bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The 3D spatial constraint on 6.1 million amino acid sites in the human proteome

Bian Li [1,2], Dan M. Roden [2,3], John A. Capra [1,4]

1. Department of Biological Sciences, Vanderbilt University, Nashville, TN 37203, USA

2. Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

3. Departments of Pharmacology and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

4. Bakar Computational Health Sciences Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143, USA

Abstract

Quantification of the tolerance of protein-coding sites to genetic variation within human populations has become a cornerstone of the prediction of the function of genomic variants. We hypothesize that the constraint on missense variation at individual amino acid sites is largely shaped by direct 3D interactions with neighboring sites. To quantify the constraint on protein-coding genetic variation in 3D spatial neighborhoods, we introduce a new framework called COntact Set MISsense tolerance (or COSMIS) for estimating constraint. Leveraging recent advances in computational structure prediction, large-scale sequencing data from gnomAD, and a mutation-spectrum-aware statistical model, we comprehensively map the landscape of 3D spatial constraint on 6.1 amino acid sites covering >80% (16,533) of human proteins. We show that the human proteome is broadly under 3D spatial constraint and that the level of spatial constraint is strongly associated with disease relevance both at the individual site level and the protein level. We demonstrate that COSMIS performs significantly better at a range of variant interpretation tasks than other population-based constraint metrics while also providing biophysical insight into the potential functional roles of constrained sites. We make our constraint maps freely available and anticipate that the structural landscape of constrained sites identified by COSMIS will facilitate interpretation of protein-coding variation in human evolution and prioritization of sites for mechanistic or functional investigation.

Introduction

The human proteome harbors millions of missense variants that could alter protein structure and function and contribute to disease risk (Karczewski et al., 2020). Strong evolutionary constraint is a hallmark of sites critical to a protein's structure or function. A common approach to identifying constrained sites in human proteins has been to align the human protein sequence to those from other species and locate amino acid residues that are conserved across multiple species (Ng and Henikoff, 2001; Pupko et al., 2002; Capra and Singh, 2007; Cooper and Shendure, 2011). When combined with protein structures, this approach can facilitate generation of testable hypotheses about the biophysical mechanisms underlying the evolutionary constraint (Landgraf et al., 2001; Halabi et al., 2009; McLaughlin et al., 2012; Huang and Golding, 2015). Such interspecific comparisons of sequences are powerful in detecting sequence conservation over long

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evolutionary timescales. Similarly, patterns of intraspecific coding variation, especially low-frequency variants, carry important information about the functional importance of proteins and variants in human development and disease (Karczewski et al., 2020).

Leveraging ever growing human genetic variation data resources (Tennessen et al., 2012; Auton et al., 2015; Lek et al., 2016; Karczewski et al., 2020), several methods have been developed to estimate gene- or region-specific constraint based on tolerance to missense or loss-of-function variants in humans (Petrovski et al., 2013; Samocha et al., 2014; Gussow et al., 2016; Lek et al., 2016; Fadista et al., 2017; Havrilla et al., 2019; Hayeck et al., 2019; Karczewski et al., 2020). These gene-level and region-level measures of constraint have been effective in identifying Mendelian disease genes, genes under strong negative selection, and genes involved in severe neurodevelopmental disorders. However, some protein sites are critical for maintaining the integrity of protein structure or function, while others can be replaced with no or only minor impact on protein structure or function (Worth et al., 2009). Metrics that yield a single score for an entire gene or a subregion do not capture the site-level variability in constraint that is essential for interpreting the effects of specific VUS. To this end, the site-specific missense tolerance ratio (MTR), which compares the observed fraction of missense variation to the expectation under a null model within a sliding window of 31 residues, was developed and shown to improve variant interpretation in epilepsy genes (Traynelis et al., 2017).

Recent analysis of the spatial distribution of missense variants in proteins showed that population-level human standing variation can be analyzed in the context of 3D protein structures to identify specific regions and domains relevant to protein function and disease (Kamburov et al., 2015; Gao et al., 2017; Sivley et al., 2018; Hicks et al., 2019; Perszyk et al., 2021; Silk et al., 2021). For example, the analyses of tumor-derived somatic mutations within the context of protein structure indicate that variants tend to form spatial clusters and that these clusters often overlap functional domains in oncoproteins and tumor suppressors (Stehr et al., 2011; Kamburov et al., 2015; Meyer et al., 2016; Niu et al., 2016; Tokheim et al., 2016). Analysis of 3D spatial patterns of both human germline and somatic variation also highlighted significant differences in the spatial constraint on different classes of mutations in protein structure (Sivley et al., 2018). Recently, amino acid residue sites that are intolerant to missense variation have been characterized by incorporating protein structures and human genetic variation from large sequencing cohorts (Hicks et al., 2019; Liang et al., 2021; Perszyk et al., 2021; Silk et al., 2021). These studies suggest that missense variant analysis at the 3D level can identify functional sites and aid in variant interpretation. However, these previous analyses are limited by the availability of high-quality protein 3D structures and generally covered less than half of the proteins in the human reference proteome. In addition, while it is well-recognized that the mutability of individual amino acid sites is influenced by nucleotide sequence context (Neale et al., 2012; Samocha et al., 2014; Lek et al., 2016; Karczewski et al., 2020) and that inter-residue spatial interactions are essential to maintaining structural and functional integrity of proteins (Worth et al., 2009), the consideration of the mutation spectrum at the resolution of native 3D interactions remains largely unexplored.

We hypothesize that connected functional sets of 3D neighboring amino acid sites, a “contact set”, collectively shape the level of constraint on each site (e.g., as quantified by the depletion of missense variation compared to the amount expected under neutral evolution). We introduce the COntact Set MISsense (COSMIS) tolerance framework, to quantify the level of

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observed vs. expected missense variation in the local 3D structural context of index sites while correcting for nucleotide sequence context-dependent mutability of amino acid sites. We applied the framework to analyze the 3D spatial distribution patterns of 4.1 million unique missense variants at 6.1 million amino acid sites in their 3D structural context. We integrated high-quality protein 3D structures from three large sources, i.e., the Protein Data Bank (PDB) (Rose et al., 2017), the SWISS-MODEL repository (Waterhouse et al., 2018), and the recently released, comprehensive database of protein 3D structures predicted by the AlphaFold2 algorithm (Jumper et al., 2021; Tunyasuvunakool et al., 2021). Collectively, our framework covers 16,533 (80.3%) of all proteins in the human reference proteome. We show that our framework captures broad missense variant intolerance at the 3D spatial level across the human proteome. We demonstrate the utility of COSMIS in variant interpretation and in revealing biophysical insights into the pathogenic mechanisms of disease-causing variants. We further demonstrate the flexibility of the framework to work with custom-built homology models of potassium channels and with proteins in their oligomeric states. We propose that our COSMIS framework will have broad applicability in answering diverse questions about variant effect and to discover new genotype-phenotype relationships.

Results

The COSMIS framework maps spatial constraint on proteins in high resolution

We developed the COSMIS framework to quantify the intra-species 3D spatial constraint at each site in a protein structure. Our framework estimates the constraint on a site of interest (index site) as the depletion of missense variants in its contact set compared to the number expected if it were evolving neutrally. We quantify this as the deviation of observed count of missense variants (m_o) from the mean (m_e) of the expected count distribution (accounting for transcript and codon missense mutability) and divided by the standard deviation of the expected distribution (m_σ) (Fig. 1 and Methods). We designate this Z score as the COSMIS score and assign it to the index site. Thus, a lower score indicates a greater depletion of missense variants in the spatial neighborhood and hence lower missense variation tolerance. We also compute an empirical p-value for each COSMIS score based on 10,000 null simulations and the resulting expected count distribution (Methods).

3D structural context differs from 1D sequence context

COSMIS scores are based on the 3D interaction context of protein-coding sites. We quantify this context using contact sets, defined as the set of amino acid residues that are in contact with the residue located at the index site (Fig. 2a). A pair of residues are considered to be in contact when the distance between their C_β atoms (or C_α atoms in the case of glycines) is less than 8 Å, a threshold commonly used to define residue contact (Shrestha et al., 2019). To compute contact sets, we collected high-quality protein 3D structures (Methods and Supplementary Fig. 1) that collectively cover 80.3% of all proteins in the human reference proteome (UP000005640, UniProt release 2021_03) from the Protein Data Bank (PDB) (Rose et al., 2017), SWISS-MODEL repository (Waterhouse et al., 2018), or the AlphaFold database of highly accurate predicted structures (AF2) (Tunyasuvunakool et al., 2021) (Fig. 2b) (Supplementary Table 1). Our framework also has high residue-level coverage. Structures from PDB, SWISS-MODEL, and AF2 have a median residue-level coverage of 78.0%, 70.9%, and 76.7%, respectively (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 1). Collectively, we computed contact sets for each of 6.1 million unique index sites in the human reference proteome (Fig. 2b). For simplicity, we report results for experimental structures and computational models together, since patterns were similar when we analyzed them separately (Supplementary Fig. 2, Supplementary Fig. 3, and Supplementary Fig. 4).

Conceptually, a contact set captures residues that are close in 3D space, even when they are far apart in sequence. Our analysis shows that contact sets include critical "long-range" (defined here as > 15 residues apart along 1D sequence) residue-residue interactions that would be missed by 1D sequence-based metrics that only consider a window around the index site. For example, for 17.9% of all 6.1 million amino acid sites, at least 50% of the 3D contacts they make are long-range, and 66.3% of all sites make at least 10% long-range 3D contacts (Fig. 2c). On the other hand, windows based on sequence context alone contain many sites that are not in 3D contact with the index site. For example, all of the 6.1 million sites in this study have at least 50% of their 30 1D sequence neighbors (15 sites on each side) not in 3D contact, and nearly half (47.5%) of all sites have least 80% of sequence neighbors not in 3D contact (Fig. 2d). Thus, long-range 3D contacts are common, as are sites that are nearby in sequence but distant in 3D. (Additional statistics about long-range 3D contacts are available in Supplementary Fig. 3.) Residues in 3D contact are likely to be essential for the structural stability and functional integrity of the residue at the index site; thus, 3D structure-based residue contact sets give a more sensitive representation of the structural and function context of coding sites than sequence-based windows.

