerated; quadrant (III) stable and tolerated and quadrant (IV) destabilized and evolutionary tolerated variants (Methods). (B) Heatmap of target versus wild-type amino acid, colored by their median values ΔΔG (top) and ΔΔE (bottom) for pathogenic variants sorted by the hydrophobicity 'transmembrane tendency' scale by ***Zhao and London*** (***2006***). The rectangles show the enrichment of variants with high (straight line) and low (dashed line) ΔΔG or ΔΔE. Median values of ΔΔG (destabilized) and ΔΔE (not tolerated) are red, while low values are blue; missing data are colored gray.

amino acid type. Among pathogenic variants that lose function due to loss of stability, substitutions to charged, polar or hydrophobic are more prominent, while we observe substitutions from more hydrophobic to less hydrophobic residues in variants that lose function due to other reasons than stability.

When analyzing the different classes of membrane proteins, we observe for transporter proteins that pathogenic variants in the transmembrane region have an AUC of 0.97 for $\Delta\Delta E$, and loss of stability does not appear to be the predominant factor in loss of function for the cell junction protein we examined. In contrast, pathogenic variants in GPCRs lose function mainly via loss of stability (AUC of 0.79 for $\Delta\Delta G$). We therefore suggest that pathogenic variants lose function via loss of stability more often in the transmembrane region of GPCRs than in the other protein classes we examined.

From a more detailed inspection of individual proteins, we found that most of the high-$\Delta\Delta E$, low-$\Delta\Delta G$ disease variants are located at positions that have known functional roles, while many of the disease variants that are not located at functional sites have both high values of $\Delta\Delta E$ and $\Delta\Delta G$, suggesting that these variants instead disrupt the stability of the folded state.

Our observations underline the importance of stability and the loss thereof in disease-causing variants of membrane proteins and thereby show how computational tools can aid in interpreting molecular mechanisms that underlie disease. Such functional understanding may help address the substantial challenge of classifying VUS (***Anderson et al., 2022***). Given the limited number of variants and proteins within this study, utilizing recent advantages like the large excess of experimental structures derived from electron microscopy or computational models from e.g. AlphaFold (***Jumper et al., 2021***) could enable a broader analysis. We include the collection of population and ClinVar variants for the entire human proteome to facilitate such studies on membrane proteins and beyond.

Limitations of this study

We note several limitations that should be considered when interpreting the results. Our general observations and conclusions on membrane proteins and their classes are limited by the available data and proteins, partly due to our choice to only analyze experimental structures with annotated pathogenic and benign variants.

Several membrane proteins, for example channels and cell junction proteins, function as (homo-) oligomers. In this study, we used structures of the individual proteins for our stability calculations and thereby may miss destabilizing variants in interfaces. Those variants are more difficult to interpret using stability calculations due to the lack of contacts that are affected by stability. In addition to missing interactions, conformational changes of the structure or different conformations might alter $\Delta\Delta G$ values, and several stabilizing variants can be explained due to missing interaction partners in these structures (e.g. R135W, R135L, and G121V in OPSD are missing either the ligand or an intracellular binding partner).

Regarding the accuracy of the $\Delta\Delta G$ calculations, our membrane protein dataset is mainly alpha-helical, which is also true for most human membrane proteins; however, the stability score function was parameterized and benchmarked on bacterial proteins, which are often beta barrels and might fold differently compared to their helical counterparts. When we compared experimental and computational $\Delta\Delta G$ values, we obtained a Spearman rank correlation coefficient of 0.46, leaving uncertainty about the predictability of the extent of loss of stability. It is worth noting that the correlation between two sets of experimental $\Delta\Delta G$ measurements in GlpG (***Paslawski et al., 2015***; ***Baker and Urban, 2012***) shows a Spearman correlation of 0.65, and when correlating all experimental datasets with at least 12 overlapping variants we obtain a mean Spearman correlation of 0.6. Preferences for specific amino acid properties in certain environments such as the membrane might be biased by their values within the respective scoring function. Our results also depend on how different protein regions are defined.

Our calculations of ΔΔE scores depend on the MSA, and we note that using a different MSA, e.g., by changing the E-value cutoff, could shift some of the ΔΔE values from tolerated to not tolerated.

Finally and as already discussed above, we combine benign and/or non-rare gnomAD variants into group B. This should be taken into account when interpreting the results and especially when investigating outliers of group B, as those could be variants of unknown significance.

Methods and Materials

Collection and processing of clinical, population and structural data

To extract all annotated human membrane proteins, we first obtained all unique proteins (UniProt-ID) of the human proteome (*organism=homo sapiens*) from the UniProt (https://www.uniprot.org/help/api) (***The UniProt Consortium, 2021***) and EMBL-EBI (https://www.ebi.ac.uk/proteins/api/doc/) database. For each UniProt-ID, we then stored its general and amino acid-based annotations (such as protein domain regions) in UniProt and further selected proteins of the type 'TRANSMEM', 'INTRAMEM', 'TOPO_DOM', or 'LIPID'. This annotation originates from assignment of structural properties or predictions by TMHMM (https://www.uniprot.org/help/topo_dom). The UniProt-ID of the first transcript is used in the further mapping and analysis.

We then further filtered the UniProt-ID list so that all remaining proteins have at least one ClinVar (***Landrum et al., 2018***) or gnomAD (***Karczewski et al., 2020***) missense variant. gnomAD data were taken from an in-house database built on exome data from gnomAD v2 and whole-genome data from gnomAD v3 (scripts available at https://github.com/KULL-Centre/PRISM/tree/main/software/make_prism_files, release-tag v0.1.1). The database was generated by first downloading the vcf files (May 2021), selecting exome GRCh38 liftover for v2 and whole-genome files for v3. The vcf files were then annotated with Variant Effect Predictor (VEP) with the GRCh38 release 100 human transcripts set from Ensembl. From the annotated vcfs we established for all protein-level variants, separately in exome and genome data, allele frequencies from the variant allele count as the sum of all DNA variants leading to the same protein-level variant. Clinvar data were obtained by parsing the following file: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz (May 2021) and only admitting entries that have a rating of at least one star, are single nucleotide variants and mapped to GRCh38 (script available at https://github.com/KULL-Centre/PRISM/tree/main/software/make_prism_files, release-tag v0.1.1). The merged dataset for the entire human proteome is provided at https://sid.erda.dk/sharelink/c3rDfqR8nn, using a UniProt-AC-based directory structure, e.g., files for GTR1 (UniProt AC P11166) can be found in subdirectory `prism/P1/11/66/`.

Next, we extracted all PDB-IDs from RCSB and PDBe with a matching UniProt-ID reference. As not every PDB-ID for a given UniProt-ID in PDBe could be found in RCSB PDB and vice versa, we searched both databases. We further included phenotype and genomic disease annotations from OMIM via mim2gene (https://omim.org/static/omim/data/mim2gene.txt) and MIM, including the proteins' chromosome information.