COSMIS score captures constraint at both protein and site levels

Consistent with our expectation and previous observations at the gene level, our framework identifies broad constraint on missense variants at the protein level and little constraint on synonymous variants (Lek et al., 2016; Karczewski et al., 2020). We computed the deviation from the expected count (observed - expected) at the protein level for both synonymous and missense variants across the entire dataset. Supporting our approach for estimating the expected variant count distributions (Methods), the deviation between the observed and expected synonymous variant count is low and centered near zero (Fig. 3a; median 1.9, standard deviation 39.4). In contrast, the difference between observed and expected is significantly shifted toward negative values for missense variants (Fig. 3a; median -29.1, standard deviation 104.8, $p < 2.2 \times 10^{-308}$, two-sided Mann-Whitney U test).

We next computed COSMIS scores for the 6.1 million unique amino acid sites across 16,533 proteins in the human reference proteome with sufficient data (Fig. 3b). As expected, the distribution of scores spans a wide range from negative (constrained) to positive (unconstrained) values, with a significant shift toward constraint (median -0.47, standard deviation 1.2). Proteins with experimentally determined structures in the PDB have a significantly lower median COSMIS score than those currently only have computationally predicted structures in SWISS-MODEL and AF2 databases (median -0.62 vs. -0.42 and 0.44, respectively, $p < 2.2 \times 10^{-308}$, two-sided Mann-Whitney U test) (Supplementary Fig. 4), suggesting that proteins with greater functional importance have historically been selected for structural characterization. In contrast, a similarly constructed score based on synonymous variants (i.e., contact set synonymous tolerance score) is centered at 0, regardless of the sources of protein 3D structures (Supplementary Fig. 4),

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistent with overall results and the hypothesis that synonymous variants are not subject to 3D spatial constraint in protein structures.

We consider sites with a COSMIS score for which the empirical p-value obtained from simulation is <0.01 as high-confidence (this is approximately equivalent to COSMIS score < -2.33). Overall, we find 313,204 sites (5.1%) with high-confidence constraint scores from 10,955 proteins (66.3%), with an average of 28.6 high-confidence constrained sites per protein (Fig. 3c, Supplementary Table 2). Overall, these findings suggest that the COSMIS score captures the depletion of missense variants in 3D structure-based contact sets resulting from varying functional constraint over protein space.

COSMIS refines gene-level constraint estimates

With the growth of large human genetic variation datasets, methods have been developed to quantify constraint on individual sites, sequence windows, and genes. Given the connection between constraint and function, constraint scores are critical components of many gene and variant interpretation tasks. For example, gene-level metrics, like pLI (Lek et al., 2016), have been extensively used to prioritize genes in which variants are likely to contribute disease risk. However, gene-level metrics are incapable of highlighting the amino acid sites within each gene that are under constraint. pLI accurately identifies gene-level constraint, but frequently classifies genes that harbor known disease associated mutations as LoF tolerant (Ziegler et al., 2019). This is not surprising, but it illustrates a weakness of gene- and region-level metrics. Site- and window-based intraspecies constraint metrics provide a higher resolution view, but as demonstrated above, sequence context is often very different from 3D structural interaction context. Since COSMIS quantifies constraint at the contact set level for each amino acid site, we hypothesized that it would provide a higher-resolution view of the clinical importance of protein regions, in addition to capturing broad constraint at the protein level.

To explore this hypothesis, we computed the distribution of per-protein COSMIS scores for 16,260 proteins stratified into the three pLI classes (Intolerant (n=2,566), Unsure (n=2,900), Tolerant (n=10,794)). (We were not able to obtain the pLI scores for 273 proteins with COSMIS scores.) As expected, on average LoF intolerant genes (pLI $\geqslant$ 0.9) have significantly lower COSMIS scores than LoF tolerant genes (pLI $\leqslant$ 0.1) (-1.1 vs. -0.12, $p < 2.2 \times 10^{-308}$, two-sided Mann-Whitney U test, Fig. 4a), indicating that sites in LoF intolerant genes are on average more constrained than LoF tolerant genes. Genes that have medium pLI scores (0.1 < pLI < 0.9) also have medium COSMIS score on average (-0.80).

While LoF tolerant genes have less evidence of spatial constraint overall, we found that 1888 (40.4%) LoF tolerant genes have at least one high confidence constrained site (COSMIS score < -2.33), with 13.6 on average. For example, the ubiquitin-like modifier-activating enzyme 5 (UBA5) is considered LoF tolerant (pLI score of 2.5×10^{-4}). However, our analysis indicates that UBA5 has many constrained sites in interfaces of UBA5 dimerization, UBA5-UFM1 binding, and UBA5-ATP interaction (Fig. 4b and 4c). Specifically, of the 30 (10%) most constrained sites in UBA5, 13 sites are located at the UBA5 dimerization interface, four sites interact with ATP, and another three are involved in UFM1 binding (Fig. 4c). This is consistent with UBA5's involvement in severe epileptic encephalopathy (Colin et al., 2016). Indeed, the three constrained sites with the strongest constraint according to COSMIS (amino acid residues 54, 57, and 58) include

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

p.Met57Val, which was found in a patient cohort to drastically reduce UBA5's catalytic activity (Colin et al., 2016). Thus, COSMIS scores of UBA5 identify amino acid sites relevant to UBA5's functions and known disease associations. This illustrates how considering constraint in spatial neighborhoods can identify genes predicted to be LoF tolerant (low pLI) that are clinically important and suggests that COSMIS can guide further investigation before discarding genes from clinical consideration.

COSMIS highlights pathogenic variants and essential proteins

To quantify the ability of COSMIS to contribute to identification of disease-associated protein variants, we compared the COSMIS scores for a total of 19,596 benign and 17,858 pathogenic missense variants with unambiguous annotations of clinical significance in ClinVar (Methods, Supplementary Table 3). Benign and pathogenic variants have significantly different COSMIS score distributions (median 0.0 vs. -1.1, respectively; $p < 2.2 \times 10^{-308}$, two-sided Mann-Whitney U test; Fig. 5a). The significant negative shift for pathogenic variants suggests strong constraint in their spatial neighborhoods, while the average neutral COSMIS score of benign variants suggests less constraint on missense variants in their contact sets. Across the COSMIS score range, the magnitude of the score correlates with enrichment for pathogenic over benign variants. High-confidence constrained sites are 13.5-fold enriched for pathogenic variants. However, only 1,706 out of the 10,955 proteins that have at least one high-confidence site have unambiguously annotated pathogenic variants in ClinVar (Supplementary Table 4), suggesting that many pathogenic variants are yet to be uncovered. Moving down the constraint spectrum, the top 10% most constrained COSMIS sites (equivalent to COSMIS score < -1.85) are 10.6-fold enriched for pathogenic variants, and the bottom 10% are 3.3-fold depleted (Fig. 5b). Our analysis suggests that constraint on missense variation in a site's 3D interaction context (as quantified by COSMIS) is strongly correlated with variant pathogenicity.

To evaluate the relationship between spatial constraint as quantified by COSMIS and function and disease associations at the protein level, we compared the COSMIS score distributions of amino acid sites in six groups of genes expected to be under various levels of constraint (and the dataset as a whole). In general, genes with essential functions and disease associations have lower COSMIS scores than genes without (Fig. 5c) and as the essentiality of a gene increases, amino acid sites in the gene have more negative COSMIS scores on average. Haploinsufficient genes (a single-copy of the functional allele is insufficient to produce the expected phenotype) (Rehm et al., 2015), genes essential in cell culture, and genes associated with dominant diseases have the lowest COSMIS distributions among all evaluated gene categories. In contrast, constrained sites are much less frequently found in nonessential genes. Not surprisingly, olfactory receptors (Mainland et al., 2015) have the least spatial constraint of any gene set considered. The abundance of high-confidence constrained sites in each gene follows the same general trend (Supplementary Fig. 5). Our analysis identifies 72 proteins with more than 50% high-confidence constrained sites (Supplementary Table 5). These proteins are likely to be under extreme purifying selection; in fact, 10 of these proteins are encoded by genes that have already been classified as haploinsufficient (Rehm et al., 2015), essential (Hart et al., 2017), or associated with diseases that follow dominant inheritance (Blekhman et al., 2008; Berg et al., 2013). Overall, our analysis indicates that the COSMIS score strongly captures functional constraint and is predictive of variant pathogenicity.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

COSMIS complements existing quantifications of intra- and inter-species constraint

To assess the relationship between COSMIS and other intra- and interspecies constraint metrics, we first compared COSMIS to four commonly used intraspecies constraint metrics that do not consider structural context (MTR, RVIS, pLI, and Missense_Z). We compared these other metrics to COSMIS in their ability to identify pathogenic variants using a total of 8,062 benign and 7,256 pathogenic missense variants from ClinVar for which all scores could be computed (Supplementary Table 6). COSMIS achieved a significantly higher area under than receiver operating characteristic curve (AUROC) than the other intraspecies constraint metrics (e.g., 0.733 vs. 0.653 for COSMIS vs. MTR, $p = 1.0 \times 10^{-65}$, two-sided DeLong's test, Fig. 6a). This suggests that 3D neighboring residues contribute critical information about the functional importance of index sites. We additionally compared COSMIS to a recently developed version of MTR that considers missense variants in 3D neighborhoods (MTR3D), but does not account for sequence context-dependent mutability (Perszyk et al., 2021; Silk et al., 2021). COSMIS also performs significantly better than MTR3D (i.e., 0.733 vs. 0.665, $p = 2.5 \times 10^{-50}$, two-sided DeLong's test, Fig. 6a), suggesting that accounting for the variability of mutability is essential to estimate constraint.

We then compiled a subset of 3.6 million amino acid sites for which the five commonly used intraspecies constraint metrics (MTR, MTR3D, RVIS, pLI, and Missense_Z) could also be computed. To summarize the relationships between these constraint scores, we computed their pairwise Spearman correlations across sites (Fig. 6b). As expected, pLI and Missense_Z have the highest Spearman's ρ (0.63), given that they both quantify gene-level constraint and were derived with similar approaches (Lek et al., 2016). Similarly, MTR and MTR3D are well correlated (Spearman's ρ 0.53). The COSMIS score has a comparable level of correlation with both MTR and MTR3D (0.41 and 0.39). The intermediate correlations suggest that the metrics capture constraint at different scales, as expected.

To illustrate the differences between the *intra*species constraint scores and *inter*species phylogenetic conservation metrics, we additionally computed the correlations for four common interspecies phylogenetic conservation metrics (GERP++, phyloP, phastCons, ConSurf) (Supplementary Table 6). Phylogenetic conservation metrics are generally more correlated with each other than with any of the intraspecies constraint scores (Fig. 6b). For example, the lowest Spearman's ρ between the four phylogenetic conservation metrics is 0.47 (GERP++ vs. phyloP), higher than the highest Spearman's ρ between a phylogenetic metric and an intraspecies constraint score (i.e., 0.38, ConSurf vs. RVIS and Missense_Z). This is consistent with previous finding that intraspecies constraint metrics are only modestly correlated with phylogenetic conservation (Gussow et al., 2016; Havrilla et al., 2019).

Given this, we hypothesized that integrating interspecies scores with COSMIS could provide additional information for pathogenicity prediction. To test this hypothesis, we combined COSMIS with ConSurf (the best performing interspecies metric on our dataset, Supplementary Fig. 6) using a logistic regression model and evaluated the resulting performances with five-fold cross validation (Methods). Our evaluation shows that integrating COSMIS and ConSurf outperformed the AUROC of both ConSurf and COSMIS alone (0.860 vs. 0.847 and 0.733, $p = 0.002$ and $p = 6.3 \times 10^{-145}$, respectively, two-sided DeLong's test, Fig. 6c). In particular, the

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

improvement from adding COSMIS to ConSurf over ConSurf alone is mainly due to better performance in the high-confidence region (Fig. 6d). However, we note that ConSurf alone outperforms COSMIS alone (0.847 vs. 0.733) (Fig. 6c). Combining all 10 scores and relative solvent accessibility in a regression model resulted in additional AUROC improvement (0.884 vs. 0.860, $p = 6.0 \times 10^{-10}$, two-sided DeLong's test, Figure 6c). Our results suggest that COSMIS score contributes additional information to phylogenetic conservation for pathogenicity prediction and that adding intraspecies constraint can improve the performance of even the best phylogenetic conservation scores.

COSMIS improves the interpretation of de novo variants in neurodevelopmental disorders

De novo mutations are often clinically relevant and are more likely to be pathogenic than inherited variation (Eyre-Walker and Keightley, 2007); however, they are difficult to interpret. *De novo* variants play a prominent role in rare and common forms of neurodevelopmental disorders (Veltman and Brunner, 2012), and *de novo* variants in neurodevelopmental disease cohorts have been used previously to benchmark the utility of constraint metrics for variant interpretation (Samocha et al., 2014). To test if considering spatial constraint could contribute to the interpretation of *de novo* variants, we compared the COSMIS distributions for 2,271 *de novo* missense variants from neurodevelopmental disorder probands (case variants) versus 541 *de novo* missense variants from unaffected siblings of autism spectrum disorder probands (Samocha et al., 2017) for which COSMIS scores can be computed (Supplementary Table 7). Control variants had a median COSMIS score significantly higher than the median COSMIS score of case variants (-0.39 vs. -0.83, $p = 3.0 \times 10^{-13}$, two-sided Mann-Whitney U test) (Fig. 7a).

For context, we compared the ability of COSMIS to enrich for case variants with the other inter- and intraspecies metrics considered previously. We did this analysis using 1,506 case and 306 control variants for which all 10 scores are available (Supplementary Table 8). For COSMIS, 24.2% case and 8.2% control variants fall within the 10th percentile of most constrained sites (i.e., COSMIS score < -1.85), corresponding to an odds ratio (OR) of 3.6 ($p = 2.6 \times 10^{-11}$, two-sided Fisher's exact test) (Fig. 7b). Both MTR and pLI achieved the next highest OR of 2.6 at the 10th percentile of most constrained sites, while being lower than COSMIS. Except for MTR3D, which has an OR of 2.1, the ORs of all the rest metrics are below 2 (Fig. 7b). At other thresholds (5th and 20th percentiles), COSMIS also has the highest ORs (Supplementary Fig. 9). We do not expect all *de novo* variants in cases to be causal; thus, the modest performance of all metrics is not surprising. Our finding is also consistent with previous observation that distinguishing case from control *de novo* mutations is challenging (Havrilla et al., 2019; Pejaver et al., 2020). Overall, we expect COSMIS to be a valuable tool to complement existing approaches for prioritizing *de novo* mutations.

Applying COSMIS to custom-built oligomeric potassium channel structures

To demonstrate the flexibility of our framework to work with custom-built protein 3D structures and to investigate whether COSMIS score could capture constraint imposed by protein-protein oligomerization (Caffrey et al., 2004; Mintseris and Weng, 2005; Li et al., 2019), we compiled a set of 41 potassium ion channel (KCN) genes (Supplementary Table 9) for which variants have been annotated in ClinVar. KCN genes encode proteins that function in homo-oligomeric states (Yu and Catterall, 2004), so we expected interface sites to be under stronger constraint than non-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interface sites. We obtained structures for these KCN proteins in their homo-oligomeric states either from the PDB or through homology modeling using the SWISS-MODEL interactive workspace (Waterhouse et al., 2018). Collectively, we structure-mapped and computed two sets of COSMIS scores, based on monomers and oligomers, respectively, for 4,762 interface and 14,331 non-interface sites in these potassium channels. As expected, we found that on average interface sites are significantly more constrained than non-interface sites (median COSMIS score -1.3 vs. -1.1, $p = 1.3 \times 10^{-16}$, two-sided Mann-Whitney U test; Fig. 8a). When computed based on oligomer structures, the scores of interface sites shift significantly to more negative values (median difference -0.13, $p = 8.1 \times 10^{-8}$, two-sided Mann-Whitney U test; Fig. 8b), suggesting that COSMIS captures additional constraint on interface sites contributed by sites in neighboring subunits. We next evaluated the performance of COSMIS in predicting the pathogenicity of missense variants in KCN oligomers. Establishing the pathogenicity of variants in KCN genes is clinically significant, because they have been associated with multiple severe neurological, psychiatric, and cardiac disorders such as epileptic encephalopathy, schizophrenia, and long QT syndrome (Shieh et al., 2000). We compiled a subset of 111 and 489 unambiguously annotated benign and pathogenic KCN missense variants from ClinVar. On this variant set, COSMIS as a single metric showed strong performance (AUROC 0.854) (Fig. 8c). These results illustrate how COSMIS can be further applied to provide constraint maps in custom use cases and oligomeric structures beyond the precomputed scores we provide for 80.3% of proteins in the human proteome.

Discussion

Establishing the clinical relevance of VUS is one of the largest challenges to genomics-enabled precision medicine (MacArthur et al., 2014; Richards et al., 2015; Starita et al., 2017; Manolio et al., 2019). In this work, we hypothesized that patterns of genetic variation at neighboring sites in 3D collectively reflect levels of functional constraint and that quantifying this constraint could aid VUS interpretation. We developed the COSMIS framework and analyzed the distribution patterns of human genetic variants in the context of their 3D protein structures. Our framework enabled us to map spatial constraint at the resolution of individual sites in 80.3% of proteins in the human proteome. We further showed that our COSMIS score is accurate in predicting gene essentiality and variant pathogenicity and in aiding in the interpretation of *de novo* variants. Furthermore, it complements information provided by other commonly used metrics like phylogenetic constraint between species. The COSMIS framework is flexible and easily expanded to new applications as illustrated by our detection of constrained sites and pathogenicity predictions in ion channels using custom-built oligomeric homology models. We expect that our framework can be applied to a wider set of genes than analyzed in this work as the structural coverage of the human proteome and other species continue to expand.

Compared to existing constraint quantification approaches, our framework has several features that are particularly valuable for variant interpretation. First, our site-specific COSMIS score quantifies the variation in constraint at a finer scale than methods that generate a single score for an entire gene or subregions of a gene. The COSMIS score is thus more precise and specific for interpreting missense variants than many common approaches. Second, our framework quantifies the constraint of sites in their local 3D structural context. Compared to metrics that are based on 1D sequences, local 3D structural context enabled the COSMIS score

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to capture native 3D interactions between residues that are far apart in sequence yet important for maintaining structural stability and functional integrity of the index site. In addition, the contact set filters out residues that are close in sequence but less likely to make contribution to the constraint of the site because they are distant in 3D. In fact, our 3D structure-based COSMIS score performed significantly better than the 1D sequence-based MTR score in predicting variant pathogenicity, while also providing important biophysical insights into the potential functional roles of constrained sites. Third, our approach accounts for variation in mutation rates in the neutral model. This led to significantly better performance than a recent 3D-aware metric (MTR3D). Fourth, COSMIS provides scores for more than 80% of the human proteome by incorporating new high quality structural models from AlphaFold, significantly more than previous structure-based analyses. Fifth, COSMIS can easily be applied to structures for new proteins or complexes, as illustrated on KCN genes. (COSMIS's strong performance on the KCN variants, suggests that it may be particularly well suited to pathogenicity prediction in ion channels.) Finally, COSMIS is complementary to other metrics. Combining the COSMIS score with phylogenetic conservation metrics yielded significantly higher performance than either approach alone in predicting variant pathogenicity. This suggests that future ensemble variant pathogenicity predictors may benefit from integrating spatial constraint as quantified by the COSMIS score.