The sequences were then aligned to the UniProt sequence using *pairwise2.align.globalds* (with BLAST defaults) from Biopython (***Cock et al., 2009***), a minimal identity of 0.6, and minimal coverage of 0.1 for alignment acceptance. All residues that do not match the UniProt sequence were discarded. The final data contained each protein sequence, its UniProt-ID, the secondary structure prediction by residue, solvent-accessible surface area for each wild-type residue, and UniProt annotations such as transmembrane region, protein modifications, total allele frequency counts from gnomAD, ClinVar significance statements, genomic disease annotations, and associated PDB-IDs.

Selection of targets used for computational predictions

To find a set of proteins for our computational sequence and structure analyses, we selected all proteins that have at least one benign and one pathogenic ClinVar annotation in an experimentally resolved transmembrane region of the protein. This reduced the number of proteins with gnomAD or ClinVar annotations from 1,504 proteins to 41 proteins. As the Rosetta membrane energy func-

tion has been developed and benchmarked on structures resolved by X-ray crystallography, we selected those, reducing the protein set to 16. The selected proteins are listed in table 2 and Supporting Data at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al.

Conservation analysis of variant effects

To calculate the effect of a variant in light of evolution, we used the GEMME algorithm (***Laine et al., 2019***) as previously described (***Høie et al., 2022***): We first construct a multiple sequence alignment (MSA) using the sequence of the first transcript of each proteins UniProt-ID as input to HHBlits (version 2.0.15) (***Remmert et al., 2011***) with the following settings -e 1e-10 -i 1 -p 40 -b 1 -B 20000 to search UniRef30_2020_03_hhsuite.tar.gz (***Mirdita et al., 2017***; ***UniProt Consortium, 2019***; ***Ruan et al., 2019***). We then further follow the GEMME algorithm that predicts the degree of conservation for all 19 substitutions ($\Delta\Delta$E). We rank-normalized the $\Delta\Delta$E values for the entire protein to allow comparison with the other proteins in the dataset. $\Delta\Delta E \approx 0$ corresponds to well-tolerated substitutions, whereas $\Delta\Delta E \approx 1$ corresponds to rare or absent variants. Additionally, we extract the sequence coverage of the MSA for each position.

Thermodynamic stability predictions

To calculate changes in thermodynamic stability ($\Delta\Delta$G), we use Rosetta version v2021.31-dev61729-0-gc7009b3115c (GitHub sha1 c7009b3115c22daa9efe2805d9d1ebba08426a54). We implemented an in-house pipeline to perform preparation, relaxation, and $\Delta\Delta$G calculations of the protein (https://github.com/KULL-Centre/PRISM/tree/main/software/rosetta_ddG_pipeline, release-tag v0.1.1). Preparation includes cleaning of the PDB structure coordinates (hereafter referred to as structure) of ligands and alternative rotamers and chains, superposing of the protein into the membrane plane as well as calculation of the membrane plane, lipid-accessible residues (***Koehler Leman et al., 2017a***) and the solvent-accessible surface area using DSSP (***Kabsch and Sander, 1983***; ***Touw et al., 2015***) (the latter is solely used for analysis purposes).

Rosetta needs membrane proteins to be centered in the internal coordinate system to calculate the membrane plane. Therefore, protein coordinate translation was performed by structural superposition of the protein to its equivalent structure obtained from the *Orientations of Proteins in Membranes (OPM) database* (***Lomize et al., 2012***), which lies already within those coordinates. Next, the membrane plane was calculated using Rosetta (***Alford et al., 2015***; ***Koehler Leman et al., 2017b***) and the protein structure was relaxed as described in ***Koehler Leman et al.*** (***2021***). Finally, $\Delta\Delta$G values for each variant were calculated as the residual energy of the variant minus the energy of the wild type.

We performed a benchmark to identify the best protocol to calculate $\Delta\Delta$Gs for membrane proteins. First, we collected 20 experimentally derived $\Delta\Delta$G datasets of in total 8 different membrane proteins (Table S3). Then, we implemented three different protocols, namely MP_repack, MP_flex_relax_ddG, and 'cart_prot', inspired by work on soluble proteins (***Park et al., 2016***) and previously published work on membrane proteins (***Alford et al., 2021***). MP_repack operates in *torsion space* and performs a simple repacking of the side chains within a defined radius after mutagenesis (following the protocol mentioned in ***Alford et al.*** (***2021***)). MP_flex_relax_ddG is inspired by ***Park et al.*** (***2016***) and allows more flexibility to accommodate the variant by allowing backbone relaxation of the variant and its sequential neighbors, in addition to repacking of side chains within a defined radius. cart_prot follows the same protocol as MP_flex_relax_ddG but is executed in *cartesian space*. For all protocols, we used the membrane protein score function 'franklin2019' (***Alford et al., 2021***) that performs comparable to older membrane scoring functions as recently evaluated in ***Koehler Leman et al.*** (***2021***). Finally, we selected cart_prot as the computed values gave the best correlation with the experimental data (0.46), and additionally, the computed values have a high reproducibility, indicated by the low standard deviation for replicates (Fig. S3). As mentioned in the limitations section, the correlation of independent experimental studies on the same protein and the same variants (***Baker and Urban, 2012***; ***Paslawski et al., 2015***) is 0.65.

Enrichment of benign variant counts by gnomAD allele frequency

To evaluate the value of our computational methods to predict variants to be benign or pathogenic using ROC analysis, we aimed for a large number of benign and pathogenic variants. In our target proteins, we have 324 pathogenic but only 122 benign variants. We aimed to supplement benign variants with variants from gnomAD. Therefore, we performed a ROC analysis of the gnomAD allele frequency on the 10,260 benign and 2,360 pathogenic ClinVar variants in the human membrane proteome that also have a gnomAD allele frequency (Fig. 1C) and obtained an AUC of 0.96 (Fig. S1). This analysis enables calculating a cutoff to separate benign from pathogenic variants using gnomAD allele frequency via the highest Youden Index (***Krzanowski and Hand, 2009***)). We thereby obtain a cutoff of $9.9 \cdot 10^{-5}$ (Fig. 1B). We define *group B* variants as the union of those that are defined by ClinVar as benign and those variants that have an allele frequency $> 9.9 \cdot 10^{-5}$ and are not pathogenic (in ClinVar). Consequently, we call pathogenic variants *group A*. This results for our target proteins in 324 *group A* and 283 *group B* variants across 16 proteins.