Our approach nevertheless has several limitations. First, the missense burden analysis and statistical identification of constrained contact sets is highly dependent upon the number and quality of variants used as references for the standing variation dataset. Current gnomAD samples carry only an average 6.3% and 10.3% of all possible missense and synonymous variants per contact set, respectively (Supplementary Fig. 10). As larger and more diverse reference genetic variation cohorts continue to increase the number of observed variants in each gene, even more accurate estimates of constraint on contact sets will be possible. It may also be possible to decrease the contact set distance threshold to capture more specific 3D interactions. Second, the COSMIS score does not directly consider the physicochemical severity of amino acid substitutions. While these patterns likely contribute to the patterns of observed variation and mutability, explicit consideration of the severity of amino acid substitution could improve estimates of site constraint. For instance, sites tolerant to both conservative and non-conservative substitutions are likely to be under less constraint than sites that are only tolerant to conservative substitutions. Third, while it is known that sites harboring variants with lower minor allele frequencies are likely under stronger selection pressure (Hartl, 1989), as with previous approaches (Samocha et al., 2014; Gussow et al., 2016; Lek et al., 2016; Traynelis et al., 2017; Havrilla et al., 2019; Karczewski et al., 2020), we counted the number of unique variants observed at each site and did not explicitly account for their MAFs. Incorporation MAFs and demographic structure into the formulation of scores such as the pLI (Lek et al., 2016), MTR (Traynelis et al., 2017), and COSMIS remains a promising topic. We anticipate that our framework can be further improved in the future by including larger human variation datasets, consideration of additional amino acid properties, and accounting for MAF.

Looking forward, we anticipate that the structural landscape of constrained sites provided by COSMIS will facilitate prioritization of sites for mechanistic or functional investigation, especially those that have not been previously associated with clinically relevant phenotypes. For example, we have demonstrated that high-confidence constrained sites have a >10-fold

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

enrichment for pathogenic variants, yet 9,249 out of the 10,955 proteins (84.4%) harboring at least one high-confidence constrained site lack any unambiguously annotated pathogenic variants in ClinVar. Variants at the constrained sites of some of these proteins are likely to be embryonic lethal, but many likely have pathogenic potential. Using COSMIS to guide investigation of the effects of variation at these sites on protein function will contribute novel insights into human health and disease.

Methods

Estimating mutation probabilities using the 1000 Genomes Project variant set

We estimated sequence-context-dependent trinucleotide synonymous and missense mutability following previous procedures (Samocha et al., 2014; Lek et al., 2016). Briefly, we retrieved all single nucleotide variants from the 1000 Genomes Project variant set (phase 3, 2504 individual genomes) (Auton et al., 2015). We filtered this initial set of variants to include only single nucleotide variants and excluded multiallelic variants, indels, and any variants with a filter tag other than "PASS". We focused on variants in intergenic regions obtained by excluding regions of the genome spanned by genes as annotated in GENCODE release 34 (Frankish et al., 2019). We did not consider variants in the coding genome since they are enriched for purifying selection. For the entire intergenic genome, we counted every instance of each of the 64 trinucleotide sequences. We then identified all variable sites in the 1000 Genomes data with an annotated ancestral allele and assumed that each variable site represents a single ancestral mutation. To compute the probability of a trinucleotide XYZ mutating to $XY'Z$, we divided the number of $Y \rightarrow Y'$ mutations in the context of XYZ by the total number of occurrences of XYZ. As described in previous work (Samocha et al., 2014; Lek et al., 2016), we scaled the probability by a proportionality constant (Neale et al., 2012) to derive the probability for one generation. In the end, we obtained a 64 by 3 matrix in which each row contains the probability of each of the three possible mutations of the central nucleotide of a given trinucleotide context. Protein-level mutability estimates obtained using our matrix agreed with previous estimates (Supplementary Fig. 7). Our mutability table is available in the source code available at our GitHub repository: https://github.com/CapraLab/cosmis. Our framework also enables the use of a custom mutation matrix.

Mapping human reference proteome to Ensembl transcripts

We started with the human reference proteome (UP000005640, UniProt release 2021_03), containing the reference amino acid sequences for a total of 20,600 proteins (Bateman et al., 2017). To determine whether the COSMIS scores for a protein can be computed, we first obtained the Ensembl stable transcript IDs for the protein through programmatic access of the UniProt database identifier mapping service (https://www.uniprot.org/help/api_idmapping). We used the transcript IDs as keys to extract coding sequence (CDS) from Ensembl CDS database. A valid CDS is necessary for the computation of COSMIS score because it is the basis for our mutation-probability-aware variant simulation procedure. A CDS is valid only if it begins with ATG, ends with a stop codon, and its translated amino acid sequence matches the UniProt reference sequence. We then used the transcript ID corresponding to the valid CDS as key to extract variant statistics from GRCh38-lifted gnomAD v.2.1.1. We used vcftools to remove all sites with a FILTER flag other than PASS gnomAD v.2.1.1 and only kept single nucleotide variants. Completing this

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

procedure for each reference protein resulted in a total of 16,533 proteins for which a "high-quality" protein 3D structure is also available in PDB, SWISS-MODEL repository, or AF2 database (see below).

Per amino acid site synonymous and missense mutability of reference protein sequence

We estimated the synonymous and missense mutability of each amino acid in the matched transcript of the protein in a nucleotide sequence context-dependent manner. In brief, the local trinucleotide sequence context was used to determine the mutability of each base in the coding region mutating to each other possible base and to determine the coding impact of each possible mutation. These mutability values were summed across the codon to determine its synonymous missense mutability. Specifically, for a given base in the codon, the trinucleotide sequence context is determined according to the coding sequence of the transcript as provided by the Ensembl CDS resource. The probability of the middle base mutating to one of the three other bases is queried in the mutation probability table and the type of change it would create is determined. The mutability is added to a running total for the type of mutation it would cause. This is repeated for the two other possible mutations for every base in the codon. In the end, we obtained a pair of per-codon synonymous and missense mutability for each codon in each transcript.

Estimating the per-protein expected number of synonymous and missense variants

We estimated the per-protein expected number of synonymous and missense variants through a fitted linear regression equation of per-protein total variant count on mutability. We first extracted 50,456 Ensembl transcripts for which SNVs were reported in gnomAD and whose CDS also met our criteria. For each of these transcripts, we then computed the synonymous and missense mutability of all codons and summed them to produce per-protein total synonymous and missense mutability (Supplementary Table 10). Our protein-level mutability estimates correlated strongly with previous estimates (Pearson's R of 0.94 and 0.95 for synonymous and missense, respectively; Supplementary Fig. 7) (Samocha et al., 2014). In parallel, we also tallied the total synonymous and missense variant counts reported in gnomAD for each of these transcripts. The total numbers of unique synonymous and missense variants from the 16,533 proteins studied in this work are 2.0 million and 4.1 million, respectively (Supplementary Table 10). To establish the relationship between mutability and expected variant count under the null hypothesis of minimal constraint, we regressed the number of synonymous variants on the total synonymous mutability per transcript. As expected, and consistent with previous observations, we found that the total synonymous variant count can be accurately predicted by total synonymous mutability with a simple linear regression model ($\hat{y} = 6.42 \times 10^{-6} \times \mu - 0.18$, where μ is per-protein total synonymous mutability, $R = 0.95$, Supplementary Fig. 8). As most synonymous variation is under minimal selective pressure, this model represents the relationship between mutability and observed variation when selection pressure is minimal. We thus estimated the expected count of missense variants t_e under minimal selection for each of the transcripts by applying this regression model to their total missense mutability.

Mapping transcripts to protein structures

Computing the COSMIS score requires projecting missense variants onto 3D protein structures. We used the Protein Data Bank (PDB) as our primary source of protein structures. We used a summary table processed by SIFTS (https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pdb_chain_uniprot.tsv.gz) (Dana et al., 2019) to obtain an one-on-one mapping between PDB chains and UniProt accession numbers. The PDB contains many cases where multiple PDB chains map to a single UniProt accession number. In these cases, we selected the PDB structure that has the most resolved amino acid residues. We also required PDB structures to have a resolution better than 5 Å and to cover as least one third of the reference amino acid sequence. If multiple PDB structures cover a protein sequence equally well, we selected the one that has the best resolution. We excluded structures for which sidechain coordinates are not resolved. For proteins for which no experimental structure in the PDB meets our criteria, we searched and retrieved structural models from the SWISS-MODEL repository (July 2021 release) (Bienert et al., 2017) and the AlphaFold protein structure database (AF2) (Tunyasuvunakool et al., 2021). The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modeling pipeline (Waterhouse et al., 2018). The AF2 database is a collection of 3D protein structures for the human reference proteome predicted using the highly accurate AlphaFold2 method (Jumper et al., 2021). We first searched the SWISS-MODEL repository for models that have a sequence identity of at least 25% and cover at least one third of the amino acid sequence of the target sequence, in consideration of increasing the number of covered proteins and maintaining a reliable level of homology model quality (Cavasotto and Phatak, 2009). In cases where multiple models satisfied these criteria, we selected the model with the highest sequence coverage to maximize the set of structure-mappable variants. For proteins with no homology models that meet our criteria, we relied on the AF2 structure database. In these cases, COSMIS scores were computed only if the AF2 predicted structure has at least one third of all residues predicted with a pLDDT > 50 (Jumper et al., 2021), and only such predicted residues were included in the computation of contact sets.

Residue-level mapping between Ensembl transcript and protein structure

The sequence of the experimental construct of a protein often does not match that of the reference sequence given in UniProt; e.g., the amino acid at position i in the PDB file might be shifted by a few positions relative to its position in the translated peptide sequence of the corresponding transcript. We thus employed the SIFTS residue-level mapping resource (Dana et al., 2019) to maintain consistency between the UniProt and PDB residue numbering for each PDB chain-UniProt sequence pair. Specifically, for each PDB entry that was used as the 3D structure for a reference amino acid sequence, we downloaded the residue-level cross-reference data in XML format. Each of these XML file serve as the reference to ensure the accuracy of the mapping of each individual variant observed in gnomAD onto its location in the protein structure. Variants at positions that were not covered by protein structures were dropped.

Construction of the COSMIS score

For a protein sequence, the COSMIS score quantifies the constraint on a spatial region centered on each site of the sequence based on a reference protein structure. Construction of COSMIS is based on the concept of a contact set S. The contact set of site r, S_r, includes all sites in the reference structure whose C_β atoms (or C_α atoms in the case of glycines) are within 8 Å of the C_β atom of r and site r itself. Conceptually, S_r encloses the local 3D spatial neighborhood surrounding site r in the reference structure and typically includes 2-3 residues that are sequence neighbors of site r and important for determining the local secondary structure of r. But more

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

importantly, S_r also captures sites that are far apart in sequence but close in space and most likely to contribute to the structural and functional integrity of r. For a protein that has L amino acid sites, we thus have L contact sets, one for each site. The spatial constraint is quantified for each contact set and assigned to the site represented by the contact set.

We represent S_r by residue indices. Our site-level mapping between amino acid sites and sites in protein structures enables us to count the observed number of missense variants within each contact set. We designate this count m_o. The COSMIS score of each contact set quantifies the deviation of m_o from a null distribution of the number of missense variants within the contact set. We derived this null distribution through simulations. We designate this expected count of missense variants t_e. To derive the null distribution of missense variants within a contact set, we first constructed a normalized missense mutability of each codon in the transcript, that is

$$p_j = \frac{u_j}{\sum_{i=1}^{L} u_i}$$

where p_j is the missense mutability of codon j normalized to the total missense mutability of the transcript, u_j is the sequence context-dependent, unnormalized missense mutability of codon j described in a previous section, and L is the total number of amino acid sites in the protein sequence.

For each permutation, we then drew the missense variants from a multinomial distribution, where the number of trials is the total expected number of missense variants, t_e, for the protein and the probability for each amino acid site is p_j. We repeated this $N = 10000$ times. After each permutation, we count the number of missense variants sampled for each contact set. We denote the mean and standard deviation of this null distribution as m_e and m_σ, respectively. The COSMIS score is then computed as

$$COSMIS = \frac{m_o - m_e}{m_\sigma}.$$

We also count the number of times out of the N permutations the permuted count of missense variants in the contact set is less than or equal to the observed number m_o. We designate this count K. We calculated the empirical permutation p value using the following formula:

$$p_{value} = \frac{(K+1)}{(N+1)}.$$

Gene lists

We obtained the lists of genes of different levels of essentiality from https://github.com/macarthur-lab/gene_lists. These include: 294 *haploinsufficient* genes with sufficient evidence for dosage pathogenicity (level 3) as determined by the ClinGen Dosage Sensitivity Map as of Sep 13, 2018 (Rehm et al., 2015), 683 *essential* genes deemed essential in multiple cultured cell lines based on CRISPR/Cas screen data (Hart et al., 2017), 709 autosomal *dominant* disease genes from OMIM (Blekhman et al., 2008; Berg et al., 2013), 1183 autosomal *recessive* disease genes from OMIM (Blekhman et al., 2008; Berg et al., 2013), 913 *non-essential* genes deemed non-essential in multiple cultured cell lines based on CRISPR/Cas screen data (Hart et al., 2017), and 284 olfactory receptors (Mainland et al., 2015). Several of these lists were also previously used to benchmark the pLI metric that quantifies intolerance to functional variation (Lek et al., 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genes in these lists are identified by their HGNC symbols. To link with COSMIS scores (indexed by UniProt access numbers), we mapped HGNC symbols to UniProt accession numbers through programmatic access of the UniProt database identifier mapping service (https://www.uniprot.org/help/api_idmapping). Collectively, our framework provides scores for 213 haploinsufficient, 622 essential, 584 dominant, 999 recessive, 721 non-essential, and 284 olfactory genes. These gene lists are available at our GitHub repository.

Intra- and inter-species constraint metrics

We focused our comparison of COSMIS with other evolutionary constraint metrics rather than ensemble variant effect prediction methods derived through machine-learning or score aggregation. In particular, our primary interest was to compare COSMIS with recently developed human-variation-based constraint metrics, i.e., the residual variation intolerance score (RVIS) (Petrovski et al., 2013), the missense Z score (Samocha et al., 2014; Lek et al., 2016), the probability of loss-of-function intolerance metric (pLI) (Lek et al., 2016), missense tolerance ratio (MTR) (Traynelis et al., 2017), and missense tolerance ratio 3D (MTR3D) (Silk et al., 2021). Additionally, we compared COSMIS with GERP++ (Davydov et al., 2010), phyloP (Pollard et al., 2010), phastCons (Siepel et al., 2005), and ConSurf (Armon et al., 2001) to investigate the potential synergistic effects of combining intra- and inter-species metrics for predicting variant pathogenicity. We computed the ConSurf scores using the Rate4Site program (Pupko et al., 2002) with default parameters and no branch length optimization. We obtained the 100-way multiple sequence alignment for each of the proteins and the tree file (hg38.100way.nh) from the UCSC Genome Browser. We computed relative solvent accessibility using DSSP 3.0 (Kabsch and Sander, 1983) within the Biopython framework (Cock et al., 2009). The sources of other scores were listed in Supplementary Table 11. The subset of amino acid sites in the human reference proteome for which all scores are available can be found at our GitHub repository: https://github.com/CapraLab/cosmis.

COSMIS score distribution of ClinVar variants

We evaluated the ability of COSMIS to predict pathogenicity of variants using ClinVar variants (Landrum et al., 2018) (retrieved in August 2021) as an evaluation set. Our evaluation set consisted of solely ClinVar missense variants that were labeled as "pathogenic" or "likely pathogenic" for true positive (pathogenic) variants and "benign" or "likely benign" for true negative (benign) variants. All variants for both sets were required to have a review status of at least one star and no conflicting interpretation. Due to the dependency of COSMIS score on 3D structures, we also required variants in the evaluation set to be mappable to our 3D structure sets. Any ClinVar variant designated as “no assertion criteria provided”, “no assertion provided”, “no interpretation for the single variant”, or not covered by protein structures was excluded from the evaluation set. Collectively, these restrictions resulted in 19,596 benign and 17,858 pathogenic variants for which the COSMIS score can be computed. These variants and their COSMIS scores are available in Supplementary Table 5.

COSMIS score distribution of de novo missense variants

The set of de novo missense variants was obtained from a previous analysis (Samocha et al., 2017). This set consists of 5113 *de novo* missense variants in 5620 neurodevelopmental disorder probands (“case” variants) and 1269 *de novo* missense variants in 2078 unaffected siblings of

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

autism spectrum disorder probands ("control" variants). Following the procedure of residue-level mapping, we were able to map 2,271 case variants and 541 control variants to protein 3D structures and compute the COSMIS scores for these variants. These variants and their COSMIS scores are available in Supplementary Table 7.

Application of COSMIS to potassium ion channels

We selected a set of 41 clinically relevant potassium ion channels for which missense variants have been unambiguously annotated in ClinVar following the same procedure as stated in the previous section. Experimental structures are available for KCNH2, KCNJ11, KCNQ1, and KCNQ2, KCNQ4 in the PDB (as of Dec. 2020). For the rest 36 potassium channels, we leveraged their high sequence identity (mean sequence identities between template and target ion channels are 56.6%) to the potassium channels with available structures and constructed homology models in their oligomeric states using the SWISS-MODEL interactive workspace (Waterhouse et al., 2018). We removed residues with a QMEAN score of < 0.3 in the intracellular intrinsically disordered regions of these ion channels. More information about templates used in our homology modeling and evaluations of model qualities can be found in Supplementary Table 9. Using these potassium channel 3D structures, we computed two sets of COSMIS scores, based on monomers and oligomers, respectively, for 4,762 interface and 14,331 non-interface sites. Collectively, we were able to map and computed the COSMIS scores for a total of 111 and 489 unambiguously annotated benign and pathogenic KCN missense variants from ClinVar. Our ion channel variant dataset, together with modeled ion channel structures and their precomputed COSMIS scores, is available at https://github.com/CapraLab/cosmis.

Data availability

Precomputed COSMIS scores that are also mapped to protein structures for 16,533 proteins from the UniProt human reference proteome, estimated sequence-context-dependent nucleotide mutability, and the expected number of synonymous and missense variants are available and can be downloaded at https://github.com/CapraLab/cosmis.

Code availability

In addition to the precomputed scores, COSMIS can also be downloaded and run as a standalone application locally. The source code of COSMIS and all scripts that can be run to reproduce the results and figures of this work are available at GitHub (https://github.com/CapraLab/cosmis).

Author contributions

B.L. and J.A.C. conceived the study. B.L. compiled the data sets, wrote the source code, and analyzed the data. B.L., D.M.R., and J.A.C. interpreted the results. B.L. and J.A.C. wrote the manuscript with edits from D.M.R. B.L. and J.A.C. acquired funding to support this work. D.M.R. and J.A.C. supervised the project. All authors approved the submitted version of the manuscript.