Filtering criteria for variant analysis of the 16 target proteins

Prior to analysis, we defined filtering criteria for the calculated ΔΔG and ΔΔE variants to obtain a set of variants with reliable scores. First, only variants for which both ΔΔG and ΔΔE calculations are available were selected. Second, *special* residues involved in disulfide bonds or known modified residues (such as those which bind covalent ligands or palmitoylated residues) were excluded. Further, variants with a low MSA sequence coverage of fewer than 50 sequences were excluded. Last, variants that have a positive wild-type Rosetta energy (E_{res}) are excluded from further analyses as those residue conformers are likely to favor any substitution in order to reduce its energy, likely due to limitations of the Rosetta energy function. By applying all filters, we obtain a final set of 15 proteins with 220 pathogenic (*group A*) and 104 *group B* variants, of which 42 are benign (see Table 2 and for the single filtering steps additional data at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al).

Definition of ΔΔG and ΔΔE thresholds and quadrant classification

To analyze the variants in terms of their ΔΔE and ΔΔG scores, we defined cutoffs for each method based on the optimal ROC value (tradeoff of high specificity versus high sensitivity) to separate *group A* and *group B* variants, in a similar fashion as done for gnomAD allele frequency above. Next, we defined four categories dependent on their position along the ΔΔE and ΔΔG axes:

(I) *Quadrant (I)* variants have high ΔΔE and ΔΔG and are likely to cause loss of function via loss of stability;
(II) *Quadrant (II)* variants have high ΔΔE and low ΔΔG and cause loss of function for other reasons than loss of stability;
(III) *Quadrant (III)* variants have low ΔΔE and low ΔΔG and are from a structural and evolutionary perspective expected to be tolerated;
(IV) *Quadrant (IV)* variants have low ΔΔE and high ΔΔG and are from an evolutionary perspective expected to be tolerated but from a structural perspective expected to cause instability;

Definition of protein regions

For analysis purposes, we assigned residues into different regions, based on their solvent accessibility, their positioning within the membrane (TM region), and their orientation toward the lipids. Relative solvent-accessibility was calculated using DSSP (***Kabsch and Sander, 1983***; ***Touw et al., 2015***) with a cutoff of 0.3, placing residues with a smaller value into the category of *buried* residues. The positioning within the membrane was obtained as described above. To assess whether a residue faces toward the lipids or not, we used a dedicated Rosetta function that returns a true or false value (***Koehler Leman et al., 2017a***). We can thereby divide the protein into six regions (with overlaps, see Fig. S4):

- *Buried*: residues with a DSSP < 0.3 and that are placed outside the membrane; this cutoff places residues within contacts as buried although they might be close to the surface of the protein
- *Solvent-accessible*: residues with a DSSP ≥ 0.3 and that are placed outside of the membrane
- *TM-region* ⋂ *buried*: buried residues that are placed within the membrane
- *TM-region* ⋂ *lipid-facing* ⋂ *buried*: buried residues within the membrane that are oriented towards the lipids
- *TM-region* ⋂ *lipid-facing* ⋂ *solvent-accessible*: solvent-accessible residues within the membrane that are oriented towards the lipids
- *Others*: combination of residues that are rare and few in number, such as *TM-region* ⋂ *solvent-accessible*, *lipid-facing* ⋂ *solvent-accessible* or *lipid-facing* ⋂ *buried*

Utilized software

- python3, incl. following third party libraries
  - adjustText
  - Biopython
  - circlify
  - matplotlib
  - numpy
  - pandas
  - seaborn
  - scipy
  - sklearn
  - squarify
  - xmltodict
  - nglview
- Rosetta version 2021.31+HEAD.c7009b3115c (c7009b3115c22daa9efe2805d9d1ebba08426a54, default.linuxgccrelease mode)
  - mp_span_from_pdb (***Alford et al., 2015***)
  - rosetta_scripts (***Fleishman et al., 2011***)
  - FastRelax (***Khatib et al., 2011***; ***Maguire et al., 2021***)
  - MembraneMover (***Alford et al., 2015***)
  - cartesian_ddg (***Kellogg et al., 2011***; ***Park et al., 2016***)
  - energy function: franklin2019 (***Alford et al., 2021***) + cart_bonded=0.5 + fa_water_to_bilayer=1.5
- Scripts available at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al
  - hMP statistics (hMP_stats.ipynb)
  - ΔΔG pipeline benchmark (MP_ddG_benchmark.ipynb)
  - Xray subset calculations (Xray_subset-calc.ipynb)
  - Xray subset analysis (Xray_subset-ana.ipynb)
- Pipelines
  - PRISM_tools (https://github.com/KULL-Centre/PRISM/software/rosetta_ddG_pipeline, release version v0.1.1)
  - PrismData, FillVariants and struc_select_sifts (https://github.com/KULL-Centre/PRISM/software/sc release version v0.2.2)
- Others
  - overleaf.com
  - Affinity Designer

Data availability

All scripts are available at https://github.com/KULL-Centre/papers/tree/main/2022/hMP-Xray-Tiemann-et-al and as stated in the 'Utilized software' section.

All parsed data from Uniprot, ClinVar, and gnomAD for the human proteome are available at https://sid.erda.dk/sharelink/c3rDfqR8nn. All parsed data from wwPDB for the human membrane proteome are available at https://sid.erda.dk/sharelink/AtoGToVaZ8. All data related to the human membrane proteome and our X-ray subset are available at https://sid.erda.dk/sharelink/ds37GLjR8U. All data related to the MP stability benchmark are available at https://sid.erda.dk/sharelink/foHTKP49jC.

Acknowledgments

We thank Matteo Cagiada for providing an automatic pipeline for GEMME calculations and Kristoffer Enøe Johansson for resourcing us with his alignment and merging implementations. Additional thanks go to Julia Koehler Leman for helpful discussion regarding membrane protein implementations in Rosetta. This study was funded by the Protein Interactions and Stability in Medicine and Genomics (PRISM) centre funded by the Novo Nordisk Foundation (NNF18OC0033950, to AS and KL-L) and a grant from the Lundbeck Foundation (R272-2017-4528, to AS). We acknowledge access to resources from the Department of Biology's core facility for biocomputing.

Author contribution

KL-L and AS conceived the original idea and supervised the project. JKST retrieved and processed all data with contribution by HZ who extracted and processed the ClinVar and gnomAD data. JKST and AS designed the Rosetta pipeline framework and JKST implemented and benchmarked it with KL-L and AS. JKST performed all calculations and processing of the data. JKST analyzed the data and interpreted the results with KL-L and AS. JKST, KL-L and AS wrote the manuscript with input from HZ.

Competing interests

The authors declare no competing interests.

References

Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, Hoffmann ER, Bernstein I, Gerdes AM, Takahashi M, Ishioka C, Lindorff-Larsen K, Hartmann-Petersen R. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. eLife. 2019 Nov; 8:e49138. doi: 10.7554/eLife.49138.

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, Shenoy A, Zhu W, Kundrotas P, Serra VR, Rodrigues CHM, Dunham AS, Burke D, Borkakoti N, Velankar S, Frost A, et al. A structural biology community assessment of AlphaFold 2 applications. bioRxiv; 2021.

del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife. 2022 Mar; 11:e75751. https://doi.org/10.7554/eLife.75751, doi: 10.7554/eLife.75751.

Alford RF, Fleming PJ, Fleming KG, Gray JJ. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. Biophysical Journal. 2021 Oct; 120(20):4635. https://www.sciencedirect.com/science/article/pii/S0006349521007530, doi: 10.1016/j.bpj.2021.09.019.

Alford RF, Leman JK, Weitzner BD, Duran AM, Tilley DC, Elazar A, Gray JJ. An Integrated Framework Advancing Membrane Protein Modeling and Design. PLOS Computational Biology. 2015 Sep; 11(9):e1004398. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004398, doi: 10.1371/journal.pcbi.1004398.

Anderson CL, Munawar S, Reilly L, Kamp TJ, January CT, Delisle BP, Eckhardt LL. How Functional Genomics Can Keep Pace With VUS Identification. Frontiers in Cardiovascular Medicine. 2022; 9. https://www.frontiersin.org/articles/10.3389/fcvm.2022.900431.

Baker RP, Urban S. Architectural and thermodynamic principles underlying intramembrane protease function. Nature Chemical Biology. 2012 Sep; 8(9):759–768. https://www.nature.com/articles/nchembio.1021, doi: 10.1038/nchembio.1021.

Booth PJ, Clarke J. Membrane protein folding makes the transition. Proceedings of the National Academy of Sciences. 2010 Mar; 107(9):3947–3948. https://www.pnas.org/content/107/9/3947, doi: 10.1073/pnas.0914478107.

Cagiada M, Johansson KE, Valanciute A, Nielsen SV, Hartmann-Petersen R, Yang JJ, Fowler DM, Stein A, Lindorff-Larsen K. Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. Molecular Biology and Evolution. 2021 Jul; 38(8):3235–3246. doi: 10.1093/molbev/msab095.

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

Clausen L, Stein A, Grønbæk-Thygesen M, Nygaard L, Søltoft CL, Nielsen SV, Lisby M, Ravid T, Lindorff-Larsen K, Hartmann-Petersen R. Folliculin variants linked to Birt-Hogg-Dubé syndrome are targeted for proteasomal degradation. PLoS genetics. 2020 Nov; 16(11):e1009187. doi: 10.1371/journal.pgen.1009187.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25(11):1422–1423.

Cournia Z, Allen TW, Andricioaei I, Antonny B, Baum D, Brannigan G, Buchete NV, Deckman JT, Delemotte L, del Val C, Friedman R, Gkeka P, Hege HC, Hénin J, Kasimova MA, Kolocouris A, Klein ML, Khalid S, Lemieux MJ, Lindow N, et al. Membrane Protein Structure, Function and Dynamics: A Perspective from Experiments and Theory. The Journal of membrane biology. 2015 Aug; 248(4):611–640. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4515176/, doi: 10.1007/s00232-015-9802-0.

Custódio TF, Paulsen PA, Frain KM, Pedersen BP. Structural comparison of GLUT1 to GLUT3 reveal transport regulation mechanism in sugar porter family. Life Science Alliance. 2021 Apr; 4(4). https://www.life-science-alliance.org/content/4/4/e202000858, doi: 10.26508/lsa.202000858.

Feinauer C, Weigt M. Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease. bioRxiv; 2017.

Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, Meiler J, Baker D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLOS ONE. 2011 Jun; 6(6):e20161. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020161, doi: 10.1371/journal.pone.0020161, publisher: Public Library of Science.

Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021 Nov; 599(7883):91–95. doi: 10.1038/s41586-021-04043-8.

Frenz B, Lewis SM, King I, DiMaio F, Park H, Song Y. Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. Frontiers in Bioengineering and Biotechnology. 2020; 8:558247. doi: 10.3389/fbioe.2020.558247.

Gaffney KA, Hong H. The rhomboid protease GlpG has weak interaction energies in its active site hydrogen bond network. Journal of General Physiology. 2018 Nov; 151(3):282–291. https://doi.org/10.1085/jgp.201812047, doi: 10.1085/jgp.201812047.

Gersing SK, Wang Y, Grønbæk-Thygesen M, Kampmeyer C, Clausen L, Willemoës M, Andréasson C, Stein A, Lindorff-Larsen K, Hartmann-Petersen R. Mapping the degradation pathway of a disease-linked aspartoacylase variant. PLoS genetics. 2021 Apr; 17(4):e1009539. doi: 10.1371/journal.pgen.1009539.

Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of Molecular Biology. 2002 Jul; 320(2):369–387. doi: 10.1016/S0022-2836(02)00442-4.

Guo R, Gaffney K, Yang Z, Kim M, Sungsuwan S, Huang X, Hubbell WL, Hong H. Steric trapping reveals a cooperativity network in the intramembrane protease GlpG. Nature Chemical Biology. 2016 May; 12(5):353–360. https://www.nature.com/articles/nchembio.2048, doi: 10.1038/nchembio.2048.

Hamel C. Retinitis pigmentosa. Orphanet Journal of Rare Diseases. 2006 Oct; 1:40. doi: 10.1186/1750-1172-1-40.

Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. Nature Reviews Drug Discovery. 2017 Dec; 16(12):829–842. https://www.nature.com/articles/nrd.2017.178, doi: 10.1038/nrd.2017.178.

von Heijne G. The membrane protein universe: what's out there and why bother? Journal of Internal Medicine. 2007 Jun; 261(6):543–557. doi: 10.1111/j.1365-2796.2007.01792.x.

Hofmann KP, Scheerer P, Hildebrand PW, Choe HW, Park JH, Heck M, Ernst OP. A G protein-coupled receptor at work: the rhodopsin model. Trends in Biochemical Sciences. 2009 Nov; 34(11):540–552. doi: 10.1016/j.tibs.2009.07.005.

Hong H. Role of Lipids in Folding, Misfolding and Function of Integral Membrane Proteins. In: Gursky O, editor. *Lipids in Protein Misfolding* Advances in Experimental Medicine and Biology, Cham: Springer International Publishing; 2015.p. 1–31. https://doi.org/10.1007/978-3-319-17344-3_1, doi: 10.1007/978-3-319-17344-3_1.

Hong H, Park S, Flores Jiménez RH, Rinehart D, Tamm LK. Role of Aromatic Side Chains in the Folding and Thermodynamic Stability of Integral Membrane Proteins. Journal of the American Chemical Society. 2007 Jul; 129(26):8320–8327. https://doi.org/10.1021/ja068849o, doi: 10.1021/ja068849o.

Hong H, Szabo G, Tamm LK. Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening. Nature Chemical Biology. 2006 Nov; 2(11):627–635. https://www.nature.com/articles/nchembio827, doi: 10.1038/nchembio827.

Huysmans GHM, Baldwin SA, Brockwell DJ, Radford SE. The transition state for folding of an outer membrane protein. Proceedings of the National Academy of Sciences. 2010 Mar; 107(9):4099–4104. https://www.pnas.org/content/107/9/4099, doi: 10.1073/pnas.0911904107.