Acknowledgments

We thank Dr. David Rinker for constructive criticism of the manuscript and members of the Capra Lab for insightful discussions. We also thank the researchers who contributed to the gnomAD resource and the DeepMind team for openly sharing AlphaFold predicted structures. This research was funded by the National Institute of Health grants R35GM127087 and R01LM013434

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(J.A.C.) and American Heart Association fellowship 20POST35220002 (B.L.). This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

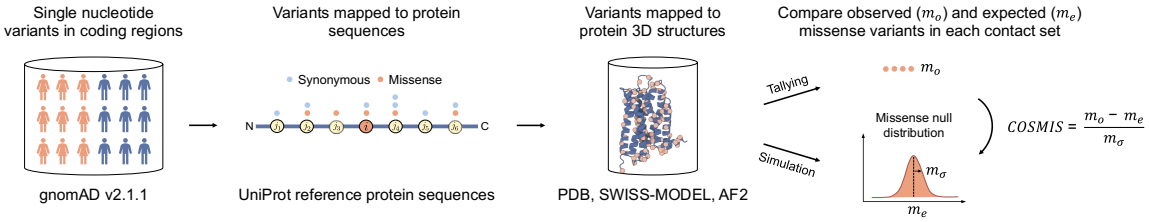

Fig 1. Schematic of the COSMIS 3D spatial constraint quantification framework. The COSMIS framework consists of mapping single nucleotide variants (SNVs) from gnomAD to human reference protein sequences and protein 3D structures, the computation of 3D contact sets, tallying of unique missense variants observed in contact sets, and comparison of observed missense variant counts to a null distribution simulated based on a mutation-spectrum-aware statistical model. We quantify the constraint on the index amino acid site with the Z score of the observed missense variant count (m_o) compared to the null distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

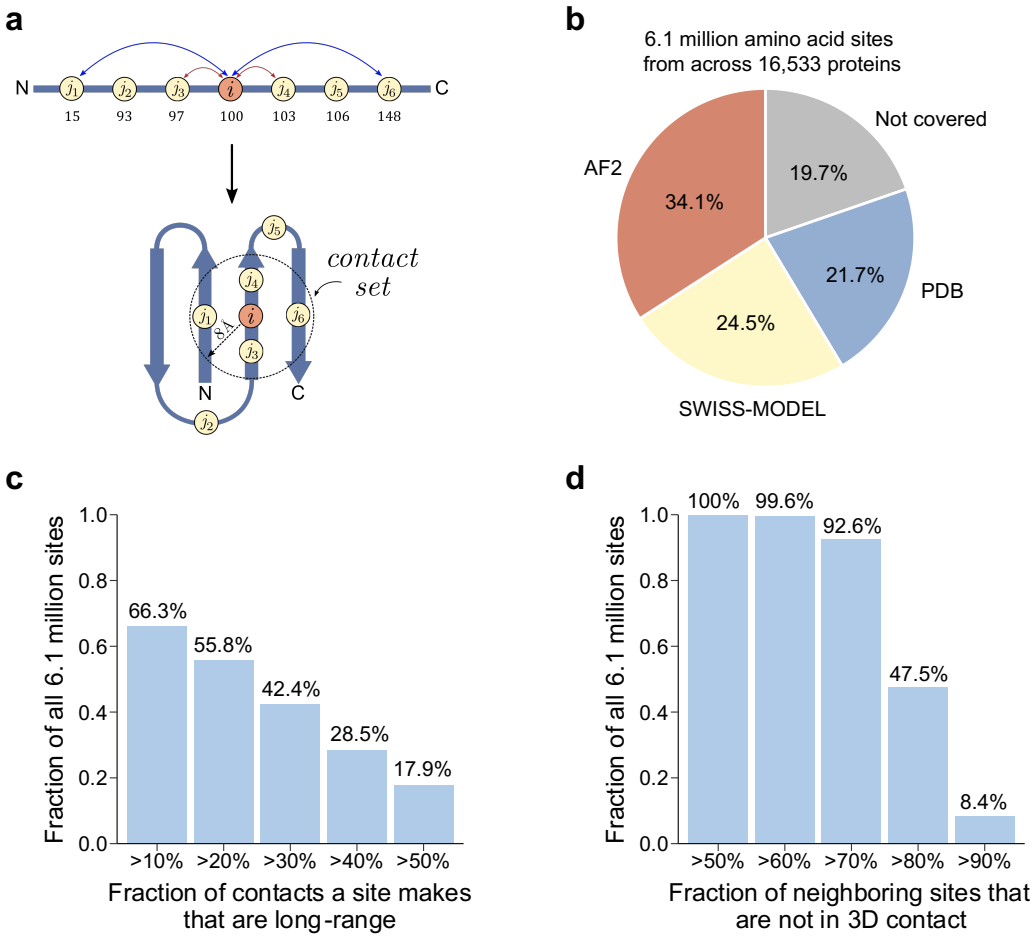

Fig 2. Protein 3D context differs from 1D sequence context. a) To quantify the 3D spatial context of each amino acid site (i), our framework defines its *contact set* as the amino acid residues that are in contact (C_β < 8 Å) with the residue. For the example index site (i), the contact set is (i, j_1, j_3, j_4, j_6). Numbers below the 1D sequence schematic represent residue sequence positions and illustrate that contact set residues may be distant in sequence from the index site. **b**) The COSMIS framework covers 80.3% of the reference human proteome. Defining the contact set of an amino acid site requires protein 3D structures. We used PDB and SWISS-MODEL as our primary sources of protein 3D structures. For proteins with no structure in the PDB or SWISS-MODEL that meet our criteria (Methods), we analyze models from the AlphaFold2 structure database (AF2). Numbers inside the pie chart represent fractions of the human reference proteome (20,600 proteins) for which we used the corresponding protein structure resource to compute COSMIS scores (Supplementary Table 1). **c**) Contact sets capture long-range sites (separated by more than 15 residues along the 1D sequence) that interact in 3D. For example, residues j_1 and j_6 in panel a) are not neighbors in 1D sequence, but nevertheless form long-range contacts with the index site i. The bar plot shows the fraction of all 6.1 million sites with at least a certain fraction of long-range 3D contacts in their contact sets. **d**) Many neighboring sites in 1D sequence do not form 3D contacts with an index site. Defining the contact set eliminates these sites from consideration. For example, residues j_2 and j_4 in panel a) are 1D sequence neighbors (within 15 residues) of the index site i but do not form 3D contacts with it. The bar plot shows the fraction of all 6.1 million sites that have at least a certain fraction of 1D sequence neighbors that do not form 3D contacts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

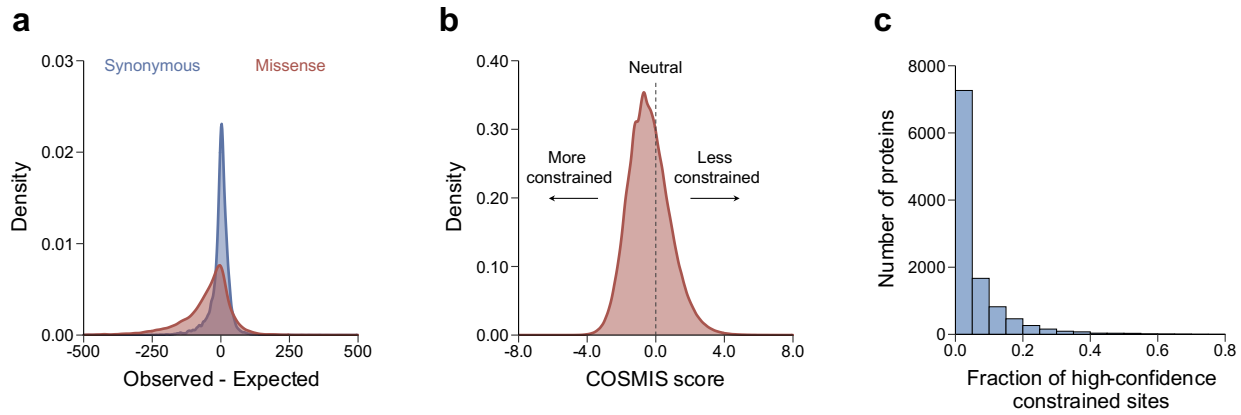

Fig 3. The COSMIS score quantifies depletion of missense variants in contact sets. **a**) Distribution of the deviation of the observed from the expected number of synonymous (blue) and missense (red) variants per-transcript computed from the mutability-aware model (Methods). The median of the deviation is roughly centered at zero (1.9) for synonymous variants but is significantly shifted towards negative values (more constraint) for missense variants (median -29.1, $p < 2.2 \times 10^{-308}$, two-sided Mann Whitney U test). **b**) Distribution of the COSMIS scores for 6.1 million unique amino acid sites of the reference human proteome. As expected, an average amino acid site in the human proteome is depleted of missense variants in its contact set (median COSMIS score -0.47) due to structural and/or functional constraint. **c**) Distribution of per-protein fraction of high-confidence constrained sites (empirical p < 0.01).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

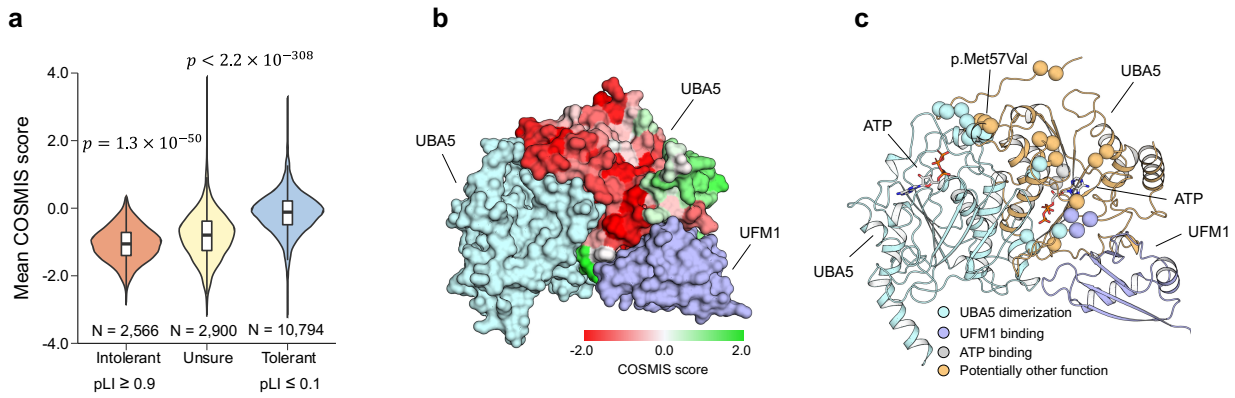

Fig 4. The COSMIS score quantifies constraint at amino acid resolution and provides structural insights into variant pathogenicity. a) COSMIS score distributions are significantly different between loss of function tolerant, unsure, and intolerant (as defined by pLI) genes. The COSMIS scores of amino acid sites in intolerant genes differ from those in tolerant genes (median -1.1 vs -0.12). **b**) COSMIS scores of UBA5 sites mapped to structure of one subunit of a dimerized UBA5 bound with the UFM1 target protein. UBA5 is predicted to be LoF tolerant, but it exhibits substantial constraint on specific spatial regions. Structures of all subunits of the complex are rendered in surface. **c)** Locations of the top 10% most constrained sites in UBA5 ranked by COSMIS score. Sites are rendered in spheres and colored according to their likely functional roles. Location of variant p.Met57Val implicated in early-onset encephalopathy is indicated. Proteins are rendered in cartoons.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

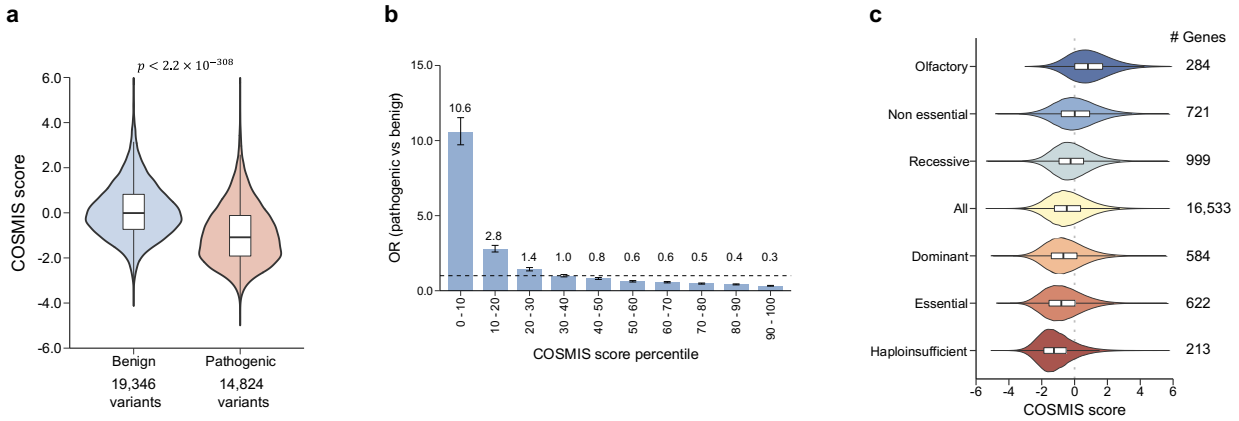

Fig 5. The COSMIS score is strongly correlated with both pathogenicity and gene constraint level. **a**) COSMIS score distributions for 19,346 benign and 14,824 pathogenic variants that are unambiguously annotated in ClinVar (Methods). Pathogenic variants have significantly more constrained 3D spatial neighborhoods (COSMIS score median -1.1) than benign variants (median score 0.0) ($p < 2.2 \times 10^{-308}$, two-sided Mann Whitney U test). **b**) Odds ratio (OR) of ClinVar pathogenic variants versus benign variants for different COSMIS score percentile bins (lower bins correspond to more constrained COSMIS scores). Amino acid sites with lower COSMIS scores are enriched for pathogenic variants whereas sites with higher scores are depleted of pathogenic variants. The horizontal dashed line indicates OR = 1. **c**) COSMIS score distributions of amino acid sites in six groups of genes with different functional annotations (and the dataset as a whole). As the anticipated functional constraint on each category increases (top-to-bottom), amino acid sites in genes in the category have more constrained COSMIS scores on average.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

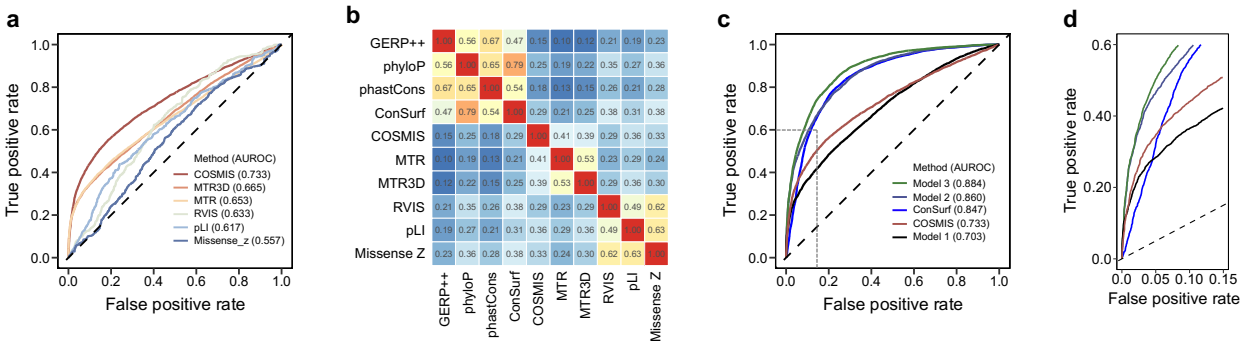

Fig 6. The COSMIS score is more predictive of variant pathogenicity than other constraint metrics. **a**) Comparison of the performance of COSMIS with five other constraint scores in predicting the pathogenicity of a total of 8,062 benign and 7,256 pathogenic missense ClinVar variants for which all scores are available. COSMIS significantly outperforms the other methods (AUROC 0.733 vs. 0.665 for MTR3D, the next best-performing method, $p = 2.5 \times 10^{-50}$, two-sided DeLong's test). **b**) A heatmap of Spearman rank correlations (absolute values) between four phylogenetic conservation scores (GERP++, phyloP, phastCons, ConSurf) and six constraint scores (COSMIS, MTR, MTR3D, RVIS, pLI, Missense Z) that are constructed based on human population genetic variants. **c**) COSMIS is complementary to phylogenetic constraint methods. ROC curves of logistic regression models integrating different combinations of the ten methods in panel b) at predicting the pathogenicity of the variants from ClinVar. Model 1: MTR + MTR3D + RVIS + pLI + Missense Z; Model 2: COSMIS + ConSurf; Model 3: all scores + relative solvent accessibility. **d**) A zoomed-in view of the high-confidence region of ROC space (bounded by the dashed lines in c). The improvement from adding COSMIS to ConSurf over ConSurf alone is mainly due to better performance in this high-confidence region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

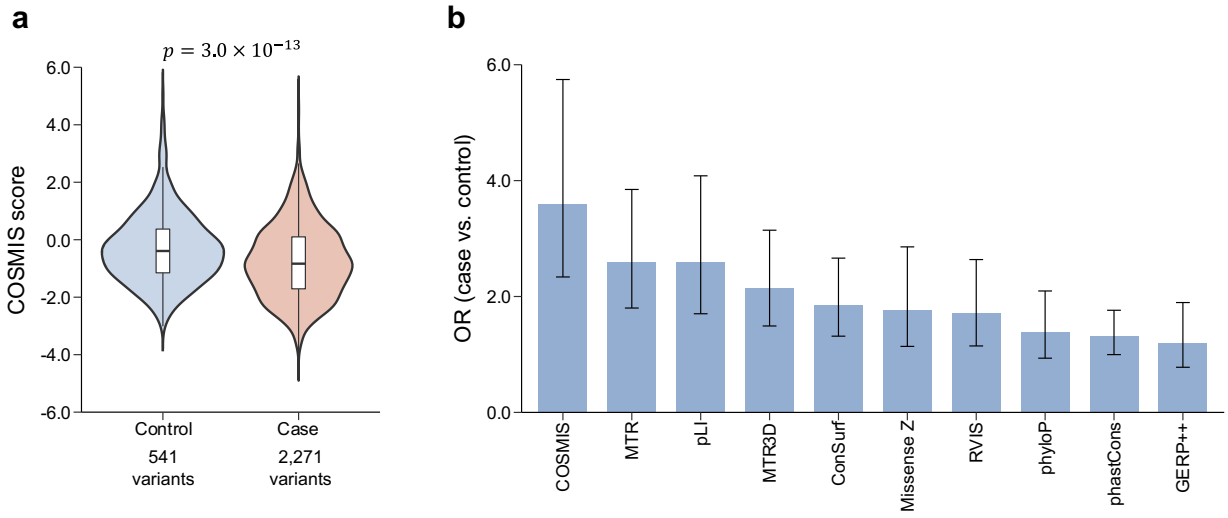

Fig 7. COSMIS score improves interpretation of *de novo* missense mutations from neurodevelopmental disorders. a) COSMIS score distributions for *de novo* missense mutations from neurodevelopmental disorder cases (Case) and from unaffected siblings of autism spectrum disorder probands (Control). Case variants have a significantly more constrained spatial neighborhoods than control variants (median COSMIS -0.83 vs. -0.39, $p = 3.0 \times 10^{-13}$, two-sided Mann Whitney U test). **b**) Case variant enrichment analysis for intra- and inter-species constraint metrics at the 10th percentile of most constrained sites. COSMIS has the highest enrichment for cases (odds ratio (OR) 3.6, 95% confidence interval [2.3, 5.7]). Error bars are 95% confidence intervals of ORs. Results of this OR analysis are consistent across thresholds other than the 10th percentile (Supplementary Fig. 9).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