Høie MH, Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. Cell Reports. 2022 Jan; 38(2):110207. doi: 10.1016/j.celrep.2021.110207.

Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Chapter 5 - Classifying disease-associated variants using measures of protein activity and stability. In: Pey AL, editor. *Protein Homeostasis Diseases* Academic Press; 2020.p. 91–107. https://www.sciencedirect.com/science/article/pii/B9780128191323000051, doi: 10.1016/B978-0-12-819132-3.00005-1.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug; 596(7873):583–589. https://www.nature.com/articles/s41586-021-03819-2, doi: 10.1038/s41586-021-03819-2, number: 7873 Publisher: Nature Publishing Group.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec; 22(12):2577–2637. doi: 10.1002/bip.360221211.

Kampmeyer C, Nielsen SV, Clausen L, Stein A, Gerdes AM, Lindorff-Larsen K, Hartmann-Petersen R. Blocking protein quality control to counter hereditary cancers. Genes, Chromosomes & Cancer. 2017 Dec; 56(12):823–831. doi: 10.1002/gcc.22487.

Kapoor K, Finer-Moore JS, Pedersen BP, Caboni L, Waight A, Hillig RC, Bringmann P, Heisler I, Müller T, Siebeneicher H, Stroud RM. Mechanism of inhibition of human glucose transporter GLUT1 is conserved between cytochalasin B and phenylalanine amides. Proceedings of the National Academy of Sciences of the United States of America. 2016 Apr; 113(17):4711–4716. https://europepmc.org/articles/PMC4855560, doi: 10.1073/pnas.1603735113.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May; 581(7809):434–443. https://www.nature.com/articles/s41586-020-2308-7, doi: 10.1038/s41586-020-2308-7.

Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011 Mar; 79(3):830–838.

Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popović Z, Baker D, Players F. Algorithm discovery by protein folding game players. Proceedings of the National Academy of Sciences. 2011 Nov; 108(47):18949–18953. https://www.pnas.org/doi/full/10.1073/pnas.1115898108, doi: 10.1073/pnas.1115898108, publisher: Proceedings of the National Academy of Sciences.

Koehler Leman J, Lyskov S, Bonneau R. Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP. BMC Bioinformatics. 2017 Feb; 18(1):115. https://doi.org/10.1186/s12859-017-1541-z, doi: 10.1186/s12859-017-1541-z.

Koehler Leman J, Lyskov S, Lewis SM, Adolf-Bryfogle J, Alford RF, Barlow K, Ben-Aharon Z, Farrell D, Fell J, Hansen WA, Harmalkar A, Jeliazkov J, Kuenze G, Krys JD, Ljubetič A, Loshbaugh AL, Maguire J, Moretti R, Mulligan VK, Nance ML, et al. Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks. Nature Communications. 2021 Nov; 12(1):6947. https://www.nature.com/articles/s41467-021-27222-7, doi: 10.1038/s41467-021-27222-7.

Koehler Leman J, Mueller BK, Gray JJ. Expanding the toolkit for membrane protein modeling in Rosetta. Bioinformatics. 2017 Mar; 33(5):754–756. https://doi.org/10.1093/bioinformatics/btw716, doi: 10.1093/bioinformatics/btw716.

Koepsell H. Glucose transporters in brain in health and disease. Pflugers Archiv: European Journal of Physiology. 2020 Sep; 472(9):1299–1343. doi: 10.1007/s00424-020-02441-x.

Krzanowski WJ, Hand DJ. ROC Curves for Continuous Data. New York: Chapman and Hall/CRC; 2009. doi: 10.1201/9781439800225.

Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. Molecular Biology and Evolution. 2019 Aug; p. msz179. doi: 10.1093/molbev/msz179.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018 Jan; 46(D1):D1062–D1067. doi: 10.1093/nar/gkx1153.

Lee E, Manoil C. Mutations eliminating the protein export function of a membrane-spanning sequence. Journal of Biological Chemistry. 1994 Nov; 269(46):28822–28828. https://www.sciencedirect.com/science/article/pii/S0021925819619800, doi: 10.1016/S0021-9258(19)61980-0.

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Research. 2012 Jan; 40(Database issue):D370–376. doi: 10.1093/nar/gkr703.

Maguire JB, Haddox HK, Strickland D, Halabiya SF, Coventry B, Griffin JR, Pulavarti SVSRK, Cummins M, Thieker DF, Klavins E, Szyperski T, DiMaio F, Baker D, Kuhlman B. Perturbing the energy landscape for improved packing during computational protein design. Proteins: Structure, Function, and Bioinformatics. 2021; 89(4):436–449. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26030, doi: 10.1002/prot.26030, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26030.

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

Marx DC, Fleming KG. Membrane proteins enter the fold. Current Opinion in Structural Biology. 2021 Aug; 69:124–130. https://www.sciencedirect.com/science/article/pii/S0959440X21000440, doi: 10.1016/j.sbi.2021.03.006.

Marx DC, Fleming KG. Influence of Protein Scaffold on Side-Chain Transfer Free Energies. Biophysical Journal. 2017 Aug; 113(3):597–604. https://www.cell.com/biophysj/abstract/S0006-3495(17)30682-3, doi: 10.1016/j.bpj.2017.06.032.

McDonald SK, Fleming KG. Aromatic Side Chain Water-to-Lipid Transfer Free Energies Show a Depth Dependence across the Membrane Normal. Journal of the American Chemical Society. 2016 Jun; 138(25):7946–7950. https://doi.org/10.1021/jacs.6b03460, doi: 10.1021/jacs.6b03460.

Meng X, Clews J, Kargas V, Wang X, Ford RC. The cystic fibrosis transmembrane conductance regulator (CFTR) and its stability. Cellular and molecular life sciences: CMLS. 2017 Jan; 74(1):23–38. doi: 10.1007/s00018-016-2386-8.

Min D, Jefferson RE, Bowie JU, Yoon TY. Mapping the energy landscape for second-stage folding of a single membrane protein. Nature Chemical Biology. 2015 Dec; 11(12):981–987. https://www.nature.com/articles/nchembio.1939, doi: 10.1038/nchembio.1939.

Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research. 2017 Jan; 45(D1):D170–D176. doi: 10.1093/nar/gkw1081.

Moon CP, Fleming KG. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. Proceedings of the National Academy of Sciences. 2011 Jun; 108(25):10174–10177. https://www.pnas.org/content/108/25/10174, doi: 10.1073/pnas.1103979108.

Nicoludis JM, Gaudet R. Applications of sequence coevolution in membrane protein biochemistry. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2018 Apr; 1860(4):895–908. https://www.sciencedirect.com/science/article/pii/S0005273617303140, doi: 10.1016/j.bbamem.2017.10.004.

Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, Kassem MM, Rasmussen LJ, Lindorff-Larsen K, Hartmann-Petersen R. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS genetics. 2017 Apr; 13(4):e1006739. doi: 10.1371/journal.pgen.1006739.

Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. Journal of Chemical Theory and Computation. 2016 Dec; 12(12):6201–6212. doi: 10.1021/acs.jctc.6b00819.

Paslawski W, Lillelund OK, Kristensen JV, Schafer NP, Baker RP, Urban S, Otzen DE. Cooperative folding of a polytopic α-helical membrane protein involves a compact N-terminal nucleus and nonnative loops. Proceedings of the National Academy of Sciences. 2015 Jun; 112(26):7978–7983. https://www.pnas.org/content/112/26/7978, doi: 10.1073/pnas.1424751112.

Pey AL, Stricher F, Serrano L, Martinez A. Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. American Journal of Human Genetics. 2007 Nov; 81(5):1006–1024. doi: 10.1086/521879.

Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2011 Dec; 9(2):173–175. doi: 10.1038/nmeth.1818.

Ruan J, Liu Z, Sun M, Wang Y, Yue J, Yu G. DBS: a fast and informative segmentation algorithm for DNA copy number analysis. BMC bioinformatics. 2019 Jan; 20(1):1. doi: 10.1186/s12859-018-2565-8.

Sanders CR, Nagy JK. Misfolding of membrane proteins in health and disease: the lady or the tiger? Current Opinion in Structural Biology. 2000 Aug; 10(4):438–442. doi: 10.1016/s0959-440x(00)00112-3.

Scheller R, Stein A, Nielsen SV, Marin FI, Gerdes AM, Di Marco M, Papaleo E, Lindorff-Larsen K, Hartmann-Petersen R. Toward mechanistic models for genotype-phenotype correlations in phenylketonuria using protein stability calculations. Human Mutation. 2019 Apr; 40(4):444–457. doi: 10.1002/humu.23707.

Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nature Reviews Genetics. 2010 Aug; 11(8):572–582. https://www.nature.com/articles/nrg2808, doi: 10.1038/nrg2808.

Stanley AM, Fleming KG. The Role of a Hydrogen Bonding Network in the Transmembrane β-Barrel OMPLA. Journal of Molecular Biology. 2007 Jul; 370(5):912–924. https://www.sciencedirect.com/science/article/pii/S0022283607006213, doi: 10.1016/j.jmb.2007.05.009.

Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends in Biochemical Sciences. 2019 Jul; 44(7):575–588. http://www.sciencedirect.com/science/article/pii/S0968000419300039, doi: 10.1016/j.tibs.2019.01.003.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021 Jan; 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100, doi: 10.1093/nar/gkaa1100.

Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, Vriend G. A series of PDB-related databanks for everyday needs. Nucleic Acids Research. 2015 Jan; 43(Database issue):D364–368. doi: 10.1093/nar/gku1028.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CAK, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, et al. Proteomics. Tissue-based map of the human proteome. Science (New York, NY). 2015 Jan; 347(6220):1260419. doi: 10.1126/science.1260419.

UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2019 Jan; 47(D1):D506–D515. doi: 10.1093/nar/gky1049.

Vanier MT. Niemann-Pick disease type C. Orphanet Journal of Rare Diseases. 2010 Jun; 5:16. doi: 10.1186/1750-1172-5-16.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2021 Nov; p. gkab1061. https://doi.org/10.1093/nar/gkab1061, doi: 10.1093/nar/gkab1061.

wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research. 2019 Jan; 47(D1):D520–D528. https://doi.org/10.1093/nar/gky949, doi: 10.1093/nar/gky949.

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. Journal of Molecular Biology. 2005 Oct; 353(2):459–473. doi: 10.1016/j.jmb.2005.08.020.

Zaucha J, Heinzinger M, Kulandaisamy A, Kataka E, Salvádor aL, Popov P, Rost B, Gromiha MM, Zhorov BS, Frishman D. Mutations in transmembrane proteins: diseases, evolutionary insights, prediction and comparison with globular proteins. Briefings in Bioinformatics. 2021 May; 22(3):bbaa132. https://doi.org/10.1093/bib/bbaa132, doi: 10.1093/bib/bbaa132.

Zhao G, London E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. Protein Science: A Publication of the Protein Society. 2006 Aug; 15(8):1987–2001. doi: 10.1110/ps.062286306.

Ó Conchúir S, Barlow KA, Pache RA, Ollikainen N, Kundert K, O'Meara MJ, Smith CA, Kortemme T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. PloS One. 2015; 10(9):e0130433. doi: 10.1371/journal.pone.0130433.