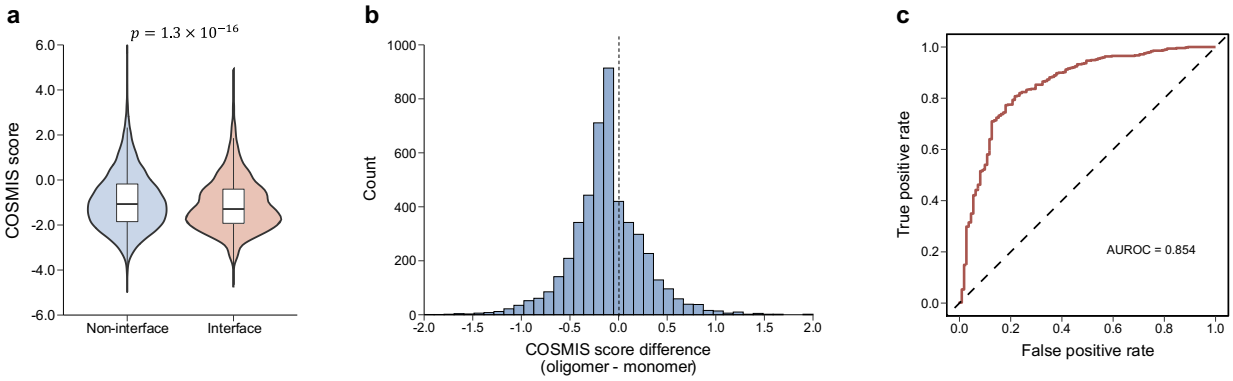

Fig 8. Applying COSMIS to custom-built oligomeric structural models facilitates interpretation of potassium channel variants. a) COSMIS score distributions for interface and non-interface amino acid sites in 41 oligomeric potassium ion channels. Overall, interface sites involved in oligomerization (making more 3D contacts in oligomers than in monomers) have a significantly lower COSMIS scores than non-interface sites (median -1.3 vs. -1.1, $p = 1.3 \times 10^{-16}$, two-sided Mann Whitney U test). **b**) COSMIS scores of interface sites computed based on oligomers are generally lower than those computed based on monomers (median difference -0.13, $p = 8.1 \times 10^{-8}$, two-sided Mann-Whitney U test). **c**) COSMIS score performs well (AUROC 0.854) at predicting the pathogenicity of 111 benign and 489 pathogenic potassium channel missense variants curated from ClinVar.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Armon, A., Graur, D., and Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol *307*, 447-463.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R.*, et al.* (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Research *45*, D158-D169.

Berg, J.S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C.P., Wilhelmsen, K.C., and Evans, J.P. (2013). An informatics approach to analyzing the incidentalome. Genet Med *15*, 36-44.

Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Research *45*, D313-D319.

Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr Biol *18*, 883-889.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci *13*, 190-202.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875-1882.

Cavasotto, C.N., and Phatak, S.S. (2009). Homology modeling in drug discovery: current trends and applications. Drug Discov Today *14*, 676-683.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B.*, et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.

Colin, E., Daniel, J., Ziegler, A., Wakim, J., Scrivo, A., Haack, T.B., Khiati, S., Denomme, A.S., Amati-Bonneau, P., Charif, M.*, et al.* (2016). Biallelic Variants in UBA5 Reveal that Disruption of the UFM1 Cascade Can Result in Early-Onset Encephalopathy. Am J Hum Genet *99*, 695-703.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet *12*, 628-640.

Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., and Velankar, S. (2019). SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res *47*, D482-D489.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP plus. Plos Computational Biology *6*.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat Rev Genet *8*, 610-618.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fadista, J., Oskolkov, N., Hansson, O., and Groop, L. (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics *33*, 471-474.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J.*, et al.* (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res *47*, D766-D773.

Gao, J., Chang, M.T., Johnsen, H.C., Gao, S.P., Sylvester, B.E., Sumer, S.O., Zhang, H., Solit, D.B., Taylor, B.S., Schultz, N.*, et al.* (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med *9*, 4.

Gussow, A.B., Petrovski, S., Wang, Q., Allen, A.S., and Goldstein, D.B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol *17*, 9.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell *138*, 774-786.

Hart, T., Tong, A.H.Y., Chan, K., Van Leeuwen, J., Seetharaman, A., Aregger, M., Chandrashekhar, M., Hustedt, N., Seth, S., Noonan, A.*, et al.* (2017). Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 (Bethesda) *7*, 2719-2727.

Hartl, D.L. (1989). Principles of population genetics / Daniel L. Hartl, Andrew G. Clark (Sunderland, Mass: Sinauer Associates).

Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of constrained coding regions in the human genome. Nat Genet *51*, 88-95.

Hayeck, T.J., Stong, N., Wolock, C.J., Copeland, B., Kamalakaran, S., Goldstein, D.B., and Allen, A.S. (2019). Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. Am J Hum Genet *104*, 299-309.

Hicks, M., Bartha, I., di Iulio, J., Venter, J.C., and Telenti, A. (2019). Functional characterization of 3D protein structures informed by human genetic diversity. Proceedings of the National Academy of Sciences of the United States of America *116*, 8960-8965.

Huang, Y.F., and Golding, G.B. (2015). FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. Bioinformatics *31*, 523-531.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A.*, et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A *112*, E5486-5495.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.*, et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. Journal of Molecular Biology *307*, 1487-1502.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., *et al.* (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062-D1067.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Li, B., Mendenhall, J., and Meiler, J. (2019). Interfaces Between Alpha-helical Integral Membrane Proteins: Characterization, Prediction, and Docking. Comput Struct Biotechnol J *17*, 699-711.

Liang, S., Mort, M., Stenson, P.D., Cooper, D.N., and Yu, H. (2021). PIVOTAL: Prioritizing variants of uncertain significance with spatial genomic patterns in the 3D proteome. bioRxiv.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., *et al.* (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469-476.

Mainland, J.D., Li, Y.R., Zhou, T., Liu, W.L., and Matsunami, H. (2015). Human olfactory receptor responses to odorants. Sci Data *2*, 150002.

Manolio, T.A., Rowley, R., Williams, M.S., Roden, D., Ginsburg, G.S., Bult, C., Chisholm, R.L., Deverka, P.A., McLeod, H.L., Mensah, G.A., *et al.* (2019). Opportunities, resources, and techniques for implementing genomics in clinical care. Lancet *394*, 511-520.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138-U163.

Meyer, M.J., Lapcevic, R., Romero, A.E., Yoon, M., Das, J., Beltran, J.F., Mort, M., Stenson, P.D., Cooper, D.N., Paccanaro, A., *et al.* (2016). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Human Mutation *37*, 447-456.

Mintseris, J., and Weng, Z.P. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America *102*, 10930-10935.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., *et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res *11*, 863-874.

Niu, B.F., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.W., Zhang, Q.Y., McLellan, M.D., *et al.* (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet *48*, 827-+.

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., *et al.* (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun *11*, 5918.

Perszyk, R.E., Kristensen, A.S., Lyuboslavsky, P., and Traynelis, S.F. (2021). Three-dimensional missense tolerance ratio analysis. Genome Res *31*, 1447-1461.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet *9*, e1003709.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18 Suppl 1*, S71-77.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L.*, et al.* (2015). ClinGen--the Clinical Genome Resource. N Engl J Med *372*, 2235-2242.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E.*, et al.* (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med *17*, 405-424.

Rose, P.W., Prlic, A., Altunkaya, A., Bi, C.X., Bradley, A.R., Christie, C.H., Di Costanzo, L., Duarte, J.M., Dutta, S., Feng, Z.K.*, et al.* (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research *45*, D271-D281.

Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. bioRxiv.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A.*, et al.* (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet *46*, 944-950.

Shieh, C.C., Coghlan, M., Sullivan, J.P., and Gopalakrishnan, M. (2000). Potassium channels: molecular defects, diseases, and therapeutic opportunities. Pharmacol Rev *52*, 557-594.

Shrestha, R., Fajardo, E., Gil, N., Fidelis, K., Kryshtafovych, A., Monastyrskyy, B., and Fiser, A. (2019). Assessing the accuracy of contact predictions in CASP13. Proteins *87*, 1058-1068.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Silk, M., Pires, D.E.V., Rodrigues, C.H.M., D'Souza, E.N., Olshansky, M., Thorne, N., and Ascher, D.B. (2021). MTR3D: identifying regions within protein tertiary structures under purifying selection. Nucleic Acids Res *49*, W438-W445.

Sivley, R.M., Dou, X., Meiler, J., Bush, W.S., and Capra, J.A. (2018). Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. Am J Hum Genet *102*, 415-426.

Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet *101*, 315-325.

Stehr, H., Jang, S.H., Duarte, J.M., Wierling, C., Lehrach, H., Lappe, M., and Lange, B.M. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Mol Cancer *10*, 54.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G.*, et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64-69.

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D.M., Kim, R., Ryan, M., Masica, D.L., and Karchin, R. (2016). Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res *76*, 3719-3731.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.15.460390; this version posted September 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res *27*, 1715-1729.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A.*, et al.* (2021). Highly accurate protein structure prediction for the human proteome. Nature.

Veltman, J.A., and Brunner, H.G. (2012). De novo mutations in human genetic disease. Nat Rev Genet *13*, 565-575.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L.*, et al.* (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research *46*, W296-W303.

Worth, C.L., Gong, S., and Blundell, T.L. (2009). Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Bio *10*, 709-720.

Yu, F.H., and Catterall, W.A. (2004). The VGL-chanome: a protein superfamily specialized for electrical signaling and ionic homeostasis. Sci STKE *2004*, re15.

Ziegler, A., Colin, E., Goudenege, D., and Bonneau, D. (2019). A snapshot of some pLI score pitfalls. Hum Mutat *40*, 839-841.