Supplementary Material

Supplementary Figures

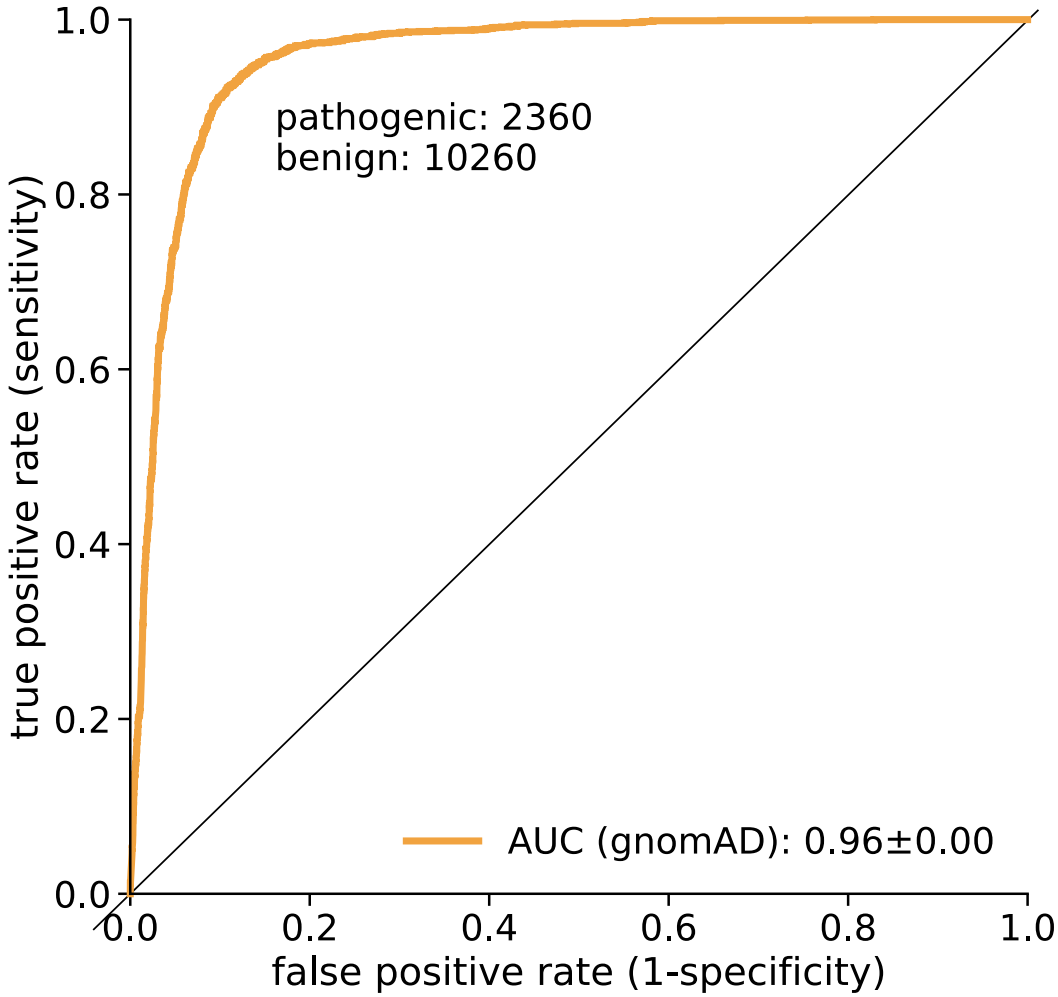

Figure S1. ROC analysis for gnomAD allele frequencies

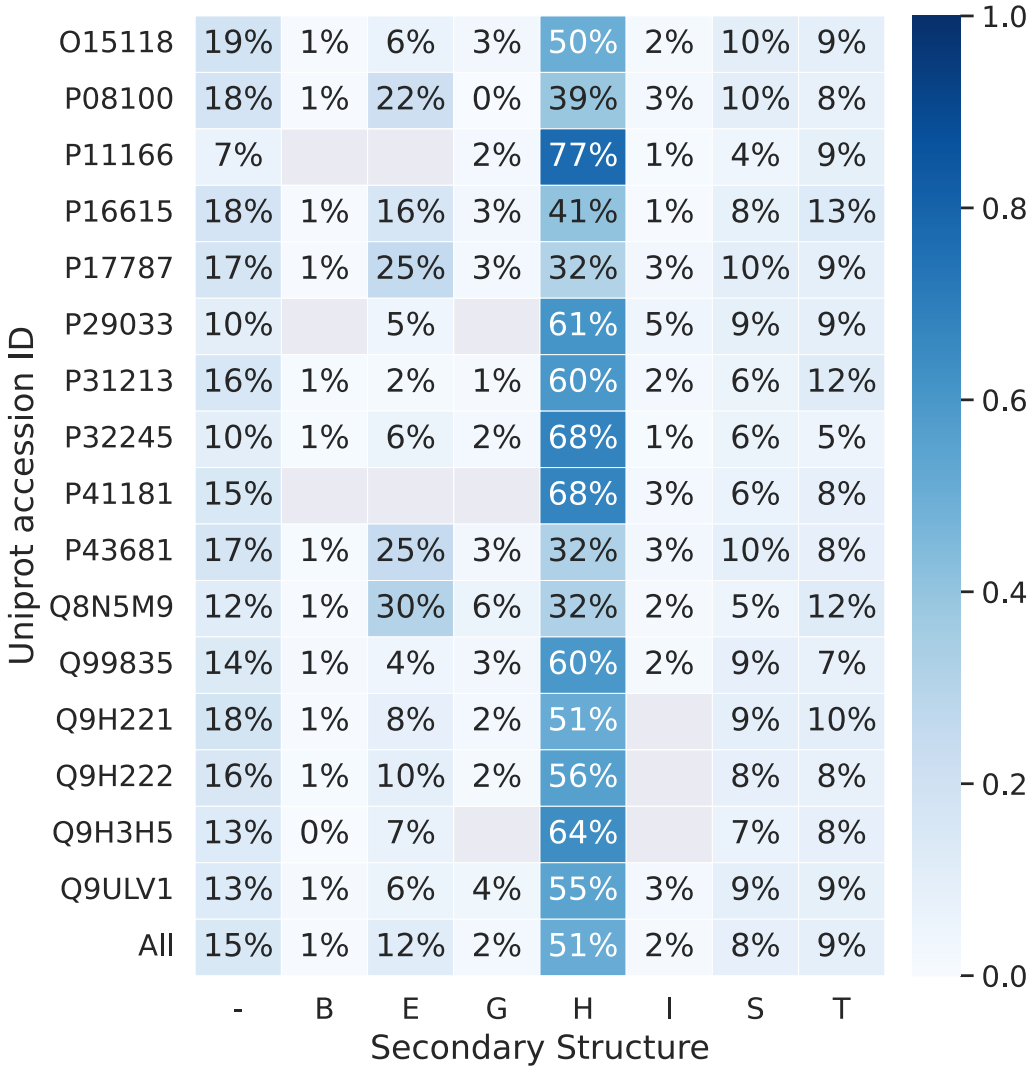

Figure S2. Secondary structure of target dataset calculated using DSSP (***Touw et al., 2015***). Abbreviation stand for H = α-helix; B = residue in isolated β-bridge; E = extended strand, participates in β ladder; G = 3-helix (3_{10} helix); I = 5 helix (π-helix); T = hydrogen bonded turn; S = bend; - = unstructured

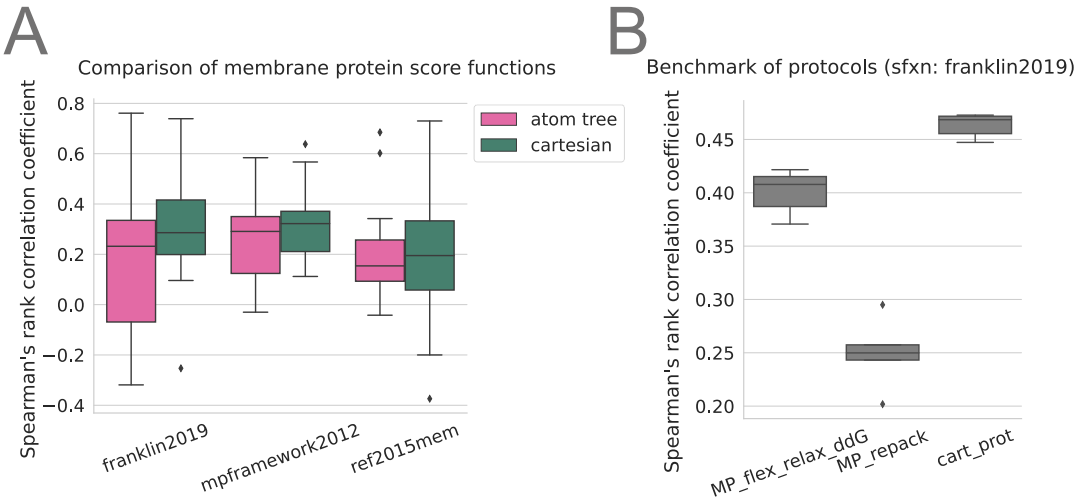

Figure S3. Benchmark of Rosetta ΔΔG calculations for MPs. (A) Comparison of accuracy of stability calculations performed with different membrane protein score functions but using the same protocol and data set. Data was extracted from (***Koehler Leman et al., 2021***). (B) Comparison using different protocols but the same score function (franklin2019; ***Alford et al.*** (***2021***)) and was conducted on the benchmark set described in table S3.

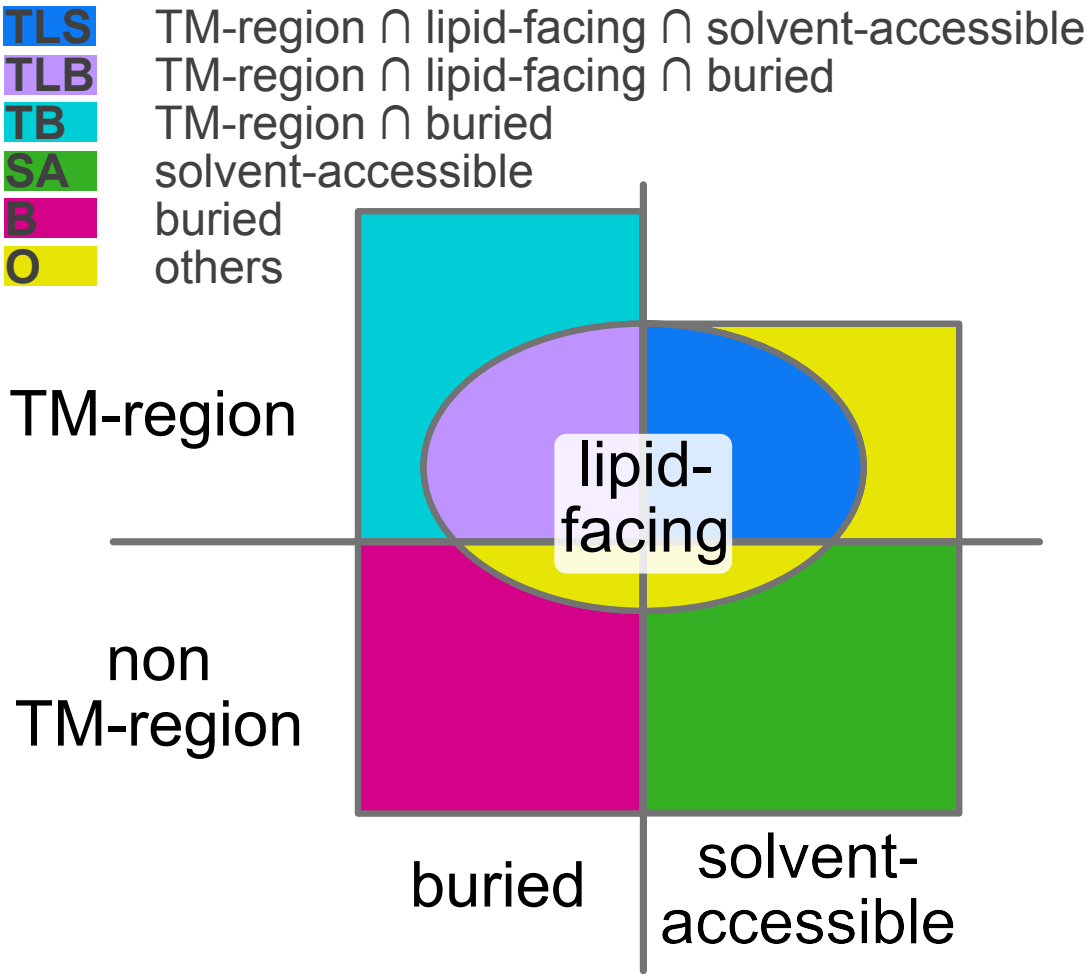

Figure S4. Protein regions and their overlaps.

Supplementary Tables

Table S1. Number of variants annotated by ClinVar or gnomAD for those proteins with at least one experimentally resolved structure per cellular compartment

	all	Extracellular	Cytoplasmic	Transmembrane	other
total	281,220	112,643	61,112	29,944	77,521
benign	2,089	957	495	169	4,029
benign ⋂ gnomAD	1,951	883	467	155	446
pathogenic	4,030	1,430	1,118	807	3,822
patho. ⋂ gnomAD	1,046	326	320	208	192
VUS (ClinVar)	18,882	7,592	5,962	1,974	3,354
VUS ⋂ gnomAD	9,685	3,777	3,190	908	1,810
only gnomAD	256,219	102,664	53,537	26,994	73,024

Table S2. Variant counts, AUC and variant counts within the quadrants for each protein class. Cutoffs are taken from the complete dataset.

protein class info			after filtering		AUC		Q I		Q II		Q III		Q IV	
class	# proteins	subselection	group A	group B	ΔΔG	ΔΔE	group A	group B	group A	group B	group A	group B	group A	group B
Cell Junction	1	all	33	16	0.42	0.84	8	1	12	0	8	11	5	4
		TM region	24	8	0.47	0.78	6	1	9	0	4	6	5	1
Enzyme	2	all	24	10	0.62	0.83	12	0	8	3	2	4	2	3
		TM region	6	33	1	0.83	4	0	0	1	0	1	2	1
GPCR	4	all	54	24	0.79	0.83	31	1	12	5	2	14	9	4
		TM region	31	11	0.81	0.87	19	1	6	2	1	5	5	3
Ion channel	3	all	11	12	0.72	0.81	6	1	3	3	2	8	0	20
		TM region	4	3	0.42	0.75	1	0	2	1	1	2	0	0
Transporter	5	all	98	42	0.63	0.82	48	3	29	9	12	20	9	10
		TM region	44	10	0.71	0.97	27	0	11	1	2	6	4	3

Table S3. Experimental ΔΔG datasets (all from E. coli) used for benchmarking. Multiple variant counts indicate different pH, labeling tags or temperatures.

uniprotID	protein	#variants	reference
P09391	GLPG	6, 3	***Gaffney and Hong*** (***2018***)
P09391	GLPG	69	***Paslawski et al.*** (***2015***)
P09391	GLPG	142	***Baker and Urban*** (***2012***)
P09391	GLPG	20, 20, 8, 8	***Guo et al.*** (***2016***)
P09391	GLPG	2	***Min et al.*** (***2015***)
P0A910	OMPA	19, 20	***Hong et al.*** (***2007***)
P0A910	OMPA	15	***Hong et al.*** (***2006***)
P0A921	PA1	36	***Moon and Fleming*** (***2011***)
P0A921	PA1	6, 6, 6	***Stanley and Fleming*** (***2007***)
P0A921	PA1	49	***McDonald and Fleming*** (***2016***)
P37001	PAGP	20	***Marx and Fleming*** (***2017***)
P37001	PAGP	19, 19	***Huysmans et al.*** (***2010***)